Term burstiness: evidence, model and applications

Thesis

Term Burstiness: Evidence, Model and Applications

by Avik Sarkar

Bachelors in Statistics,
Calcutta University, India (1999)

Masters in Applied Statistics and Informatics,
Indian Institute of Technology, India (2001)

Submitted to the Computing Department
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computing

at THE OPEN UNIVERSITY

August 2006

DATE OF SUBMISSION 30 AUGUST 2006
DATE OF AWARD 16 JANUARY 2008

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Acknowledgements

Behind every successful effort, there is an unfathomable sea of gratitude to those who inspire and encourage. I am grateful to those who have provided support and guidance during my PhD. This thesis would be incomplete without mentioning those people whose help has made this thesis possible.

First and foremost I would like to express my deepest sense of gratitude to my supervisors, Prof. Anne De Roeck and Prof. Paul H Garthwaite, for providing me with valuable guidance, esteemed suggestions and constant encouragement, which led to the successful completion of this thesis. Also I would like to thank my supervisors for reading through my thesis several times and suggesting me corrections. I am grateful to my external examiner Prof. Rob Gaizauskas for his valuable time in reading my thesis and comments to improve the thesis. I would also like to thank my internal examiner Prof. Darrel Ince for this help and support.

I wish to thank the academic staff of the Computing Department. Special regards are due to Dr. Marian Petre for her help and support in understanding the meaning and essence of doing a PhD and research work through her postgraduate forums. I would like to convey my thanks to Dr. Robin Laney as my third party supervisor. I would like to convey special thanks to the members of the NLP group in *The Open University* for interesting readings and discussions on various NLP related topics. Thanks to Dr. Alistair Willis and Prof. Donia Scott for stimulating questions and discussions on my research work in departmental seminars.

My colleagues and office mates of the Computing Department all gave me the feeling of being at home at work. Dr. Francis Chantree, Dr. Katerina Tzanidou, Dr. Andrea Capiluppi, Dr. Geke vanDijk, Debra Haley, Mohammad Salifu and Armstong Nhlabatsi: many thanks for being your colleague. Special thanks are due to my colleague David Jenkinson from the Statistics Department,

who had provided me initial help in getting started with the WinBUGS software. I would like to thank Gaston Burek and Dileep Damle for stimulating discussions and help regarding NLP.

I wish to thank the secretaries of Department of Computing, especially Debbie, Jennifer and Merisa for the assistance and co-operation received during I the PhD. Many of the experiments reported in this thesis would not have been possible without the constant support of the staff from the TBT IT Support Group: David Clover, Maja Dunn, Allan Thrower and Robin Goodman. I would like to convey special thanks to Allan Thrower from TBT for help with the Linux clusters and heavy computations that were required for my thesis.

I would like to convey special gratitude and thanks to my parents Mr Jadu Nath Sarkar and Mrs Bharati Sarkar for their constant support and encouragement while in India. I am very grateful for my dear wife Tannishtha, for her love and patience during the PhD period. I have been promising her I would finish my thesis a long time now!

Preface

Certain portions of this thesis have been published. The work on homogeneity experiments for detecting the effect of term burstiness on a dataset has been published [DRSG04a, DRSG04b, SDR04, DRSG05]. The model for term burstiness is published [SGDR05]. Applications based upon the proposed model are also published [SGDR05, SDRG05, DRSG05]. Recently our proposed model of term burstiness has been used for the purpose of word sense disambiguation in Machine Translation at the *Stockholm University* [Arg06]. Researchers at the *University of Tübingen* are currently using the proposed model of term burstiness to obtain burstiness profiles of terms in a corpus.

Avik Sarkar

August, 2006

Term Burstiness: Evidence, Model and Applications

by

Avik Sarkar

Submitted to the Computing Department on August 2006,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computing

Abstract

The present thesis looks at the phenomenon of term burstiness in text. Term burstiness is defined as the multiple re-occurrences in short succession of a particular term after it has occurred once in a certain text. Term burstiness is important as it aids in providing structure and meaning to a document. Various kinds of term burstiness in text are studied and their effect on a dataset explored in a series of homogeneity experiments. A novel model of term burstiness is proposed and evaluations based on the proposed model are performed on three different applications.

The "bag-of-words" assumption is often used in statistical Natural Language Processing and Information Retrieval applications. Under this assumption all structure and positional information of terms is lost and only frequency counts of the document are retained. As a result of counting frequencies only, the "bag-of-words" representation of text assumes that the probability of a word occurring remains constant throughout the text. This assumption is often used because of its simplicity and the ease it provides for the application of mathematical and statistical techniques on text. Though this assumption is known to be untrue [CG95b, CG95a, Chu00], but applications [SB97, Lew98, MN98, Seb02] based on this assumption appear not to be much hampered.

A series of homogeneity based experiments are carried out to study the presence and extent of term burstiness against the term independence based homogeneity assumption on the dataset. A null hypothesis stating the homogeneity of a dataset is formulated and defeated in a series of experiments based on the χ^2 test, which tests the equality between two partitions of a certain dataset. Various schemes of partitioning a dataset are adopted to illustrate the effect of term burstiness and structure in text. This provided evidence of term burstiness in the dataset, and fine-grained information about the distribution of terms that might be used for characterizing or profiling a dataset.

A model for term burstiness in a dataset is proposed based on the gaps between successive occurrences of a particular term. This model is not merely based on frequency counts like other existing models, but takes into account the structural and positional information about the term's occurrence in the document. The proposed term burstiness model looks at gaps between successive occurrences of the term. These gaps are modeled using a mixture of exponential distributions. The first exponential distribution provides the overall rate of occurrence of a term in a dataset and the second exponential distribution determines the term's rate of re-occurrence in a burst or when it has already occurred once previously. Since most terms occur in only a few documents, there are a large number of documents with no occurrences of a particular term. In the proposed model, non-occurrence of a term in a document is accounted for by the method of data censoring. It is not straightforward to obtain parameter estimates for such a complex model. So, Bayesian statistics is

used for flexibility and ease of fitting this model, and for obtaining parameter estimates.

The model can be used for all kinds of terms, be they rare content words, medium frequency terms or frequent function words. The term re-occurrence model is instantiated and verified against the background of different collections, in the context of three different applications. The applications include studying various terms within a dataset to identify behavioral differences between the terms, studying similar terms across different datasets to detect stylistic features based on the term's distribution and studying the characteristics of very frequent terms across different datasets. The model aids in the identification of term characteristics in a dataset. It helps distinguish between highly bursty content terms and less bursty function words. The model can differentiate between a frequent function word and a scattered one. It can be used to identify stylistic features in a term's distribution across text of varying genres. The model also aids in understanding the behaviour of very frequent (usually function) words in a dataset.

Thesis Supervisor: Anne De Roeck
Title: Professor

Thesis Supervisor: Paul H Garthwaite
Title: Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, a punctuation or space delimited string that appears in text is referred to as a "term". A term tends to be reused in the same text after it has been used once. The property of multiple occurrences of a particular term in a close vicinity is called "term burstiness". This thesis discusses the issue of "term burstiness" in text. Burstiness is an inherent characteristic of text because as the text progresses certain terms tend to re-occur and provide structure and flow to the document, which in turn provides meaning to the text. The following section describes ways in which term burstiness is currently handled, and gives an overview of the thesis and its main contributions.

1.1 Present Scenario

Term burstiness is a inherent characteristic of text. It means that a particular term tends to re-occur several times in the same text after it has been recently mentioned [Kat96]. Term burstiness is also known as clumpiness [Kil01, NSJ64].

The current scenario in fields like statistical Natural Language Processing (NLP) and Information Retrieval (IR), including text classification, is to treat term occurrences in a document or text as

independent events. This is embodied in the "bag-of-words" assumption. The "bag-of-words" model assumes that terms occur independently: i.e. that the probability of a term occurring is the same throughout the text. In other words, it assumes that the terms are spread homogeneously. This independence assumption of text supports the Vector Space model, a frequency based representation of textual data. Text is represented as a n-dimensional vector where each independent dimension of the vector represents a term and the magnitude of that vector in each dimension is proportional to the term's frequency in the text. The model is immensely popular, due to its simplicity and the ease with which mathematical and statistical techniques can be applied to it. Text represented using the "bag-of-words" model or "bag-of-words" representation is also referred to as text following the "bag-of-words" assumption or independence assumption.

On the other hand, the consequences of this model are far reaching. It assumes that once a term occurs in a document, its overall frequency in the entire document is the only useful measure that associates a term with a document. It does not take into consideration whether the term occurred in the beginning, middle or end of the document. Neither does it capture whether the term occurs many times in close succession or whether it occurs uniformly throughout the document. It also assumes that additional positional information does not provide any extra leverage to the performance of the NLP and IR applications based on it. This assumption has been shown to be invalid, for instance in applications of automatic ambiguity resolution [Fra97].

There is a growing literature which investigates "burstiness" in the distribution of content words in documents - i.e. the fact that repeated occurrences of an informative word in a document tend to cluster together (e.g. [Chu00]). By and large, however, function words are ignored, or assumed to distribute evenly throughout text, to the point of becoming uninformative. Several words used in English act as connectors to add lexical cohesion or readability to the underlying text [HH76, Hoe91]. Such words include co-ordinating/subordinating conjunctions such as *and, but,*

because and *if*, and adverbials such as *however*, *consequently* and *therefore* can also be clue items to understand the lexical relationships in discourse. In addition to this, reference words such as *this*, and *that* can also fall in the same category of words. In this thesis this category of words will be loosely referred to as function words as they individually do not describe the topic or content of the text. Here, in this thesis, a term related to the main topic of the text is referred to as "content term", whereas terms that aid in the lexical cohesion of text is referred to as "function word".

Indeed, Katz [Kat96] develops a model for bursty distributions of "content terms", and distinguishes between function words and content words on the grounds that function words are distributed more homogeneously throughout text. However, once text is represented using the "bag-of-words" model, positional information is lost and it is not possible to distinguish between bursty and non-bursty occurrences of terms: the independence assumption inherent in the "bag-of-words" model will only be capable of capturing homogeneous distributions. This is also true for the Vector Space model that represents text in a form that has lost a lot of information from the original representation. If terms were in reality homogeneously distributed, then this independence assumption would incur no great loss, and indeed the success of "bag-of-words" and Vector Space models suggests that perhaps distributional information is irrelevant for many applications. On the other hand, it is clear that such representations loose information that has been shown to be relevant in distinguishing between topical and non-topical terms. So, if terms are not homogeneously distributed (which we know they are not), then does this fact constitute some extra information which text processing applications might use? Also, how should we measure to what extent terms are burstily and not homogeneously distributed in a particular collection, and is this useful information? This thesis aims to answer this question based on a series of homogeneity experiments at various levels of granularity, and conducted on different collections.

Since term burstiness is an inherent characteristic of text it will be helpful to capture this phe-

nomenon in terms of a model. An adequate model of term burstiness would provide insight about the different characteristics and bursty usage of various terms. This would in turn benefit applications where different term behaviours are to be exploited. Existing models for term distribution are based on the independence assumption (discussed in Chapter 3), and are unable to fully capture the term burstiness phenomenon.

This poses the question as to whether it is appropriate to model term burstiness based only on frequency counts collected from individual documents. Since a large portion of information about a term is lost by using frequency counts alone, how well would such models represent burstiness information? A new method for modeling term burstiness is proposed based on the gaps between individual occurrences of a particular term. This method is not based on frequency counts and hence is not subject to some of the drawbacks of frequency based approaches. Having described the model, a measure of term burstiness is proposed.

The effectiveness of the proposed model is verified and discussed against the backdrop of a series of applications where the findings from the model are compared with other findings reported in the literature.

1.2 Contributions

The main contributions of the thesis are:

- Providing large scale experimental evidence about term burstiness; developing a methodology for exposing the extent to which the independence assumption is not capturing term distribution characteristics in different documents and collections.

- Developing a fine grained methodology for measuring degrees of heterogeneity of term distributions in different collections.

- Developing a model of term burstiness based on the gaps between successive occurrences of the term. This model retains positional and structural information about the term.
- Modeling the above mentioned gaps based on the mixture of exponential distributions and using Bayesian statistics for parameter estimation based on this complex model.
- Evaluating findings based on the model by comparing them to findings using frequency based methods alone.
- The model can be used to differentiate between different term behaviours, which cannot be captured by frequency information alone.
- The model may be used for identifying stylistic differences across terms of different genres.
- The distributional characteristics of very frequent function words of a dataset are studied based on the proposed model to detect any evidence of burstiness among these terms.

The other contributions of the thesis and the above stated main contributions of the thesis are discussed in detail in the rest of this section.

The thesis looks at the phenomenon of term burstiness and differentiates between the various kinds of burstiness in text, viz. term burstiness or within-document burstiness, concept burstiness and document-level term burstiness (Chapter 2).

The thesis looks into ways of finding the evidence of term burstiness in a dataset as a whole. This is based on a series of homogeneity experiments on datasets at various levels of granularity. The homogeneity experiments are an extension of the basic framework proposed by Kilgarriff [Kil97]. The homogeneity measure or self-similarity measure of the dataset is calculated by dividing the dataset into two halves and then measuring the similarity between the two halves. In this thesis the similarity is measured by using the χ^2 test, with the χ^2 statistic as a measure of similarity between

the two halves. The probability value of this χ^2 test provides evidence about the similarity of the two halves at a level of statistical significance (Chapter 5).

Various schemes for dividing the dataset into two halves are adopted. These schemes were aimed at capturing the effect of term burstiness at document level, within a document and further into various text chunks. The effect of term burstiness is more prominent in longer text with more structure. The homogeneity measures at various levels of granularity provide indications of significant differences between datasets and may be of use for dataset profiling (Chapter 5). The homogeneity measures for the frequent terms in particular may be used for the purpose of dataset profiling. The performance of any machine learning, information retrieval or natural language technique is very much dependent on the dataset to which the technique is applied. Dataset profiling aims at capturing differences among various datasets and identifying characteristics of a dataset that may influence the performance of techniques based upon it.

Review of a series of methods aimed at modeling term distribution in a dataset showed that the methods based on frequency counts cannot fully capture the phenomenon of term burstiness (Chapter 3). This chapter also provides motivation for a better model that addresses the issue of term burstiness.

A model of term burstiness and term re-occurrence in a text collection is proposed. It is based on the gaps between successive occurrences of a term in the dataset. These gaps are modeled using a mixture of exponential distributions. Parameter estimation is achieved using a Bayesian framework that allows the fitting of a flexible model. The model provides measures of a term's re-occurrence rate and *within-document burstiness*. The parameters capture the mean distance between bursts, and the mean distance between terms occurring within a burst. The model allows a distinction to be made between different distribution patterns which may fit with the behaviour of different types of words, as suggested in the literature ([CG95b, Kat96]). The model can estimate after how

many words a particular term is likely to occur in the overall dataset and after how many words the term will tend to re-occur in the close vicinity. The model will be used to verify whether different term behaviours can be captured by using it. In this thesis, the particular interest will be on the following types of distribution. In line with Katz [Kat96], a content bearing term is identified as one that occurs rarely in the dataset, but, re-occurs multiple times in close vicinity once it has occurred. In contrast, the hypothesis is that a frequent function word like *the* or *and* are expected to occur at a more constant rate at least throughout collections with large amounts of running, connected text (e.g., not lists). Less frequent function words like *expect* are expected to occur rarely and that too in a scattered manner. In many applications, term importance is determined based on the amount of content bearing information in it, so a measure is proposed to account for the term's importance based on its distribution pattern in the dataset (Chapter 6).

It will then be verified whether the results returned by the model reflect actual term behaviours. Due to the computational cost of running the model, it is not possible to achieve this by setting up evaluations based on applications that involve large amounts of terms, or fast calculation of parameters. Instead, relying on alternatives the model is checked and verified manually against the backdrop of various applications which are sensitive to term distribution information, such as genre identification (Chapter 8). A range of different terms across various datasets is studied based on the proposed model and compared with findings from frequency based models reported in the literature for the same datasets. The term burstiness model is found to be capable of modeling frequent function words, rare content words and medium frequency terms. The model can be used to understand term characteristics within a dataset. The model may also be used to detect stylistic variations across datasets of different genres by modeling a particular term behaviour across different genres. The distributional characteristics of very frequent function words, which are removed as background noise in various applications, are studied.

1.3 Structure of thesis

The thesis starts (Chapter 2) by discussing the issue of term burstiness in text and differentiating the different types of burstiness, namely term burstiness or within-document burstiness, concept burstiness and document-level burstiness. The focus of the current thesis is set on "term burstiness" with occasional references to the other mentioned types of burstiness.

Chapter 3 takes a look at term distribution models, in particular those aimed at the phenomenon of term burstiness. Many of these models are based on frequency counts and have shortcomings when capturing term burstiness. Some models aim to deal with certain special characteristics of text, like large numbers of zero counts, or the properties of texts by different authors or different genres etc. Some of the existing models of term distribution are based on frequency counts of terms in a document and in the document collection. Models that are solely based on frequency counts assume independence of terms in text, hence these models are unable to capture some key aspects of the distribution of terms and the inherent structure of text. This motivates the development of models of term distribution that will account for term burstiness within a text and also across the whole document collection.

Chapter 4 introduces the various datasets used for experiments in this thesis. Basic summary statistics for each of these datasets are provided. A series of basic profiling steps are carried out to identify any inherent discrepancies in these datasets. Manual sampling of the dataset was carried out along with studying the most frequent terms for each dataset. Any issues with regards the sparseness of the datasets were verified based on the type-to-token ratio and evaluating the fit of Zipf's curve to the dataset.

Chapter 5 presents the homogeneity experiments based on the χ^2 methodology. Various methods for measuring homogeneity are studied and the χ^2 method is chosen for homogeneity experiments.

An approach to measure heterogeneity is developed in this thesis by (a) postulating homogeneity as a null-hypothesis and (b) defeating that null-hypothesis in a progressive sequence of steps which allows one to gain an understanding of how easy it was to defeat the hypothesis for a given collection. The χ^2 methodology is used as a statistical test to judge self-similarity of a dataset at a level of statistical significance. The homogeneity measures at various levels of granularity revealed several complex differences between datasets. The experiments also showed heavy dependence of these measures on the inherent structure of the documents and the document collection that is under consideration. The experiments defeated the homogeneity assumption. The homogeneity values for the top terms may be used for profiling datasets. The findings in Chapter 5 indicated that term burstiness is an important characteristic of text that ought to be taken into consideration while working with text based applications.

Chapter 6 proposes a term burstiness model of a term's re-occurrence pattern in a dataset. It models gaps between successive occurrences of a particular term, and not the term frequency count. The model retains structural information about a term's distribution in the dataset. It can capture first and last occurrence and the gaps are modeled by a mixture of exponential distributions. Non-occurrence of a term in a document is modeled by the statistical concept of *censoring*, which states that the event of observing a certain term is censored at the end of the document, i.e. the document length.

Chapter 7 describes the Bayesian framework that is used for deriving the parameter estimates based on the proposed model. Deriving parameter estimates from a model that is a mixture of two distributions is quite complicated and a closed form solution cannot be derived. Bayesian statistics is used for the purpose of parameter estimation. Gibbs sampling based on Markov Chain Monte Carlo (MCMC) methods are used for parameter estimation. Various sampling techniques used for deriving the parameter estimates are discussed in that chapter.

Chapter 8 describes some applications of the term re-occurrence burstiness model. A range of terms of various characteristics is chosen to demonstrate the applicability of the burstiness model. Based on the model, term characteristics in a dataset can be determined. Another application uses the burstiness model for differentiating between text of various genres. In this application, certain terms are chosen and modeled across datasets representing different genres. The model helps in differentiating text across genres. The model is used to study the characteristics of very frequent terms in a dataset. These are often treated as background noise, as they are assumed to be non-informative and homogeneously distributed. But studying these frequent terms based on the burstiness model revealed bursty characteristics of several function words.

Conclusions and directions for future work are presented in Chapter 9. The achievements and contributions of the thesis are also summarized in that chapter. The effect of term burstiness and evidence using the homogeneity experiments are discussed. The lack of an appropriate model for term burstiness motivated the development of the proposed model of term burstiness based on term re-occurrence. Findings from the model are summarized. The limitations of the present model are discussed along with proposed research directions to overcome the limitations. Application areas where the the model might be successful are also discussed in that chapter.

Chapter 2

Term Burstiness

This chapter introduces the concept of "term burstiness" and attempts a definition. It discusses and contrasts different notions of burstiness in text. The importance of term burstiness is highlighted.

2.1 Burstiness

The term "burstiness" derives from the noun "burst", and the meaning of burst relevant to this thesis is *"a sudden increase in something, especially for a short period"* [1]. Burstiness is a phenomenon in nature observed in various events around us. In this thesis the discussion will revolve around the sudden increase in the occurrence of a certain term, i.e. "term burstiness".

One may observe burstiness in various events and activities in our daily life. If one looks at events casually, they might appear to be random; e.g. whether it rains the next day, sales in a high street shop on a certain day, or whether a certain team will win the football match today. On closer inspection across many similar events, trends may emerge out of them. Rain is more likely during the rainy season, sales of commercial goods are high during festival months or special

[1] *Cambridge dictionary*

events, match results over a long period can tell us about the likely winner of a forthcoming match. We also expect certain events to exhibit some uniform pattern or behavior. One hopes buses will arrive at the bus-station at regular intervals, or expects a uniform load on a certain server that is processing jobs, or a fixed level of traffic in a data transfer network. However, in many cases one may observe that events do not occur at regular intervals; instead they occur in a "burst". This characteristic deviation of a steady inflow of certain events is characterized as "burstiness". For example, a bus does not come for a long time and then once one arrives, more buses arrive soon afterwards. Burstiness is also observed in other scenarios, like transmitting data over a network which is expected to be uniform but is bursty at times, and video/audio transmission between two devices also exhibits burstiness in the transmission pattern. In such scenarios there may be very minimal or uniform data transmission over a time period, and then an abrupt increase in the amount of data being transferred [WC93].

2.2 Types of burstiness in text

The phenomenon of multiple occurrences of a particular term in a close neighborhood is called *term burstiness*. The issue of term burstiness is the primary focus of this thesis. There are some other types of burstiness in text, viz. *document-level term burstiness* and *concept burstiness* which are discussed briefly in the following sections along with the discussion on *term burstiness*.

2.2.1 Term Burstiness

The phenomenon of multiple occurrences of a particular term in a close neighborhood is called *term burstiness*. Burstiness is acknowledged as a fundamental characteristic of text and document streams, since the mention of a particular term in a document indicates that the term is being discussed in that document, and so it tends to occur some more times in the span of the discussion.

Burstiness is also often referred to as *clumpiness* based on the re-occurrence of different multi-word terms [Kil01, NSJ64]. The phenomenon of term burstiness has also been observed by other researchers in the past.

Simon [Sim55] observes burstiness in text, stating "*as a text progresses, it creates a meaningful context within which words that have been used already are more likely to appear than others*".

Church and Gale [CG95b] commented on terms in text being "like an contagious disease" that is likely to occur again, in contrast to lightning which is extremely unlikely to strike again. They stated:

"*Under standard independence assumptions, it is extremely unlikely that lightning would strike twice (or half a dozen times) in the same document. But text is more like a contagious disease than lightning. If we see one instance of a contagious disease such as tuberculosis in a city, then we would not be surprised to find quite a few more.*"

Term burstiness can easily be observed for content bearing words. If a text discusses a certain person or topic, certain terms referring to the subject of discussion will tend to re-occur several times as the discussion on that person or topic progresses. The bursty behaviour of terms is not captured by approaches to text processing that use the "bag-of-words" model. The "bag-of-words" in which only frequency information of a term in a document is retained and other positional information about a term is lost, makes an independence assumption. The independence assumption postulates that a term is equally likely to occur in any position in the document. It is known that the independence assumption is incorrect. If this assumption was correct, and terms were equally likely to appear at any position in the document, the distribution of terms would be highly homogeneous - indeed random. Such an occurrence pattern would curtail the ability of text to convey meaning, i.e. burstiness is an inherent characteristic of language, and hence text, precisely because it is the deviation from random noise that imparts meaning and ensures communication.

Church and Gale observed that content words have a far more bursty occurrence pattern than non-content words. The table in Figure 2.1 is taken from their paper [CG95b]. In this table they study a list of low frequency terms across the entire collection, i.e. terms that occur $15-19$ times in the entire corpus. They then divide these terms into two groups, one in which the term occurs in only a single document and in the other in which the terms occur across at least two documents. The table shows the values of frequency and document frequency (number of documents in which the term occurs) for these terms. All the words in any row occur the same number of times and it can be seen that the occurrences of most likely non-content words are scattered across many documents in contrast to the comparatively content bearing terms that occur many times in one document only. These terms that occur in only one document are mostly all content bearing terms, and tend to have a bursty characteristic in documents. This is in contrast to the likely function words (which includes obvious function words, such as "the", and "of", as well as occurrences of terms which are acting as function words, such as the adverb "current", and which will be called function bearing word in this thesis) which tend to scatter across many different documents.

In another study [Chu00] Church observed that topical content words (i.e. a content bearing word associated with the main topic of the document), especially person names and surnames, tend to occur many more times after first mention in a document. Church observed that the actual probability of multiple occurrences of a term did not obey the "bag-of-words" independence assumption.

Now, let us turn to burstiness in the behaviour of function and function bearing words. In many applications, it is common practice to remove stop words by using a fixed list. These lists are either compiled to include function words with some linguistic knowledge, or are harvested from frequency counts, using some threshold, to exclude very frequent terms. In the latter case also, function words are usually well represented in the list. The practice of removing stop-words by using fixed lists,

freq	df = 1 (bursty)	df = freq (diffuse)
15	Blackman, Dandy, Drug's, Eugenia, Fromm's, Hardy's, Juanita, Selden, Ulyate, collage, tappet	Naturally, Norman, Otherwise, Somehow, Thank, cease, claiming, clue, confident, indispensable, landed, originated, plunged, restricted, sweep, termed
16	Gilborn, Handley, Hanford, Nicolas, Styka, Willis, clover, leveling, secants, thyroglobulin	Already, Back, None, Right, absurd, appearing, collect, delighted, deserves, devised, discussing, faster, inherited, legitimate, lined, link, men's, persuade, piled, praise, refuse, severely, shops, sole, spreading, thereafter, unnecessary, waved
17	Angie, BOD, Giffen, Krim, Lalaurie, Lizzie, Moreland, Nadine, TSH, Trevelyan, accelerometer	35, Go, K., artificial, capture, consistently, designated, expecting, formally, grasp, lit, obscure, pushing, respective, spontaneous, surprisingly, vitality
18	Andrei, Barco, Helion, Keys, Kitti, Langford, Madden, Saxon, Stevie, Upton, effluent, nonspecific	Beyond, avoided, birthday, emphasized, escaped, gather, instantly, packed, proceed, repeatedly, sixty, submit, surrounded
19	Haney, Killpath, Letch, tetrachloride, tsunami	Which, alike, amazing, bold, happily, notable, overwhelming, remainder, rid, rush, savage, whereby

Figure 2.1: Figure showing that content words have a bursty distribution compared to non-content words. Terms with a fixed number of occurrences in the dataset (document frequency, df) are considered. The terms are divided into two groups; in the first group the term occurs in only a single document and terms in the second group occur across multiple documents. This table is taken from [CG95b].

suggest that there are at least some function words which are assumed never to act as content words. On the other hand, performance in information retrieval can be enhanced by calculating stop-word status on the fly, depending on the contexts provided by the query and the document [WS92, YW96]. Hence, from the information retrieval perspective there is evidence that even very frequent function words may contribute to content in significant ways, and some words, which are not usually associated with stop-word lists may behave like function words in the context of a particular query. Note that the terms stop-word and function word are not used interchangeably here, but we built on the verifiable assumption that stop-word lists and very frequent function words share a considerable overlap.

The same work uses distributional characteristics to determine whether or not a word is content bearing (or topical) or not [YW96]. This echoes the approaches taken by Katz [Kat96] and by Church [CG95b, Chu00], who specifically associate burstiness with status as topical content word. This is nonetheless hard to quantify. Whether a distribution is a bursty one or not depends on several factors, including the length of the text. For instance, in a 50 word text, it may be difficult to decide whether a content word appearing 7 times is distributed burstily or diffusely. Also, in some very specialized circumstances, it will be the case that some topical content bearing words become stop-words because they are "noise" - for instance the words "rugby union" are not useful search terms for a the website of the UK rugby union federation or the word "carbon" will not be useful in searching a collection on organic chemistry.

For the purpose of this thesis, it is necessary to adopt some convention to differentiate between content word and function word. A first attempt at articulating the distinction between content and function word has been made in chapter 1. This thesis also shares the position articulated elsewhere [CG95b, Kat96], that a bursty distribution is an indicator of status as content word in the context of a particular text and document collection (chapter 8). Since function words are

linguistic "glue", it also seems reasonable to accept that they will tend to occur more diffusely through text, though this claim will be later evaluated in chapter 8. It is clear from the above discussion, that the assumption is that some words may on some occasions behave like "noise" words and sometimes as topical content words. The kinds of question we will be interested in, for instance, is whether it is possible to determine from the distributional characteristics of a term, on which occasion it is a topical content term, and on which occasion it is not.

The phenomenon of burstiness in text can be observed also for multi-word terms like 'New York', a collocation [MS99] where several terms together exhibit a bursty characteristic as distinct from that of the individual terms. Studies on term burstiness have focused on both individual terms [CG95b, Chu00] and multi-word terms [NSJ64, Kat96]. But in this thesis, the burstiness patterns of such collocations and multi-word terms are not studied and the focus is on individual terms only.

2.2.2 Document-level term burstiness

As a phenomenon term burstiness stretches beyond the confines of a single text. Consider news articles related to a certain topic or event. Once the event has occurred, terms describing the event continue to occur in the news for the next few days after the event. Once a topic of discussion arises, one might receive several emails related to the particular topic over a span of few days or weeks and then nothing afterwards. In other words, one may observe a burst of terms *across documents* in a short time span. This phenomenon of a term occurring multiple times but across documents is referred to as *document-level term burstiness* or *document-level-burstiness* as described by Katz [Kat96].

Katz [Kat96] divides the phenomenon of term burstiness in text into two areas:

- *document-level-burstiness* denotes multiple occurrence of a content word or phrase in a single

text document, which is contrasted with the fact that most other documents contain no instances of the word or phrase at all.

- *within-document-burstiness* denotes the close proximity of all or some individual instances of a content word or phrase within a document exhibiting multiple occurrence.

The focus in this thesis will be on *within-document-burstiness* of terms, and will often simply be called *term burstiness*. Study of *document-level-burstiness* is beyond the scope of this thesis, but some existing research in that field is discussed in section 3.8.

2.2.3 Concept burstiness

Often different terms are used in text to refer to the same concept. For example, if a topic of discussion in a certain text is about *elections*, a specific reference is made using the term *election*. Again the process can be referred to as *electing* or after the process has ended, someone might be *elected* or to *elect* someone. So the different terms *elections*, *election*, *electing*, *elect* and *elected* are related to the same "concept" of **election** and in such a case their occurrence will not be independent within the document. Also, the author of the article might choose to use various synonymous terms to refer to the same "concept" of **election** by terms like *vote*, *ballot* or *polls*. Such a bursty behaviour of different terms related to a similar concept is referred to as *concept burstiness*. This type of burstiness in text can also be observed for proper names, where a person say *Saddam Hussein* may be referred by various terms or phrases like *Saddam*, *Hussein*, *Mr Hussein*, *president of Iraq* or using pronouns like *he* or *his* (see Appendix A for an example document with this type of burstiness). This characteristic of different terms relating to the same concept was noted by researchers [NSJ64] where they refer to these as clumps. This thesis is aimed at the study of *term burstiness* so studying *concept burstiness* is outside the scope of the thesis.

This chapter introduces the various types of burstiness in text and their relation to term burstiness which is the focus of this thesis. Then, we will investigate how term burstiness is handled in models of term distribution. Chapter 3 looks at various models of term distribution and how these models tackle the issue of term burstiness. Knowing that terms are bursty, in particular content terms, one would be inquisitive about the effect of term burstiness on the entire dataset. The effect of term burstiness on a dataset is verified by formulating a hypothesis that burstiness does not exist and that terms in a document are homogeneously distributed. This homogeneity hypothesis is then defeated through a series of similarity based experiments using the χ^2 statistic (Chapter 5).

Chapter 3

Term distribution modeling

Term burstiness is an inherent characteristic of text as it provides structure and context to a document. Term distribution models are used to encode information about a term in certain applications where accurate encoding of information about term behaviour would be helpful for performance. In this chapter several term distribution models are studied. How accurately these model encode certain characteristics of text, in particular term burstiness, is reviewed. Analysis and identification of gaps and deficiencies in the existing models are discussed. This is followed by a review of some literature that illustrates the use of positional information of terms in various applications.

3.1 Introduction

Term distribution modeling engages different research communities. A primary reason for modeling term distribution is to obtain concise information about a term's characteristic behaviour in a dataset and to make inferences based upon it. For instance, in machine learning, in the context of text classification, the learner algorithm learns the distribution of terms in the training data

associated with the class labels. The testing algorithm then predicts the class label of a newly arrived document based on the learned term distribution [Mit97].

The distribution of a term (or terms) has various features that may be focused on. These may include the presence or absence of a term in a document, the number of occurrences of the term within a document [CG95b, RW94, Kat96], co-occurrences of the term with other terms [MS99], occurrence positions of the term within a document, and re-occurrence pattern within and across documents. Moving onto another area of authorship attribution which often focusses on the distributional characteristics of the various parts-of-speech instead of the original terms [DKLP03, CS03].

The characteristic that will be explored in detail in this thesis is "term burstiness". A common approach to representing text is to treat terms in a document as a "bag-of-words" and drawing up frequency counts to represent term occurrence information, and from there to infer the relative importance of words in a document. This approach does not capture positional and structural coherence information in text. Term burstiness is a potentially useful characteristic of text because there is evidence that burstiness relates to the importance of a term in a document that can be exploited in several applications. Importantly, simple frequency count based approaches cannot capture term burstiness because they represent text in a way that makes an independence assumption, that a word is equally likely to occur in any position in a text. This chapter starts by discussing term distribution modeling across various applications and the term independence assumption inherent in most models. Then existing term distribution models are reviewed, each of which aims at capturing certain characteristics of text. These models are discussed and applications that rely on the positional information of terms are reviewed.

3.2 Different approaches to Term distribution modeling

The application area of machine learning motivated usage of models of term distribution in various different research communities. Each community models certain aspects of term distribution that are relevant for applications in that area.

Term distribution modeling might look at the distribution of different terms in a single document or might look at the distribution of a particular term across different documents. In this chapter the distribution of all (different) terms in a single document will be referred to as **Document Term Models**, whereas the distribution of a particular (single) term across different documents will be referred to as **Individual Term Models**. Such a differentiation of term distribution modeling has been discussed by Katz [Kat96], who states that *inter-document distribution* is the distribution of words along an extent of documents within the dataset (*Document Term Models*) and *within-document distribution* is the distribution of words along the text extent of the individual documents (*Individual Term Models*).

The Information Retrieval (IR) community is interested in modeling a term's distribution to differentiate between relevant and irrelevant documents in a retrieval task based on certain keywords [Rij79, BYRN99]. Information retrieval applications are primarily concerned with inter-document dependencies based on Document Term Models, focusing on the differences between different documents as relating to the differences in the documents' relevance to a query. Such models are usually based upon the presence or absence of particular keywords in the document collections.

The task in Machine Learning (ML), especially in text classification is to learn about a term's distribution reflecting a certain topic based on some labeled documents, and later use this knowledge to classify un-labeled documents [Mit97, Lew98, Joa98, MN98, YL99, Seb02]. Here a term's characteristic behaviour across many documents on a particular topic is studied to evaluate fea-

tures that may be used to identify the specific topic. These models too are based on comparing the Document Term Models for the documents in each category. Usually the occurrence of a term in a document along with its frequency are of particular interest. Terms with very different characteristics across various topics are also of interest. Highly topical content words that occur in only a few documents tend to be a better discriminative factor for topics compared to commonly used (function) words. The latter (words like *the*, *of*, *and*), are assumed to occur evenly across different topics and are not of much interest to this community. Machine Learning research is closely related to research in topic tracking or detection, which aims to identify the topics of documents that are arriving in a continuous stream. They too learn about features in the term's distribution that are useful for distinguishing topics [Gil90, Low99a].

Research in the fields of genre identification, authorship attribution and stylistic analysis focuses on the study of term distribution and various other features of text to differentiate between various genres of document or between text written by different authors [Aar99, AKFS03]. Here the focus is on various stylistic features, such as the distribution of function words, word-length and sentence length. In these fields, function words or parts-of-speech of words might prove helpful for differentiating across genres, author and styles, and Document Term Models are used but draw on a different set of features than those used for topic based text classification. Verbs, adverbs and adjectives play a more important role in the field of genre identification, authorship attribution and stylistic analysis as compared to nouns.

In speech recognition, the aim is to predict a spoken word given the previous experience of the system. Information about the ordering of words, is of particular interest [JM00]. Speech recognition applications focus on word distributions within individual documents (based on Individual Term Models) since the probability distribution of words in the next word position, is their primary concern. In such studies, collections of two or more consecutive words are of particular interest

rather than just a single word at a time.

In the Computational Linguistics community, researchers aim to understand better the structure of text and documents [CG95b, RW94, Kat96]. They also try and understand the implications of the models of term distribution, and the effect these models will have on applications based upon them. This field looks at Individual Term Models for the terms of interest based on within-document term distributions.

Though the models developed are aimed at diverse applications, several of them have features in common. Some are based on statistical distributions [CG95b, RW94, MS99], in which the aim is to emulate the process of text generation by some external phenomenon or some mathematical equation, whose properties are well understood. Document Term Models focus on per document term characteristics and how these helps in differentiating between documents whereas Individual Term Models focus on term relations and term co-occurrences. Term burstiness, the relationship between two or more occurrences of the same term also falls in the within-document term distribution. The term burstiness model proposed in this thesis (Chapter 6) falls into the class of Individual Term Models.

3.3 Statistical Modeling and Independence Assumption

Approaches to modeling term distribution discussed in this chapter are based on statistical modeling principles. In this tradition, because of the immense complexity of language, varying simplifying assumptions are made about language and the underlying text. In making these assumptions, many characteristics of text are lost in the process.

For instance, the independence assumption is often inherent in the "bag-of-words" representation of text that build on frequency of occurrences of the terms. This assumption leads to the loss

of document structure along with losing information about the term's burstiness. When used, for instance in IR, the model assumes that the overall frequency of a term in an entire document is the only useful measure that associates a term with a document. It does not capture whether the term occurred in the beginning, middle or end of the document. Neither does it capture whether the term occurs many times in close succession or whether it occurs homogeneously, at a more steady rate throughout the document.

This chapter discusses different term distribution models found in the literature. All of them, first make the simplifying independence assumption while collecting information from the documents, and then try and develop sophisticated models to predict actual term distribution characteristics based on the limited information. The following sections discuss models in the two broad groups: first, **Individual Term Models** that deal with the distribution of a particular term across many documents in the dataset and, second, **Document Term Models** that simultaneously model the occurrences of different terms within a document. Within these sections, models that capture the phenomenon of burstiness are also discussed.

3.4 Simple Individual Term Models

3.4.1 Binomial Model

One of the most widely used and well understood statistical distribution is the *Binomial* distribution. Hence it is not surprising that it has been used in the modeling of term frequency distribution [MW84, CG95b]. The Binomial model assigns the probability for k[1] occurrences of a particular

[1]Here k denotes the random variable. In statistical literature x, y or z are commonly used to refer to random variables. In the NLP community often k is used to refer to the number of number of occurrences of a term and that convention will be followed in this chapter.

term in a document by the following equation:

$$Pr_{Bin}(k; p, N) = \begin{pmatrix} N \\ k \end{pmatrix} p^k (1-p)^{N-k}, \tag{3.1}$$

for $k = 0, 1, \ldots, N$, where N is the document length and p is the word rate or probability of success, i.e. the probability that any randomly chosen term in the document is the one of interest.

Church and Gale [CG95b] modeled the number of term occurrences based on the Binomial model and observed poor fit to the observed data. One possible reason for the poor fit of the Binomial model is that it assumes a constant rate of occurrence of the term across different documents, which cannot be true for most terms. The assumption that the number of occurrences of a term in a document is dependent on the document length cannot always be true.

3.4.2 Poisson Model

The Poisson model is often used for modeling occurrences of rare events. It is used extensively in the modeling of the number of defective items in a manufacturing process, or the number of typing or printing mistakes per page. The number of occurrences of a particular content word in a document is also a rare event and is similar to the above phenomena, so the Poisson model is often used in the modeling of term occurrences [MS99, CG95b]. If the number of occurrences of a term in a document is denoted by k, then the model assumes:

$$Pr_{Pois}(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \tag{3.2}$$

for $k = 0, 1, 2, \ldots$ and $\lambda > 0$ is the average number of occurrences of the term per document. A feature of the Poisson distribution is that the mean and the variance of this distribution are both

equal to λ, which reduces the number of parameters to be estimated. This very fact also reduces the flexibility of the model.

The Poisson distribution is a limit to the Binomial distribution discussed in Section 3.4.1. The Binomial distribution tends to the Poison distribution if one lets $N \rightarrow \infty$ and $p \rightarrow 0$ in such a way that $N.p$ is constant at some value $\lambda > 0$. If the Poisson model is appropriate for the data being modeled, then the following conditions [MS99] should hold:

- the probability for one occurrence of the term in a (short) piece of text is proportional to the length of the text.
- the probability for more than one occurrence of a term in a short piece of text is negligible.
- the occurrence of events in non-overlapping intervals of text are independent.

Estimates based on the Poisson model have been found to be good for non-content, non-informative terms, but not for the more informative content terms [MS99]. This may be because the Poisson model assumes independence between term occurrences. This assumption holds loosely for words acting as non-content words and function words in a text, but not for content words.

Church and Gale [CG95b] experimented by modeling the number of term occurrences using Poisson distribution. The experiments revealed that the Poisson model did not fit the data well. They suspected that both the Binomial and the Poisson models "systematically underestimated the variance of term occurrences as they assumed no dependencies on hidden variables such as genre, author, topic, etc, and these factors always tend to inflate the variance". The variance of term occurrences across documents was observed to be larger than the predicted value by the Poisson model. This indicated an over-dispersed characteristic of term occurrences across documents.

3.5 Individual Term Models for burstiness

3.5.1 Discrete Poisson Mixture

Bookstein and Swanson [BS74] proposed using a multiple mixture of Poisson distributions to model the number of occurrences of a term in a document. This was used in the context of information retrieval to account for the fact that certain terms have different distributions in relevant documents as compared to the irrelevant ones. The model has the following general form:

$$Pr_{DiscMixPois}(k; \alpha_i, \lambda_i) = \sum_{i=1}^{n} \alpha_i \cdot e^{-\lambda_i} \frac{{\lambda_i}^k}{k!}, \tag{3.3}$$

where, n is the number of mixture components, such that $\sum_{i=1}^{n} \alpha_i = 1$.

A direct adaptation of the above model is the **Two-Poisson** model ($n = 2$) [Har75a, Har75b, MS99, RW94]. The model assumes that there are two classes of documents associated with a term, one class with a low average number of occurrences (the non-privileged class) and the one with a high average number of occurrences (the privileged class):

$$Pr_{2Pois}(k; \alpha, \lambda_1, \lambda_2) = \alpha \cdot e^{-\lambda_1} \frac{{\lambda_1}^k}{k!} + (1 - \alpha) \cdot e^{-\lambda_2} \frac{{\lambda_2}^k}{k!}, \tag{3.4}$$

where α and $(1-\alpha)$ denote the probabilities of a document in each of these classes. The Two-Poisson model postulates that a content word plays two different roles in documents. In the non-privileged class, its occurrence is accidental or by chance and does not relate to the main content of the document. In the privileged class, the word relates to the main content of the documents. This model tackles the issue of term burstiness by differentiating the documents in which the term occurs into two classes; first the privileged class in which the term has topical behavior identified by multiple bursty occurrences and in the second class the term does not have a topical behavior

identified by non-bursty nature of the term. Often this model under-estimates the probability that a term will occur exactly twice in a document; the model incorrectly predicts that the number of documents with two occurrences of a term are less likely than the number of documents with three or four occurrences. This characteristic is considered a disadvantage of the model [MS99, CG95b].

3.5.2 Continuous Poisson Mixture

The Continuous Poisson Mixture model captures term burstiness by allowing the rate of occurrence of a word to vary across documents and thus identifying multiple occurrences of a term associated with its bursty nature. Church and Gale [CG95b] observed that the rate of occurrences of a term in a collection does not obey the "term independence" assumption introduced by the "bag-of-words" model. They also observed that Binomial and Poisson distributions underestimated the variance of term occurrences because these distributions do not assume any dependencies on hidden variables such as genre, author and topic which lead to inflation of the variance. Their study also observed the drawbacks of the two-Poisson model based on the discrete mixture of Poisson distributions.

To tackle these deficiencies, Church and Gale [CG95b] proposed the generalized form of the **Continuous Poisson Mixture**. This can be thought of as a generalization of the Discrete Poisson Mixtures where the mixing parameter, α, is replaced with an arbitrary density function, ϕ. The general form of the Continuous Poisson Mixture is:

$$Pr_{ContPoisMix}(k;\phi,\lambda) = \int_0^\infty \phi(\lambda) \cdot e^{-\lambda}\frac{\lambda^k}{k!}d\lambda, \tag{3.5}$$

for $k = 0, 1, \ldots$ and ϕ is an arbitrary density function that is intended to capture the variation across documents, where,

$$\int_0^\infty \phi(\lambda)d\lambda = 1$$

Some special cases can be derived based on the choice of the function ϕ. The **Two Poisson** model (Equation 3.4) can be derived by choosing

$$\phi_{2Pois}(\lambda) = \alpha \cdot \delta(\lambda - \lambda_1) + (1 - \alpha) \cdot \delta(\lambda - \lambda_2),$$

where $\delta(x)$ is Dirac's delta function acting as an indicator function; the density (or value) is ∞ when $x = 0$, and otherwise, 0.

The **Negative Binomial** model which is an infinite mixture of Poisson distributions can also be derived from Equation 3.5. The Negative Binomial model allows the word rate to vary across documents. The Negative Binomial distribution has been successfully used for modeling term occurrences across documents. [MW84, CG95b, Jan03].

The **K Mixture** Model described in section 3.5.3 can also be derived from the Continuous Poisson Mixture:

$$\begin{aligned} \phi_{KMix}(\lambda) &= (1 - \alpha) \cdot \delta(\lambda) + \frac{\alpha}{\beta} \cdot e^{-\frac{\lambda}{\beta}} \\ Pr_{Kmix}(k) &= (1 - \alpha) \cdot \delta_{k,0} + \frac{\alpha}{\beta + 1} \cdot \left(\frac{\beta}{\beta + 1}\right)^k \end{aligned}$$

where $\delta(x)$ is Dirac's delta function; the density function is ∞ when $x = 0$, and otherwise, 0. $\delta_{x,y}$ is 1 when $x = y$ and is 0 otherwise.

The Continuous Poisson Mixture is an improvement over the Binomial and Poisson models and provides a better fit to data with varying word rates across documents [CG95b]. This is because the Continuous Poisson Mixture allows word rates to vary over documents and even accounts for variable document length. Church and Gale suggested applications of their model in the areas of Information Retrieval, Authorship Identification and Word-Sense Disambiguation.

Though this model was a significant improvement over the other existing models. It captures word rate variability across documents, but does not model the degree of burstiness of terms, for example. Multiple occurrences of a term are associated with term burstiness. But if a certain term occurs in two separate bursts in a long document, such a characteristic of the term cannot be captured based on this model.

3.5.3 K Mixture Model

Katz [Kat96] suggested that most approaches to modeling term distribution in text are motivated by statistical principles or stochastic processes. He wished to model term distribution based upon linguistic phenomenon, as he argued that formation of text is not based upon statistical principles, but rather on linguistic intuitions. Katz discusses some drawbacks in the assumptions made in the modeling of term distributions and how the issue of burstiness is ignored. He goes on to propose a model to account for these shortcomings.

Katz [Kat96] stated that a content term's occurrence in a document may be classified as *topical* or *non-topical*. A *non-topical* occurrence describes a content word when it occurs only a few times in a document, but does not determine the main content of the document. Katz considers one occurrence of a term related to non-topical behaviour of the term and more than one occurrences related to topical behaviour. In contrast, when a concept named by a content word describes and determines the topic of the document, the occurrence is termed *topical* and can be identified by bursty multiple occurrences of the specified content word. Katz observed that the length of a burst or the number of occurrences of a term in a burst is not related to the document length. Also the document length is claimed to be related to the number of different content words in it, rather than multiple occurrences of a particular content word. Katz argues that once a term occurs topically in a text, the number of occurrences of the content word in a text is only related to the

intensity with which the concept word is treated in the document. Katz also observed the fact that content words with multiple occurrences are not very likely to occur in all documents. Also, the probability of multiple occurrences of a content word is much higher than would be expected under the independence assumption, i.e. the probability of repeat occurrences is higher than what would be expected using relative frequency alone.

Katz proposed the linguistically motivated **K Mixture Model** to model term distribution in text. The model is based on the following parameters that should reflect the term's characteristics in the document collection:

- the probability that a term occurs in a document at all (relative frequency), denoted by α.
- the probability that it will occur a second time in a document given that it has occurred once, i.e. the probability of the term being used topically given that it has already occurred once before, denoted by γ
- the probability that it will occur a further time, given that it has already occurred k times (where k $>$ 1), i.e. the average intensity of topical usage of the term (B_k), denoted by B.

These three parameters provide an elegant way to account for term occurrences. $Pr(1|0) = \alpha$ is the probability of at least one occurrence of the term, $Pr(2|1) = \gamma$ is the conditional probability of entering a burst, and $Pr(k+1|k)$ for $k \geq 2$ are the conditional probabilities of repeat occurrences within a document-level burst. Katz observed that α and γ were negatively correlated in the terms studied by him, which indicates that occurrence of a term in a document does not provide any evidence of the term having a bursty characteristic. However, as expected, there was a positive correlation between γ and B, indicating that if a term is used twice in a document, it is likely to be used again in the same document. It was observed that the probability of having at least one occurrence, α, relates to the document length, while the other two parameters did not depend on

document length. It was also observed that the conventional definition of relative frequency based on the independence assumption was only valid in cases where the term did not have a bursty behaviour.

Based on these parameters, Katz proposed the following model for k occurrences of a content term in a document as:

$$\begin{aligned} Pr_{K-Mix3}(k;\alpha,\gamma,B) &= (1-\alpha)\cdot\delta_{k,0}+\alpha\cdot(1-\gamma)\cdot\delta_{k,1} \\ &\quad +\frac{\alpha\gamma}{B-1}\cdot\left(1-\frac{1}{B-1}\right)^{k-2}\cdot(1-\delta_{k,0}-\delta_{k,1}), \end{aligned}$$

for $k = 0, 1, 2, \ldots$ and $\delta_{i,j} = 1$ when $i = j$ and 0 otherwise.

Katz further defines P^* as the average probability of repetition of a term within a burst as:

$$P^* = 1 - \frac{1}{B-1} \tag{3.6}$$

At times it was observed that the data were quite sparse which led to difficulties in the estimation of three parameters. Hence Katz suggested a simplified model based on more assumptions by considering the condition for the equality of the probability of entering a burst, γ, with the average probability of repeats in the burst, P^*, i.e. $\gamma = P^*$. This leads to simplification of the model and reduction in the number of independent parameters. The simplified model is stated as:

$$Pr_{K-Mix2}(k;\alpha,B) = (1-\alpha)\cdot\delta_{k,0}+\frac{\alpha}{B-1}\cdot\left(\frac{1}{B-1}\right)^{k-1}\cdot(1-\delta_{k,0}), \tag{3.7}$$

for $k = 0, 1, 2, \ldots$ and $\delta_{i,j} = 1$ when $i = j$ and 0 otherwise. But the suggested simplification of the model was not very accurate, as it was observed by Katz that γ was smaller than P^* in most cases, with an average value of 1.52 for the ratio P^*/γ. They were nevertheless of the same order

of magnitude and difficult to distinguish for sparse data.

The original 3-parameter model provided a good fit to the multi-word terms chosen from technical documents. The method gave accurate estimation for 0 or 1 occurrence of the term, and statistical tests were used to validate the fit for multiple repeat occurrences.

Some limitations of the model are:

- it can handle only content terms, and is not suitable for high frequency function words or medium frequency terms; and
- the rate of re-occurrence of the term or the gaps between the terms within a burst cannot be used to provide information.

3.5.4 Other modeling approaches

Term burstiness is an inherent property of text that is not captured by "bag-of-words" representations. Applications have used heuristic based methods to account for this phenomenon. A concept of cache memory is used in the speech recognition task [KM90, JMRS91, RH92]. These techniques remember the terms recently encountered in the document and assign additional probability for further occurrences of the recently observed terms. The application remembers the recently observed terms, as they are more likely to re-occur than unseen terms. Thus it exploits the bursty nature of terms. A queuing theory modeling approach was used by Munro [Mun03] to capture the phenomenon of burstiness.

Kwok [Kwo96] proposed a measure of burstiness as a binary value reflecting the magnitude of average-term frequency of the term in the corpus and applied it in information retrieval. This measure takes the value 1 (bursty term) if the average-term frequency value is large and 0 otherwise. However, the measure is too naive and incomplete to account satisfactorily for term burstiness. Francis and Kucera [FK82] suggested adjusting the term's frequency to reflect the effect of term

burstiness. The term's frequency was adjusted based on the overall distribution of the term across many different documents and the adjusted frequency used instead of the actual term frequency value.

3.6 Simple Document Term Models

The term distribution models discussed up to this point capture the number of occurrences of a particular term in a document. Hence these models deal with one term at a time. Such models are mostly used in applications that require analysis about individual terms, as with word-sense disambiguation, for example.

There are other application areas that look at a term as a part of an entire document. Some of these application areas include text classification, authorship attribution and information retrieval. In text classification the application judges the category of a document based on all the terms within the document. Even in authorship attribution the author of a text should ideally be determined by taking into consideration the entire document and all the terms within it. These studies are based on *Document Term Models* of the text where different terms in a document are considered simultaneously.

The following sections discuss some Document Term Models.

3.6.1 Vector Space Model

The Vector Space Model is very popular in information retrieval [SL68, Sal71]. Unlike the other models described in this chapter, the Vector Space model is not a mathematical model capturing some statistical distribution. Rather, it represents and visualize a document in a vector space based on the terms in the document.

According to this model, there exists an N-dimensional vector space, where N is the number

of available distinct terms. The direction of each vector is determined by the presence or absence of a particular term, and the magnitude is proportional to the term's frequency in the document. This model allows an elegant way of determining a term's similarity with a document based on the sum of element-wise dot-product of the vector and hence it is used for *keyword retrieval.* This representation also allows documents with similar occurrences of terms to appear closer to each other. Similar dot-products between two such vectors representing different documents may be used for measuring the similarity between two documents and hence this representation is often used for *document similarity* and even *document clustering* based on the similarity measure [BYRN99].

There is a terminology associated with this document representation. The first is the process of **document indexing** in which the key terms in the document are extracted and indexed for future use. Usually function words or stop words are removed from the list as they do not bear much of the document's content. Then comes the process of **term weighting**, in which each indexed term is assigned a weight that is proportional to the term's relevance to the document. This is determined by some combination of the term's occurrence in the document (Term Frequency, TF) [Luh58] and its spread across the whole document collection (Inverse Document Frequency, IDF) [SJ72]. This combination produces the immensely popular **TF-IDF** measure [SB88, Aiz03] for term indexing and relevance. A term that occurs in many documents is not very useful in identifying a particular document, as compared to another term that occurs in only a few documents. This inverse concept of the term's spread is captured by the IDF component.

Various forms and combinations of the TF-IDF measure have been explored. The measure proposes a weight for a term, t, in a particular document, d, based on the term's frequency in the document, tf, and the number of documents in which the term occurs, document frequency, df.

$$weight(t, d) = f(tf) * g(df) \tag{3.8}$$

where, $f(.)$ and $g(.)$ are functions. Table 3.1 lists some of the commonly used functional forms for $f(tf)$ and $g(df)$. In this table N denotes the number of documents in the collection. Document similarity based on this representation suffers from the problem of document length, as longer documents have many terms in them and hence exhibit more similarity than shorter documents. This effect of document length is nullified by normalizing the term weights with respect to the document length to work with a unit length document.

$f(tf)$	$g(df)$
tf	N/df
1 + log(tf)	log (N/df)

Table 3.1: Various functional forms of the TF and IDF functions.

3.6.2 Multivariate Binomial model

The Binomial Model is a direct mathematical formulation that carries the independence assumption inherent in "bag-of-words" representations of text. Let us suppose there are T possible terms represented in terms of a vector as $\widetilde{t} =< t_1, \ldots, t_T >$. $\widetilde{x}$ is a binary vector representing the terms in a particular document. The length of $\widetilde{x}$ is equal to the length of $\widetilde{t}$. $x_i = 1$ if term t_i is present in the document, otherwise $x_i = 0$. It is assumed that the occurrences of terms in a document are independent, so that the occurrence of one term in a document does not affect the probability of other terms occurring in that document. The Binomial model provides a likelihood for a document as:

$$Pr(\widetilde{X}) = \prod_{i=1}^{T} \theta_i^{x_i} \cdot (1 - \theta_i)^{1-x_i} \tag{3.9}$$

where, θ_i is the proportion of documents in the entire collection where the term t_i is present. If the term t_i is present in the document the $\theta_i^{x_i}$ component contributes for the term's presence (positive evidence) in the document, otherwise the $(1 - \theta_i)^{1-x_i}$ term contributes towards the term's absence

in the document (negative evidence). Just the presence or absence of a term in a document is considered and the number of occurrences of a term in a document is not considered. This is possibly the only model that accounts for a term's absence in a document [Sch04]. Other models only provide information about terms present within the document.

This model is often used for the purpose of text classification and information retrieval. In text classification [Mar61, MN98, Seb02] the documents in a certain category or topic are used for deriving the parameter estimates θ_i for that category. Then for a new document the likelihood for each category is calculated. And the document is classified into the category with the maximum likelihood. The Binomial model has also been used for relevance feedback in information retrieval [Har92b, SB97, RJ76]. In relevance feedback, a user query is given to a search engine, which produces an initial ranking of its document collection by some means. The user examines the initial top ranked documents and gives feedback to the system as to which are relevant to their interest and which are not. The search engine then applies the binomial model to re-rank the documents.

3.6.3 Multinomial model

The Multinomial Model is possibly the most popular model used for text classification, topic detection and various other applications. This model takes into account a term's frequency in a document and carries the independence assumption inherent in "bag-of-words" representations of text [Mit97]. A document of length L is generated by tossing a T-sided dice for each term position in the document, where T is the number of possible terms. According to this model, the probability of a particular term occurs in one word position is the same as its probability of occurring in any other word position. Multiple occurrences of a particular term are independent of each other. Let $\widetilde{t} =< t_1, \ldots, t_T >$ denote the terms in the collection. Also let $\widetilde{n} =< n_1, \ldots, n_T >$ denote the

frequencies of the corresponding term in that document, such that $n_1 + \ldots + n_T = L$, the document length. If a term does not occur in a document, $n_i = 0$ for that term, so it does not add to the document length, L. Hence the likelihood for that document based on the frequencies $\widetilde{n}$ is denoted as:

$$Pr(\widetilde{n}) = \frac{L!}{\prod_{i=1}^{T} n_i!} \prod_{i=1}^{T} \theta_i^{n_i}$$

where, θ_i denotes the probability of seeing term t_i in the document and is estimated by the proportion of its occurrences across the entire collection, i.e. total number of occurrences of term t_i across all documents divided by the total document length for all documents.

The Multinomial model is often used for the purpose of text classification. A text classification method based on this model is the Naive Bayes which is based upon *Bayes Theorem* [Lew98, MN98, CDAR97]. The Naive Bayes classifier applies to learning and classification tasks where each document is described by a conjunction of many attribute values (term frequencies) and one would assign the document to an element (category) in a finite set V. A set of training documents with the term frequency values are provided for training. When a new document with attribute values $(t_1, t_2, \ldots, t_T)$ is provided, it is assigned to target category v which has the maximum probability of having those attributes.

$$\begin{aligned}
v &\equiv \arg\max_{v_j \in V} \Pr(v_j | t_1, t_2, \ldots, t_T) \\
&= \arg\max_{v_j \in V} \frac{\Pr(t_1, t_2, \ldots t_T | v_j) \Pr(v_j)}{\Pr(t_1, t_2, \ldots, t_T)} && \text{[Using Bayes rule]} \\
&= \arg\max_{v_j \in V} \Pr(t_1, t_2, \ldots, t_T | v_j) \Pr(v_j) && \text{[common denominator]} \\
&\approx \arg\max_{v_j \in V} \Pr(v_j) \prod_i \Pr(t_i | v_j) && \text{[independence assumption]} \\
&= \arg\max_{v_j \in V} \Pr(v_j) \prod_{i=1}^{T} \theta_i^{n_i} && (3.10)
\end{aligned}$$

Various other algorithms have also been based on the term frequency information such as k-Nearest Neighbors (kNN), decision trees, logistic regression, neural networks, Support vector Machines(SVM) [YL99, Seb02]. Support vector Machines(SVMs) are the most successful machine learning techniques to date for the purpose of text classification [Joa98].

The range of learning algorithms that have been described all assume independence between term occurrences, not just between different words, but also between multiple occurrences of the same word, even though this is known not to be the case in real text [Chu95]. The likelihood of the Multinomial model is the product of the probabilities of the terms in the document, so the likelihood is dependent on the document length. This is a disadvantage of the Multinomial model since it is biased for document length and hence it is very poor for the purpose of ranking documents (of different lengths) in a search engine. This is known as the document length bias of the Multinomial model. This problem does not occur in text classification as a single document is evaluated across many categories or topics. This is also a possible reason for the huge popularity of this model in text classification but not in text based information retrieval. Researchers have also demonstrated that the Multinomial model performs well for the Naive Bayes text classifier [DP96, DP97].

3.7 Document Term Models for Burstiness

The term distribution models discussed in the previous section were elementary, so these models failed to account for term burstiness in text. The Multinomial model is the most popular among the models discussed in the last section. In this section various models are discussed. Most of them are improvements, generalizations or variations of the Multinomial model to tackle term burstiness.

3.7.1 Variations of the Multinomial Model

The Multinomial model assumption for text is a little simplistic as multiple occurrences of a term are not independent of each other. Heuristic solutions to overcome the deficiencies of the model have been suggested in the context of the Naive Bayes text classifier [RSTK03]. Naive Bayes is seen as a linear classifier and this work aimed at improving the decision boundary weights.

The probability of multiple occurrences of a term in a document is smaller than that predicted under the independence assumption [Chu00, RSTK03]. This is due to the effect of *term burstiness* in text [CG95b, Kat96, SGDR05]. [RSTK03] suggested a heuristic transformation of term frequencies to account for term burstiness. According to this heuristic, the term frequency, n_i, for term i is replaced by $log(n_i + 1)$ in the Multinomial model equation 3.10. Based on this transformation, the frequencies are modeled in a more realistic way than with the Multinomial, while still retaining the form and simplicity of the Multinomial model.

Further heuristics for improving the model have also been suggested. Inverse Document Frequency (IDF) is a popular method in information retrieval (discussed in section 3.6.1) and it has been suggested that it reduces the effect of common terms on the model. The Multinomial model is also dependent on document length, which introduces bias between documents of varying length. Document length normalization was suggested to tackle this issue in the model. A Naive Bayes classifier based on these suggested heuristics provided better classification accuracy as compared to the one without these heuristics[RSTK03].

3.7.2 Exponential family

Modeling approaches provide an elegant way to handle text, but they make some underlying assumptions about the nature of text. The limitations that ensue from these assumptions are rarely discussed. The Multinomial model is a popular generative model for many applications, including

that of text classification. It has been observed that the Multinomial does not model text quite accurately [TK03] because its ability to represent a range of features associated with text is limited by the independence assumption inherent in the model. There is a trade-off between the simplicity of the model, and its ability to capture phenomena that involve the dependence between term occurrences. This also raises the issue of making the appropriate choice of distributions for modeling and representing text, and then, which model to fit to it.

The Naive Bayesian framework is an example of a model used for capturing term distributions, that makes the independence assumption. Term occurrences are assumed to be independent of each other and interactions between them are not captured. [TK03] argued that the term occurrences in a document are not completely independent given document length. This is because once the document length is fixed, a large number of occurrences of one term leaves fewer spaces for other terms to occur. The paper also states that a single distribution is not sufficient to model text across different collections. Instead they propose an approach that draws on the exponential family of distributions. For a term t_i which occurs n_i in the document with document length l, the general form of the model is:

$$Pr(n_i|\phi_i) = f(n_i).g(\phi_i).e^{\phi_i h(n_i)} \tag{3.11}$$

where, $f(.)$ and $g(.)$ are functions and $g(\phi_i)$ is a normalizing constant (to ensure the function integrates to 1) equal to the inverse of $\int f(n_i).e^{\phi_i h(n_i)} dn_i$. The model for a particular data collection is specified by the choice of the functions f and h. A range of distributions fall in this family such as the Poisson, Binomial, Uniform and Normal distributions. For example, for a Binomial model, $f(n_i) = \begin{pmatrix} l \\ n_i \end{pmatrix}$ and $h(n_i) = n_i$. For the Poisson model $f(n_i) = 1/(n_i!)$ and $h(n_i) = n_i$. In the learning phase the values of the ϕ_i parameters are estimated so as to fit the model. This is done by maximum likelihood by maximizing over various choices of f, h, ϕ_i. The learned model

provided a better fit to the original text by considering the multiple occurrences of terms. The model was applied in the task of keyword retrieval and produced better results than those based on the Multinomial model [TK03].

3.7.3 Dirichlet Compound Multinomial model

The previous sections noted some of the limitations of the Multinomial model that follow from its inherent assumption that terms occur independently. Improvements to the Multinomial model have been suggested, some based on heuristics and others based on empirically searching for an exponential family model that fits the document collection. In this section the *Dirichlet Compound Multinomial (DCM) model* [Min03] is discussed as an alternative to the Multinomial model [MKE05]. The DCM model has an additional degree of freedom which allows it to capture burstiness.

In one experiment, the Multinomial model was used to model the occurrences for common terms, average terms and rare terms [MKE05]. The Multinomial model performed reasonably well for commonly occurring terms, but not for average and rarely occurring terms. This suggests that the Multinomial model captures the occurrences of common terms which will include many very frequent function words whose distributions are not very bursty in nature, but is less suited to capturing the inherent burstiness in the behaviour of average and rare terms. This is in line with what one would expect to be the limitations of the Multinomial model that ensue from its adherence to the independence assumption. In any case, this study again showed the limitations of the Multinomial model for modeling different types of term behaviours.

Madsen et. al. suggested the DCM as a generative model for the documents [MKE05]. The documents are not generated directly based on it, instead the Dirichlet distribution is used to generate a Multinomial distribution; and this Multinomial is then used to generate the document. Looking at it another way, the Dirichlet distribution is used to add prior knowledge to the parameters θ_i of

the Multinomial model (equation 3.10). This allows the parameter to have more degrees of freedom and the parameter value can vary over documents. The Dirichlet distribution is defined as:

$$Pr(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{T} \alpha_i\right)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \cdot \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} \tag{3.12}$$

where, $\widetilde{\alpha}$ are the parameters of the Dirichlet distribution. Based on the Dirichlet distribution, the DCM can be represented as:

$$Pr(\widetilde{n}) = \frac{(\sum_{i=1}^{T} n_i)!}{\prod_{i=1}^{T} n_i!} \cdot \frac{\Gamma\left(\sum_{i=1}^{T} \alpha_i\right)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \cdot \prod_{i=1}^{T} \frac{\Gamma(n_i + \alpha_i)}{\alpha_i} \tag{3.13}$$

The DCM model provided a better fit for term occurrences for average and rare terms in the collection. This was because the Dirichlet has an extra degree of freedom to account for a word's burstiness. (Though both the Dirichlet and the Multinomial distributions have the same number of parameters, the Multinomial parameters are constrained to sum to one, unlike the DCM parameters, so the DCM has one extra degree of freedom.) Smaller α_i parameters indicate more bursty behaviour of the word involved. This form of the model has been used for the purpose of text classification [MKE05].

The DCM model can be used to measure the similarity between two documents [Elk05]. It can account mainly for two aspects in a similarity measure that are presently not captured in representations of text that make the term independence assumption. These are (a) (related to burstiness phenomena), repeated appearances of one word in the same document are of decreasing informativeness and (b) words that appear across a large number of different documents in a collection are less informative. The DCM model contains components similar to the log-term-frequency and inverse-document-frequency components of the TF-IDF measure (Section 3.6.1).

3.7.4 Other techniques

In the above sections, models were discussed that attempt to account for term burstiness, and some limitations were identified that can be traced back to some models' fundamental assumption that terms occur independently. Some other studies have attempted to relax the stringent independence assumptions of the Multivariate Binomial model and the Multinomial model. These models also aim at tackling other assumptions based on term independence.

A characteristic of a content term is that it typically occurs in only a few documents and does not occur in the rest of the document collection. So there is a very big probability mass on the 0 occurrence of a term. But according to the independence assumption every word is equally likely, and so does not assign "extra" probability mass for 0 occurrence of a term. Jansche [Jan03] proposed **Zero Inflated distribution** [JK69] models of frequency counts to tackle this issue. Jansche studied a range of terms and observed that extra bias on 0 occurrences is more prominent among terms bearing some content information than in terms behaving as function words. He proposed modeling term frequency using two-component mixtures, where one component is a degenerate distribution (always has the same value) whose entire probability mass is assigned to the outcome 0, and the other component is any standard distribution.

$$Pr_{ZI\Phi}(k; z, \widetilde{\theta}) = z(k \equiv 0) + (1 - z) \cdot \Phi(k, \widetilde{\theta}) \tag{3.14}$$

where, $k \equiv 0$ is 1 if $k = 0$ and 0 otherwise and $\Phi(k, \widetilde{\theta})$ is any standard distribution with the parameter set $\widetilde{\theta}$. It reflects the view that, whether a given word appears at all in a document is one thing; and how many times it appears, if it does, is another thing. Jansche obtained better results using the Zero Inflated Binomial (ZIBinom) model in comparison to the Multivariate Binomial model for the task of automatically classifying news articles.

The binomial distribution has often been used to model term occurrence counts (section 3.4.1). But the binomial distribution model assumes a constant rate of occurrence of a term across all documents. This is not often true, as the occurrence of a term is usually dependent on its role in the document and the collection and hence the rate of occurrence varies across documents. The **Beta-Binomial model** [Gil90] was proposed as a means of accounting for the variation in the rate of term occurrences across documents. The rate of occurrence of a term p for term i in equation 3.1 is allowed to vary across documents, even when the documents arise from the same source. The word generation probability p_i is characterized by a beta probability density function with parameters α_i and β_i.

$$\begin{aligned} Pr_{BetaBinom}(k|N,\alpha_i,\beta_i) &= \int_0^1 Pr_{Binom}(k|p,N) \cdot Pr_{Beta}(p|\alpha_i,\beta_i)dp \\ &= \left(\begin{array}{c} N \\ k \end{array} \right) \frac{(\alpha_i)_k \cdot (\beta_i)_{N-k}}{(\alpha_i+\beta_i)_k} \end{aligned}$$

where $(x)_m$ is the ascending product:

$$(x)_m = \prod_{j=1}^{m-1}(x+j) = \frac{\Gamma(x+m)}{\Gamma(x)}$$

The Beta-Binomial model has been used successfully in information retrieval for the purpose of topic tracking and detection [Low99b, Low99a], and has produced better classification accuracy as compared to the Multinomial model. A similar concept of mixture distribution was used by Burrell [Bur80], who used the Gamma-Poisson distribution to model the process of library loans.

3.8 Models of Document level burstiness

The issue of "burstiness" is one of the central issues of this thesis. Burstiness is the phenomenon of multiple occurrences of a term within a document, with some of the occurrences close together. Now consider the case of streaming newswire that arrives daily or of daily emails coming to a certain person. At times a certain topic is discussed in many news articles or emails over a few days or even months. Hence one can observe a burst of activity with respect to the topic over the time period. This is the other type of *burstiness*, i.e. burstiness at a document level, called *document-level burstiness* or *document-level term burstiness.* Here one is not concerned about the number of times a particular word occurs within each of the documents. This document-level burstiness is not the core discussion of this thesis, but in this section some research on document-level burstiness is discussed to derive similarity and understanding of within-document burstiness. Also in this section, talking about a burst or burstiness will mean bursty behaviour of a certain topic with respect to time.

Kleinberg [Kle02] develops a formal approach for modeling document level "bursts" which draws analogy from queuing theory and bursty network traffic. According to the model, a sequence of message arrival times is based on an exponential distribution. Messages are emitted in a probabilistic manner, so that the gap x in the time between messages i and $i+1$ is distributed according to the exponential density function:

$$f(x) = \alpha e^{-\alpha x} \tag{3.15}$$

where the parameter $\alpha > 0$ refers to the rate of message arrivals. Based on this, a "bursty" model would exhibit periods of lower rates interleaved with periods of higher rate. This is based on a probabilistic automaton A with two states q_0 and q_1 corresponding to the "low" and "high" arrival rates. When A is in state q_0, messages are emitted at a slow rate, with gaps x between consecutive

messages distributed independently according to a density function $f_0(x) = \alpha_0 e^{-\alpha_0 x}$. And when A is in state q_1, messages are emitted at a faster rate, with gaps distributed independently according to $f_1(x) = \alpha_1 e^{-\alpha_1 x}$, where $\alpha_1 > \alpha_0$. Also, between messages, A changes state with probability $p \in (0, 1)$ and remains in the current state with probability $1 - p$.

This two state model is further extended to multiple stages with a cost attached for moving from a state of lower burst to a state of higher burst. The cost function is introduced to prevent identification of a large burst as many smaller bursts. Then the identification of a burst is a process of optimizing the sequence of inter-arrival times to the one with the minimum cost. This would create a hierarchy of many small bursts within a larger burst which helps in understanding the structure of the burst. Klienberg applied this algorithm successfully in identifying bursts in emails. He also applied this algorithm to academic articles published in computer science conferences and journals and identified bursts of topics against time.

Documents were assumed to arrive at a uniform rate in Klienberg's algorithm. But it was argued that when an event occurs, there is an increase in activity with respect to that topic and hence the rate of arrival of documents increases. A model was proposed for extracting bursts in a word without the assumption that the documents arrive at a uniform rate. The model was applied to web-blogs and bulletin board articles [FNSO04]. Klienberg's burst detection algorithm was used to generate maps for the identification of major research topics and trends in the research articles published in *Proceedings of the National Academy of Sciences (PNAS)* [MB04]. Klienberg's algorithm was extended to study the bursty evolution of blogs across time [KNRT03].

3.9 Beyond individual term distribution

The thesis looks at individual term distributions for text based applications. A document is made up of multiple terms, and they interact together to give a meaning to the document. Hence, a combination of terms or term sequences will convey more information than a single term. Markov models are aimed at finding the next term in a sequence given the present term [JM00]. Also, n-gram based methods explore multiple consecutive terms as features in text instead on considering isolated terms [MS99]. These techniques are not discussed in this thesis, as the thesis deals with single term distributions and the study of their re-occurrence patterns.

3.10 Positional information of terms

The models that have been discussed in this chapter mostly have a built-in assumption that terms occur independently (for instance, by adopting a "bag-of-words" representation of text). Hence they cannot encode fine grained structural information about a document or positional information of a term. Other researchers in various fields have also identified the drawbacks of this assumption. The lack of structural information and the independence assumption were shown to be harmful for the purpose of part-of-speech tagging [Fra97]. Also the issue of capturing term burstiness is ignored in most of these models. It will be shown experimentally in Chapter 5 that the independence assumption does not hold for text. Chapter 6 introduces a model of term re-occurrence and term burstiness, which draws on positional information about terms occurring in text.

Previous work has attempted to use and encode positional structural information of terms in various applications. The area of *Text Tiling* aims at automatically detecting sections and subsections within text [Hea94], using positional information of terms. The method was extended for the purpose of subject boundary detection [RSA97] within text based on the burstiness pattern

of terms and their positional information in a document. Lin and Hovy [LH97] use a term's position in text to identify the topic of the document based on Edmundson's *Position Method* [Edm69]. Differences between important and unimportant sentences in a document have been used for improved performance on the text categorization task [KPS04]. Heuristics were applied for encoding the positional information of terms for the purpose of sentiment classification [PLV02] by dividing a document into four quarters and tagging each word with the quarter it occurred in, but such approaches did not yield much of a hike in performance of the application.

Chapter 4

Datasets and their Basic Profiling

4.1 Introduction

This chapter discusses the datasets that have been used in the various experiments throughout the thesis. It describes the seven different datasets in the TIPSTER collection. Other datasets used include pages from "The Open University" intranet and extranet collected and constructed for this research; and a dataset in the Bengali language obtained from the Central Institute of Indian Languages.

A basic profiling of these datasets is carried out to detect any evident discrepancies with respect to sparseness and quality in these datasets. The profiling framework adopted from [GDR01] involves manual sampling of the dataset to identify any evident quality problems, followed by obtaining basic statistics for each of the datasets. Absence of sparseness and overall dataset quality are verified roughly using Zipf's law validation and type-to-token ratio.

4.2 Datasets

The aim was to study the effect of term burstiness and homogeneity on datasets with varying characteristics, various types and with stylistic differences. In a first step, we need to ascertain that we have such datasets available. We considered the TIPSTER dataset collection [Har92a] as it contains data from seven different areas. Moreover, the TIPSTER collection is easily available and widely used for language research, so that it is well understood and easily interpretable. The TIPSTER dataset also contains collections that have been used by other researchers in this specific area (i.e. term burstiness modeling - e.g. Church [CG95b, CG95a, Chu00]) and so choice of this reference corpus will allow us to compare our results with some published by others.

The TIPSTER collection provides a range of dataset collections across various genres; each of these were artificially compiled and edited by humans (possibly removing any discrepancies) from a narrow base of similar text types, or from a particular domain. To contrast our results and findings experiments were also conducted on data collected from *The Open University (OU)* intranet and extranet web pages. This data was crawled automatically from The Open University domain (*open.ac.uk*) and, so as to understand the characteristics of real live data, no manual cleaning or selection was done on it. This dataset is more diverse in terms of document type and domain content than the TIPSTER datasets.

Though the datasets from TIPSTER and the OU dataset provided a range of datasets from various domains and genres, all these datasets were in English. We were interested also in studying whether the manifestation of term burstiness is similar across languages. Hence dataset of a non-English language, *Bengali* was added to the list of datasets.

Bengali is one of the ten most spoken languages in the world, with almost 200 million speakers. There is a rich literature, but little is available electronically. However, online textual resources

Data Set	Contents of the documents
AP	Copyrighted Associated Press Newswire stories from 1989.
DOE	Short abstracts from the Department of Energy.
FR	Issues of the Federal Register (1989), reporting source actions by government agencies.
PAT	U.S. Patent Documents for the years 1983-1991.
SJM	Copyrighted stories from the San Jose Mercury News (1991).
WSJ	Stories from Wall Street Journal 1987-89
ZF	Computer Select disks 1989/90, Ziff-Davis Publishing Co.
OU	The Open University intranet and extranet web-pages.
BEN	Bengali corpus from the Central Institute of Indian Languages.

Table 4.1: Description of contents of each of the datasets

are growing and a clear need for Bengali language applications and retrieval systems is emerging. As with other languages with little prior NLP history, development of robust language based applications will require applied Language Engineering and Information Retrieval research and a collection of reasonably balanced textual datasets. For profiling, we choose the Bengali corpus from the Central Institute of Indian Languages (CIIL) which was developed as a part of the Technology Development of Indian Languages (TDIL) Programme of the Ministry of Information Technology, Government of India. We made this choice because the corpus is freely available on-line, it was constructed in the context of developing language applications, and, on cursory investigation, it seemed of good quality.

The contents of each of the dataset, i.e. the TIPSTER dataset, the OU dataset and the Bengali (BEN) dataset are listed in table 4.1.

4.3 Dataset overview

Pre-processing was carried out on each of the datasets. TIPSTER datasets have all been tagged to provide specialized information about the article, like author, date, location, publication house, etc. These tags were all stripped off retaining only plain text. Numbers and special characters were removed during the tokenization process and no further linguistic processing was done. Hence a term like *model-based* would be tokenized as *model* and *based.* Also, *it's* and *he'd* would be tokenized as "it and s" and "he and d" respectively. Due to this tokenization methodology, the vocabulary contained a high number of occurrences of some words like "s", "d", "th", etc which strictly do not have any meaning.

4.3.1 Manual Sampling

The first task for gaining an overview of the datasets is manual sampling of them. Manual sampling of the dataset is carried out to identify any outliers in the dataset. This includes certain anomalies or peculiar behaviour in the dataset, which may include occurrence of a large number of numerical entries in the text, or a large incidence of acronyms or text in some language or sub-language, or even specific formats or data types - e.g. a phone book or other types of listings. Manual sampling does not result in the removal of documents or certain terms. The aim was to look for non-textual elements in the documents, like tables, numeric data, long lists, content in other language, codes in some programming language, because these would not display characteristics associated with representative running text. The TIPSTER collection is an acknowledged reference collection and its seven datasets are of high quality and manual sampling did not reveal unexpected ideosyncracies, although some of the collections are quite distinctive (e.g. the FR (Federal Register) which contains large lists, the DOE dataset which concerns one specific domain (energy), etc. The OU and the BEN dataset saw a more extreme mix of characteristics in this respect.

The OU dataset contained many documents that were forms to be filled by people. The set included much raw numeric data from the "Physics and Astronomy Department", which as a result of the pre-processing were left with no textual data and hence removed from the analysis. There were documents like meeting minutes or reports which were in the form of bullet points and are not in plain English, but they were retained in the dataset.

Manual sampling of the frequency lists of the BEN dataset showed that, to our surprise, this corpus contains a substantial number (8,791) of English words, noted in English script. These constitute a mere *0.29%* of the entire dataset, but *3.9%* of the distinct (or unique) terms. Most occur only once, and none occur with a frequency higher than 4, so whilst worth noting, their presence is unlikely to skew the statistical profile of the corpus and were retained in the BEN dataset. The top ten frequent English terms in the collection are listed in Table 4.2. The reason for the incidence of the English terms is that many of the documents in the dataset consist of educational or course materials, and a lot of technical terminology is not so well established in Bengali so authors have to resort to the alternative English representation.

Term	Freq.	Term	Freq.
the	4	ultimate	3
ph	4	types	3
world	3	transport	3
wind	3	th	3
war	3	system	3

Table 4.2: 10 most frequent English terms in the BEN dataset, with their frequency

4.3.2 Basic Summary Statistics

The next step in gaining insight about a dataset is to collect some basic overall statistics of the dataset. These statistics will include the number of documents in the dataset, their lengths and some measure of average document lengths in the dataset, along with the length of the longest and

shortest documents, to provide a range for the document length. The size of the dataset, i.e. the number of terms in the entire collection and the number of distinct terms will also be important information to collect while planning an application. These numbers will have direct connection to the amount of hardware and memory requirements that a computer should have for automatically processing the data. The vocabulary size can provide insight about the data structure that could be used for the application. Basic summaries for each of the datasets are presented in Table 4.3.

Data Set	No. of Docs.	Length of Data Set	Average Doc. Length	Vocab-ulary Size	Average Vocabu-lary Size per Doc.	Short-est Doc. Length	Longest Doc. Length
AP	242,918	114,438,101	471.1	347,966	238.25	9	2,944
DOE	226,086	26,882,774	119.0	179,310	72.90	1	373
FR	45,820	62,805,175	1,370.70	157,313	292.65	2	387,476
PAT	6,711	32,151,785	4,790.91	146,943	653.05	73	74,964
SJM	90,257	39,546,073	438.15	178,571	223.60	21	10,393
WSJ	98,732	41,560,108	420.94	159,726	204.26	7	7,992
ZF	293,121	115,956,732	395.59	295,326	168.42	19	75,030
OU	53,681	39,807,404	744.36	304,468	219.87	1	15,430
BEN	1,270	3,052,522	2,403.60	192,007	1,149.50	160	4,742

Table 4.3: Basic profiling statistics for each of the datasets

Table 4.3 shows substantial differences between the different datasets. ZF is the largest dataset with respect to total number of documents and total corpus length. On average, PAT has very large documents and DOE the smallest. This is not surprising as PAT consists of patent articles, which are long and contain descriptive technical details, whereas DOE consists of short abstracts from the Department of Energy. Note that though the size of the datasets vary greatly, the vocabulary size (number of distinct terms in the dataset) across all the datasets are of the same order of magnitude. The number of documents in the Bengali dataset, BEN is quite small as compared to the others, but average document length and the vocabulary size are in comparable order of magnitude. The high value of the "average vocabulary size per document" is due to the fact that Bengali is an inflected

language and morphologically informed tokenization has not been performed.

4.3.3 Study of the most frequent terms

The next activity in the basic overview of the datasets consisted of listing and inspecting the most frequent terms in the collection. The most frequent terms in a dataset are studied to identify any evident effect of anomalies in the dataset. This is because a text document like a story or article that is talking about a certain concept will have a few terms referring to the concept and many other terms (such as function words) that help in binding the concept, providing structure and readability to the document. So a collection of many different texts will possibly discuss many different concepts, but still use the same set of function words to bind concepts to create sentences and hence a document. So if all the terms are ranked according to their frequency in the collection, the most frequent terms in a collection of "well-behaved" documents should all be function words. Studying the most frequent terms in a collection will help in detecting any such anomalies in the dataset. A set of common function words is compiled and the frequent terms of the dataset are compared against them. If a term which is not in the function word list features as a frequent term, manual inspection of the dataset is done to understand the reason for the term's appearance.

However if a collection of texts from a telephone directory or product catalogs are considered, which are not "well-behaved" texts, they will not have a high occurrence of function words. For example, a collection of product records from *Amazon*[1] or *ebay*[2], or a collection of telephone directories will not have the properties of plain text. In such collections, function words will not always show as frequent terms and in many cases there may be no occurrence of function words at all. So a dataset of good quality that draws on representative sample of plain textual language is likely to have a list of very frequent terms that overlaps considerably with obvious function words

[1] `http://www.amazon.com`
[2] `http://www.ebay.com`

(though not all function words will occur very frequently). Table 4.4 presents the 10 top terms of the datasets.

Data Set	10 Most Frequent Terms
AP	the, of, to, a, in, and, said, s, for, that
DOE	the, of, and, in, a, to, is, for, with, are
FR	the, of, to, and, a, in, for, or, that, be
PAT	the, of, a, and, to, in, is, for, said, as
SJM	the, a, of, to, and, in, s, for, that, is
WSJ	the, of, to, a, in, and, s, that, for, is
ZF	the, m, p, and, to, of, a, in, is, for
OU	the, of, to, a, and, j, in, k, is, report

Table 4.4: 10 most frequent terms for each of the TIPSTER datasets and the OU dataset.

Unsurprisingly, the frequent terms for all the TIPSTER datasets are function words or morphenes, with "the", the most frequent word in English language having the first place for all the datasets. Due to the tokenization methodology, "s" appears several times in the table. "m" and "p" in ZF and "j" and "k" in OU arise from alpha-numeric characters, the numeric part of which has been stripped off leaving mere *meaningless* single characters.

The frequent terms in the OU dataset differ from the other datasets. In particular; the content term "report" is in the list of 10 most frequent terms. Looking beyond the 10 most frequent terms leads to the discovery of terms like "section" in position 19 in the list of frequent terms for FR, "software" in position 21 in ZF and "invention" in position 26 of PAT. All these terms reflect the domain and topic of the document collection. Again, there are applications that drop frequent terms in a collection arguing that they just add noise [Lew98]. Also, certain applications eliminate frequent function words based on a pre-defined list of so called "stop-words". But, the list of frequent function words is not always fixed and the characteristics of a term are dependent on the domain under consideration. Further discussions about the appropriate selection of function words and content words specific to a domain can be found Chapter 8.

4.4 Dataset quality and sparseness

Dataset sparseness is a known problem for experimental NLP, and if a dataset does not include sufficient information to be representative of the language then the dataset may need to be extended synthetically until it does. Rose and Haddock [RH97] synthetically extended a technical collection of emails by incorporating similar documents from the *British National Corpus* (BNC)[3].

The following sections investigate a couple of measures for judging the quality and sparseness of a dataset. Each section describes the suitability of the method, along with some experiments based upon it, followed by discussions for each of the datasets.

4.4.1 Type-to-token ratio

Type-to-token ratio, the rate of incidence of a term in running text, can also be used as very rough indicator for the measure of sparseness of a dataset [GDR01]. The measure shows on average how much evidence there is in a text for the behaviour of each term. Table 4.5 and Table 4.6 present the values of type-to-token ratios for the various datasets. Type-to-token ratio is calculated as the ratio of the number of tokens (number of terms in the running text) to the number of types (number of unique terms in the mentioned running text). Hence in the following tables a type-to-token ratio of 2 for the running *text length* of 100 means that there are $100/2 = 50$ unique terms in the running text of 100 terms. Table 4.5 presents the type-to-token ratio values for the seven different collections of the TIPSTER dataset. Here, while reporting values on the type-to-token ratio, the value of the ratio at text length $1,000,000$ is reported. To gain a fine grained picture of each of the datasets, the type-to-token ratio for various text lengths are reported also. The ratio for a particular dataset increases with the text length at quite a steady rate. This reflects a natural property of plain text, that however large the dataset is, one is always likely to encounter new

[3] `http://www.natcorp.ox.ac.uk/`

Text Length	AP	DOE	FR	PAT	SJM	WSJ	ZF
100	1.33	1.52	1.49	1.32	1.43	1.28	1.47
200	1.63	1.56	1.67	1.54	1.61	1.55	1.68
400	1.88	1.76	2.05	2.26	1.87	1.89	1.94
800	2.14	2.07	2.57	3.07	2.04	2.07	2.31
1600	2.80	2.32	3.05	4.27	2.48	2.58	2.76
3200	3.06	2.82	3.84	5.17	3.01	3.23	3.29
6400	3.56	3.58	5.44	6.01	3.56	3.83	4.24
16000	4.56	4.74	8.58	9.74	4.15	4.57	5.29
20000	4.97	5.20	9.20	11.03	4.46	4.99	5.38
1000000	30.57	30.16	50.57	62.64	26.38	30.91	38.11

Table 4.5: Type to token ratio for the datasets in the TIPSTER collection

terms once a new document is added to the collection. This rate of seeing new terms decreases as the length of the sample increases.

The highest value of the ratio at text length 1,000,000 is observed for the PAT dataset (62.64), followed by that of the FR dataset (50.57). The measure suggests that one is comparatively less likely to encounter new terms for some sample of for running text in these collections than in the other collections. Looking at Table 4.3 of the basic statistics for these datasets, it is evident that these two datasets contain large documents, and possibly due to *term burstiness*, even in a large document the same set of terms tends to be used many times repeatedly, leading to a high value of the type-to-token ratio. Again one of the lowest values of the ratio is for the DOE dataset (30.16). This can again be possibly justified based on similar arguments, as DOE is a collection of short abstracts, with an average document length of 119.0. The start of a new abstract leads to the introduction of new terms leading to a lower value of the type-to-token ratio. The lowest values of the ratio are for SJM (26.38), DOE (30.16), AP (30.57) and WSJ (30.91). Of these, AP, SJM and WSJ are about newswire or reported news articles, in which many topics tend to be discussed across different documents, lowering the value of the type-to-token ratio.

Text Length	TIPSTER (overall)	Brown Corpus	OU	BEN	Arabic Corpus
100	1.41	1.45	1.47	1.20	1.19
200	1.61	1.61	1.69	1.39	1.34
400	1.95	2.42	2.25	1.67	1.42
800	2.32	2.44	2.62	1.86	1.58
1600	2.89	2.58	3.05	2.29	1.77
3200	3.49	3.67	3.67	2.78	2.08
6400	4.32	4.70	4.31	3.31	2.36
16000	5.95	5.93	6.24	4.66	2.77
20000	6.46	6.34	6.94	5.21	2.88
1000000	38.48	20.41	36.13	10.81	8.25

Table 4.6: Type to token ratio for the overall TIPSTER collection, the OU and BEN datasets, the Brown Corpus and for a corpus of Arabic text.

Table 4.6 presents the type-to-token ratio for the overall TIPSTER collection, the OU and BEN datasets, the Brown Corpus and for a corpus of Arabic text (obtained from [GDR01]). The TIPSTER dataset represents an English collection of the 1990's and for which a measure of sparseness, the average type-to-token ratio is 38.48. This may be compared with the type-to-token ratio of the Brown Corpus [FK82], which is a collection of English text compiled in the 1960's. One may observe that the ratio for the Brown Corpus [GDR01] at text length $1,000,000$ is 20.41, much lower than the average of the TIPSTER collection and any of it's datasets. So the English in the Brown Corpus would appear to be sparser than that of the TIPSTER collection.

The OU dataset consists of documents of a mixed genre (meeting minutes, annual reports, departmental homepages, personal homepages of students and faculty, research information, guidelines, administrative documents containing rules and procedures of certain events, data files, etc). In spite of the huge diversity of genre in the OU dataset, the type-to-token ratio for text length $1,000,000$ is 36.13, which is close to the average value for the overall TIPSTER dataset.

The Bengali dataset, BEN is sparser as compared to the different collections of the TIPSTER dataset and the the Brown Corpus. Though the values for BEN look much lower than the datasets

in English, the values are quite consistent with the type-to-token ratio values for an Arabic corpus obtained from [GDR01]. These values do not provide any indication of evident discrepancies in the various datasets.

4.4.2 Zipf's Law

Zipf's Law [Zip49] describes the term distribution behaviour in a dataset in a compact form. It draws a relationship between the frequency of a word (f) and its position in a list sorted for rank order (r). The law states that, for a reasonably representative sample of a language, the relationship between rank order and frequency is constant i.e.:

$$f = c.r^a$$

where, c is a constant and a is very close to -1. So, if rank order is plotted against frequency on a logarithmic scale, it should be a straight line with slope -1 if it obeys Zipf's Law. This may be viewed as a solution of the differential equation $df/f = a.dr/r$, which says that if the rank changes from r_1 to r_2 ($r_2 > r_1$), then the frequency drops by a factor of $(r_2/r_1)^a$. In simple terms, suppose all text in a certain dataset or document is considered and the number of occurrence of each term is counted. Then, if these counts were sorted by rank, with the most frequently occurring word first and so on, then the shape of the curve is a "Zipf's Curve". Zipf's Law is based on the principle of least effort, which states that since languages are means of transmitting information, their structure should be optimal, thus allowing transmission with the least effort. Zipf's Law is a characteristic of term occurrence in human languages, and many other natural and human phenomena, such as city sizes, incomes, corporate wealth and earthquake magnitude. The law implies that the most common word in a dataset is a hundred times as common as the hundredth most common word, a

thousand times as common as the thousandth most common word, and a million times as common as the millionth most common word. Hence, it is a very skewed distribution.

Powers [Pow98] summarizes some linguistic areas that are affected by Zipf's Law. In Statistical Learning Methods, Zipf's Law helps in predicting the amount of text to look at and the preciseness of the statistics to achieve the target level of expected error [Fin93, Pow96]. In Information Retrieval and Semantics, Zipf's Law provides a base-line model for the expected occurrence rate of terms, based on which one can determine the term's importance in the document collection [SP98]. Zipf's Law provides a basis for evaluating parsers and taggers by evaluating the transferability of results across datasets [EP98]. In Computational Psycholinguistics, the law provides a distributional foundation for models of the language learner's exposure to segments, words and constructs, and permits evaluation of learning models [Bre97].

Because of these characteristics, Zipf's Law is useful as a rough description of the frequency distribution of words in human languages [MS99]. Zipf's Law validation is a process of plotting the word frequency against the sorted rank order in a logarithmic scale for the dataset and visually the judging its closeness to an ideal Zipf curve with slope -1. Obtaining frequency distribution in a dataset is cheap, and Zipf's Law validation is a rough-and-ready way of detecting some obvious problems with a dataset [GDR01]. A dataset may not be a suitable source for developing a language model if it does not fit Zipf's distribution, as non-compliance would almost certainly be an indicator of excessive sparseness, or of an idiosyncratic term distribution pattern, as might be caused by a particular sub-language or document type. An unusual distribution pattern may render the corpus unsuitable as the basis for developing general language resources. Figure 4.1 shows an example of the plot of rank versus frequency on a logarithmic scale that violates Zipf's Law as the slope is step-wise instead of being a smooth one, with data obtained from a very sparse collection of text for the Arabic language [GDR01]. Such a step-wise Zipf curve is typical of sparse data because in

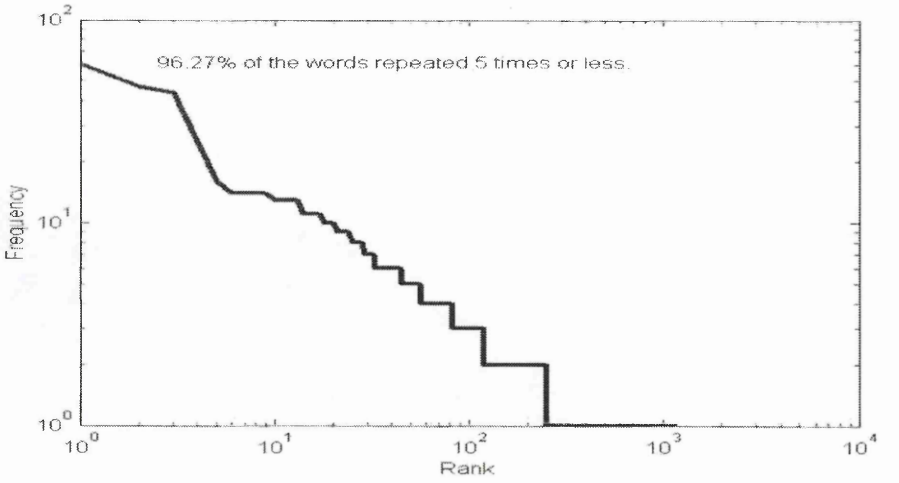

Figure 4.1: Figure showing the Zipf's curve plot of a sparse Arabic dataset. (obtained from [GDR01])

sparse datasets, it is more likely that gaps occur in the frequency counts.

For English text, it was observed that the law tends to hold for the more frequent terms, but breaks down for less frequent terms [Pie80]. Others have observed that the law does not hold for the most common function words, but holds for other frequent words [TST97]. They state that the law will hold for English language if it had three times more *the*'s, twice as many *of*'s etc. This would make the language quite awkward and not an efficient means of communication. The Zipf-Mandelbrot's Law was suggested as an extension to Zipf's Law [Man54, Man83]. This stated that:

$$f = c.(r + d)^a$$

where, d is a constant added to the model. Mandelbrot also observed that the value of a was generally slightly smaller than -1. Lavalette's nonlinear Zipf's Law was also suggested as a variation of the original law [Pop02]. This states:

$$f = c.\left(\frac{r.N}{N - r + 1}\right)^a$$

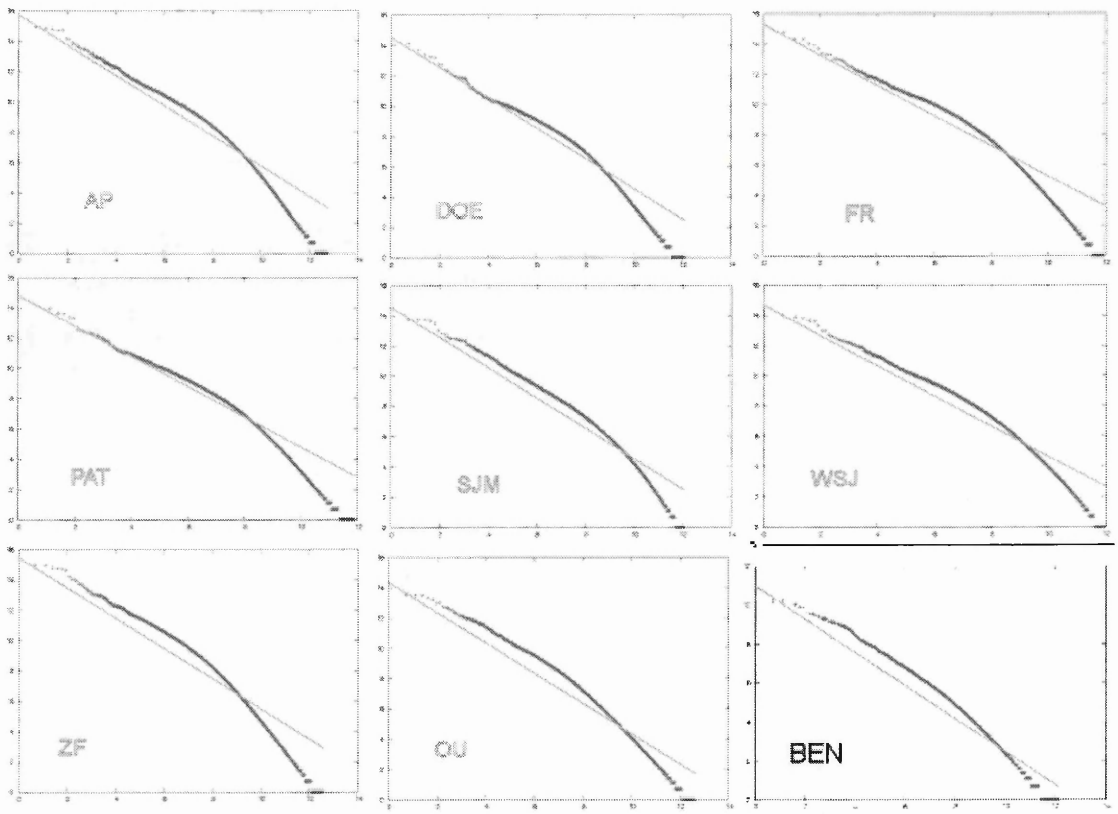

Figure 4.2: Figure showing the Zipf's curve plot of the **Log Rank Order** against **Log Rank Frequency** for all the datasets.

where the newly introduced variable N denotes the length of the document or the dataset on which the law is validated. Other studies show that Zipf's Law falls in the general category of a power law distribution which also includes the Pareto Law, which is mostly used in economics [Ada01, Baa01]. It might be possible to divide the entire set of terms into two regimes, such that terms in one set obey Zipf's Law with the exponent $a = -1$ and in the second regime with $a = -2$.

Figure 4.2 shows the Zipf's curve plot of the rank order versus rank frequency on a logarithmic scale for all the datasets involved in our experiments. The plots for all the seven TIPSTER datasets do not reveal problems. These datasets are manually compiled, hence any document that might prove to be an outlier may have been removed from the set. The OU and the BEN dataset also obey

the Zipf's Law, and there is no evidence of sparseness in these datasets.

Further checks and manual sampling of these datasets did not reveal any noticeable discrepancies in the TIPSTER datasets. The OU dataset in contrast had files containing experimental data, which have been removed manually. The BEN dataset had certain proportion of words in English script. The basic statistics derived from these datasets show the data covers a vast range, like very short documents in DOE as compared to large patent documents in PAT. These datasets did not reveal any evident discrepancies with regarding sparseness based on the type-to-token ratios and the Zipf's Law validation.

Chapter 5

Homogeneity Experiments

5.1 Introduction

Mainstream techniques and approaches in the statistical NLP and IR literatures tend to make a "homogeneity assumption" about the distribution of terms by adopting a "bag-of-words" representation for text as their starting point. The "bag-of-words" representation of text essentially extracts from text information on the frequency with which terms occur, thereby discarding information on dependency between occurrences of (the same or different) terms, and uses these frequency counts as the basis point for further processing. One consequence for approaches that use such representations, is that they make a built-in assumption, that terms occur independently from each other, and that therefore the probability of a word occurring is constant throughout the text. In other words, such approaches assume that the distribution of a term is spread homogeneously throughout a text. Because positional information is lost in such representations, they do not capture the bursty behaviour of terms.

A different, indirect version of a "homogeneity assumption" is evident in the treatment of very frequent function words. (The relationship between frequent words and function words is discussed

in Chapter 2, but for the current purpose, the distinction is not important because in most balanced text, the overlap between very frequent words and function words is substantial.) Very frequent function words are usually considered to constitute background noise and in many applications, they are routinely removed without any damaging side effect. They are either removed by means of a stop-list, or by removing the most frequent terms. The point here is that function words are assumed to be "noise" and to be uninformative because they are distributed more homogeneously throughout the text than other kinds of words. If terms were actually homogeneously distributed, then picking techniques that make such underlying assumptions about homogeneity in term distributions would be no great loss. However, since it is known that terms do not occur independently from each other, and are not normally homogeneously distributed, then there is extra information in text that such techniques do not capture, and that text processing applications might be able to use. The first step would be to try and estimate how heterogeneous term distributions actually are.

This chapter describes a series of homogeneity experiments intended to demonstrate, in a quantifiable way, to what extent terms are not uniformly distributed in the collection. The experiments are constructed in the following way. It is known that terms do not occur independently from each other, that the probability of a term occurring is not constant throughout text, and hence that terms do no not distribute homogeneously. Hence we will articulate, as our null hypothesis, that terms do distribute homogeneously, and we will seek to defeat that hypothesis. The extent to which we manage to defeat it will give us an indication of how heterogeneous term distributions actually are.

We shall pay some special attention to very frequent words, because by definition, they occur many times, and so they present a lot of evidence. They are also interesting because they overlap substantially with function words, that are routinely assumed to distribute more homogeneously that other types of terms. The homogeneity experiments were conducted on structured text and

are designed to destroy, to varying degrees, evidence of dependencies between term occurrences, and to measure the effects of doing so.

This chapter develops a notion of homogeneity detection to a level of statistical significance. The experiments show that the homogeneity assumption does not generally hold, and that it is defeated at different levels for different types of collections. It is also shown that the homogeneity assumption does not hold for function words. The homogeneity assumption is defeated substantially for collections known to contain similar documents, and more drastically for diverse collections, so there is a correspondence between the outcomes of the experiments, and what we expect to find given what we know about the collections (see Chapter 4). This chapter concludes that it is statistically unreasonable to assume that word distribution within a dataset is homogeneous. Because homogeneity findings differ substantially between different collections, it may be argued for the use of homogeneity measures are a suitable means of profiling datasets.

Portions of the reported work in this chapter have been previously published [DRSG04a, DRSG04b, SDR04, DRSG05].

5.2 Homogeneity Measures

Kilgarriff [Kil97] describes a basic method for using measures of similarity to gauge homogeneity in a corpus. In the corpus literature, measuring this particular flavour of homogeneity has been linked to gauging the distance between corpora and to genre detection [KR98]. Starting from the position that no corpus can be more similar to another corpus than it is to itself, Kilgarriff casts homogeneity as internal similarity of distributions, between two halves of a document collection. Clearly, distributions of different features can be checked for similarity. The methodology begins by dividing the entire dataset randomly into two halves. Kilgarriff chooses consecutive text chunks

of 5000 words and assigns the chunks at random to either of the dataset partitions. Once the two partitions are obtained, a series of similarity measures are applied to judge the closeness between the two randomly created partitions. Based on the two partitions of the dataset, a frequency list for every term in the dataset is generated, which provides the frequency information about every term in each partition. This gives an $N*2$ (Read as N by 2) contingency table of numerical values; where N is the number of unique terms in the entire dataset and each column of the contingency table corresponds to one of the dataset partitions. This is illustrated in Table 5.1. Similarity is determined on the basis of the data in this table. The following sections describe some of the usual similarity measures, along with a discussion of the strengths and weaknesses of each measure.

Term	First Partition	Second Partition
$Term_1$	$tf_{1,1}$	$tf_{1,2}$
$Term_2$	$tf_{2,1}$	$tf_{2,2}$
$Term_3$	$tf_{3,1}$	$tf_{3,2}$
⋮	⋮	⋮
$Term_i$	$tf_{i,1}$	$tf_{i,2}$
⋮	⋮	⋮
$Term_N$	$tf_{N,1}$	$tf_{N,2}$

Table 5.1: Table showing a (N*2) contingency table representing the two halves of the datasets for calculating homogeneity. Here $tf_{i,j}$ denotes the frequency of $Term_i$ in the j^{th} partition.

5.2.1 Chi-Square, χ^2

The Chi-square, χ^2 measure [Kil96b, Kil96a, Kil97, KR98, Kil01] can be used to determine whether the two dataset partitions have been generated from the same population. The measure investigates any significant difference between the term frequency distributions of the two partitions, and also evaluates if the difference between the two partitions is random or systematic in nature. The χ^2 measure is calculated from the difference between the observed and expected values for a certain

cell in the contingency table and is stated as:

$$\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{2} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

where, $O_{i,j}$ is the observed value in a certain cell, and $E_{i,j}$ is the expected value calculated from the entire dataset. Specifically, $E_{i,j}$ for a certain cell is calculated as the product of the row sum and column sum divided by the overall sum for all elements in the table.

The $\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$ provides a measure of the difference of the term's frequency between the two partitions. The χ^2 measure is appropriate only if the expected value for every cell meets some conditions. A common rule-of-thumb is that each $E_{i,j}$ must be greater than 5 [Dun93]. For detecting statistically significant homogeneity between the two partitions, a null hypothesis stating the equality of the two partitions is postulated. As Kilgarriff [Kil97, Kil05] states, since language is not random, the null hypothesis cannot be sustained. The χ^2 measure, however, was used to reflect the degree of difference between British English (LOB corpus) and American English (Brown corpus) [HJ82]. It has also been used to identify distinctive terms in a dataset [Kil96b]. In addition, it has been used to judge the similarity between two different datasets [KR98], or different language varieties [Cav02b] or measuring inter-document distance [Cav02a].

5.2.2 Log Likelihood, G^2

The log likelihood measure was suggested by Dunning [Dun93] as a better measure for dealing with rare events, which he casts as events of surprise and coincidence. Dunning points out that rare events, such as the occurrence of many words and most n-grams in most datasets, do not follow the normality assumptions. The log likelihood measure is calculated for one particular term at a time, and the overall log likelihood measure for the entire dataset is obtained by summation

of the measures obtained for each term. To obtain the measure for a particular term, the $N * 2$ contingency table (Table 5.1) is reduced to a $2 * 2$ contingency table as shown in Table 5.2, where one row corresponds to the frequency of a particular term, and the other row corresponds to the of all other terms. Based on the table the observed value for a certain cell, O_i, is the observed

Term	First Partition	Second Partition
$Term_i$	$tf_{i,1}$	$tf_{i,2}$
Not $Term_i$	$tf_{i',1}$	$tf_{i',2}$

Table 5.2: Table showing a (2*2) contingency table for term i used for the log likelihood calculation.

frequency in that cell and the expected value for that cell, E_i is calculated as the product of the row sum and the column sum for that cell divided by the overall sum of all the cells in the table. Based on these, the log likelihood measure, G^2, is calculated as:

$$G^2 = 2\sum_{i=1}^{N} O_i \times ln\left(\frac{O_i}{E_i}\right)$$

where, $ln(x)$ denotes the natural logarithm of x. The log likelihood measure is also based on the null hypothesis that the two partitions are drawn from the same population and determines whether the variation between the partitions is random or systematic. Systematic variation between partitions would indicate lower levels of homogeneity. G^2 has proved a useful measure for rare events in terminology extraction [Dai96] and it was used to expand a vocabulary list in a speech processing application [RH97]. It was also deployed for measuring inter-document distances [Cav02a] and for comparison between two datasets [RG00].

5.2.3 Spearman Rank Correlation, S

The Spearman rank correlation measure, S, computes the similarity between the two dataset partitions, based just on the rank ordering of terms within the dataset partition, rather than on their actual frequencies. The frequency lists are sorted, a rank list is created, and the difference in rank order for a term is used for calculating the measure. Suppose that the difference in rank order for a particular term, i, in the two partitions is denoted as d_i. Then the Spearman rank correlation, S, is calculated as:

$$S = 1 - \frac{6 \times \sum_{i=1}^{N} {d_i}^2}{N(N^2 - 1)}$$

where N is the total number of terms in each of the rank lists. One advantage of this measure is that it is non-parametric and hence it does not make any normality assumptions. The S measure has been used for comparing differences within datasets and between two datasets [Kil96a, Kil01] The measure can be used to compare datasets of different sizes and has been used to automatically expand a technical email collection using documents from the British National Corpus (BNC) [RH97].

This approach has disadvantages. Since frequency values are converted into ranks, useful information may be lost in the process. This is because a term occurring high up in the rank list is given similar weight to a term occurring much lower down in the rank list. For example, a difference in rank from 1 to 3 will be assigned identical weight to a term occurring much lower down the frequency list and having a similar rank difference (e.g. 1001 to 1003). The absolute difference between the rank gets higher weight irrespective of the ranks that produced it. So a difference in rank from 1000 to 1050 with rank difference $10050 - 10000 = 50$ will get more weight than a difference of 2 $(3 - 1)$ for a term occurring much higher up in the rank list. This is an obvious disadvantage of this measure.

5.2.4 Information Theoretic measures

Most similarity measures for comparing two partitions of a dataset or two different datasets are statistically motivated. In contrast, this section discusses a measure motivated by information theory and widely used in statistics. The **Kullback-Leibler (KL) divergence** (or distance), also known as the **relative entropy** or **cross entropy**, is a well established measure in the field of information theory [Mac03, MS99]. It gauges the similarity between two probability distributions. The individual frequency values from each partition are divided by the sum total of the frequencies in that partition to obtain corresponding probability values for the term in that partition. Suppose the probability distributions obtained from each partition of the dataset are denoted by p_i and q_i, where index i indicates the term under consideration. Then the Kullback-Leibler (KL) divergence measure is denoted as:

$$KL(p||q) = \sum_{i=1}^{N} p_i \times log\left(\frac{p_i}{q_i}\right)$$

The KL divergence is not a symmetric measure, so $KL(p||q) \neq KL(q||p)$. This criterion makes it a poor choice for measuring homogeneity, as the two halves of the dataset are not treated equivalently. At times the frequency count for a term in a certain partition may be zero, leading to a probability value of zero. The measure will not be valid if the denominator is zero, so at times smoothing is carried out by adding a small value to both the probability distributions. The KL-divergence has been used in measuring homogeneity between and across different language varieties [KR98, Cav02b].

5.2.5 Choosing the appropriate measure

Given this range of established similarity measures across various disciplines, one would need to determine which is the best suitable measure for the purpose of measuring homogeneity of a dataset.

Clearly each method has its own strengths and weaknesses, so no single measure can be stated as the best one. Some of the measures work with the actual frequency values (χ^2 and G^2), while the KL divergence transforms the frequency values into probabilities, and the rank based measure only considers the rank of the elements in the frequency list. It is interesting to note the similarity between KL divergence, χ^2 and G^2 similarity measures, as they are all based on some form of ratio either between the actual observed value and the expected value, or the ratio of values between the two partitions. Some of these measures have been used for expanding a small technical dataset based on document-level similarity with other easily available datasets [RH97], or for determining which terms are characteristics of certain text [Kil96b]. Cavaglia [Cav02b] defines a homogeneous corpus as one that belongs to the same sub-language. In all this, the work focus tends to be on similarity as a means of establishing that two collections belong to the same genre or sub-language, by measuring lexical and syntactic features such as term frequency or POS (part-of-speech) distributions. A departure from this theme is [Cav02a], who uses term frequency and part-of-speech (POS) distributions together with un-supervised learning to generate corpora. Cavaglia [CK01] uses homogeneity measures on web documents to judge the spread of documents based on certain keyword searches.

Kilgarriff [Kil97, KR98] and Rose and Haddock [RH97] partition their corpus by placing successive chunks of 5000 words in each half. This basic technique of comparing two halves of a corpus has been used with different similarity measures. Kilgarriff and Rose [KR98] compare Spearman's S with χ^2. Rayson and Garside [RG00] deploy log-likelihood on different features, to expose different aspects of similarity. Cavaglia [Cav02a] uses relative entropy (Kullback-Leibler divergence), χ^2 and G^2. χ^2 is found to perform well in comparative experiments [RH97, Cav02b], as long as certain conditions are met. Notably, each of the individual frequency values must be greater than or equal to 5. Dunning [Dun93] states that most statistical tests assume some underlying distribution (usually

either normal or Chi-Square (χ^2)). He also shows through experiments that these assumptions can only be made if the sample size is large. Since, comparative experiments provided better results for the χ^2 measure [Kil97, KR98] and the constraints specified [Dun93] are fulfilled as described later, it was decided to use the χ^2 measure for the homogeneity experiments.

5.3 χ^2 for Measuring homogeneity from frequency data

Kilgarriff [Kil97] outlined a methodology for measuring homogeneity in a dataset using measures of similarity. Specifically, Kilgarriff casts homogeneity as internal similarity between two halves of a document collection as measured by the χ^2 statistic. The basic method involves the following steps:

(1) Split the dataset into two halves by randomly placing text chunks of 5000 words in one of two halves.

(2) Produce a word frequency list for each half.

(3) Calculate the χ^2 statistic for the difference in term frequency distributions between the two halves.

(4) Normalize for corpus length.

(5) Iterate and average over successive random partitions.

Kilgarriff uses the χ^2 statistic as a measure of homogeneity. The aim here, however, is to gauge the extent to which terms distribute burstily. This will be done by postulating a null-hypothesis that terms distribute in a homogeneous way - the homogeneity assumption - and by investigating under which conditions this hypothesis, or assumption can be defeated. We will verify the extent to which the homogeneity assumption for term distributions is valid. This will be approached using

Kilgarriff's outline methodology, by two-ways partitioning of a textual dataset, and by comparing the distribution of terms in the two halves. To increase the granularity of the experiments, a statistical significance test is required to examine whether the null hypothesis has been defeated. For this a more fine-grained tool than simple use of the χ^2 statistic as a measure. In this context, the relationship between the χ^2 test and the χ^2 statistic is explored.

5.3.1 The χ^2 test and the χ^2 statistic

The χ^2 test is a standard statistical method to test the null hypothesis that two or more samples are homogeneous, i.e. that they are drawn from the same population at random. In this case the null hypothesis would imply that any difference between the two dataset partitions is entirely random. When implemented (using the `SPlus` software[1] on a Linux platform), the χ^2 test produces three values in the output. First, the χ^2 statistic is calculated based on equation 5.1, which tests the difference between expected ($E_{i,j}$) and observed ($O_{i,j}$) occurrences of events. It is calculated with $(N-1)$ degrees of freedom (this is the second output). Here, N is the number of terms in the frequency list under consideration.

$$\chi^2 = \sum_{i=1}^{N}\sum_{j=1}^{2}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \tag{5.1}$$

Table 5.3 shows a $N*2$ contingency table to demonstrate the calculation of the χ^2 statistic. For each term, two frequency values are obtained from each half of the partitioned dataset. The χ^2 statistic is calculated based on the difference between the observed and expected values following equation 5.1. The observed value for each cell in the table is the term frequency value of that cell:

$$O_{i,j} = tf_{i,j}$$

[1] `http://www.insightful.com/`

Term	Half-1	Half-2	Row Sum
$Term_1$	$tf_{1,1}$	$tf_{1,2}$	$\mathbf{sum_{1,-}}$
$Term_2$	$tf_{2,1}$	$tf_{2,2}$	$\mathbf{sum_{2,-}}$
$Term_3$	$tf_{3,1}$	$tf_{3,2}$	$\mathbf{sum_{3,-}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Term_i$	$tf_{i,1}$	$tf_{i,2}$	$\mathbf{sum_{i,-}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Term_N$	$tf_{N,1}$	$tf_{N,2}$	$\mathbf{sum_{N,-}}$
Column Sum	$\mathbf{sum_{-,1}}$	$\mathbf{sum_{-,2}}$	$\mathbf{sum_{all}}$

Table 5.3: Table showing a (N*2) contingency table representing the two halves of the datasets used for the χ^2 calculation.

For calculating the expected value of each cell, the sum total of each row is denoted by $sum_{i,-}$, and the column total by $sum_{-,j}$. Also, the overall sum total of all the elements in the table is denoted sum_{all}. Based on these, the expected value of each cell is calculated as:

$$E_{i,j} = \frac{sum_{i,-}.sum_{-,j}}{sum_{all}}$$

These expected values should satisfy certain criteria. As mentioned earlier, the χ^2 test only produces proper results if certain conditions are meet. Criteria are that the terms should be independent of each other, the observed frequencies cannot be too small, and the expected values must be at least 5. The settings in which the experiments were conducted satisfy these criteria. The expected value of each cell is based on the sum total of the row and the sum total of the column. The calculated value of χ^2 from equation 5.1 follows a χ^2 distribution with $(N-1)*(2-1) = N-1$ degrees of freedom.

The third output value of the `SPlus` software is the p-value, a measure of statistical significance. The p-value is obtained by comparing the calculated χ^2 statistic with the value obtained from the χ^2 distribution with $(N-1)$ degrees of freedom. Being a probability, p-value lies in the range 0

to 1. A value close to 0 indicates that, based on the sample size, the null hypothesis of similarity between two samples should be rejected. The χ^2 statistic has been seen as a similarity measurement [Kil97]. In the case of perfect similarity (i.e. homogeneity in this case), one would expect the observed and expected occurrences to be close. Hence a lower χ^2 value would indicate greater similarity as compared to a higher χ^2 value. As a consequence, the χ^2 value may be viewed as a measure for comparing the similarity of two corpora, provided the degrees of freedom (N-1) is kept constant. This is due to the fact that a χ^2 value is calculated by summation over all the terms under consideration, which leads to a higher value if more terms are considered. The effect of the number of terms can be approximately nullified by dividing the χ^2 value by the degrees of freedom (N-1). The measure is called **Chi-square By Degrees of Freedom (CBDF)**. This is the corpus homogeneity measure used by Kilgarriff [Kil97]. Most other work [RH97, RG00] on corpus homogeneity that uses the χ^2 statistic sees it as an independent measure, without a statistical test of significance.

This basic use of χ^2 statistic is inappropriate for our purposes. Even a small departure from homogeneity can be detected if a sample's size is large enough. The question is whether the evidence is statistically significant. The p-value will get closer and closer to 0 as the sample size increases. One would like a measure of homogeneity that is not affected greatly by sample size, so that datasets of different lengths can be compared. Also, it is preferable if the similarity measure is compatible with a test of homogeneity: if two datasets are of similar size, the one with the larger value on the similarity scale should also have the smaller p-value for the test of homogeneity. Using CBDF as the similarity measure and the χ^2 test as the test of homogeneity fulfills these desirable properties.

5.3.2 Statistically significant Homogeneity detection

The task of investigating the homogeneity assumption requires a more fine-grained tool than simple use of the χ^2 statistic as a homogeneity measure. We are interested in understanding the conditions under which non-homogeneity is detected. The p-value in statistics tests the validity of a null hypothesis, and in this case the null hypothesis is that homogeneity will be found between the two halves of a two-ways partition, i.e. the hypothesis that the term distributions in the two halves come from the same population. In this case, the two halves are generated from the same dataset, but that does not imply by any means that the term distributions come from the same population; indeed for burstily distributed terms, they will not.

p-value	Interpretation about the null hypothesis
< 0.1	weak evidence against
$< \mathbf{0.05}$	**significant (moderate) evidence against**
< 0.01	strong evidence against
< 0.001	very strong evidence against

Table 5.4: Table for judging the null hypothesis based on the p-value.

The results will be differentiated in two ways, by reporting the p-value and by giving the CBDF statistic. Given a null hypothesis (in this case, an assumption that homogeneity will be detected between two halves), the p-value allows one to estimate the strength of the evidence offered by the data. Table 5.4 presents the criteria that are used for judging evidence for or against the null hypothesis. A p-value < 0.1 is usually interpreted as constituting weak evidence against the hypothesis, a p-value < 0.01 as strong evidence against, and p-value < 0.001 as very strong evidence against the hypothesis. Normally, a p-value < 0.05 is considered significant (moderate evidence against the hypothesis). In this case, a p-value < 0.05 will be taken to indicate that the homogeneity assumption (the null-hypothesis) has been defeated with statistical significance. The CBDF measure then relates to the text data and indicates the level of heterogeneity. Defeating the null

hypothesis about the term distribution similarity between the two partitions of a dataset renders the homogeneity assumption invalid, and provides evidence against the independence assumption inherent in several mainstream approaches, including the "bag-of-words" representation. It points up the presence of bursty term behaviour and presents some measure of the extent to which it occurs.

5.4 Frequent term distribution measures for Dataset Profiling

Dataset Profiling aims at characterizing different datasets using dimensions that are particularly relevant to applications in Natural Language Processing and Information Retrieval. Section 4.3.3 highlighted the frequent terms of the different datasets we are using. At times certain non-function words did appear among the top terms of the dataset. This indicated certain characteristics of the domain and genre of the dataset. Since the frequent terms in a dataset occur several times across many documents, the distribution pattern of the frequent terms can provide information about the dataset as a whole. The homogeneity experiments contribute new insights because they aim at gauging the extent to which actual data depart from the term independence assumption inherent in many NLP techniques. The results are reported for the N frequent terms of the dataset. The homogeneity measure and the statistical significance value for the frequent terms are studied to profile aspects of the dataset. Nonetheless, the method by which the dataset is split has an effect on whether the homogeneity null hypothesis can be defeated, and different schemes of partitioning can be used to capture the presence of term burstiness at different intensities.

5.5 Schemes for dividing a dataset

The basic methodology by Kilgarriff [Kil97] requires a dataset to be split into two halves, by randomly placing text in one of two halves. The obvious question is how to execute this division? One way might be to dissolve document boundaries and split the corpus halfway. Kilgarriff [Kil97] and others [RH97] dissolve document boundaries, but place consecutive chunks of 5000 words in each partition. Why chunks of size 5000 were chosen, rather than some other size, is not explained. The method of partitioning a document collection raises important questions that may affect the outcome of similarity measures based upon them. The various schemes of partitioning a dataset are aimed at capturing the effect of term burstiness at different levels and intensities. A chunk size of 1, for example, destroys all evidence in the data about inter-dependence between term occurrences and would amount to mimicking a random distribution. This will be confirmed in the experimental results. Increasing the chunk size (or window) will decrease the random element and allow the dependencies between term occurrences to impact on the experiment. Different methods of partitioning can shed light on whether terms distribute differently between documents in a collection, between different parts of the same document, or, as in Kilgarriff's case, in the language sample constituted by the whole collection. They may also shed light on the relationship between burstiness and, say document length or genre.

To answer some of these questions, alternative ways of partitioning a dataset were examined, with different ways of handling document boundaries. Each of the following schemes of dividing a dataset aims at partitioning at document-level, within-document level and chunk level. Briefly, the following three experiments were conducted:

- Choose a document and assign it at random to either of two partitions (**docDiv** experiment).
- Divide each document in the middle, and randomly assign one half to either of the partitions,

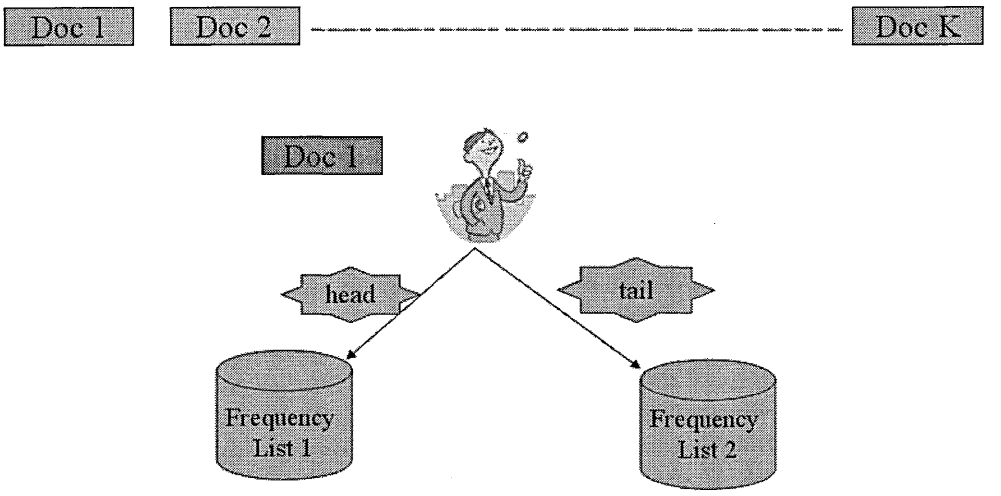

Figure 5.1: Figure showing the scheme of dividing a dataset in the docDiv experiment.

and the other half to the other partition (**halfdocDiv** experiment).

- Remove document boundaries and repeat the experiments of Kilgarriff [Kil97] with various chunk sizes, from 5 to 5000, and observe the homogeneity measure (**chunkDiv** experiment).
- The document is assigned at random to either of the partitions, but only the presence or absence of a term in a document is noted and the frequency is ignored (**binomialDiv** experiment).

5.5.1 docDiv

The **docDiv** experiment is designed to reveal the extent to which terms distribute homogeneously across documents in a collection. This experiment comes up with a measure of homogeneity of a dataset by retaining the document boundaries, and heterogeneity in a dataset caused by documents with different term distribution characteristics may be captured by this experiment.

Figure 5.1 describes the scheme for dividing the dataset into two halves, retaining the docu-

ment boundaries. Here, documents arrive in a sequence and a coin is tossed for each document, determining which partition it is allocated to. Tossing of the coin to generate a decision is based on a random number generator. Now a frequency list of terms is generated from each of the halves and a χ^2 test is performed to test the null hypothesis that the the two halves are homogeneous.

	N most frequent terms							
DataSet	**10**	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**2.107**	**1.576**	2.583	2.290	2.732	2.601	2.441	2.435
	0.122	**0.214**	0.001	0	0	0	0	0
DOE	**1.172**	**1.450**	1.755	1.983	1.838	1.786	1.795	1.872
	0.463	**0.160**	0.026	0	0	0	0	0
FR	54.524	41.715	72.093	66.787	51.387	61.266	39.043	23.534
	0	0	0	0	0	0	0	0
PAT	21.074	29.315	62.494	55.353	50.265	44.824	32.056	22.468
	0	0	0	0	0	0	0	0
SJM	**3.595**	2.768	3.231	2.976	3.012	2.959	2.560	2.511
	0.119	0.007	0	0	0	0	0	0
WSJ	**2.358**	2.663	2.364	2.335	2.623	2.749	2.831	2.917
	0.178	0.002	0	0	0	0	0	0
ZF	11.947	8.133	6.907	6.576	6.122	5.634	4.595	4.576
	0	0	0	0	0	0	0	0
OU	232.913	158.520	94.749	67.293	32.663	25.181	14.224	8.297
	0	0	0	0	0	0	0	0

Table 5.5: **docDiv** Results. Average CBDF and p-value per dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has **not** been defeated (p-value > 0.05)

Table 5.5 presents the results from the **docDiv** experiment. Each cell in the table reports the average Chi-square By Degrees of Freedom (CBDF) value (top value in the cell) and the average p-value (bottom value in the cell) for the particular dataset for the N most frequent terms, averaged over five successive iterations. Datasets have different vocabulary sizes, hence for comparability, results are reported for the N most frequent terms, with varying values for N. Table 5.5 shows that the experiment finds significant evidence of heterogeneity in almost all cases. The homogeneity null hypothesis is defeated in almost all cases; only very frequent words in some of the datasets behaved

in such a way that the homogeneity hypothesis could not be defeated. The exceptions are the AP and the DOE datasets when the 10 and 20 most frequent terms are considered, and the WSJ and SJM datasets for the 10 most frequent terms. All the other datasets show statistically significant evidence of non-homogeneity, with p-values close or very close to 0 (very strong evidence against the homogeneity hypothesis).

The CBDF values for the frequent terms provide further insight about the datasets. In most cases, they are quite large, indicating high levels of non-homogeneity. As expected, the experiment shows more evidence of heterogeneity in datasets which are known to be very diverse. The OU dataset has large values of CBDF and p-values of 0 indicating large variation in the contents across documents, as expected for a collection with documents of different genres, types and purposes. The OU dataset has a large degree of document-level-burstiness, i.e. large variability in the distribution of terms across documents. Among the TIPSTER datasets, large CBDF values are observed for the FR, PAT and ZF datasets which also have comparatively high evidence of document-level burstiness of certain terms. This is in line with our intuitions about the kinds of texts they contain. For instance, the FR dataset contains long, stylistically diverse government documents on varying topics, also the documents are quite large, adding a large bias in the frequency list of the half it belongs to. Large values of CBDF can be observed in PAT which has large patent documents of different subject areas adding to the bias just explained. In contrast to these, the newswire articles tend to discuss similar topics across different documents and have a similar reporting style, leading to small values of CBDF and even statistically insignificant p-values in certain cases. DOE has the smallest values of CBDF, possibly because it contains short abstracts, leading to less bias due to document length, and also because each document is dealing with a narrow range of topics covered by the dataset.

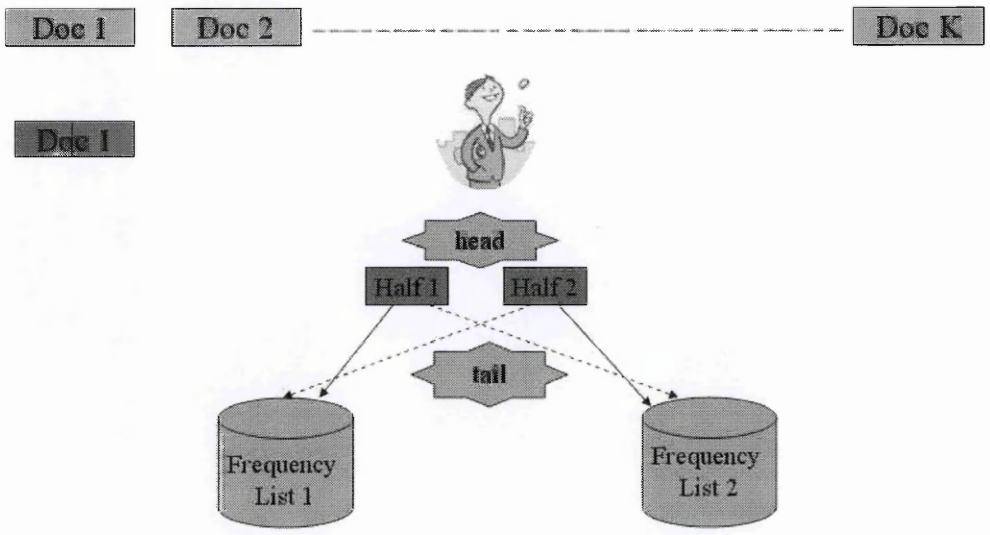

Figure 5.2: Figure showing the scheme of dividing a dataset in the halfdocDiv experiment.

5.5.2 halfdocDiv

The **halfdocDiv** experiment aims at investigating the within-document variation of term distribution in a dataset. Fine grained information on within-document term distribution is lost in text represented by the "bag-of-words" representation. The experiment aims to evaluate the extent to which terms distribute homogeneously within document boundaries. The experiment is sensitive, for instance, to whether a particular term distributes evenly across the document or whether it is clustered in a certain region. This experiment also aims at examining the term independence homogeneity assumption.

Figure 5.2 illustrates the **halfdocDiv** experimental scheme for dividing the entire dataset into two halves. Here each document is split into two parts at the mid-point of the document based on the document length, and each half of the document is assigned to a random partition of the dataset depending on the outcome of tossing a coin. For instance, if the coin turns tails, the first half is allocated to the second partition and the second part of the document to the first partition.

	N most frequent terms							
DataSet	**10**	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**1.774**	**1.473**	**1.369**	**1.271**	1.171	1.187	1.147	1.136
	0.087	**0.117**	**0.057**	**0.066**	0.021	0.001	0	0
DOE	**0.728**	**0.931**	**1.054**	**1.043**	**1.061**	**1.027**	**1.014**	**1.01**
	0.655	**0.533**	**0.438**	**0.372**	**0.195**	**0.285**	**0.271**	**0.182**
FR	7.905	9.549	11.627	11.642	8.847	8.166	6.543	5.336
	0.001	0	0	0	0	0	0	0
PAT	20.360	15.568	16.017	11.886	7.694	6.243	5.102	4.611
	0	0	0	0	0	0	0	0
SJM	**1.323**	**1.569**	**1.320**	**1.469**	1.332	1.297	1.24	1.242
	0.386	**0.392**	**0.444**	**0.107**	0	0	0	0
WSJ	**1.563**	**1.618**	**1.342**	**1.298**	1.236	1.210	1.178	1.150
	0.279	**0.248**	**0.203**	**0.260**	0.017	0.007	0	0
ZF	**1.948**	**1.858**	**1.709**	1.609	1.559	1.598	1.536	1.556
	0.129	**0.116**	**0.028**	0.024	0	0	0	0
OU	7.721	6.103	8.091	8.216	6.366	5.502	4.223	3.087
	0.033	0.002	0	0	0	0	0	0

Table 5.6: **halfdocDiv** Results. Average CBDF and p-value per dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has **not** been defeated (p-value > 0.05)

Table 5.6 presents the results of the **halfdocDiv** experiment for the N most frequent terms in the dataset across all the datasets. Five random partitions of the dataset were carried out (to reduce the variability effect in different random partitions), and each cell in the table provides the average value of the Chi-square By Degrees of Freedom (CBDF) and the p-value. Values in bold indicate cases where the homogeneity assumption has not been defeated: i.e. there is no statistically significant evidence to support the view that the terms in the partitions are not drawn from the same population.

Compared to Table 5.5 for the docDiv experiment where the homogeneity assumption remained undefeated, Table 5.6 for the halfdocDiv experiment has many more cases where statistically significant evidence of non-homogeneity cannot be found (values in bold). Also, comparing cell values in the same table position reveals lower CBDF values than the corresponding values from the docDiv

experiment. This lower CBDF value is consistent with the fact that document level heterogeneity is removed in this experiment as both the partitions contain equal parts of every document. The effect of document-level-burstiness is reduced in the halfdocDiv experiment. If we look at the cell for the 10 most frequent terms, the drop in value for the OU dataset is drastic from 232.9 in the docDiv experiment to 7.7 in the halfdocDiv experiment. In comparison, for PAT, the value for the 10 most frequent terms drops merely from 21.0 in docDiv to 20.3 in halfdocDiv. These figures reflect aspects of the document structures within a dataset, for instance where patent documents in PAT are made up of different sections or parts, each of which discusses a different topic area with few overlapping terms. The results are consistent with a view like Katz' [Kat96], that content (concept) terms behave burstily, and that on occasions where a topic of discussion is stated, the concept named by the topic will tend to be mentioned many times in a small neighborhood, rather than being spread across the entire document for longer texts. Burstiness is a general characteristic of textual data, but the clear within-document heterogeneity in long documents in PAT supports the case for homogeneity measures as ways of gaining some useful insights into datasets. Among the TIPSTER datasets, FR and PAT have the largest documents; and for these datasets there is strong evidence of heterogeneity. Again, the DOE dataset with very short documents and limited opportunity for covering several topics provides little evidence of heterogeneity across the cells in the halfdocDiv experiment.

5.5.3 chunkDiv

The docDiv and the halfdocDiv experiments described earlier study the homogeneity of a dataset between different documents and within documents. Hence, in these two experiments the amount of text that goes into each dataset partition is dependent on the length of the documents within the dataset. The aim of the **chunkDiv** experiment is to externally control the amount of randomness

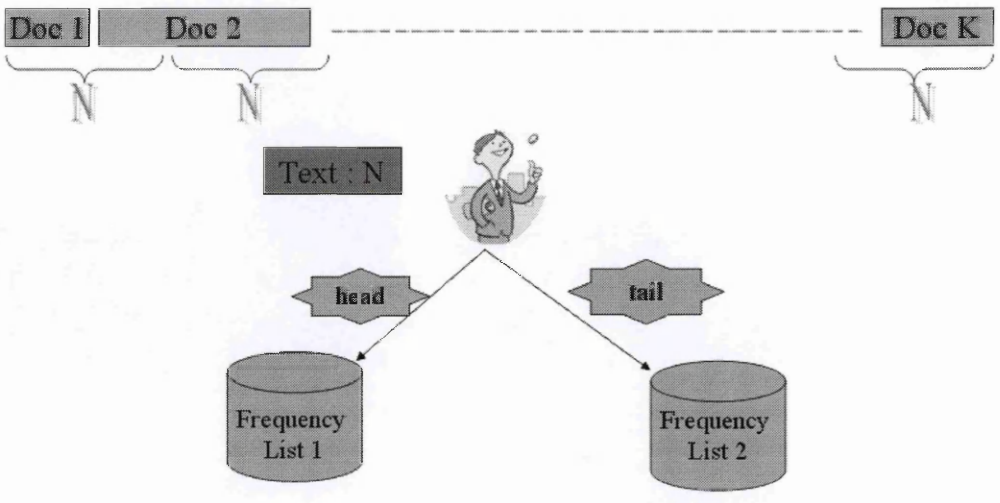

Figure 5.3: Figure showing the scheme of dividing a dataset in the chunkDiv experiment.

induced in the dataset by choosing chunks of text of varying lengths and placing them at random into either of the dataset partitions. This allows us to gain some view sensitive to the language variety used in the collection overall, and constitutes a refinement on Kilgarriff's original approach.

Figure 5.3 describes the **chunkDiv** strategy for dividing the dataset into two partitions. All document boundaries are dissolved by joining all the documents into a single sequence of text. Then for a chosen chunk size N, the next N consecutive terms are linked together in a sequence to form a chunk of text. Then a coin is tossed, and the chosen chunk of size N is assigned to one of the partitions, based on the outcome of tossing the coin.

The question arises, how the chunk sizes are chosen, as this will have a substantial effect on the generated partitions. One extreme way of dividing a dataset is by randomly assigning each consecutive word to one of two partitions. Such a division would introduce maximum level of randomness in the partitioning, and would destroy any document structure and any evidence of dependencies between terms. In such a case, one would not expect the experiments to register statistically relevant evidence of non-homogeneity. This is indeed the case, as we verified exper-

imentally. On the other hand, Kilgarriff [Kil97] reports non-homogeneity in partitions assigning chunks of 5000 words. Two questions arise. For a particular dataset, how large must the chunks be before non-homogeneity in the distribution of terms is statistically significant (p-value < 0.05)? Is this level dataset dependent? The experiment aims at answering these questions.

	N most frequent terms							
DataSet	**10**	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**0.628**	**0.836**	**0.871**	**0.984**	**0.990**	**1.007**	**1.018**	**1.012**
	0.752	**0.638**	**0.677**	**0.484**	**0.535**	**0.523**	**0.160**	**0.179**
DOE	**1.141**	**1.225**	**1.151**	**1.050**	**1.038**	**1.002**	**1.008**	**1.008**
	0.395	**0.346**	**0.251**	**0.354**	**0.423**	**0.462**	**0.431**	**0.367**
FR	**0.754**	**0.961**	**0.967**	**1.033**	**1.016**	**1.025**	**1.022**	**1.013**
	0.650	**0.504**	**0.540**	**0.405**	**0.417**	**0.335**	**0.228**	**0.211**
PAT	**1.284**	**1.457**	**1.255**	**1.153**	**1.051**	**1.007**	**1.008**	**1.020**
	0.245	**0.091**	**0.227**	**0.186**	**0.226**	**0.429**	**0.330**	**0.077**
SJM	**1.204**	**1.175**	**1.226**	**1.127**	**0.979**	**1.004**	**1.012**	**1.010**
	0.429	**0.375**	**0.293**	**0.268**	**0.608**	**0.454**	**0.262**	**0.181**
WSJ	**0.834**	**1.008**	**0.778**	**0.924**	**0.957**	**0.984**	**1.000**	**1.010**
	0.573	**0.492**	**0.822**	**0.679**	**0.682**	**0.620**	**0.498**	**0.252**
ZF	**0.861**	**0.791**	**0.939**	**0.913**	**0.994**	**1.012**	**1.007**	**1.016**
	0.578	**0.704**	**0.636**	**0.703**	**0.525**	**0.394**	**0.393**	**0.126**
OU	**1.242**	**1.257**	**1.165**	**1.023**	**1.081**	**1.054**	1.042	1.033
	0.340	**0.271**	**0.234**	**0.424**	**0.118**	**0.142**	0.034	0.005

Table 5.7: **chunkDiv** Results for **chunk size 5**. Average CBDF and p-value per dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has **not** been defeated (p-value > 0.05)

Studies were conducted on various chunk sizes, ranging from $5 - 5000$. Here full details of the experiments for chunk sizes 5 (Table 5.7) and 100 (Table 5.8) are reported, to provide an understanding of the effect of chunk size on the homogeneity measures. Most values in Table 5.7, for chunk size 5, are in bold, indicating the fact that there is no statistically significant evidence to indicate that the dataset partitions are heterogeneous. Also, the values of CBDF are quite low, and in many cases below 1, providing evidence of a high degree of homogeneity between the two partitions.

	N most frequent terms							
DataSet	**10**	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**0.824**	**1.105**	**1.412**	1.607	1.471	1.372	1.300	1.303
	0.602	**0.356**	**0.074**	0.002	0	0	0	0
DOE	**1.102**	1.864	1.646	1.511	1.354	1.414	1.401	1.424
	0.394	0.028	0.023	0.032	0.030	0	0	0
FR	**1.006**	**1.441**	**1.608**	1.803	1.924	1.834	1.782	1.746
	0.507	**0.229**	**0.076**	0.025	0	0	0	0
PAT	4.181	3.051	2.682	2.420	2.252	2.104	1.977	1.876
	0.023	0.003	0.001	0	0	0	0	0
SJM	**0.995**	**1.117**	**1.146**	**1.180**	1.410	1.402	1.317	1.291
	0.472	**0.385**	**0.320**	**0.246**	0	0	0	0
WSJ	**1.112**	**1.213**	**1.198**	**1.230**	1.196	1.283	1.290	1.319
	0.374	**0.324**	**0.243**	**0.094**	0.038	0	0	0
ZF	**1.576**	**1.283**	1.709	2.190	1.410	1.673	1.315	1.884
	0.415	**0.366**	0.011	0	0	0	0	0
OU	6.231	5.657	4.870	4.278	3.310	2.733	2.261	1.865
	0	0	0	0	0	0	0	0

Table 5.8: **chunkDiv** Results for **chunk size 100**. Average CBDF and p-value per dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has **not** been defeated (p-value > 0.05)

Table 5.8 gives experimental results for chunk size 100. As expected, compared to Table 5.7 this table has far fewer numbers in bold because the partitioning retains more inherent structure within running text. The distribution of terms within a document creates a meaningful content of the document. By taking smaller chunk sizes the inherent structure in documents and the impact of it are destroyed, resulting in a high degree of homogeneity between partitions. As chunk size increases, more structure is retained in partitioning the datasets, leading also to larger CBDF values.

There also appears to be a relationship between registering non-homogeneity and a combination of document length and diversity of domain coverage. Where a dataset contains many very short documents, even small chunks are likely to cross document boundaries (DOE is an example, which would explain why, at chunk size 100, this collection is less homogeneous than reported in the halfdocDiv experiments - in spite of the average document length exceeding 100 words). Where such

collections also cover diverse domains, documents are more likely to contain a higher proportion of distinct terms for the same amount of text. This is consistent with the OU data starting to register non-homogeneity at smaller chunk sizes than the other collections (Table 5.8), as it combines a high incidence of short documents with diverse domain coverage. This suggests that care must be taken when interpreting these measures in isolation and that it is important to rely on a range of complementary homogeneity experiments when investigating term distribution characteristics.

The **chunkDiv** experiment shows how evidence of homogeneity is related to chunk size. It shows how the picking of increasingly smaller chunk sizes is in fact similar to choosing representations that capture increasingly fewer dependencies between term occurrences, and so it is harder to defeat the homogeneity hypothesis. The question arises whether there is a language dependent element in where the breaking point lies between the chunk size where the homogeneity hypothesis can no longer be defeated, depending on morpho-syntactic characteristics (eg use of compounds, case marking, etc). To investigate this briefly, and also to test the methodology for a language other than English, the **chunkDiv** experiment was conducted for the BEN (Bengali) dataset. Only the 200 most frequent terms were looked at because it will be sufficient to show that the relationship between chunk size and evidence to defeat the homogeneity assumption is also present in a language other than English. In order to draw any conclusions about the relative behaviours of English and Bengali, it would be necessary to conduct a range of experiments on parallel or similar types of texts which we did not have at our disposal.

Table 5.9 presents results of the chunkDiv experiment at various chunk sizes for Bengali, and reveals that the BEN dataset behaves similarly to the other English datasets with respect to the relationship between chunk size on the one hand, and statistically significant evidence in support of non-homogeneity: larger sections of text contain more information on dependencies between term behaviours. There is no point in comparing the actual values returned in this experiment, so we

are only reporting results to chunk size 1000, beyond which there is clearly overwhelming evidence that even very frequent terms do not distribute homogeneously.

	N most frequent terms				
Chunk Size	**10**	**20**	**50**	**100**	**200**
5	**0.603**	**0.747**	**0.789**	**0.866**	**0.949**
	0.775	**0.733**	**0.812**	**0.799**	**0.669**
10	**1.170**	**1.049**	**0.987**	**1.032**	**0.997**
	0.343	**0.453**	**0.504**	**0.396**	**0.497**
50	**1.660**	**1.611**	**1.587**	**1.356**	1.261
	0.453	**0.297**	**0.136**	**0.126**	0.033
100	**1.735**	**1.485**	**1.402**	**1.352**	1.425
	0.178	**0.101**	**0.099**	**0.089**	0.009
1000	4.899	3.874	3.662	2.933	3.041
	0	0	0	0	0

Table 5.9: **chunkDiv** experimental Results for the BEN dataset. Average CBDF and p-value per dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has **not** been defeated (p-value > 0.05)

To gain a pictorial understanding of the homogeneity measures with respect to variable chunk sizes, two very diverse but very different datasets AP and PAT were chosen, and values of the CBDF values were plotted. Figure 5.4 present the plots for the AP dataset and Figure 5.5 for the PAT dataset. These figures plot the values of CBDF for the **docDiv** experiment, the **halfdocDiv** experiment and the **chunkDiv** experiment at chunk sizes 5, 50, 100, 1000 and 5000. Table 5.10 shows the ordering of CBDF values across the different experiments. These plots intersect each other at certain points, hence the ordering does not hold for all data points. So, the ordering presented in the table is a rough overview based on human judgement taking into consideration major portion of the plot. It is interesting to note how the **docDiv** and **halfdocDiv** values figure higher in the ordering for PAT as compared to AP, possibly due to large document lengths and huge variability between the documents. Also, there is a systematic aspect to the **chunkDiv** experiment, as smaller chunks always yield smaller CBDF values, indicating greater degree of homogeneity between

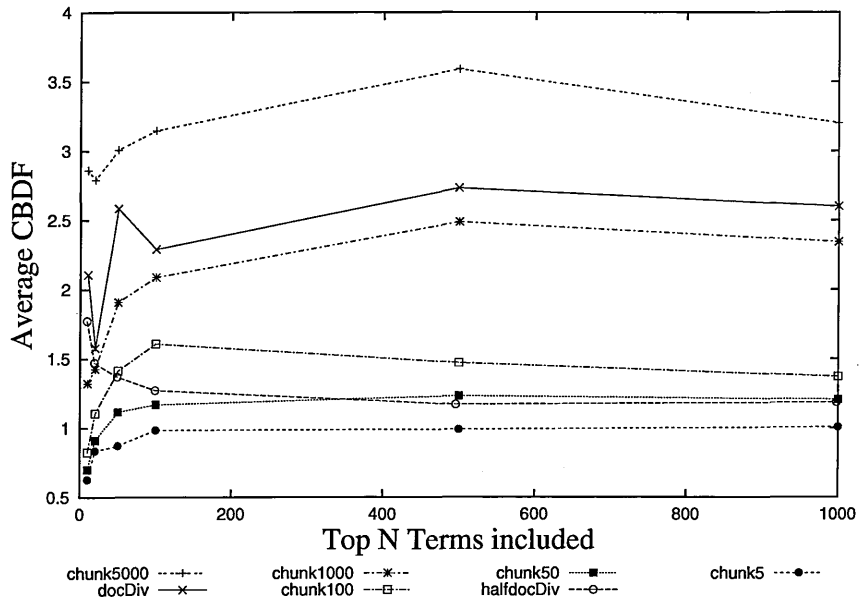

Figure 5.4: Figure showing plots from the **docDiv**, **halfdocDiv** and **chunkDiv** experiment for various chunk sizes on the AP dataset.

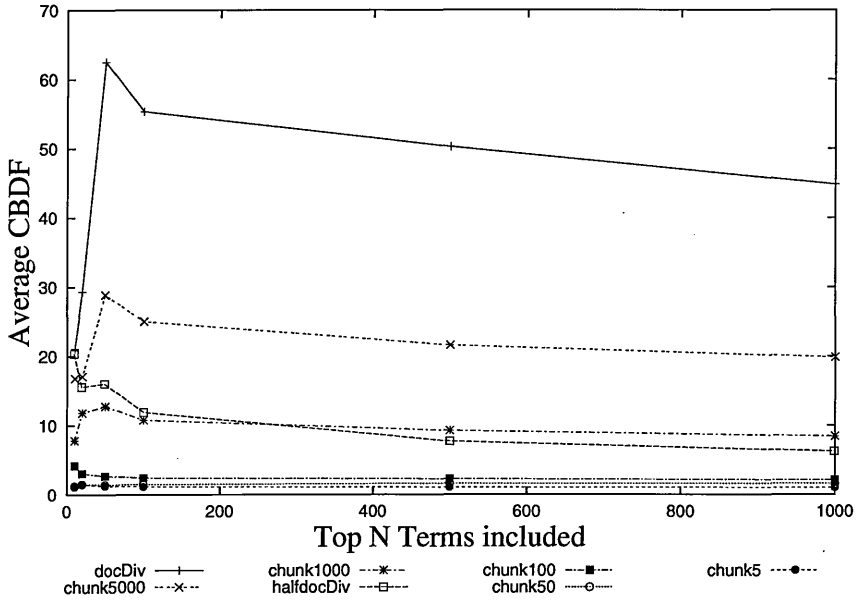

Figure 5.5: Figure showing plots from the **docDiv**, **halfdocDiv** and **chunkDiv** experiment for various chunk sizes on the PAT dataset.

partitions.

Rank (descending)	AP	PAT
1	chunkDiv 5000	docDiv
2	docDiv	chunkDiv 5000
3	chunkDiv 1000	halfdocDiv
4	chunkDiv 100	chunkDiv 1000
5	halfdocDiv	chunkDiv 100
6	chunkDiv 50	chunkDiv 50
7	chunkDiv 5	chunkDiv 5

Table 5.10: Table showing ordering of the Chi-square By Degrees of Freedom (CBDF) values for the docDiv, halfdocDiv and chunkDiv experiments conducted on the AP and PAT dataset

5.5.4 binomialDiv

In the **binomialDiv** experiment, the binomial model of terms is considered and only one occurrence of a term in a document is noted irrespective of its frequency. The halfdocDiv and the chunkDiv experiment aim at introducing randomness in a dataset by dividing the documents into halves and by breaking up a dataset into consecutive chunks of text. Also, there are a range of machine learning and text classification applications where only the presence or absence of a term in a document is considered and the term's frequency in a document is completely ignored [Mit97, BYRN99]. This motivates investigating the homogeneity of a dataset where each document in the dataset is represented by a binary representation indicating the presence or absence of the term. To this end, each document was selected at random and assigned to either of the partitions. Then homogeneity was measured between the two partitions obtained, based on presence/absence of a term.

Table 5.11 presents the results of the binomialDiv experiment. As expected, these two partitions revealed no heterogeneity. This suggests that in isolation, terms do not convey much information about a document, and it is the positioning of these terms in the document that determines the

	N most frequent terms							
DataSet	**10**	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**0.050**	**0.095**	**0.273**	**0.407**	**0.740**	**0.807**	**0.946**	**0.976**
	0.999	1	1	1	1	**0.999**	**0.901**	**0.934**
DOE	**0.111**	**0.406**	**0.605**	**0.774**	**0.917**	**0.924**	**0.979**	**0.987**
	0.999	**0.958**	**0.968**	**0.944**	**0.849**	**0.899**	**0.769**	**0.801**
FR	**0.060**	**0.124**	**0.226**	**0.279**	**0.584**	**0.705**	**0.949**	**0.912**
	0.999	1	1	1	1	1	**0.761**	**0.781**
PAT	**0.014**	**0.044**	**0.090**	**0.206**	**0.526**	**0.660**	**0.904**	**0.975**
	1	1	1	1	1	1	**0.976**	**0.999**
SJM	**0.075**	**0.113**	**0.253**	**0.396**	**0.705**	**0.816**	**0.963**	**0.986**
	0.999	**0.999**	1	1	1	1	**0.861**	**0.838**
WSJ	**0.100**	**0.198**	**0.265**	**0.356**	**0.708**	**0.840**	**0.962**	**0.987**
	0.999	**0.999**	1	1	1	**0.998**	**0.793**	**0.795**
ZF	**0.109**	**0.254**	**0.370**	**0.505**	**0.744**	**0.803**	**0.945**	**0.979**
	0.999	**0.998**	**0.999**	**0.999**	**0.999**	1	**0.998**	**0.960**
OU	**0.376**	**0.425**	**0.478**	**0.518**	**0.637**	**0.695**	**0.902**	**0.962**
	0.931	**0.980**	**0.986**	**0.996**	1	1	**0.999**	**0.847**

Table 5.11: **binomialDiv** experimental Results. Average CBDF and p-value per dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has **not** been defeated (p-value > 0.05)

information content. This experiment supports the view that it is the incidence and the arrangement of the terms in documents that is of importance. More precisely it is the particular structure of a document that distinguishes it from another document. These results also raise serious questions about the applications based on the Binomial model where only presence or absence of a term in a document is noted. Such applications tend to lose a lot of information in longer documents by not even taking frequency into consideration.

5.5.5 Behaviour of frequent terms

The homogeneity experiments were conducted under certain conditions [Dun93], which include that the expected values in the frequency list should be greater than or equal to 5. To fulfill these conditions, many rare terms that occur only a few times in the entire dataset were removed from

the frequency list for the χ^2 calculation. Also, if the frequency lists were sorted in descending order, the top terms would be all function words, as they occur in many documents across the dataset, as compared to content terms that may appear in only a few documents. Also, it has been observed that the top 1% of frequent terms account for almost 70% of the entire vocabulary. Hence, the above studied homogeneity measures are mostly based on the characteristics of frequent terms in the dataset.

Very frequent terms, which will tend to include most obvious function words, require bigger chunk sizes before non-homogeneity is apparent, when compared to experiments with a greater number of less frequent terms. Also CBDF values are lower when only high frequency terms are considered. To some extent, these results confirm Kilgarriff [Kil96b] and Katz [Kat96] who anticipate that the more frequent function words have distributions that are more similar among documents than less frequent (content) terms. Importantly, however, there are clear differences between the behaviour of very frequent (mostly function) words in different datasets of the TIPSTER collection. (Results for the OU dataset are consistent with the conjecture of Kilgarriff and Katz, because the OU most frequent terms contain non-function words). The present homogeneity experiments also reveal similar findings. Less evidence of heterogeneity between the partitions is noticed when dealing with few top terms, but evidence of heterogeneity increases as less frequent terms are considered. So the values in bold in Tables 5.5, 5.6, 5.8 decrease as less frequent terms are considered.

5.6 Findings and Summary

This chapter aimed to investigate the extent to which textual data diverge from the term independence assumption. If terms did occur independently from each other, then the probability of a term occurring would be constant throughout a text, and the distribution of terms would be homo-

geneous. We approached our aim by postulating as a null hypothesis that terms would distribute homogeneously, and then defeating the hypothesis in a series of homogeneity experiments based on the χ^2 statistical test. Homogeneity in term distribution of the N most frequent terms was investigated in this section. Starting from Kilgarriff's work, a notion of detecting non-homogeneity with a level of statistical significance was developed, and experiments conducted with different partitions of a range of datasets. The primary result is that the homogeneity hypothesis does not generally hold, even for function words. The experiments also showed that different datasets will exhibit different homogeneity properties, and these appear to correlate with a range of characteristics of the dataset. Thus, the above study provides experimental evidence that the frequently used independence assumption as, for instance, inherent in the "bag-of-words" representation of text, does not hold. Term burstiness is an important characteristic of text responsible for providing structure to text affecting even very frequent function words. Losing information about a terms's burstiness in a document and in a dataset leads to loss of information about a document. In analysis of a corpus it is often convenient to treat term distribution as homogeneous, and whether results would be biased to an important extent will depend on the analysis being performed and the purpose for which it is required. For example, an application on *Text Tiling* to detect boundaries and sections within documents will benefit from information about the term distribution within and across documents. This sort of information might be obtained from the discussed homogeneity experiments. As the degree of non-homogeneity differs substantially between different collections, one may also argue for the use of homogeneity measures as a means of deciding if an assumption of homogeneity is likely to lead to serious error for a specified dataset.

Chapter 6

Modeling Term Re-occurrence

6.1 Introduction

The model for term re-occurrence in a document and document collection is introduced in this chapter. The chapter starts by explaining the terminology and notation used in the model. The next section describes the model. Some properties of the model are discussed in the later sections. The term re-occurrence and burstiness model has been published [SGDR05].

As discussed in Chapter 3, most models of term distribution and term burstiness are based on the independence assumption and representation of multiple occurrences of a term in a document is confined to a frequency count. The model of term burstiness proposed in this chapter is not based on the independence assumption. The model accounts for multiple occurrences of a term in a document by looking at gaps between occurrences of the term. The position of the term's occurrence in a document is noted and, based on the positions, the gaps between occurrences of a term are obtained.

The gaps between successive occurrences of the term are modeled based on a mixture of exponential distributions. The model assumes that gaps are either generated from the exponential

distribution with larger mean which indicates the start of a burst for the term, or the gaps are generated from the exponential distribution with smaller mean which indicates the term's re-occurrence within a burst. In other words, the exponential distribution with larger mean dominates the rate of occurrence of term between bursts, the second exponential distribution dominates the rate of immediate re-occurrence within bursts.

6.2 Terminology and Notation

We aim to build a single model to capture the behaviour of a particular term in a given dataset or corpus. Let us suppose the term under consideration is x as shown in Figure 6.1. We describe the notation for a particular document, i, in the dataset.

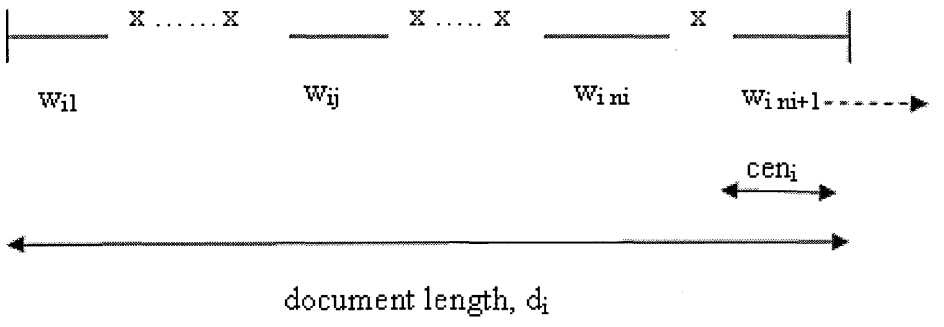

Figure 6.1: The document structure and the gaps between terms

- d_i denotes the number of words in document i (i.e. the document length).
- n_i denotes the number of occurrences of term x in document i.
- w_{i1} denotes the position of the first occurrence of term x in document i, by counting the term offsets.
- $w_{i2}, \ldots, w_{in_i}$ denote the successive gaps between occurrences of term x in document i. For

two occurrences of the term x in the document, the gap is measured by subtracting the term offset of the first occurrence of term x from the term offset of the later occurrence of term x.

- w_{in_i+1} denotes the gap for the next occurrence of x, somewhere after the document ends.
- cen_i denotes the gap after the last occurrence of the term to the end of the document and it is the value at which observation w_{in_i+1} is censored (explained in section 6.3.2). The value of cen_i is calculated by subtracting the term offset of the last occurrence of term x in the document from the document length.

6.3 The Model

The proposed model of term burstiness captures the re-occurrence pattern of the term within documents and in the dataset. The behaviour of gaps between occurrences of the term is modeled to identify bursts in the term's usage in the dataset. For example, in a long document, a term might occur with short gaps in the beginning of the document and then again occur after a very long gap within the same document, along with a few more occurrences in short gaps. This might be identified as two distinct bursts of the term within the document which can only be accounted by looking at gaps between the term's occurrences. If only the frequency count of a term was available, then such findings could not be made about the term.

Let us suppose we are looking through a document, noting when the term of interest occurs. The model assumes that the term occurs at some low underlying base rate $1/\lambda_1$ but, after the term has occurred, then the probability of it occurring soon afterwards is increased to some higher rate $1/\lambda_2$. Specifically, the rate of re-occurrence is modeled by a mixture of two exponential distributions. So, each gap is generated from both the exponential distributions simultaneously. Each of the exponential components is described as follows:

- The exponential component with larger mean (average), $1/\lambda_1$, dominates the rate with which the particular term will occur if it has not occurred before or it has not occurred recently.

- The second component with smaller mean (average), $1/\lambda_2$, dominates the rate of re-occurrence in a document or text chunk given that it has already occurred recently. This component captures the bursty nature of the term in the text i.e. the *within-document burstiness.*

The mixture model is described as follows:

$$\phi(w_{ij}) = p\lambda_1 e^{-\lambda_1 w_{ij}} + (1-p)\lambda_2 e^{-\lambda_2 w_{ij}} \tag{6.1}$$

for $j \in \{2, \ldots, n_i\}$. Where p and $(1-p)$ denote, respectively, the probabilities of membership for the first and the second exponential distribution.

The gaps between successive occurrences of a term are modeled based on a mixture of exponential distributions. The exponential distribution with larger mean dominates the rate of occurrence of the term across a large span of text. The second exponential distribution with smaller mean dominates the rate of occurrence of the term within a burst, i.e. after it has occurred recently.

The start of a document brings in a larger possibility of occurrence of new terms. The end of a document indicates an abrupt end to the process of a term's re-occurrence. These boundary conditions are handled separately in the model. Each of these cases are discussed below:

6.3.1 First occurrence

Both the λ parameters model re-occurrence of term, so the first occurrence of a term in a document cannot be generated in this way. Hence, the model treats the first occurrence of a term differently from the other gaps. The position of the first occurrence of a term in a document provides valuable information about the term's relevance to the document. The role of the second exponential

component measures the rate of term re-occurrence and so it is not related to the first occurrence. Hence the distribution for the first occurrence of the term in the document involves only the first exponential distribution:

$$\phi_1(w_{i1}) = \lambda_1 e^{-\lambda_1 w_{i1}} \tag{6.2}$$

6.3.2 Censoring

Document boundaries are handled by a novel technique in this model. Here we discuss the modeling of two cases that also require special attention, corresponding to gaps that have a minimum length but whose actual length is unknown. These cases are:

- The last occurrence of a term in a document.
- The term does not occur in a document at all.

In both cases there is some information available about a portion of the document where the term does not occur. The information is incomplete, but can add to the inference the model makes about rare terms that occur in only a few documents, or terms that occur only a few times within a document. Nonetheless, since these terms occur very few times, there are not enough term re-occurrence gaps available to build the model as set out before.

To deal with this, a standard technique from clinical trials is adopted here, where a patient is observed for a certain amount of time and the observation of the study is expected in that time period (the observation might be the time until death, for example). In some cases it happens that the observation for a patient does not occur in that time period. In such a case it is assumed that the observation would occur at sometime in the future but time for observing that event has abruptly ended or has been censored due to some reason. This is called **censoring** at a certain

point [Col03].

Adapting this approach to the case of term occurrence, it is assumed that a particular term would eventually occur, but the document has ended before it occurs so we do not observe it. In our notation we observe the term n_i times, so the $(n_i+1)^{th}$ time that the term occurs is *after* the end of the document. Hence the distribution of w_{in_i+1} is censored at length cen_i. Where, cen_i is the gap between the last occurrence of the term and the end of the document. If cen_i is small, so that the n_i^{th} occurrence of the term is near the end of the document, then it is not surprising that w_{in_i+1} is censored. In contrast if cen_i is large, so the n_i^{th} occurrence is far from the end of the document, then either it is surprising that the term did not re-occur, or it suggests the term is rare. The information about the model parameters that is given by the censored occurrence is,

$$\begin{aligned} Pr(w_{in_i+1} > cen_i) &= \int_{cen_i}^{\infty} \phi(x)dx \\ &= pe^{-\lambda_1 cen_i} + (1-p)e^{-\lambda_2 cen_i} \\ \text{where,} \quad cen_i &= d_i - \sum_{j=1}^{n_i} w_{ij} \end{aligned}$$

Also when a particular term does not occur in a document, the model assumes that the term would eventually occur had the document continued indefinitely. In this case the first occurrence is censored and censoring takes place at the document length. If a term does not occur in a long document, it suggests the term is rare. The w_{in_i+1} occurrence of the term is censored, and the its distribution falls into the general form of a truncated distribution, where the event of observing a certain event is cut at a certain point. The distribution of the w_{in_i+1} occurrence is $\phi_c(w_{in_i+1})$,

where:

$$\phi_c(w_{in_i+1}) = \begin{cases} 0 & \text{if } w_{in_i+1} \leq cen_i \\ \frac{\phi(w_{in_i+1})}{\int_{cen_i}^{\infty} \phi(x)dx} & \text{if } w_{in_i+1} > cen_i \end{cases}$$

6.4 Density plots of the Mixture distribution

This chapter discusses the mixture of exponential distributions that is used to model the gaps between term occurrences. It is assumed that the generation of larger gaps is dominated by the exponential distribution with the larger mean and that the generation of smaller gaps is dominated by the exponential distribution with the smaller mean. This section aims to provide insight into the mixture distribution by looking at the density plots of the individual exponential distributions and comparing them to the density plot of the exponential mixture distribution.

Here the term *church* from the AP dataset is chosen to demonstrate the density plots and the shape of the exponential mixture distribution. The model parameter values for this term are $\widetilde{\lambda_1} = 11174.43$, $\widetilde{\lambda_2} = 70.13$ and $\widetilde{p} = 0.9229$ (Chapter 7 describes the detailed methodology for obtaining these parameter estimates).

Figure 6.2 shows the density plots of the exponential mixture and the individual exponential distributions that generate the mixture distribution. In this plot (and also in Figure 6.3 and Figure 6.4) the label *first-exponential* refers to the plot of the first component of the exponential mixture, i.e. p times the single exponential distribution with parameter $\widetilde{\lambda_1} = 11174.43$, i.e. $p \cdot \lambda_1 e^{-\lambda_1 w}$ and similarly the label *second-exponential* refers to the plot of $(1-p)$ times the single exponential distribution with parameter $\widetilde{\lambda_2} = 70.13$, i.e. $(1-p) \cdot \lambda_2 e^{-\lambda_2 w}$. The label *exponential-mixture* refers to the plot of the mixture of exponential distributions (Equation 6.1) proposed in this chapter with parameters $\widetilde{\lambda_1} = 11174.43$, $\widetilde{\lambda_2} = 70.13$ and $\widetilde{p} = 0.9229$, i.e. the plot of

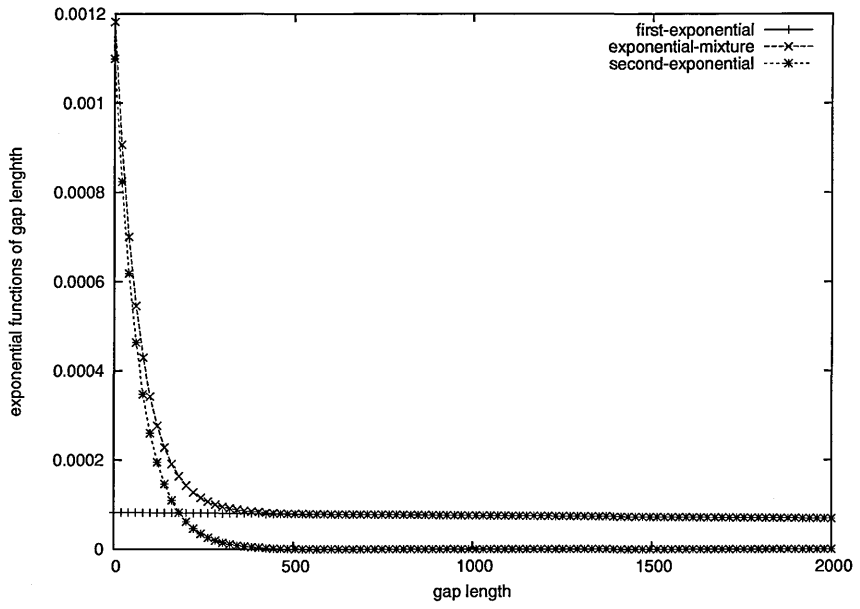

Figure 6.2: Figure showing the probability density plot of the exponential mixture distribution along with the plots of the individual exponential components.

$p \cdot \lambda_1 e^{-\lambda_1 w} + (1-p) \cdot \lambda_2 e^{-\lambda_2 w}$, where w denotes the gap length in the above equations.

The top line in Figure 6.2 is the exponential mixture distribution, the flat line is the first exponential distribution with the larger mean and the line that starts in the middle and drops to the bottom is the second exponential distribution with smaller mean. It can be seen that the individual exponential distributions are very different from each other due to the large difference in their parameter values.

In Figure 6.3 we focus on large gaps to understand which of the two exponential distributions dominates the generation of these gaps. Here we focus on the portion of the density plots with gap lengths between 1000 and 2000. Based on the understanding of the model, the generation of such gaps are likely to be dominated by the exponential distribution having the larger mean. It can be observed in Figure 6.3 that the probability density plots of both the *first-exponential* and the *exponential-mixture* have merged with each other indicating the clear dominance of the first

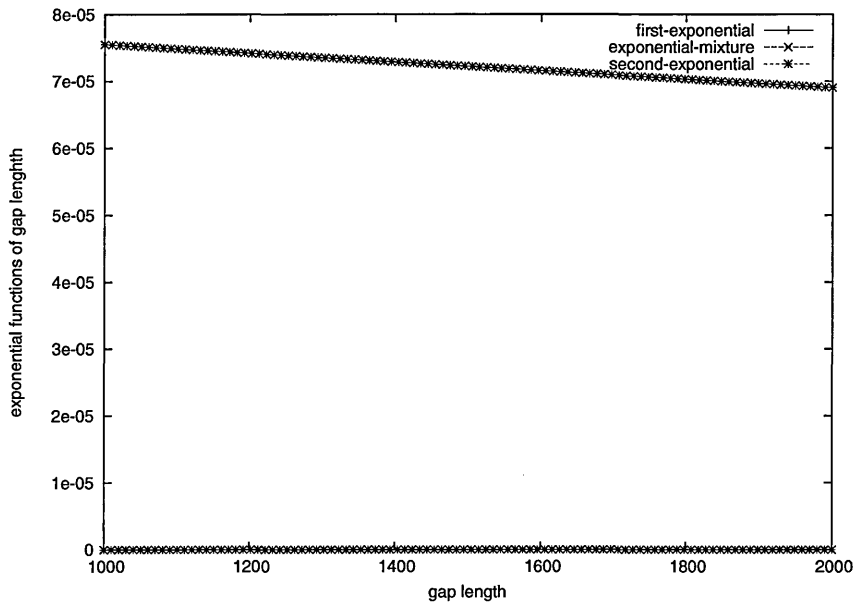

Figure 6.3: Figure showing the probability density plot of the exponential mixture distribution along with the plots of the individual exponential components. Here the focus is on larger gap lengths whose generation is dominated by the first exponential distribution with the larger mean.

exponential distribution (with the larger mean) in the generation of large gaps. The plot of the *second-exponential* distribution (with the smaller mean) has very low values as compared to the first exponential distribution in this plot and hence does not have much effect in the generation of large gaps.

In Figure 6.4 we focus on small gaps of length less than 200. Based on the understanding of the model, the generation of smaller gaps is likely to be dominated by the second exponential distribution with smaller mean. In Figure 6.4, the probability density values of the *second-exponential* distribution are much larger than the values of the *first-exponential* distribution at the bottom. Also the plot of the *exponential-mixture* is very close to that of the *second-exponential* distribution in that gap range. So the generation of smaller gaps is dominated by the *second-exponential* distribution with smaller mean. It can also be observed that as the gap length becomes large, the dominance of the *second-exponential* distribution diminishes.

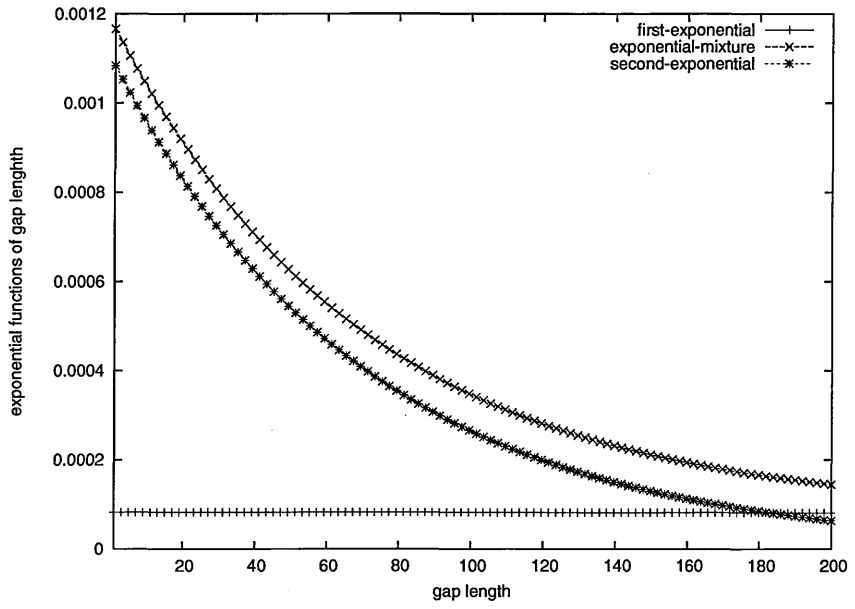

Figure 6.4: Figure showing the probability density plot of the exponential mixture distribution along with the plots of the individual exponential components. Here the focus is on smaller gap lengths whose generation is dominated by the second exponential distribution with the smaller mean.

6.5 Summary

This chapter has presented a model of term burstiness based on a term's re-occurrence pattern in the document and in the entire dataset. The gaps are modeled as a mixture of exponential distributions. In doing so, the model is capable of retaining positional and structural information about the term. This model would seem a significant improvement on standard methods that model term behaviour based only on their frequency in the document alone. This claim will be verified by applying the model to various applications and comparing it with results using frequency count methods (Chapter 8). Before the model can be applied, we need to be able to fit the model to data. The mixture model we have proposed is complex and so Bayesian statistics is used to fit the model and estimate parameters. Chapter 7 introduces Bayesian statistics along with the `WinBUGS` statistical software that will be used for obtaining parameter estimates.

Chapter 7

Bayesian Statistics

Bayesian statistics has been used for parameter estimating and fitting the model described in chapter 6. This chapter provides motivation and reasons for using Bayesian statistics and compares the Bayesian approach with frequentist methods. Parameter estimation techniques and convergence criteria based upon the Bayesian approach are then discussed. This is followed by a discussion of the data augmentation techniques that have been particularly useful for fitting the term re-occurrence model, which is a mixture of exponential distributions.

For the term re-occurrence model (equation 6.1), which is based on a mixture of exponential distributions, a model is built for every individual term of interest in the dataset. For each model, data is collected about the length of gaps between successive occurrences of a particular term of interest in the entire dataset. The parameters and the data are defined as follows:

$\vec{\Theta} = \langle p, \lambda_1, \lambda_2 \rangle$ denotes the parameters of the model.

The data are the observed gaps between successive occurrences of the term and they are denoted by $\vec{W} = \langle w_{ij} \rangle$, for $j = 1, \ldots, n_i, n_i + 1$ and $i = 1, \ldots, N$ where, N denotes the number of documents in the dataset and n_i denotes the number of occurrences of the particular term of interest in the i^{th} document.

7.1 Parameter estimation based on the Frequentist approach

The task is to estimate model parameters from the observed data. The frequentist approach treats parameters as unknown constants and only the data varies. One of the most popular and widely used approaches for parameter estimation is **Maximum Likelihood Estimation (MLE)**. The MLE estimator is the most likely values of the model parameters given the data values. The joint probability density of all the data points $\vec{W}$ is denoted as:

$$f(\vec{W}|\vec{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i+1} w_{ij}$$

When the data $\vec{W}$ are given, the joint probability density may be looked upon as a function of the model parameters $\vec{\Theta}$. This function is called the **likelihood function** of $\vec{\Theta}$ and denoted by $L(\vec{\Theta})$. The Maximum Likelihood method consists of taking the value as an estimate of $\vec{\Theta}$ for which $L(\vec{\Theta})$ is maximum. That is, if $\widehat{\vec{\Theta}}$ is the MLE of $\vec{\Theta}$, then:

$$\widehat{\vec{\Theta}} = \max_{\vec{\Theta}} L(\vec{\Theta})$$

This is obtained by simultaneous solving the following equations:

$$\begin{aligned} \frac{\delta}{\delta p} L(\vec{\Theta}) &= 0 \\ \frac{\delta}{\delta \lambda_1} L(\vec{\Theta}) &= 0 \\ \frac{\delta}{\delta \lambda_2} L(\vec{\Theta}) &= 0 \end{aligned}$$

where, $\frac{\delta}{\delta x}$ is the partial derivative with respect to x, such that terms in the expansion of $L(\vec{\Theta})$ which are independent of x are considered constant.

The problem with this approach is that the functional form of w_{ij} is a mixture of exponential distributions (along with the last observation being censored), leading to a complex likelihood function. Also, the exponential distribution has parameters λ_1 and λ_2 in the exponent which further increases the complexity of the likelihood function. Thus, the equations obtained from the partial derivatives all contain terms which are a combination of $\langle p, \lambda_1, \lambda_2 \rangle$, with λs in the exponent. It is impossible to find a close form solution for these complex equations, so alternate methods are needed to solve these equations and obtain the parameter estimates.

7.2 Bayesian Vs Frequentist Statistics

Two philosophically different approaches to statistical inference are classical or *frequentist* statistics and Bayesian statistics. The essential difference between them is that the Bayesian approach allows any unknown quantity to have a probability distribution while the frequentist approach only allows random variables to have probability distributions. For instance, an unproven conjecture in mathematics is that any positive even number can be written as the sum of two prime numbers (12=7+5; 26=3+23; etc). A Bayesian might state that 0.9 is the probability that this conjecture is true. Frequentist statistics says that there is no random uncertainty, so the probability that the conjecture is true is either 0 or 1, and can be nothing in between. The distinction means that the parameters of a model, such as λ_1 and λ_2, can have probability distributions with the Bayesian approach but not with the frequentist approach.

In the Bayesian approach, a *prior* distribution is used to convey the information about model parameters that was available before data were gathered. This is combined with the information

supplied by the data, which is contained in the *likelihood*, to yield a *posterior distribution*. Formally,

$$\text{posterior} \propto \text{prior} \times \text{likelihood},$$

where $\propto$ means 'is proportional to'. Prior distributions are a strength of Bayesian statistics in that they enable background knowledge to be incorporated into a statistical analysis. However, they are also a weakness because a prior distribution must always be specified, even if no useful background information is available or if one does not wish to use background knowledge, perhaps to ensure the analysis is transparently impartial, or because it can be difficult and time-consuming to specify a prior distribution that provides a good representation of the available prior knowledge. Consequently, in practice a prior distribution is almost always chosen in a mechanical way that yields a distribution designed to be non-informative.

Bayesian methods have sprung to prominence over the last fifteen years. This is because of the development of good computational techniques, notably **Markov Chain Monte Carlo (MCMC)** methods, that have solved many of the numerical problems formally associated with the practical application of the Bayesian approach. With these new techniques, Bayesian methods can now analyze complex problems that frequentist methods cannot handle. This has led to the general acceptance of Bayesian methods.

7.3 Bayesian formulation of the term re-occurrence model

In the Bayesian approach, prior distributions are assigned to the parameters of a model. For the term re-occurrence model (equation 6.1), non-informative priors were chosen, as is common practice

in Bayesian applications. So the following prior distribution was used. For λ_1,

$$\lambda_1 \sim Uniform(0,1)$$

To tell the model that λ_2 is the larger of the two λs, we put $\lambda_2 = \lambda_1 + \gamma$, where

$$\gamma > 0, \text{ and } \quad \gamma \sim Uniform(0, 1-\lambda_1)$$

As a non-informative prior distribution for p we put,

$$p \sim Uniform(0,1)$$

The term re-occurrence model also contains known quantities that specify features of a document. For the i^{th} document these are cen_i, d_i and n_i: cen_i depends on the document length d_i and the number of occurrences of the term in that document, n_i. The dependencies between these quantities and the model parameters is shown in the graphical model in Figure 7.1. Fitting mixture techniques is tricky and requires special methods. Data augmentation techniques are used to to make it feasible to fit the model using MCMC techniques discussed in section 7.4.

7.4 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are a form of simulation technique that generate values of the parameters $\vec{\Theta}$ from a density function $f(\vec{\Theta}|\vec{W})$. The methods can be used when $f(\vec{\Theta}|\vec{W})$ is complicated and often, in fact, $f(\vec{\Theta}|\vec{W})$ itself cannot be specified. MCMC methods

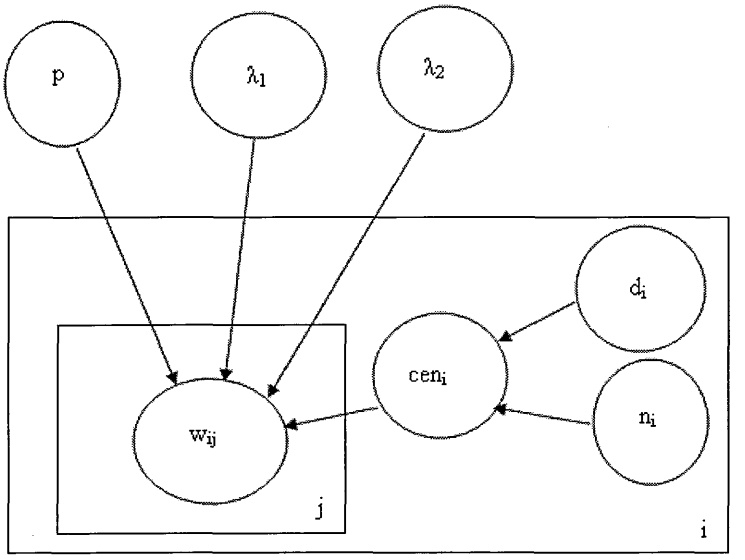

Figure 7.1: Bayesian dependencies between the parameters

only require the joint distribution of the parameters $\vec{\Theta}$ given the data $\vec{W}$. The samples generated from the MCMC method may be used to calculate the mean, median and other quantiles of the distribution, or any other features of the distribution.

MCMC methods work by generating a large sample of observations from the joint distribution $f(\vec{\Theta}|\vec{W})$. The integrals of the complex distributions can be approximated from the generated data. The values are generated based on the Markov chain assumption, which states that the next generated value only depends on the present value and does not depend on the values previous to that. Based on mild regularity conditions, the chain will gradually *forget* its initial starting point and will eventually converge to a unique *stationary distribution.*

The software package `WinBUGS` [STBL03] has been used for running the MCMC sampling and obtaining the parameter estimates. In the following sections, some techniques commonly used by MCMC methods are discussed, followed by the approach adopted by `WinBUGS` for sampling parameter values. The convergence criteria adopted during the sampling process and the analysis

of the output from the `WinBUGS`.

7.4.1 Gibbs Sampling

Gibbs Sampling is the simplest and most widely used MCMC method. It was first used for analyzing Gibbs distribution on lattices, and derives its name from there. The Gibbs sampler iteratively draws samples from the full conditional distribution of each parameter in the distribution. The *full conditional distribution* for a particular parameter is a function of the parameter derived from the likelihood given all other parameter values.

For the term re-occurrence model, $\vec{\Theta} = \langle p, \lambda_1, \lambda_2 \rangle$ denotes the parameters of the model. $\vec{W} = \langle w_{ij} \rangle$, for $j = 1, \ldots, n_i, n_i + 1$ and $i = 1, \ldots, N$ is the data, where N denotes the number of documents in the dataset and n_i denotes the number of occurrences of the particular term of interest in the i^{th} document. Hence the joint probability density of likelihood function is defined as:

$$f(\vec{W}|\vec{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i+1} w_{ij}$$

Since, each w_{ij} is a mixture of exponential distributions with parameters, $\langle p, \lambda_1, \lambda_2 \rangle$, the likelihood function is also a function of these parameters, $\langle p, \lambda_1, \lambda_2 \rangle$. In principle, the marginal distribution of each parameter is obtained by integrating the likelihood function over the other parameter values. However, the integration cannot be performed analytically and instead, Gibbs sampling is used to generate a large sample of values from each marginal distribution.

In Gibbs sampling, initial random values are assigned to each of p, λ_1 and λ_2 and these values are then iteratively updated to obtain new values of $\langle p, \lambda_1, \lambda_2 \rangle$. To obtain the updated value for a certain parameter, the existing values of all the other parameters are put into the conditional distribution of the parameter of interest. For example the initial values of λ_1 and λ_2 are put in the

conditional distribution of p to obtain the updated value of p. This process is repeated for every parameter in turn. Since, this iterative process takes some time to settle down or to achieve a stationary state, some early values are discarded until the parameters reaches a stable state. After stability is reached, further sampling in a similar manner is carried out and these parameter values are used for estimating the parameters.

Straightforward Gibbs sampling requires conditional distributions of parameters. It is not possible to obtain a closed form conditional distribution for the current model. This is because the likelihood function is a product of mixture of exponential distributions and added to that is the censoring of certain observations. This leads to a complex form of likelihood function, with many intermingling terms of parameters. To overcome this problem, data augmentation is used.

7.4.2 Data Augmentation for Mixture Models

Working with mixture models, like the term re-occurrence model in this case, one often ends up with a very complex conditional distribution from which sampling is not convenient and straightforward. Hence for the purpose of Gibbs Sampling, often dummy variables are introduced to create conditional distributions which are convenient for sampling [Rob96]. This practice of introducing additional dummy variables to the model is known as *data augmentation.*

For the term re-occurrence model, a dummy integer value $M_{i,j}$ is introduced to identify which of the two exponential distributions generated a certain observation w_{ij}, with the following probabilities:

$$M_{ij} = \begin{cases} 1 & \texttt{with probability} \quad p \\ 2 & \texttt{with probability} \quad 1-p \end{cases}$$

Suppose w_{ij} is generated from the first exponential distribution with parameter λ_1, then information about that particular w_{ij} is used for estimating value of λ_1. So the w_ij's that originate from the

first exponential distribution can be separated out conditioned on the dummy indicator variable M_{ij}. Similarly, some other w_{ij} values are used for estimating the value of λ_2. p estimates the proportion of cases w_{ij} is generated from the first exponential distribution.

Conditioned on the newly introduced dummy variables $M_{i,j}$, the likelihood function can be represented in the following form:

$$f(\vec{\Theta}|\vec{W},\vec{M}) \equiv \pi_1(p) \times \pi_2(\lambda_1) \times \pi_3(\lambda_2) \times \pi_4(\vec{M})$$

where, $\pi_2(\lambda_1)$ is some functional form that involves only the parameter λ_1 and none of the other parameters. The introduction of the dummy variable means there is now an extra parameter $\vec{M}$ to be estimated in the Gibbs sampling process. Factoring the likelihood function in this form, leads to a more manageable approach to sampling parameters. Values are sampled from the dummy variable $M_{i,j}$ also, and these values might be helpful for classifying the observations with respect to the mixture components.

7.4.3 Using `WinBUGS` for MCMC Sampling

`WinBUGS` [STBL03], the Windows version of the statistical package, **BUGS (Bayesian inference Using Gibbs Sampling)**, has been used for estimating parameters of the model. The term re-occurrence model is represented by means of a BUGS routine, which is then compiled and the observed gap values for a certain term are supplied to the model. Initial starting parameter values also have to be given.

Based on the supplied data and the model in hand, `WinBUGS` decides upon the sampling strategy, from a collection of available techniques, to be adopted for every node in the model. `WinBUGS` adopts the following steps to decide upon the sampling strategy (obtained from the `WinBUGS` manual):

- If the functional form of the prior distribution and its conjugate prior is known, sampling is done directly from that distribution.
- Otherwise, if the functional form of the conditional distribution is derived and is log-concave, then Gibbs sampling is done from the full conditional when it is possible to sample from that distribution.
- Slice sampling is used for non log-concave densities on a restricted range.
- In the case of complex unrestricted range densities the Current point Metropolis method is used.

For discrete target distributions, an inversion based method is used for densities with a finite upper bound and direct sampling using a standard algorithm is used for shifted Poisson densities.

For the term re-occurrence model, the unknown parameters of the model are $\{p, \lambda_1, \lambda_2\}$. The random variable p, used to decide which of the two exponential distributions generated the data has a Bernoulli distribution, and hence its conjugate prior is the well known Beta distribution [GCSR95]. Hence to generate samples from the distribution p, `WinBUGS` samples from a Beta distribution. The parameters λ_1 and λ_2, which are the parameters of the exponential distributions, do not have any closed form conditional distribution. Hence polynomial based approximation methods are used to derive the full conditional distributions. And according to `WinBUGS`, Derivative Free Adaptive rejection Sampling [Gil92] is used for sampling from the distributions of λ_1 and λ_2.

7.4.4 Analyzing `WinBUGS` Output and Convergence Criteria

The Bayesian sampling software that has been used for parameter estimation for the present model is `WinBUGS`. The model has to be specified to `WinBUGS` in a language specific to the software. After following the steps in Table 7.1, `WinBUGS` presents output statistics for each unknown node in the

model. The steps followed in WinBUGS are specified in the following order:

1. Specify the model to WinBUGS in the software specific language.
2. WinBUGS checks and validates the supplied model and reports errors, if any.
3. The observed data are supplied to the model.
4. WinBUGS compiles the model, along with the supplied data and reports discrepancies, if any. At this point, WinBUGS forms an internal graph based on the supplied data and the model parameters.
5. Initial starting values for each of the unknown parameters are specified.
6. If an initial value for a node in the graph is not specified, WinBUGS chooses a value.
7. WinBUGS updates the parameter values, based on the user specified number of updates.
8. To aid inference, WinBUGS provides a range of plots, which have to be examined to decide if the model parameters have converged.
9. Parameter estimates are obtaining by averaging over the parameter values obtained during the update process. These values are used as model parameter estimates for further analysis.

Table 7.1: Steps followed in WinBUGS to obtain parameter estimates given the observed data.

The term re-occurrence model is represented in a language specific to WinBUGS (see Appendix B). Ideally, one would wish to specify some convergence criteria that would decide the amount of *burn-in* simulations required for the model to converge, and collect data on the model parameters after convergence has been achieved. However there is no available and established methodology for determining the convergence criteria. A range of convergence diagnostics [CC96] have been proposed, but due to the complications and inaccuracies of the methods, visual techniques are used for the term re-occurrence model. Also, these methods are mainly based on the MCMC output values along with other theoretical and practical assumptions. Visual inspection of the MCMC plots are the most obvious and commonly used techniques for determining convergence diagnostics. Hence, in the following sections visual inspection procedures of various MCMC plots from the WinBUGS software are discussed.

The model is fitted to various terms in different datasets and parameter estimates are obtained.

The model fitting first requires the convergence of the model (*burn-in* simulations). As discussed above the exact convergence point is best judge based on visual analysis of the various plots. But for the proposed term re-occurrence model, it has been observed that about 1000 simulations are enough for the model to converge. So, the first 1000 simulation values are discarded as *burn-in* simulations. A further 6000 simulations are performed, and the parameter values from these 6000 simulations are used to estimate the parameter values.

In total 7000 (1000+6000) simulations are performed for each term of interest. The simulations were performed on a *Windows* desktop computer with $1G$ of *RAM* (Random Access Memory) using the `WinBUGS` software package. The time required for fitting a model for a particular term was dependent on the frequency of the term's occurrence in the dataset and the size of the dataset. Due to hardware limitations it was not possible to consider all the documents in a certain dataset for fitting the model. Hence, depending on the size of the dataset, some $1\% - 5\%$ of the unique documents were picked at random and used for model building. Since the documents were picked at random from the entire pool of all the documents, it is assumed that documents of different characteristics will be represented in the random sample. For a rarely occurring term it is especially ensured that some of the randomly picked documents contains the term of interest. To consider more documents for model building we would require much larger *RAM*, for the `WinBUGS` software to store and update values in the dependency graph (Figure 7.1). Based on this random document selection, fitting the model for a very rarely occurring term in a dataset requires about $1 - 5$ minutes, for a mid-frequency term the modeling requires about $5 - 15$ minutes, whereas very frequently occurring top function words would require about $60 - 90$ minutes (i.e. $1 - 1.5$ hours) for the completion of the 7000 simulations in fitting the model. Overall, the time required to fit the model is directly related to the number of times (including frequency) the term occurs in the dataset and in the documents used for the modeling.

The computational cost of the approach has consequences. First of all, it is impossible to run the model on all terms in a sizeable collection, and this limits the methods we can use for verifying the model. We will address this in Chapter 8. Second, the computational cost also puts limits on possible practical applications, depending on whether they need to run in real time, on the amount of data coverage required, and on the computational resources available. These are not limitations inherent in the model, but they are associated with the software and methodologies available for estimating the parameters.

To aid better understanding and clarity of the visual inspection mechanism, the discussion will center around an example for the term "*church*" in the Associated Press (AP) dataset. To clarify the notion of convergence, two sets of MCMC simulations were run; in the first run 7000 simulations were run to obtain the parameter estimates and the plots, whereas in the second set 1000 simulations were used as *burn-in* and a further 6000 simulations are used for parameter estimates and obtaining the plots. The first run of 7000 simulations is expected to provide poor estimates as no *burn-in* has been set. This claim is validated based on the output obtained from the `WinBUGS` software. The various plots and parameter estimates from the two separate runs are compared in subsequent sections. In all the plots the four model parameters are plotted, viz $P[1]$ and $P[2]$ corresponding to p and $1-p$, $lambda[1]$ corresponding to λ_1 and $lambda[2]$ corresponding to λ_2.

7.4.4.1 History Plots

History plots are plots of the parameter values obtained in each simulation along a time series, which is the number of iterations. This section highlights how the history plots might be used to judge convergence of the `WinBUGS` output. A parameter is assumed to have reached convergence when there is no visible trend in the time series plot and any variation observed in the parameter value across iterations seems purely random. Figure 7.2 shows plots for parameter values based on

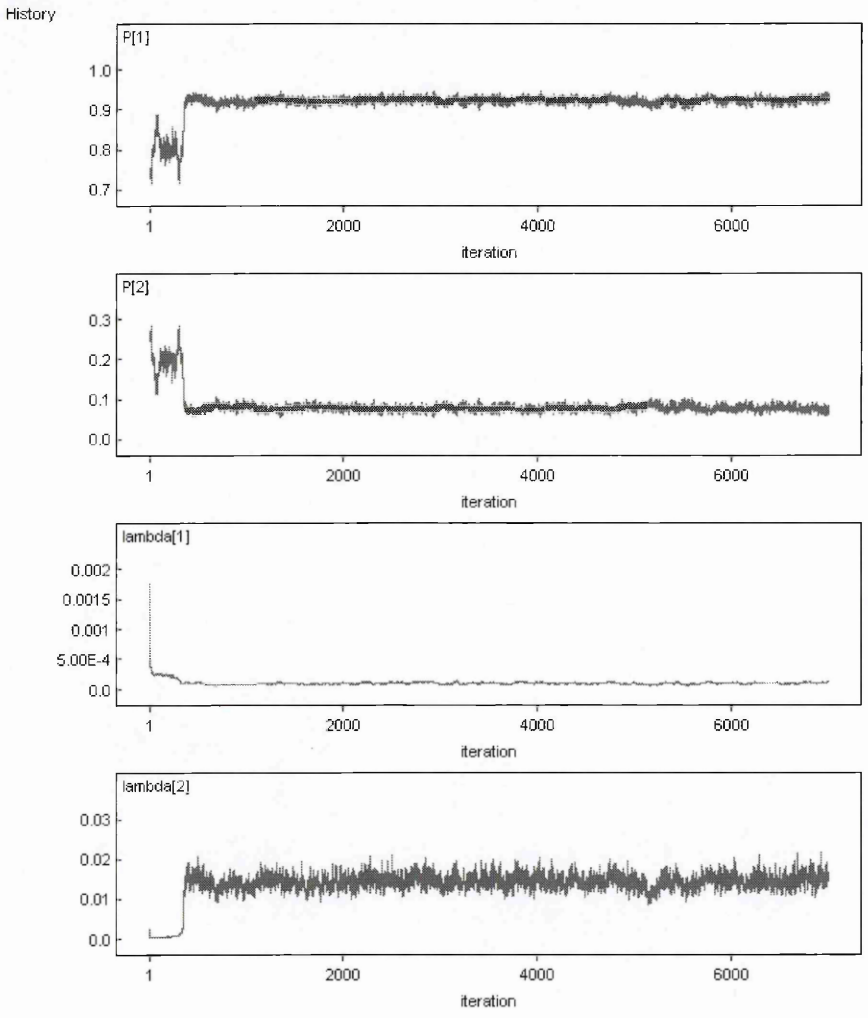

Figure 7.2: Plot history for the first run of 7000 simulations.

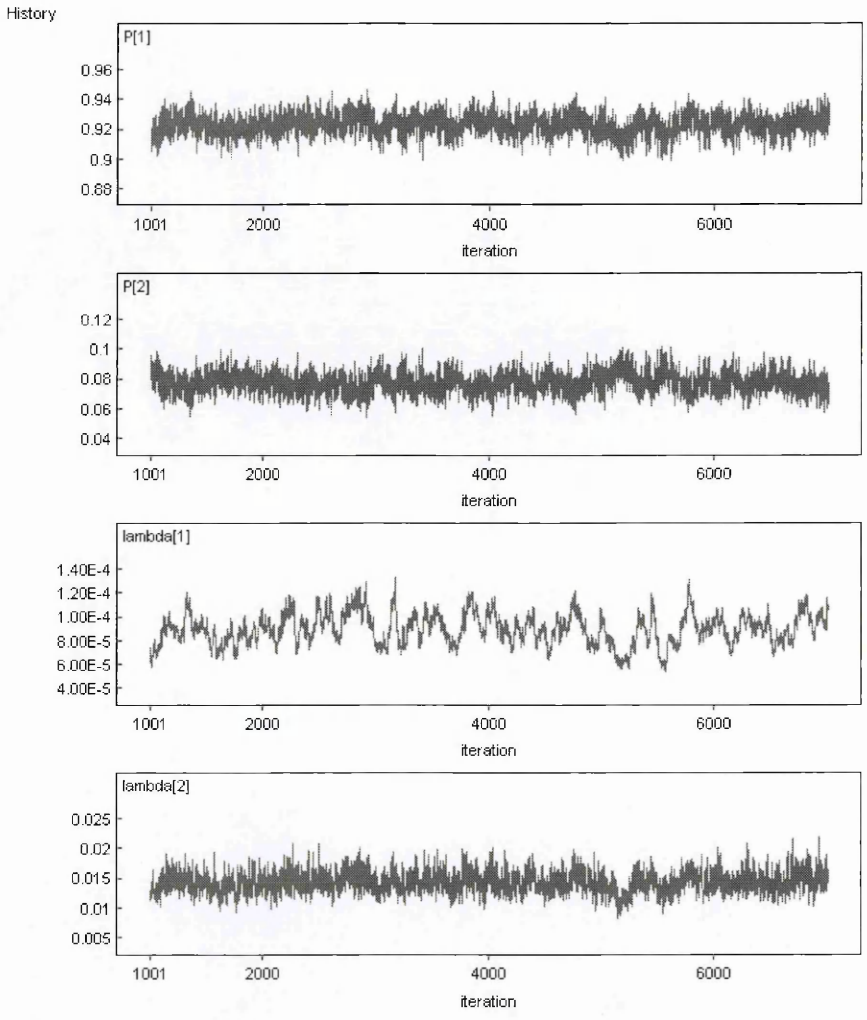

Figure 7.3: Plot history for the second run of 6000 simulations.

the first 7000 iterations. One may observe a trend in the first few values. In the second experiment, 1000 iterations are used as burn-in and are discarded, then a further 6000 iterations are run and the history plot is obtained (Figure 7.3). It may be observed that there are no visible trends in the plot, and any variation that occurs in the parameter values appear to be purely random. So, the plots in Figure 7.3 represents a case where convergence has been reached before the start of the plotted values.

7.4.4.2 Auto Correlation Plots

Auto Correlation plots measure the correlation of the time series obtained during the simulation with another similar time series starting at a previous time. The difference in the present time and the previous time based on which the auto correlation is calculated is called the "lag". The auto correlation plot provides evidence about the presence of cyclic and seasonal trends in the obtained simulation pattern. For example, if a certain trend in the simulation values is observed after every 30 simulations, then an auto correlation of lag 30 would detect that pattern. It is called *Auto Correlation*, because the value obtained is a "correlation" between a term and its own value at a previous simulation, hence the term "auto". `WinBUGS` provides Auto Correlation plots for the present time series with a lag ranging from 1 to 50. Auto correlation values are in the range -1 to 1, where a high value close to $|1|$ (absolute value, i.e. either close to -1 or 1) indicates high correlation between the two time series, whereas a value close to 0 indicates low correlation. If an auto correlation function plot (Figures 7.4, 7.5) drops quickly it indicates that convergence is not a problem. If it drops slowly, it means that convergence is slow. A possible reason for this might be inappropriate mixing of the parameter values, which means that the parameter values are not being picked by the sampler from the entire possible range. In cases where convergence is slow it is recommended not to record every picked value but to record values after a fixed interval, say every 10^{th} value. This method is called *thinning* and saves storage by eliminating many similar values that do not add much extra information.

The Auto Correlation plots of the model parameters for the first 7000 simulations (Figure 7.4), indicates high values close to 1. This is because a *burn-in* phase is needed, as was shown by the history plots.

The Auto Correlation plots for the second set of experiments (Figure 7.5) provides low auto

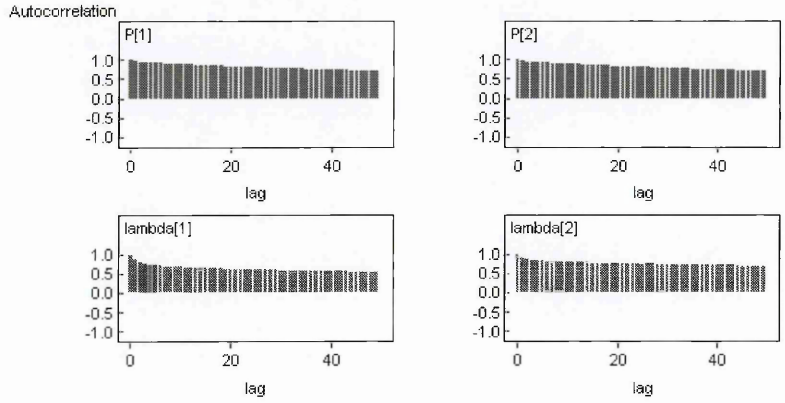

Figure 7.4: Auto Correlation plots of model parameters for the first run of 7000 simulations.

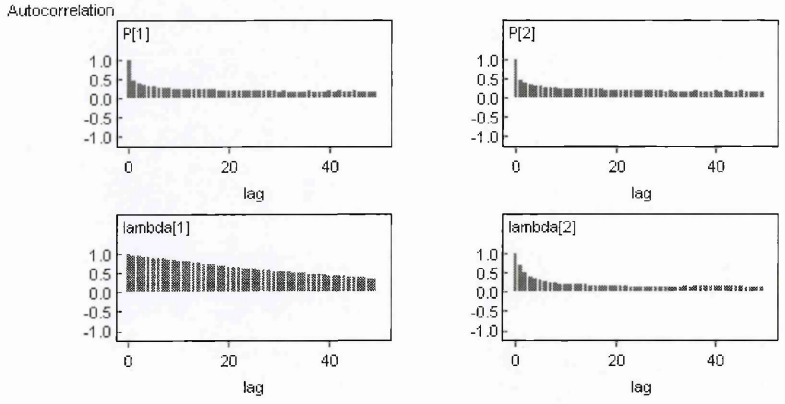

Figure 7.5: Auto Correlation plots of model parameters for the second run of 6000 simulations.

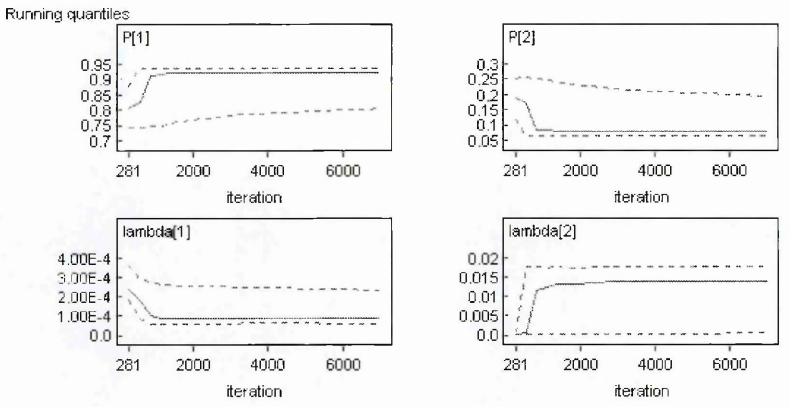

Figure 7.6: Quantiles plots of model parameters for the first run of 7000 simulations.

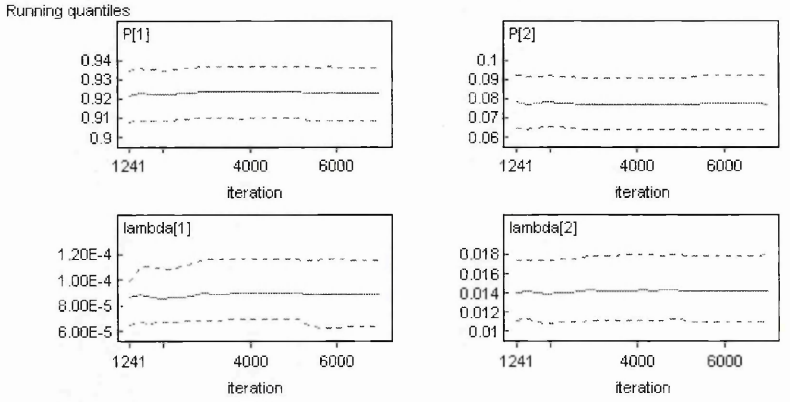

Figure 7.7: Quantiles plots of model parameters for the second run of 6000 simulations.

correlation values for the parameters P[1], P[2] and lambda[2] indicating that the chain is mixing well and the length of the *burn-in* phase is adequate. The auto correlation values for lambda[1] are in the moderate range, indicating a slow rate of convergence for this parameter as compared to the others.

7.4.4.3 Quantiles Plots

Running quantiles plots give the running mean for the parameter estimates with running 95% confidence intervals against iteration number. These plots are not used for checking the parameter convergence of the model, although their values should be stable after convergence.

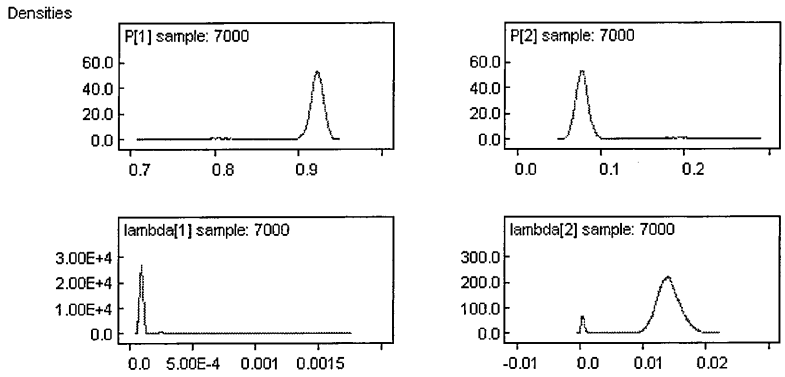

Figure 7.8: Density plots of model parameters for the first run of 7000 simulations.

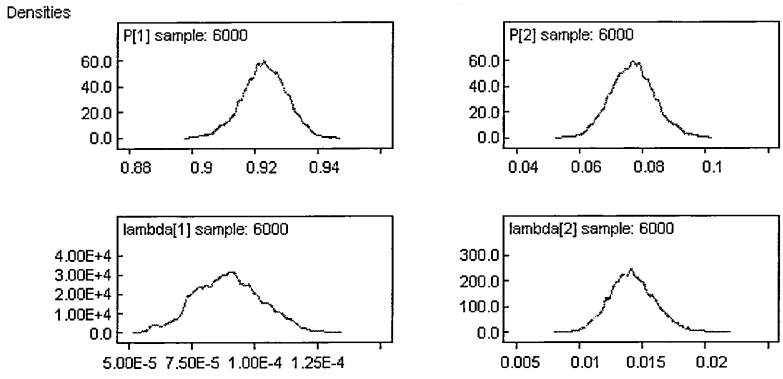

Figure 7.9: Density plots of model parameters for the second run of 6000 simulations.

The quantiles plots for the model parameters for the first 7000 simulations where convergence has not been reached (Figure 7.6) show initial instability. On the contrary, the quantiles plots for the second set of experiments where convergence has been reached (Figure 7.7) show much greater stability, although a slight deviation in the lambda[1] plot is observed possibly suggesting that a longer *burn-in* would provide better estimates.

7.4.4.4 Density Plots

In Bayesian statistics, unlike frequentist statistics, each of the model parameters follows some distribution. The density plots allow us to visualize the distribution obtained from the sampling procedure. They provide an understanding of the shape of the distribution and might be bene-

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
P[1]	0.9169	0.02765	0.002823	0.8046	0.9224	0.9361	1	7000
P[2]	0.08314	0.02765	0.002823	0.06396	0.0776	0.1955	1	7000
lambda[1]	9.572E-5	4.164E-5	3.814E-6	6.142E-5	8.96E-5	2.342E-4	1	7000
lambda[2]	0.01354	0.003443	3.322E-4	4.372E-4	0.01403	0.01789	1	7000

Figure 7.10: Summary Statistics of model parameters for the first run of 7000 simulations.

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
P[1]	0.9229	0.007038	3.806E-4	0.9084	0.923	0.9362	1001	6000
P[2]	0.07709	0.007038	3.806E-4	0.0638	0.07698	0.09165	1001	6000
lambda[1]	8.949E-5	1.308E-5	1.195E-6	6.426E-5	8.928E-5	1.153E-4	1001	6000
lambda[2]	0.01426	0.001773	9.156E-5	0.01098	0.01418	0.01793	1001	6000

Figure 7.11: Summary Statistics of model parameters for the second run of 6000 simulations.

ficial for improving upon the specified prior distribution. These plots are not used for checking convergence. This density function for each parameter in the model is based on the parameter's values during the simulation process. The density plots from the first experimental run of 7000 simulations are shown in Figure 7.8.

Contrast these with the density plots of 6000 simulations (Figure 7.9), where the first 1000 simulations have been discarded as *burn-in*. The long tails of the distributions and the second peak for lambda[2] have disappeared.

7.4.4.5 Statistics or Parameter Estimates

In the previous sections various `WinBUGS` plots were discussed, along with ways of detecting convergence based upon some of them. In practice either the history and auto correlation plot or a combination of both is used to check convergence of the simulations.

Once convinced that the simulation has converged, summary statistics of the parameters of interest are obtained and used for further studies. The example summary statistics of the two experiments are reported in Figure 7.10 for the first experiment with 7000 simulations where convergence has not been reached and Figure 7.11 for the second experiment with 1000 iterations as

burn-in and a further 6000 iterations have been used for obtaining the parameter estimates. There is also a visible difference in the parameter estimates obtained on the two experiments based on similar data sets.

7.5 Summary

This chapter introduces the Bayesian statistics theory and its advantages and flexibility for model fitting as compared to traditional frequency based methods. Data augmentation was used for the ease of fitting the mixture model. Markov Chain Monte Carlo based simulation methods were used to obtain the parameter estimates. `WinBUGS` software was used for fitting the model and obtaining the estimates. Time taken to fit the model is discussed in this chapter. Techniques for analyzing the output from the `WinBUGS` software and judging the convergence of the parameter values based on these outputs were also discussed in this chapter. Once the parameter estimates are obtained, the model can be applied to various applications to judge its quality compared to frequency based methods. In Chapter 8, three different applications are described that use the term re-occurrence model. Performance of these applications is also compared using frequency based methods alone. Further application areas where this model might be applied are suggested in Chapter 9.

Chapter 8

Applications based on the term Re-occurrence Model

This chapter discusses ways in which the term re-occurrence model may be applied. The applications aim to answer a common question i.e. what useful information can the term burstiness model tell us about the term's character as compared to frequency based measures?

We require a methodology for verifying or evaluating the model developed in Chapter 6, and gauging the improvements the model can bring in the context of an application domain. One option would be to build the model into an application that might benefit from information on burstiness, and run a standard evaluation exercise, comparing implementations with and without the burstiness information. Several established application domains like information retrieval or stylistic analysis have established methodologies for evaluation, as will be touched upon in subsequent sections. However, this option is not viable for us because of the computational load of calculating the parameters for large number of terms from sizeable datasets (section 7.4.4). As a consequence, it is not possible to follow standard routes for evaluation.

The alternative evaluation route adopted in this thesis is to pick a small set of terms from the datasets and evaluate the effect in the context of the application on the chosen terms. Such a practice is observed in other work of term distribution modeling [CG95b, Kat96, Chu00], where the computational limitations had limited the established evaluation of the technique. The model developed in chapter 6 is aligned with the other existing models of term burstiness [CG95b, Kat96], which we seek to align with and extend. The approach we take is further supported by the fact we sought to use datasets, and data on terms, which were also explicitly used in other, published, work, so that we can compare our outcomes with that of earlier related research. Hence, the thesis will investigate the behaviour of the proposed model in the context of three application areas, and will focus on the terms and datasets that have been explored in the literature, especially in the work of Church [CG95b, Chu00] and Katz [Kat96]. The evaluation based upon the application areas will be discussed in the following sections along with the methodology adopted in this thesis for evaluating the model.

For better understanding of the applications, interpretation of the parameter values are discussed. In the first application, a range of terms from the Associated Press (AP) dataset are studied with the term re-occurrence model, with the burstiness characteristics of the terms reflected in the model parameters. The model aids in understanding the usage pattern of the terms within the dataset. In the second application a set of style indicative terms were hand picked and the characteristics of these terms compared across datasets of different genres. The model parameters for these terms were used to identify variation across the different genres. In the third application the very frequent terms for each of the TIPSTER datasets were studied. Very frequent terms are usually considered to be more evenly distributed and less informative because they occur copiously in all documents, and hence are removed as background noise. This study investigates whether the very frequent function words are indeed homogeneously distributed or have a bursty character. In many

cases, the findings from the model were verified by manually screening through the documents in which the term occurred. Appendix A contains some of the example documents that shall be used to support various facts/arguments in this chapter.

Much of the work in this chapter has been previously [SGDR05, SDRG05, DRSG05].

8.1 Interpretation of Parameters

The parameters of the model lend themselves to interpretation in the following manner:

- $\widetilde{\lambda_1} = 1/\lambda_1$ is the mean of an exponential distribution with parameter λ_1. $\widetilde{\lambda_1}$ measures the rate at which this term is expected to come up in a running text corpus, or the rate at which a term is expected to enter a burst. $\widetilde{\lambda_1}$ measures the distance between bursts between documents or the distance between bursts within a longer document. $\widetilde{\lambda_1}$ might be used to determine the rarity of a term in a corpus, as it is the average gap at which the term occurs if it has not occurred recently. Thus, a large value of $\widetilde{\lambda_1}$ tells us that the term is very rare in the corpus and vice-versa.

- Similarly, $\widetilde{\lambda_2}$ measures the *within-burst term re-occurrence rate*, i.e. the rate of re-occurrence of a term given that it has occurred recently or within a burst. Typically, it measures the term re-occurrence rate in a burst within a document. Small values of $\widetilde{\lambda_2}$ indicate a bursty nature of the term.

- $\widetilde{p}$ and $1 - \widetilde{p}$ denote, respectively, the probabilities of the term occurring with rate $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ in the entire dataset. $\widetilde{p}$ denotes the proportion of time a gap between a term's occurrence is generated from the first exponential distribution with parameter $\widetilde{\lambda_1}$. Non occurrence of a term in the entire document is also considered as a gap in the model with the observation censored. There will be many such large gaps for rarely occurring terms in dataset which

accounts for high values of $\widetilde{p}$ for those terms.

Table 8.1 presents some heuristics for drawing inference based on the values of the parameter estimates.

	$\widetilde{\lambda_1}$ small	$\widetilde{\lambda_1}$ large
$\widetilde{\lambda_2}$ small	very frequent occurring like a function word	word occurring in bursts spaced at large intervals, possibly a topical content word
$\widetilde{\lambda_2}$ large	comparatively frequent but well-spaced possibly function word	infrequent and scattered word

Table 8.1: Heuristics for inference, based on the parameter estimates.

Based on the parameters and their interpretation, the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic seems useful for inference.

8.2 Behaviour within a collection: Informing Retrieval

In the context of information retrieval, we might seek to evaluate whether the term re-occurrence model can provide improved term weighting techniques for indexing. This might be evaluated in the following manner. The distribution of all terms in the entire dataset are modeled by the term re-occurrence model and indexed using some variation of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic. As a base-line comparison, all terms in the dataset are also indexed using a standard term weighting method, such as TF-IDF [Aiz03, SB88] measure. An *ad-hoc* retrieval task could be performed using a fixed set of queries from say, TREC [TRE] evaluation track. Outcome from the two indexing approaches could then be compared to evaluate performance of the term re-occurrence model. This approach is not open to us given computational constraints. Instead we propose to carry out an alternative evaluation using a small set of terms whose burstiness characteristics we can calculate given out

method.

This approach aims also at evaluating the term re-occurrence model for analyzing term distribution characteristics in a certain dataset [SGDR05]. We choose terms from the *Associated Press (AP)* newswire articles, a standard corpus for language research which had been used previously in the literature [CG95a, Chu00, MS99, UC00] with respect to modeling different term distributions. These terms provide a comparative picture with respect to the findings by other researchers. To verify the claim that the model can handle both frequent function terms and rare content terms, terms were chosen accordingly. Some medium frequency terms were also chosen to demonstrate their characteristics.

8.2.1 Parameter estimates

Table 8.2 shows the parameter estimates for the chosen terms (it does not show the values of $1 - \widetilde{p}$ as they can be obtained from the value of $\widetilde{p}$). Some other basic statistics of these terms' behaviours are also noted. These include the total frequency of a term in the dataset (`Total Freq` in the Table), number of documents where the term occurs (`Num docs` in the Table) and Inverse Document Frequency which is frequently used for judging a term's importance [SJ72](`IDF` in the Table). The value $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ is included since it gives an indication of the differences in burstiness that can be inferred from $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ for each term, and the rows in the table containing terms are sorted on the basis of that value.

The top part contains mostly words that are usually classed as frequent (function) words. The middle region contain terms that are obviously not very high frequency ones, but that are not necessarily immediately recognizable as topical. The bottom part contains terms that occur rarely and mark content. Based on traditional term frequency data alone, `IDF`, overall frequency and number of documents the term occurs in are the only means to evaluate the term's importance in

the dataset. These measures can correctly identify terms like *the*, *of*, *and* as frequent (function) words and terms like *noriega*, *boycott*, *vietnam* as rarely occurring content words. However, using frequency information alone suggests that, rarely occurring terms like *except*, *follows*, *yet* and *somewhat* are equally important terms carrying content information. Term burstiness information makes it possible to identify such terms as rarely occurring and scattered, possibly function words.

Term	Total freq	Num docs	IDF	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
the	6990645	242329	1.00	0.82	16.54	16.08	1.03
and	2500696	236957	1.03	0.46	46.86	45.19	1.04
of	3039809	239180	1.02	0.58	38.85	37.22	1.04
except	7648	7205	33.72	0.67	21551.72	8496.18	2.54
follows	2342	2280	106.54	0.56	80000.00	30330.60	2.64
yet	17706	15863	15.31	0.51	10789.81	3846.15	2.81
he	757301	163884	1.48	0.51	296.12	48.22	6.14
said	1448540	218399	1.11	0.03	895.26	69.06	12.96
government	192934	78016	3.11	0.60	1975.50	134.34	14.71
somewhat	3011	2899	83.79	0.84	75244.54	4349.72	17.30
federal	88847	46235	5.25	0.84	2334.27	102.57	22.76
here	45808	34684	7.00	0.94	3442.34	110.63	31.12
she	164030	48794	4.98	0.73	1696.35	41.41	40.97
george	27562	21135	11.49	0.88	17379.21	323.73	53.68
bush	138290	30577	7.94	0.71	3844.68	53.48	71.90
soviet	124836	28570	8.50	0.71	4496.40	59.74	75.27
kennedy	12201	5911	41.09	0.78	14641.29	99.11	147.73
church	29685	10484	23.17	0.92	11291.78	70.13	161.02
book	17121	8354	29.08	0.92	17143.84	79.68	215.16
vietnam	14174	4971	48.87	0.92	32701.11	97.66	334.86
boycott	3427	2076	117.01	0.98	105630.08	110.56	955.42
noriega	18238	2936	82.74	0.91	86281.28	56.88	1516.82

Table 8.2: Parameter estimates of the model and basic statistics for some selected terms, sorted by the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ value (ascending)

The top part of the table consists of the very frequently occurring function words occurring several times throughout the corpus. Along with the overall statistics and IDF their status as function word is supported by the low values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$. These values are quite close, indicating that the occurrence of these terms shows low burstiness in a running text chunk because average

gap length between-bursts and within-bursts is similar. This supports our heuristics about the value of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$, which is small for such terms. Moderate, not very high values of $\widetilde{p}$ also support this statement, as the term is then as likely to be generated from either of the exponential distributions (*the* has high value of $\widetilde{p}$, but since the values of λ are so close, it doesn't really matter which distribution generated the observation). We observe large values of both $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ for terms like *yet*, *follows* and *except* which indicates large gaps between and within bursts. The large value of $\widetilde{\lambda_1}$ can be explained, as these terms are rarely occurring words in the dataset. They do not occur in bursts and their occurrences are scattered, so values of $\widetilde{\lambda_2}$ are also large (Table 8.1). Interestingly, in our heuristic, these large values nullify each other to obtain a small value of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$, because the within-bursts and between-bursts difference between these terms is minimal, as a result of that the terms like *yet*, *follows* and *except* do not have a bursty characteristic.

The middle part of the table contains mostly *non-topical content terms* as defined in the literature [Kat96]. These terms do not describe the main topic of the document, but some related aspect of the document or a nearby topical term. For each of the terms listed, manual inspection of the documents containing these terms was carried out and some examples are in Appendix A. For instance, in a document about *George Bush* the term *bush* occurs many times but *george*, also referring to *bush* occurs only a few times. The complete name is mentioned in the beginning and further references to *George Bush* are made using the word *bush*. Our intuition is to class *bush* as a topical term, but not *george*. The $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ casts *bush* as a comparatively more content bearing term as compared to *george*. Similarly, the term *government* is a general concept, and refers in some newswire articles to some government in any state or any country, future references to which are made using this term. Similarly the term *federal* is often used to make references to the *US Government* (see Appendix A for an example document with words *government* and *federal*). As the words *federal* and *government* are used frequently for referencing, they exhibit comparatively

small values of $\widetilde{\lambda_2}$ (within-burst gaps are small). We were surprised by the occurrence of terms like *said*, *here* and *she* in the middle, as they are commonly considered as function words (Appendix B). Closer examination revealed that *said* has some dependence on the document genre in the dataset in question, with respect to the content and reporting style. Articles reporting a statement or some conversation would have a higher incidence of the term *said* as compared to other documents. The data were based on newswire articles about important people and events. The majority of such people appearing in AP documents are male, hence there are more articles about men than women (*he* occurs 757,301 times in 163,884 documents as the 13^{th} most frequent term in the corpus, whereas *she* occurs 164,030 times in 48,794 documents as the 70^{th} frequent term). This explains why *he* has a smaller value of $\widetilde{\lambda_1}$ than *she*. But the $\widetilde{\lambda_2}$ values for both of them are quite close, showing that they have similar usage pattern. Again, newswire articles are mostly about people and events, and less often about some location, referenced by the term *here*. This explains the large value of $\widetilde{\lambda_1}$ for *here*. Again, because of its usage for referencing, it re-occurs frequently while describing a particular location, leading to a small value of $\widetilde{\lambda_2}$. Possibly, in a collection of "travel documents", *here* might have a smaller value of $\widetilde{\lambda_1}$ and thus occur higher up in the list, which would allow the model to be used for characterizing genre.

Terms in the third part, as expected, are *topical content terms*. Frequency information alone can pick these terms as being rare in the dataset and hence possibly carrying content information. The fact that a topical content word tends to re-occur several times in close vicinity [Kat96] cannot be detected based on the frequency counts alone. An occurrence of such a term defines the topic or the main content word of the document or the text chunk under consideration. These terms are rare in the entire corpus, and only appear in documents that are about this term, resulting in very high values of average between-burst gaps, $\widetilde{\lambda_1}$. Also low values of average within-burst gaps, $\widetilde{\lambda_2}$ for these terms mean that repeat occurrences within the same document or within different bursts in a

long document are quite frequent; the characteristic expected from a topical content term. Because of these characteristics, our heuristics assign these terms very high values of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$, and hence class them as informative terms in the dataset (in the context of retrieval or topic based classification).

8.2.2 Case Studies

Here the behaviour of pairs of terms are contrasted and studied using the term re-occurrence model. We chose these term pairs because they have been studied before by other researchers. These terms were chosen to compare our approach with previous work and also to explore the range of inferences that may be derived from the model.

8.2.2.1 somewhat *vs* boycott

These terms occur an approximately equal number of times in the AP corpus, so frequency cannot be used to distinguish between them. *Inverse document frequency* (IDF) was used to set them apart [CG95a]. In Table 8.2, the frequency based measure of IDF provides some indication that *somewhat* is less important in the dataset with respect to *boycott*, but the IDF values do not set them apart very distinctly, and would not be able to separate their behaviour from that of *follows* very clearly. The term re-occurrence model highlights extra information about these terms' behavior that would fall beyond the scope of the other frequency based measures. It gives approximately similar rates of occurrence ($\widetilde{\lambda_1}$) for these two terms as shown in Table 8.2, but the re-occurrence rate (within-burst gaps), $\widetilde{\lambda_2}$, is 110.56 for *boycott*, which is very small in comparison with the value of 4349.72 for *somewhat*. The term re-occurrence model and the heuristic class *somewhat* as a rare word occurring in a scattered manner over the entire dataset. The term *boycott* is identified as a topical content word, as it should be.

8.2.2.2 follows *vs* soviet

These terms were studied in connection with fitting Poisson distributions to their distribution [MS99], and with a view to determining their characteristics[1]. Fit to the Poisson distribution [MS99] could differentiate between these terms with *follows* as a function bearing term and *soviet* as a content bearing term. The term re-occurrence model, shows that *follows* has large values of both $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ (Table 8.2), so that it has the characteristics of a scattered term. *Soviet* has a large $\widetilde{\lambda_1}$ value and a very small $\widetilde{\lambda_2}$ value, so that it has the characteristics of a topical content word. The findings from our model agree with the original work on the assumption that the terms have similar behavior in the *New York Times* articles as compared to the *Associated Press* articles.

8.2.2.3 kennedy *vs* except

Both these terms have nearly equal *inverse document frequency* for the AP dataset [Chu00, UC00] and will be assigned equal weight using standard frequency based approaches in any retrieval or classification based application. [Chu00, UC00] used a measure based on average-term frequency to determine the nature of the term. This measure (originally proposed by [Kwo96]) takes the value 1 (bursty term) if the average-term frequency value is large and 0 otherwise. We used the term re-occurrence model to investigate the bursty nature of these terms, as compared to the frequency based methods alone. According to the term re-occurrence model, the $\widetilde{\lambda_2}$ value of *kennedy* is very small as compared to that for *except*. Hence the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic correctly identifies *kennedy* as a topical content term and *except* as an scattered and infrequent function word. These findings are an improvement on the analysis by [Chu00, UC00], which provide a 0 or 1 indicator value about the bursty nature of the term, whereas the term re-occurrence model assigns a quantitative value to the degree of burstiness of the term. The term re-occurrence model can differentiate between the terms

[1]The original study was based on the *New York Times*, ours on the *Associated Press* corpus

with high confidence based on the different parameters and by assigning a quantitative measure to indicate the degree of difference between the importance of these terms. This importance value assigned to a term based on the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ measure can be used for keyword indexing for retrieval and classification based applications.

8.2.2.4 noriega *and* said

[Chu00] previously investigated the behaviour of these terms in the context of an adaptive language model to demonstrate the fact that the probability of a repeat occurrence of a term in a document defies the term independence assumption inherent in "bag of words" representations. The deviation from independence is greater for content terms like *noriega* than for general terms like *said.* The term re-occurrence model also captures this distinction: *said* has small values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$, and their values are quite close to each other (Table 8.2). Hence *said* is distributed more evenly, (or more homogeneously) in the corpus than *noriega*, which has a bursty characteristic in the dataset. Therefore, *noriega* defies the independence assumption to a much greater extent than *said.* The findings of [Chu00] are well explained by the term re-occurrence model.

8.3 Analyzing terms across different datasets: Exploring Genre

The second background application explores stylistic differences across different datasets, through the distribution characteristics of a small chosen set of terms.

Having explored the behaviour of different terms in a single dataset with the term re-occurrence model, we want to investigate term characteristics across different datasets. In this case, established evaluation methodologies would involve modeling all terms across all the datasets using the term occurrence model. The terms could then be clustered according to similarity of the model's parameter values. Terms with significantly varying parameter values across datasets would serve as

possible indicators of stylistic variation between the datasets (with respect to those terms). Due to computational limitations, only a small set of terms can be explored, and their behaviour evaluated across datasets of different genres. We chose terms that are likely to be indicative of genre and explore their behaviour using the term re-occurrence model as a means of validating the model for analysis of term behaviours across different genres [SDRG05].

The aim of this analysis is to investigate whether term burstiness patterns can contribute usefully to the analysis of style - i.e. whether burstiness models based on term re-occurrence patterns can uncover useful information about the behaviour of terms that is not typically available from methods using frequency counts alone. The study will investigate the behaviour of different kinds of terms, because it seems reasonable to expect that the contribution made by function words, for instance, may be of a different nature than that of content or rare terms. Also, some words may be closely associated with a particular style and may therefore display a different behaviour in documents representative of a genre, as compared with their "standard" behaviour. We realize that we cannot be exhaustive in our study of style and genre for the purpose of validating our approach. As a consequence, our main purpose will be to sample the datasets in question, and relate what we find as explanations of the model's findings.

For this study, a series of experiments is performed. First of all, five different datasets were selected, each associated with a different genre. One should be careful not to make the assumption that any collection automatically amounts to a style or genre. Rather, these are initial experiments and we picked standard, high quality collections from the TIPSTER dataset, which can be reasonably argued to represent different genres. Table 8.3 gives a short description of each. Basic profiling and quality check for these datasets were discussed in Chapter 4. Note, on the basis of the descriptions of these datasets, we have ensured that the selection includes genres that might be identified with a particular domain (short energy reports: DOE), the medium of publication (texts associated

with press releases and newspapers in AP and SJM), as well as "outlier" collections that are highly specialized and long documents (patents: PAT) or very diverse in content (diverse government documents: FR). The selection should challenge our ability to draw conclusions.

Dataset	Contents of the documents
AP	Copyrighted Associated Press Newswire stories from 1989.
DOE	Short abstracts from the Department of Energy.
FR	Issues of the Federal Register (1989), reporting source actions by government agencies.
PAT	U.S. Patent Documents for the years 1983-1991.
SJM	Copyrighted stories from the San Jose Mercury News (1991).

Table 8.3: The contents of chosen datasets from the TIPSTER collection that are used for analyzing style

Four different types of terms were identified for modeling. Choice was made of candidates for very frequent and less frequent function words, some terms that are used in connection with reported speech and reporting styles (which are relevant to two of the datasets: AP and SJM), and some terms that behave like content words in at least one of the collections. For each of these, frequency based information across the different datasets were collected, as shown in Table 8.4. *Relative document frequency* tells us about the proportion of documents in the collection that contain the particular term. *Rate of incidence* measures the relative frequency of a particular term in the entire dataset, and it provides a measure of the term's distributional density across the entire corpus (rate of incidence = (total number occurrences of the term in the corpus) / (corpus length)). In this case the rate of incidence is expressed as the incidence of the term per 100,000 words in the collection.

These terms were analyzed using the term re-occurrence and burstiness model and the estimates of the model parameters were obtained. The burstiness patterns that emerged across different

datasets were then compared. Special attention was paid to differences between terms displaying similar behaviour according to the frequency based measures, to identify where our model adds information.

In the following sections, one group of terms shall be chosen at a time and our model's findings compared to those based on relative document frequency and rate of incidence. Here Inverse Document Frequency (IDF) has not been used for comparison as IDF is suitable for comparing terms within a dataset but when comparing terms across datasets with varying document lengths and varying number of documents, IDF figures will not be comparable.

Dataset Term	AP	DOE	FR	PAT	SJM
are	0.595 331.5	0.556 861.4	0.639 463.0	0.927 587.6	0.534 384.2
as	0.728 459.7	0.382 543.4	0.689 639.2	0.999 769.3	0.620 468.3
associated	0.482 110.5	0.041 39.9	0.106 23.4	0.323 30.5	0.210 55.5
called	0.214 58.5	0.013 11.8	0.031 4.3	0.208 9.5	0.151 44.3
could	0.316 103.0	0.045 45.1	0.136 41.7	0.322 19.6	0.272 101.2
current	0.074 19.2	0.059 70.6	0.190 47.9	0.238 65.7	0.057 15.9
data	0.028 8.8	0.154 202.1	0.207 89.9	0.241 159.3	0.029 11.0
energy	0.041 16.2	0.165 258.6	0.151 48.6	0.173 27.3	0.024 9.5
except	0.030 6.6	0.008 7.1	0.154 37.1	0.235 12.1	0.032 8.0
in	0.980 2189.7	0.863 2361.9	0.898 1840.4	0.951 2039.3	0.944 1842.3
of	0.985 2656.3	0.978 5022.3	0.975 4081.9	1.000 4267.3	0.946 2328.7
report	0.152 63.0	0.062 65.4	0.182 53.7	0.036 1.9	0.136 47.3
said	0.899 1265.7	0.003 4.7	0.070 10.4	0.812 867.5	0.602 635.2
should	0.185 57.4	0.037 37.6	0.464 153.2	0.494 30.5	0.169 59.9
the	0.998 6108.6	0.986 7703.9	0.957 6814.6	1.000 8388.1	0.977 5271.1

Table 8.4: Table showing values of relative document frequency (proportion of documents where the term occurs) and rate of incidence ((total occurrences of the term in the corpus)x(10^5) / (corpus length)) for the chosen terms across all the datasets. The top value in each cell is the relative document frequency and the lower value is the rate of incidence (x10^5)

8.3.1 Very frequent function words

The very frequent function words *the*, *of* and *are* (Table 8.5), were selected because they are ubiquitous. They are often subject to stop word removal because they are thought to behave like background noise in any collection (Appendix C). Certainly, their frequency based profiles show that *the* and *of* occur in almost all documents in each collections, and at approximately equal rates of incidence. *Are* is less ubiquitous compared to the other terms in this section.

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
are	AP	0.45	47.39	45.48	1.04
	DOE	0.32	31.85	30.24	1.05
	FR	0.07	101.30	33.70	3.01
	PAT	0.27	423.73	84.32	5.03
	SJM	0.01	473.26	45.13	10.49
in	AP	0.13	94.79	42.55	2.23
	DOE	0.17	91.74	36.81	2.49
	FR	0.02	359.07	50.68	7.08
	PAT	0.08	137.17	41.70	3.29
	SJM	0.10	141.48	48.17	2.94
of	AP	0.53	38.65	36.63	1.06
	DOE	0.62	21.10	19.72	1.07
	FR	0.01	200.28	24.05	8.33
	PAT	0.02	86.06	21.54	3.99
	SJM	0.04	204.37	39.45	5.18
the	AP	0.59	16.58	16.11	1.03
	DOE	0.29	20.49	12.72	1.61
	FR	0.01	194.89	13.47	14.47
	PAT	0.02	68.07	10.36	6.57
	SJM	0.02	168.52	17.80	9.47

Table 8.5: Parameter estimates of very frequently occurring function words

The terms *the* and *of* have low values of $\widetilde{\lambda_1}$, indicating frequent usage of these terms in most datasets. In FR and SJM, $\widetilde{\lambda_1}$ values are higher. This is consistent with the possibility that some documents contain notices or instructions, which are not plain English hence *the* and *of* do not occur in them (see Appendix A for such example documents from the FR dataset). Similarly, news

documents often adopt telegraphic styles which have reduced incidence of some function words. Small values of $\widetilde{\lambda_2}$ indicate frequent re-occurrence. The $\widetilde{\lambda_2}$ values for the different datasets are in a close proximity. In FR and SJM, however, the behaviour of *of* and, surprisingly, *the* is clearly much burstier than it is in the other datasets, with long gaps separating close bursts. The term *in* displays quite similar behaviour across the board, using both types of measures, and in all collections. This term is rarer in most datasets compared to the terms *the* and *of*. The FR dataset has large $\widetilde{\lambda_1}$ values for the term *in* as it contains many documents in the form of notices or instructions which are not plain text.

The term *are* has very close values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ for the AP and DOE datasets, indicating the fact that this term occurs evenly across these two datasets. The behaviour of *are* in the other datasets is quite different. It has high values of $\widetilde{\lambda_1}$ for FR, PAT and SJM. This is combined with low $\widetilde{\lambda_2}$ value for the SJM dataset, suggesting a relative bursty behaviour. The model for *are* again shows a distinctive behaviour in SJM, particularly as compared to AP, even though the associated frequency based profiles are indistinguishable.

8.3.2 Less frequent function words

Not all function words are as frequent as the examples in the previous section. Though they are often removed as stop words, less frequent function words tend to be associated with certain types of syntactic structure, and hence may be indicative of style. Some less frequent function words we study are *could*, *should*, *as* and *expect* (Table 8.6). These terms also feature in a standard English stop-word list (Appendix C).

Syntactically speaking, *could* and *should* are both modals[2] and have some comparable usage in English. Table 8.4 shows that they have different relative document frequency values, but

[2]modals are special verbs which behave very differently from normal verbs

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
could	AP	0.52	1631.85	539.37	3.03
	DOE	0.61	3095.02	1078.98	2.87
	FR	0.74	3810.98	293.17	13.00
	PAT	0.74	9174.31	273.90	33.50
	SJM	0.53	2741.23	450.86	6.08
should	AP	0.45	2890.17	1020.30	2.83
	DOE	0.62	4677.27	1715.56	2.73
	FR	0.48	1423.08	101.50	14.02
	PAT	0.73	5065.86	268.96	18.83
	SJM	0.58	5063.29	627.35	8.07
as	AP	0.93	241.55	7.60	31.76
	DOE	0.93	215.52	6.85	31.47
	FR	0.45	287.85	72.46	3.97
	PAT	0.27	300.75	68.54	4.39
	SJM	0.90	256.61	6.30	40.72
except	AP	0.82	19755.04	3650.97	5.41
	DOE	0.60	17908.31	3593.24	4.98
	FR	0.49	7668.71	1056.97	7.26
	PAT	0.83	13622.12	192.31	70.84
	SJM	0.67	29120.56	6309.15	4.62

Table 8.6: Parameter estimates of some less frequent function words

that their rate of incidence across the different collections is almost equivalent. Hence, even using linguistic knowledge these terms cannot be differentiated on that basis. The term re-occurrence model, on the other hand, shows a consistently bursty behaviour in the FR and PAT collections, indicating a different usage pattern in government reports and in patent documents. Both these sets use comparatively formalized styles and document structures that are not uniform throughout the document.

The term *as* is quite interesting. It occurs in quite a high proportion of documents in all the datasets, with a uniform rate of occurrence. This is borne out by the $\widetilde{\lambda_1}$ values which show a uniform distance between bursts. However, in FR and PAT within-burst distance is larger, and it behaves like a relatively scattered function word, whereas in AP, DOE and SJM the very low values of $\widetilde{\lambda_2}$ depict a very bursty behaviour. In the AP and SJM datasets *as* has a bursty behavior as certain

documents refer frequently to some reference object using the term *as* (see Appendix A for such an example document). In the DOE dataset there are many documents that refer to the chemical *Arsenic, As* and due to the lowercase transformation *As* the chemical symbol is conflated to the function word *as* (see Appendix A for such an example document from DOE). This would suggest that a text based application on the DOE dataset should be careful while making case changes.

The frequency based profile of *except* is diverse. Based on the term re-occurrence model, it has large values of $\widetilde{\lambda_1}$ for all the datasets, and also has quite large values of $\widetilde{\lambda_2}$. Hence based on Table 8.1 and on the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics this term appears to behave as a scattered function word. The exception is the PAT dataset, where the term occurs much more burstily.

8.3.3 Style indicative terms

Some terms may be associated with particular styles or genres such as verbs indicating reported speech, or a specific way of attributing sources or information. The behaviour of three such terms: *called, report* and *said* was investigated (Table 8.7).

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
called	AP	0.48	3780.72	997.01	3.79
	DOE	0.72	12861.74	1826.82	7.04
	FR	0.82	38804.81	68.78	564.22
	PAT	0.79	32637.08	656.60	49.71
	SJM	0.71	8237.23	489.96	16.81
report	AP	0.85	4472.27	94.61	47.27
	DOE	0.97	2474.63	5.56	444.69
	FR	0.68	4315.93	71.74	60.16
	PAT	0.94	259875.26	303.49	856.29
	SJM	0.85	8264.46	112.20	73.66
said	AP	0.04	687.76	68.97	9.97
	DOE	0.67	61349.69	12224.94	5.02
	FR	0.84	26385.22	392.62	67.20
	PAT	0.06	2080.30	13.43	154.94
	SJM	0.16	2460.63	92.34	26.65

Table 8.7: Parameter estimates of terms related to the style of reporting

Table 8.4 shows that *report* occurs in a smaller proportion of documents in DOE and PAT as compared to the other collections. At the same time, DOE exhibits the smallest value of $\widetilde{\lambda_1}$ and PAT has the largest (Table 8.7). The value of $\widetilde{p}$ is much larger for DOE and PAT than the other collections. The term also has hugely differing values of $\widetilde{\lambda_2}$, though the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics indicates similar overall behaviour when comparing between-burst and within-burst gaps. Based on the model parameters, the term *report* behaves like a rarely occurring content word in the DOE (few documents in the form of reports where the term occur several times) and the PAT (patents about report generation systems) dataset. The term re-occurrence model helps to determine that *report* may be an important term in the stylistic analysis of these collections, something that simple frequency based measures do not reveal.

The term *called* has close values of document frequency for both the AP and PAT collections, and the values are large enough to be comparable to a function word. However, Table 4.3 shows the average document length for PAT is much larger than that of AP dataset. The term re-occurrence model presents the term as more bursty in PAT than in AP. Using the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic, *called* does not behave like a function word in PAT. In the PAT dataset, *called* has different meanings in various scenarios, like person who *called* on the telephone, referencing to some object using a name, function *called* in a computer program or some system. Also, the rate of incidence is of the same order of magnitude in the FR and PAT collections, supported by close values of the $\widetilde{\lambda_1}$ parameter. In the FR collection, however, this term is of a much more bursty nature, having comparatively smaller $\widetilde{\lambda_2}$ values. Hence based on our $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics, *called* behaves like a bursty content term for FR. In the FR dataset also *called* has different usage in various documents with several documents where the term is used as an adjective *so-called* and by tokenizing based on non-alpha-numeric characters these two terms are separated.

Probably the most interesting term in Table 8.7 for analyzing style is *said* because it directly

indicates a document or a collection referring to a conversation. This term has high values of relative document frequency and rate of incidence for the AP, PAT and SJM datasets. This is perhaps unsurprising for the AP and SJM as they are about news. This is supported by the parameter estimates of the term re-occurrence model for these datasets. For PAT however, *said* has a rather bursty nature due to large value of $\widetilde{\lambda_1}$ combined with small $\widetilde{\lambda_2}$ value and has the characteristic of a content bearing word. Examining the patent documents in the PAT dataset reveals that *said* is used mostly as an adjective to refer to some object, for e.g. *the said invention.* In such scenarios the term tend to re-occur several times within-burst to maintain the continuity of the discussion.

8.3.4 Content terms

A selection of content bearing terms were looked at, or terms that might refer to a topic in the collections. Those terms were chosen which were expected to show different behavior in the various datasets. The terms studied in this section are *associated*, *current*, *data* and *energy* (Table 8.9). The terms *data* and *energy* are frequently used terms in some of the datasets. *Current* is an ambiguous word which can be used both as a noun and adjective. *Associated* though not a noun by itself, is sometimes used as a proper noun in the *Associated Press* articles of the AP dataset.

Data Set	Noun	Verb	Adjective	Adverb
AP	2.64	0.01	92.25	5.10
DOE	15.48	0.02	82.28	2.22
FR	4.33	0.04	69.94	25.69
PAT	19.09	0.08	77.08	3.74
SJM	4.09	0.00	90.41	5.50

Table 8.8: Percentage distribution of different part-of-speech assigned to the term *current* for the TISPTER datasets.

The term *current* is interesting, because it is ambiguous between an adjective or adverb (in "present" time period) and a noun (in "electricity"). Table 8.4 shows uniform rate of incidence and relative

document frequency values for each of the collections. The percentage distribution of different part-of-speech assigned to the term *current* is listed in Table 8.8. The term *current* occurs as a noun with higher percentage in the DOE and PAT datasets. These figures will be evaluated using the term re-occurrence model. The term re-occurrence model also records bursty behaviour with low $\widetilde{\lambda_2}$ values in the DOE and PAT collections. The $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics show that there is a higher incidence of bursty behaviour of the term, and so perhaps for the DOE and PAT datasets, the word *current* behaves more often like a bursty content term and as a scattered function word for the other collections. This is consistent with the nature of these collections: DOE and PAT contain technical documents. Whilst this requires further investigation, it would appear that the term re-occurrence model may be useful in disambiguation with an approach along these lines.

Term	Dataset	$\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
associated	AP	0.68	8019.25	36.44	220.05
	DOE	0.40	7968.13	2461.24	3.24
	FR	0.83	8928.57	522.47	17.09
	PAT	0.50	13104.44	330.91	39.60
	SJM	0.93	2453.99	6.69	366.87
current	AP	0.32	17445.92	4027.39	4.33
	DOE	0.92	3642.99	69.78	52.20
	FR	0.69	4299.23	366.17	11.74
	PAT	0.36	7189.07	60.68	118.48
	SJM	0.86	14039.03	884.96	15.86
data	AP	0.76	24673.08	243.49	101.33
	DOE	0.82	1591.85	67.11	23.72
	FR	0.58	2833.66	64.77	43.75
	PAT	0.23	5336.18	48.50	110.03
	SJM	0.90	46468.40	188.89	246.00
energy	AP	0.96	10711.23	78.74	136.03
	DOE	0.77	1548.71	43.18	35.87
	FR	0.67	14863.26	107.46	138.32
	PAT	0.49	11195.70	86.88	128.86
	SJM	0.71	28457.60	1258.34	22.62

Table 8.9: Parameter estimates of terms with some dependence on topic and genre

If one were to calculate the inverse document frequency for *associated*, the term would have

maximum weight for DOE (lowest relative document frequency) and least weight for AP (highest relative document frequency). The rate of incidence is similar for all the datasets. In the term re-occurrence model, the $\widetilde{\lambda_1}$ values have similar distances between bursts in FR, DOE and AP, but in DOE, the term is very scattered across the whole collection, and hence has the characteristics of a rare and scattered word. For AP and SJM, though the rate of occurrence is quite high, the re-occurrence rate is quite small, which leads to large values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio. Here, the term has the characteristics of a bursty content word. Investigation of these datasets reveal that in the AP dataset the term *Associated Press* is used in articles to refer to events and the organization. Even in the SJM dataset, there are several articles that refer to news articles form the *Associated Press* using these terms. This explains why a term like *associated*, which is a verb or an adjective has the properties of a content bearing terms in these datasets.

The behaviour patterns of the term *data* helps in identifying a drawback of frequency based measures. The values for document frequency and the rate of incidence for this term are quite low for AP and SJM when compared to the other collections. A pure frequentist approach would have reason to treat this term as an informative content word in AP and SJM, and as a uninformative word in the other collections. Doing so would ignore the issue of burstiness. The term re-occurrence model shows small values of $\widetilde{\lambda_2}$ for DOE, FR and PAT as compared to those of AP and SJM, and the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics would class this term as a content word for DOE, FR and PAT; and as a scattered rarely occurring non-informative term in AP and SJM. Though they would need some means of confirmation, these findings are plausible given the content of the datasets.

The term *energy* is an interesting term. It is a content word in general but, like all content words, could behave as a non-informative and non-topical word in an appropriate specialized domain - in this case about energy, such as the DOE collection. This term has a low rate of occurrence, $\widetilde{\lambda_1}$, in all the datasets except DOE, and bursty nature as indicated by the $\widetilde{\lambda_2}$ value for most of the

collections. Because of this, the term will be considered as an informative content term in the AP, FR and PAT datasets. The two lowest values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic are from DOE, where the within-burst and between-burst gaps are smallest, and SJM where the within-burst and between-burst gaps are largest. The heuristic suggests that in both collections, the term behaves like a scattered non-topical word, but in DOE it has the characteristics akin to a frequent non-topical word. The patterns associated with this term demonstrate how the term re-occurrence model can be used for differentiating the behaviour of a pervasive content word behaving like a non-topical word in a collection.

8.4 Studying characteristics of frequent terms in a dataset

This application explores the behaviour of very frequent terms in a dataset and investigates how homogeneously they are distributed and whether their distribution characteristics add any valuable information about the document collection [DRSG05]. In this section, we contrast "burstiness" in the behaviour of a single term, with the notion of "homogeneity" in its distribution pattern, where homogeneity is equated with an even rate of occurrence. Here we will take a different approach from the discussion in Chapter 5, where we postulated, and defeated a homogeneity null-hypothesis as part of our methodology, in order to establish the extent to which term occurrences in documents, collections and language varieties deviate from the term-independence assumption. Here we will model distribution characteristics of some specific very frequent terms in different collections using our term re-occurrence model and then verify the results of the model against our data.

The very frequent terms in a dataset are usually function words. It is possible for a very frequent term in a collection not to be a function word, but under a view where very frequent terms are unlikely to contribute useful information about a text, such terms may be removed for some

applications together with function words. Function words remain function words even if they do not occur very frequently, and may be removed by stop-word lists. It is also possible for some terms to behave like content words in one document, and like a function word in another. Under the approaches by Katz [Kat96] and Church [CG95b, Chu00] informative content words behave burstily, and so it would seem to follow that on occasions where a term behaves as a function word, this would be characterized by a less bursty (and more homogeneous) distribution pattern. The problem is that this discussion space is occupied by distinctions made along three different dimensions: frequency (very frequent words), linguistic function (function words) and distribution (content words).

We are working with raw data and so have limited ability to explore issues surrounding burstiness and function words. Nonetheless, by confining ourselves to modeling very frequent function words, we may be able to gain an impression of the degree of burstiness in their behaviour, and to get some insight into the ability of our model to capture the behaviour of function words. Very frequent function words also have the advantage that they bring large amounts of evidence into the modeling.

This view of function words as general background noise is consistent with their removal through stop lists or frequency thresholds in many applications. More sophisticated approaches, however, show that stop word removal based on collection specific distribution patterns leads to improved performance in text categorization [WS92, YW96]. This constitutes some evidence that function words and stop-words show significant behavioral differences between texts, in the sense that taking account of their behaviour affects the effectiveness of categorization techniques. The current experiment highlights the behavior of very frequent terms in the datasets using the term re-occurrence model and examines in some more detail whether they distribute homogeneously or have bursty characteristics.

8.4.1 Experimental Framework

An experiment was conducted to study the gaps between successive occurrences of some very frequent function words. Two alternatives were examined for modeling these gaps. Both used models formed from exponential distributions. The first alternative is based on the "bag-of-words" representation, which has inherent in it the term independence assumption, under which the gaps between successive occurrences of a particular term are generated from a single exponential distribution. This is in contrast to the second alternative, which assumes that terms occur in bursts and the gaps between successive occurrences of a term are generated from a mixture of two exponential distributions, the one with the larger mean reflecting the overall rate of occurrence of the term in the corpus and the one with the smaller mean reflecting the rate of re-occurrence after it has occurred recently. Since function words, including very frequent function words are believed to be distributed homogeneously, the first approach to modeling their behaviour should not loose significant information. It is this belief that we intend to investigate based on the alternatives of the exponential mixture model.

Now based on the term re-occurrence model discussed in Chapter 6, if terms do distribute homogeneously throughout the text, then the mixture model will be over-parameterized, as the gaps will be generated from a single exponential distribution. In that case, one of the following conditions must hold so as to dissolve one of the mixture components and end up with a single exponential distribution. These conditions are:

- $p = 0$
- $p = 1$
- $\lambda_1 = \lambda_2$

First the gaps between successive occurrences of the frequent terms are modeled using a mixture of

exponential distributions and then the above claims are investigated with respect to the obtained model parameters.

The 10 most frequent terms from each of the TIPSTER datasets are studied in this experiment. The 10 most frequent terms (Table 8.10), show a high degree of overlap, and are clearly function words. The datasets were tokenized removing space and punctuation, hence we had tokens like *s*, *o*, *m*, *p* occurring among the top 10 terms (Table 4.4). These have been removed from the table as they are not real terms. Only the term *san* in this table features as a very frequent term in the SJM dataset, and is not a function word. Focusing on these ten terms in experiments across all the different collections of the TISPTER dataset (section 4.2) should yield information on the behaviour of a small collection of very frequent function words.

Data Set	10 Most Frequent Terms
AP	the, of, to, a, in, and, said, for, that, on
DOE	the, of, and, in, a, to, is, for, with, are
FR	the, of, to, and, a, in, for, or, that, be
PAT	the, of, a, and, to, in, is, for, said, as
SJM	the, a, of, to, and, in, for, that, is, san
WSJ	the, of, to, a, in, and, that, for, is, said
ZF	the, and, to, of, a, in, is, for, that, with

Table 8.10: The 10 most frequent terms for each of the TIPSTER datasets

8.4.2 Experimental Results

As before, the model provides estimates of the mean of each of the exponential distributions ($\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$) and estimates of the probability of a gap being generated from each of these distributions ($\widetilde{p}$ and $1-\widetilde{p}$).

In a homogeneous distribution, one of the following conditions have to hold:

$\widetilde{p}=0$ or $\widetilde{p}=1$ or $\widetilde{\lambda_1}=\widetilde{\lambda_2}$.

The validity of any of these statements would reduce the mixture model to a single component exponential distribution, which would be consistent with an assumption of homogeneity in a distribution. We constructed the mixture models for the terms in Table 8.10, and provide a full list of the parameter estimates for three of those terms *the* (Table 8.11), *of* ((Table 8.12)) and *said* (Table 8.13).

Data Set	$\widetilde{p}$	$1-\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$
AP	0.59	0.41	16.58	16.11
DOE	0.29	0.71	20.49	12.72
FR	0.01	0.99	194.89	13.47
PAT	0.03	0.97	58.96	10.61
SJM	0.02	0.98	168.52	17.80
WSJ	0.70	0.30	17.46	17.00
ZF	0.10	0.90	67.80	18.39

Table 8.11: Parameter estimates for the term *the* for all the TISPTER datasets.

For the term *the* (Table 8.11), $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ are very similar in the AP and WSJ datasets and $\widetilde{p}$ is close to 0 in the FR, PAT and SJM datasets, so in these datasets *the* may distribute homogeneously. In the DOE and ZF datasets, however, $\widetilde{p}$ is near neither 0 nor 1, and $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ differ markedly, so *the* does not appear to distribute homogeneously in these two datasets. Similarly, the term *of* (Table 8.12) has very similar values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ for AP, DOE and WSJ datasets and $\widetilde{p}$ is close to 0 for the FR, PAT, SJM and ZF datasets. Hence, for the term *of*, each of the datasets provide little evidence against it having a homogeneous distribution from the values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ or $\widetilde{p}$. In contrast, according to our model, the distribution of the term *said* (Table 8.13) shows evidence of homogeneity only for the AP and PAT datasets, for which the value of $\widetilde{p}$ is close to 0.

To investigate whether the model assigns homogeneous distribution characteristics to the other common terms, the ratio between the two $\widetilde{\lambda}$s, $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ is calculated and the closeness of this ratio to 1 is studied. A $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio of 1 indicates that the two exponential distributions have equal means,

Data Set	$\widetilde{p}$	$1-\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$
AP	0.65	0.35	38.37	36.44
DOE	0.62	0.38	21.10	19.72
FR	0.02	0.98	106.25	24.01
PAT	0.03	0.97	73.42	21.82
SJM	0.04	0.96	205.38	39.45
WSJ	0.42	0.58	36.91	35.39
ZF	0.01	0.99	262.47	46.51

Table 8.12: Parameter estimates for the term *of* for all the TIPSTER datasets.

Data Set	$\widetilde{p}$	$1-\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$
AP	0.04	0.96	696.38	69.01
DOE	0.67	0.33	61349.69	12224.94
FR	0.84	0.16	26385.22	392.62
PAT	0.06	0.94	2167.32	13.10
SJM	0.16	0.84	2499.38	92.42
WSJ	0.12	0.88	1608.49	72.62
ZF	0.42	0.58	8810.57	177.21

Table 8.13: Parameter estimates for the term *said* for all the TIPSTER datasets.

and hence reduce to a single exponential distribution. A large deviation of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio from 1 reveals the presence of two very distinct exponential distributions and would indicate evidence against a hypothesis of the homogeneity in the term's distribution in the corpus provided the value of $\widetilde{p}$ is close to neither 0 or 1. If the value is very close to 0 or 1 (a difference of less than 0.05) it is argued that one of the exponential distributions have negligible effect and there is little evidence against the term being homogeneously distributed. Table 8.14 provides the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio and the values of $\widetilde{p}$ for the most frequent terms of each of the datasets.

In the table ratios of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ that are less than 1.2 are given in bold-face type, as are values of $\widetilde{p}$ that are below 0.05 or above 0.95. For clarity, combinations are underlined when one (or both) of the values are in bold. According to the model, these terms are distributed non-burstily, whereas it suggests that bursty distributions are present for those instances that are not underlined in the

table. Only the term *of* show signs of being homogeneously distributed, according to the model, across all the datasets either based on the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio or values of $\widetilde{p}$ being close to 0. The terms *and*, *are*, *the* and *to* also seem to be homogeneously distributed across many of the datasets. The other 12 terms in Table 8.14 only appear to be homogeneously distributed in at most 2 of the 7 datasets.

Term	AP	DOE	FR	PAT	SJM	WSJ	ZF
a	1.98 (0.17)	3.62 (0.46)	3.88 (0.23)	3.34 (0.15)	2.33 (0.10)	2.00 (0.14)	2.59 (0.10)
and	**1.05** (0.30)	**1.05** (0.53)	2.74 (0.11)	3.62 (**0.02**)	3.62 (0.07)	**1.06** (0.14)	**1.05** (0.10)
are	**1.06** (0.69)	**1.05** (0.32)	3.01 (0.07)	5.09 (0.33)	10.43 (**0.01**)	**1.05** (0.69)	**1.16** (0.47)
as	31.64 (0.93)	31.47 (0.93)	3.97 (0.45)	4.46 (0.24)	40.73 (0.90)	65.09 (0.90)	56.38 (0.91)
be	3.03 (0.73)	1.30 (0.49)	6.06 (0.13)	5.71 (0.27)	3.48 (0.64)	2.13 (0.33)	2.23 (0.27)
for	2.04 (0.29)	3.19 (0.54)	4.40 (**0.05**)	4.09 (0.26)	2.61 (0.55)	1.89 (0.31)	15.94 (**0.01**)
in	2.23 (0.13)	2.49 (0.17)	7.08 (**0.02**)	4.01 (**0.05**)	2.94 (0.10)	1.93 (0.22)	2.83 (0.08)
is	2.93 (0.58)	4.67 (0.35)	3.87 (0.19)	5.34 (0.07)	4.04 (0.34)	2.43 (0.34)	5.76 (**0.02**)
of	**1.05** (0.65)	**1.07** (0.62)	4.43 (**0.02**)	3.37 (**0.03**)	5.21 (**0.04**)	**1.04** (0.42)	5.64 (**0.01**)
on	1.99 (0.31)	5.73 (0.72)	4.72 (0.21)	5.95 (0.25)	2.59 (0.46)	1.95 (0.55)	2.58 (0.22)
or	41.50 (**0.95**)	3.69 (0.48)	6.87 (0.36)	9.98 (0.28)	8.63 (0.81)	4.58 (0.71)	3.66 (0.78)
said	10.09 (**0.04**)	5.02 (0.67)	67.20 (0.84)	165.39 (0.06)	27.04 (0.16)	22.15 (0.12)	49.72 (0.42)
san	112.37 (0.92)	21.19 (0.74)	579.36 (0.93)	855.81 (0.93)	14.43 (0.46)	149.67 (0.92)	90.67 (0.80)
that	2.78 (0.16)	1.23 (0.59)	4.89 (0.15)	4.69 (0.21)	3.42 (0.20)	2.47 (0.11)	4.34 (**0.04**)
the	**1.03** (0.59)	1.61 (0.29)	14.47 (**0.01**)	5.56 (**0.03**)	9.47 (**0.02**)	**1.03** (0.70)	3.69 (0.10)
to	3.18 (0.10)	**1.15** (0.56)	12.45 (**0.01**)	3.81 (**0.05**)	6.78 (**0.02**)	**1.13** (0.41)	3.24 (**0.04**)
with	2.62 (0.40)	2.40 (0.28)	3.54 (0.23)	3.55 (0.24)	2.70 (0.64)	1.29 (0.45)	2.53 (0.12)

Table 8.14: Table showing values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio and values of p for the frequent terms for all the datasets. $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratios close to 1 are marked in bold (and underlined) and values of $\widetilde{p}$ close to 0 or 1 are also marked in bold (and underlined), providing evidence of the term being uniformly distributed in that dataset.

Said is an interesting term in the table. It has very high values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio and the values vary over a large range. Also, the value of $\widetilde{p}$ for *said* is close to 0 for the AP dataset. This is because the term *said* has a huge dependence on the document's content and style, and these characteristics can be explored and studied by modeling the gaps.

The term *san* is an outlier in the list. It is not a function word, but it featured in the list of top 10 terms in the SJM (stories from San Jose Mercury newswire) collection. Bearing in mind our discussion about the relationship between very frequent, function and non-informative words, it would appear that although very frequent, *san* is not a function word, but may be non-informative

Data Set	$\widetilde{p}$	$1-\widetilde{p}$	$\widetilde{\lambda_1}$	$\widetilde{\lambda_2}$	$\widetilde{\lambda_1}/\widetilde{\lambda_2}$
AP	0.92	0.08	12495.31	111.20	112.37
DOE	0.74	0.26	82644.63	3900.16	21.19
FR	0.93	0.07	25933.61	44.76	579.36
PAT	0.93	0.07	345781.47	404.04	855.81
SJM	0.46	0.54	298.06	20.65	14.43
WSJ	0.95	0.05	10258.51	68.54	149.67
ZF	0.80	0.20	13080.44	144.26	90.67

Table 8.15: Parameter estimates and the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratios for the term *san* for all the TIPSTER datasets.

in the sense that many articles in SJM will be about *San Jose*, and hence perhaps likely to include the term *san*. Note the list does not include the term *jose*. According to the model, *san* is a very rare term in all the datasets other than SJM, as indicated by large rate of occurrence $\widetilde{\lambda_1}$, and it is either bursty in nature or scattered as indicated by small or medium range values of $\widetilde{\lambda_2}$, leading to large values of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio (Table 8.15). The term has characteristics of a rarely occurring and scattered one in the DOE dataset due to large values of $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$, and hence a low value of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio. However, in contrast, the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio for the SJM collection, is relatively small when compared to the values of the other collections. This is consistent with our view on the data that the term *san* is a less informative term in SJM, compared to the other collections. Further examination reveals that the SJM dataset contain numerous articles referring to city names starting with *san* (e.g. *San Francisco*, *San Jose*, *San Andres*). This makes the term *san* one of the most frequently occurring terms in SJM, but not the term *jose*, which refers to only one of such cities.

The term *as* exhibits large values of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio in all the collections other than FR and PAT. PAT has comparatively large values of $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio for most of the other terms. The model suggests that the term has dependence on the content, style and structure of the document and collection. Section 8.3.2 contains a detailed explanation for the behaviour of term *as* in the various datasets.

Our model, and our verification against the datasets, suggest that many frequently occurring

function words have bursty characteristics and are not homogeneously distributed across datasets. Indeed, there is some evidence that suggests that the behaviour of very frequent function words bears some relationship to the characteristics of certain types of documents. This is consistent with other approaches that use function word behaviour in the detection of genre [Cha01, AKFS03], and with the conclusions by [WS92, YW96] that a more fine grained approach to stop-word removal benefits the effectiveness of applications such as text classification.

8.5 Overall Discussion

In previous chapters, we have developed a term distribution model that captures term re-occurrence by means of a mixture of two exponential distributions. Due to computational limitations, it was not possible to evaluate the behaviour of the model in the context of live applications, and contrast it in that way with the performance of other (mostly frequency based) approaches. In this chapter, we therefore proceeded to validate the model by checking whether the term behaviours it highlights can be verified against the data, bearing in mind the requirements of a background application.

Since our model aims to capture bursty term distribution behaviour, we developed a heuristic based on the ratio between the values of our model's two parameters (where $\widetilde{\lambda_1}$ measures the mean gap distance between bursts and $\widetilde{\lambda_2}$ the mean distance within a burst). Throughout this chapter statements have been made about the $\widetilde{\lambda_1}$ and the $\widetilde{\lambda_2}$ value being large or small, but no cut-off or guideline has been provided for making such decisions. Whether the value of $\widetilde{\lambda_1}$ or $\widetilde{\lambda_2}$ is high or low is a comparative measure. There is no fixed cut-off value that has been used to assign values as high or low. Instead, the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio criteria provides a quantitative measure to determine the order of magnitude difference between the between-burst gaps and the within-burst gaps. This value captures some of our intuitions and might then be used as a guideline. For frequently occurring

function words, one might expect their distribution to be comparatively non-bursty, with both $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ fairly small (say less than 100) and hence a $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio close to 1. Scattered words, such as rarely occurring function words, would have a similar behaviour, but with both $\widetilde{\lambda_1}$ and $\widetilde{\lambda_2}$ fairly large (in order of magnitude 100s, 1000s or even more) and again a $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio value remaining fairly small. Several approaches [CG95b, Kat96] have associated topical, content bearing terms with rare (and hence informative) terms that display a bursty behaviour, which the model would capture as a comparatively large $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio (100 or more) with $\widetilde{\lambda_2}$ fairly small (round 100 or even less), showing a long distance between occurrences of burst, within which the re-occurrence distance is small. It is not currently possible to provide fixed cut-off values of $\widetilde{\lambda_1}$ or $\widetilde{\lambda_2}$ being small or large as these values are dependent on factors like dataset size and document lengths in the dataset. Further research will have to investigate the relation between the parameter values for a term in comparison to some basic statistics for a dataset. Also note that our measures would associate large values of $\widetilde{p}$ with most content terms, because they take account of documents where the term does not occur, and the complete document length adds to the $\widetilde{\lambda_1}$ estimate. Since a content word is rare in a dataset, large proportions of documents where the term does not occur account for the large value of $\widetilde{p}$.

This chapter looked at three applications of the term re-occurrence model, and in each attempted a comparison with a standard frequency based approach. We verified the behaviour of our model against the data in our collections, both where the model confirmed our expectations, and where we found values that did not correspond to our intuitions. We found that the model performed better at highlighting burstiness in term distributions than traditional approaches, and was more fine-grained at detecting burstiness and making distinctions between usage characteristics beyond the scope of frequency based measures. The $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio heuristics of the model parameters can provide a useful quantitative measure of a term's usage and importance in the dataset, particularly

when used with detailed knowledge of the values of contributing parameters and the value of $\widetilde{p}$. Extending this heuristic for improved feature selection and keyword indexing would be a desirable next step if computational limitations allowed it. Some of these possible future work are discussed in Chapter 9.

In the second application, the model was used to bring out differences in stylistic features by modeling a term across datasets of varying genres. A range of function words, style indicative terms and content bearing terms were studied in this application. Frequency based approaches failed to highlight differences in behaviour of terms across genres. Behavior of the model parameters were verified by manual inspection of the documents. This includes unusual behavior of the terms *the* and *of* in the FR dataset as they contain documents in the form of notice and instructions which are not plain text. This application revealed that stylistic differences across datasets could also be used for word sense disambiguation.

The third application provides evidence that the frequent function words in the dataset are not homogeneously distributed. The frequent function words for all the TIPSTER datasets were studied based on the model and barring a few terms, most were found to display a bursty character. This provides support to the findings made in Chapter 5 that even frequent terms in a dataset are not homogeneously distributed.

Chapter 9

Conclusion and Future Work

This chapter discusses the contributions of the thesis and directions where the model might be applied. It also highlights limitations of the model and suggests approaches to tackling those limitations.

9.1 Main contributions

In a snapshot, the most significant contributions of the thesis are:

- Experimental evidence is provided that the term independence assumption inherent in the "bag-of-words" representation of text is indeed invalid and that by and large all terms display bursty behaviour.
- A methodology is developed to provide some measure of the extent to which term distribution behaviours in different collections diverge from a null hypothesis of homogeneity, as a placeholder for term independence.
- A model of term re-occurrence and burstiness is proposed based on the gaps between individual occurrences of a term, thus retaining positional information about the term's behaviour.

- The term re-occurrence model is instantiated and verified against the background of different collections. This includes understanding term characteristics within a dataset, comparing terms across different genres and analyzing bursty characteristics of frequent function words. The ability of the model to capture term burstiness characteristics is compared with that of traditional frequency based models and found to be able to make distinctions beyond the scope of those.

9.2 Contributions discussed

The thesis looks into the issue of term burstiness in text. Term burstiness is the phenomenon of multiple occurrences of a term in close vicinity to each other. Terms related to the topic of discussion tend to re-occur, which helps provide flow and structure in the document. The different types of burstiness in text were discussed: *term burstiness* is the multiple occurrences of a term in close vicinity to each other within a text, *document-level term burstiness* is where the term tends to occur a few times in nearby documents (possibly with respect to time) and does not occur in other documents, and *concept burstiness* is the re-occurrence of multiple terms that refer to the same concept.

Traditional frequency based approaches to term distribution modeling have limited ability of capturing bursty term behaviour because they lose positional information and carry an inherent assumption that terms occur independently of each other. The extent to which term burstiness is evident in a dataset is gauged through a sequence of homogeneity experiments, where the term independence assumption is set as a null hypothesis. Various applications of Natural Language Processing, Information Retrieval and Machine Learning use a "bag-of-words" representation with the inherent independence assumption, to represent text. This representation uses the term's

frequency in the document and loses any positional information about a term. This assumption postulates that a term is equally likely to occur anywhere in the document and is hence considered to be homogeneously distributed. A null hypothesis stating term distribution homogeneity between two random partitions of the dataset was formulated. The null hypothesis was defeated convincingly based on a series of χ^2 based homogeneity experiments. As expected, it was experimentally shown that the independence assumption made in text is invalid. Importantly, as a side effect, it was also shown that different collections have different degrees of heterogeneity, across the collection, within each document, and as an aspect of language variety, possibly genre.

Having identified the effect and importance of term burstiness in a dataset, the quest to model the term burstiness information began. Previous proposed approaches for modeling term burstiness have been based on frequency counts and handled under the independence assumption. Consequently, the decision about a term exhibiting bursty behaviour was judged from the rate of occurrence of the term in a document. These approaches lose information on the distance between the terms within a burst, and the presence of multiple bursts in a long document.

A model of term burstiness was proposed to capture the re-occurrence patterns of terms in the document and in the entire dataset. The model is based on the gaps between successive occurrences of the term and thus retains structural information about a term's distribution in the dataset. The gaps are modeled by a mixture of exponential distributions. Non-occurrence of a term in a document is modeled by the statistical concept of *censoring*, which is appropriate when the event of observing a certain term is censored at the end of the document. This is the only model of term distribution we are aware of that retains positional information about a term. It enables valuable information about the term's occurrence to be used for making inferences. The model is based on statistical principles, is language independent and does not require any prior information.

Fitting a complex mixture model is not straightforward, so Bayesian statistics methods with

data augmentation were used to estimate the model parameters. Data augmentation introduces additional parameters to the model and, surprisingly, this simplifies the fitting. Markov Chain Monte Carlo based simulation methods were used to obtained the sample values.

The proposed term re-occurrence model was applied in three contexts, and validated against our intuitions about term behaviour and evidence in our datasets. The model can be used for all kinds of terms, be they rare content words, medium frequency terms or frequent function words. Our findings show that the proposed model can identify bursty occurrence patterns of a term in the dataset that cannot be captured based on frequency information alone. The model can differentiate between distribution patterns associated with content terms having a bursty character, frequent and evenly distributed function words, and words scattered throughout the dataset. Another application investigated the burstiness model's ability to bring out differences between text of various genres. Here too, the model captures characteristics of a term's usage across different collections that would be inaccessible if only frequency data were used. Finally, the model was also used to study the characteristics of very frequent terms in a dataset. Many techniques treat them as non-informative background noise. The burstiness model revealed bursty characteristics of several function words which supports recent findings on the effectiveness of stop-word removal.

The thesis presents a novel method of modeling term burstiness by studying the gaps between term occurrences. By doing so the model retains positional information about the term's occurrence in the dataset. The model can discover certain properties of a term's usage in the dataset that would otherwise be ignored if only frequency information were used.

9.3 Future work

The current research has opened up new areas of research. The term re-occurrence model can be applied in a range of settings. Some of these application areas are now discussed. The model has some limitations in comparing a term's characteristics across two different datasets. Certain extensions to the model are suggested to overcome these limitations. Extension to the model is also discussed.

9.3.1 Applications based on the model

Feature selection is an important step in machine learning so that the most important features in each category are retained for the classification task [YP97, Kil96b, Mit97]. In Chapter 8 of the thesis it was shown that the model parameters along with the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristic are a good criterion for judging a term's importance in a dataset. In machine learning, in the context of **text classification**, the terms in each category could be modeled based on term burstiness and then the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics might be used to pick the most important features of each class. The union of all the important features from each class could then be considered as the feature set for the text classification task. Improved feature selection might provide better performance for **authorship attribution** and **genre classification**.

Often, **stop-word removal** is performed based on a pre-compiled stop-word list for various text classification and information retrieval tasks. Certain methods have been suggested for domain specific noise reduction [WS92, YW96]. Stop-word removal based on the term burstiness information and the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics may prove effective in detecting the noise words in the dataset, sensitive to dataset characteristics.

Keyword indexing is a method of assigning weights to terms in a dataset for the purpose of **Information Retrieval** [BYRN99]. A popular approach to keyword indexing looks at the rarity

of a term in the dataset based on frequency data using TFIDF [SB88]. The ability of the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ heuristics to capture burstiness characteristics, as compared to the IDF measure was demonstrated in Chapter 8. Hence it is worth evaluating the $\widetilde{\lambda_1}/\widetilde{\lambda_2}$ ratio or some transformation (possibly logarithm) of this ratio for keyword indexing.

In **Corpus Linguistics**, the arrival of a new dataset involves hours of manual work to identify the important terms in the corpus. The term re-occurrence model might be used to automatically rank the terms in a dataset with respect to their importance using burstiness information.

In **Machine Translation**, often one word in a certain language translates to multiple words in another language and the challenge is the task of **word sense disambiguation** in the context to which the word is translated. The term in the source language and the multiple terms in the target language can be modeled based on term burstiness and re-occurrence. Then the target term in the translated language with maximum distributional similarity is chosen as the correct translation of the source term. This approach has been adopted with success in the translation of text from Swedish to English [Arg06].

The thesis discusses the issue of individual term burstiness but in reality various single and multi-word terms might refer to a similar concept. This is referred to as **Concept burstiness**. For instance, an article about *Saddam Hussein* might refer to the same person using varying terms and phrases like *Saddam*, *Hussein*, *Mr Hussein*, *former president of Iraq* and even referring to him using terms like *he*, *his* or *him*. Such a pattern of concept burstiness can also occur for company names like *International Business Machines* referred as *IBM*, *Intl Bus Mac* or some other variants. A set of single and multi-word terms referring to the same concept can be determined using methods of anaphora resolution and coreferencing [BDM+02, MY03, BB98, GA04, PNH06, Mit99]. Studying the burstiness pattern of these various single and multi-word terms would provide an understanding of the concept's overall burstiness in the dataset, extending the reach of any such approach which

looks at terms only.

9.3.2 Limitations and extension to the model

Consider the following scenario, where a certain term, say *electricity*, occurs exactly 5 times in two different documents (both quite long). In one document all five occurrences of the term are within a few paragraphs and hence it is a bursty content word. In the other document the term occurs in a scattered manner throughout the document and does not reveal a bursty characteristic. Our current model of term burstiness provides a single set of parameters for the entire dataset, and there are no document specific parameters to judge a term's usage in the document. Also, as Katz [Kat96] states, the occurrence of a content word is classified as *topical* or *non-topical* depending on whether it occurs once or many times in the document. The current model cannot account for alternative behaviour of a term across different documents.

In this thesis the proposed model has been applied to the task of comparison between text of genres (Chapter 8). This required modeling the term distribution for each dataset separately and then comparing them. A quantitative measure of the difference in the term's characteristic across different datasets was not proposed. Such a quantitative measure could bring out measurable differences between the datasets. This might further be extended by replacing each dataset with a category in the text classification task. And the improved model may bring out differences between different categories with respect to individual terms.

Frequency based methods for text classification would compute the similarity between two documents with respect to a certain term by computing the similarity between the frequency counts. There is no way in which a term's positional information can be encoded to measure document similarity for text classification. A document similarity measure that can encode positional information about a term would be useful for document clustering, as documents with similar distribution of

terms will be assigned closer to each other. But the current model of term burstiness cannot provide such a document specific measure for comparison purpose.

Hierarchical modeling might address these deficiencies and limitations. In Bayesian approaches, hierarchical models are useful when a population can be broken down into sub-populations that have broad similarities, but are not identical. In such a case, each subpopulation is an individual dataset or category (topic/author/genre) and the population consists of all the text across all datasets or all categories. At the lowest level of the hierarchy, models with similar structure but with different parameter values are fitted to each subpopulation. However, although the parameter values differ from one subpopulation to another, it is assumed that they are unlikely to differ dramatically. At the next level of the hierarchy, a probability distribution is given to the values taken by a parameter in the different sub-populations. This approach enables the information in one sub-population to "borrow strength" from the information provided by other sub-populations. It also helps to identify a subpopulation that has unusual features. Hierarchical models thus allow direct comparison between term characteristics across different topics, genres, datasets or authors.

As an example of a hierarchical model, suppose we are interested in the term distribution of the word "grant" in a dataset of many texts. In a model we have proposed for term burstiness, the number of words between successive occurrences of the word "grant" would follow a mixture-exponential distribution with parameters p, λ_1 and λ_1. It is probably unreasonable to suppose that the distribution of "grant" is identical in the different texts, so p, λ_1 and λ_1 will take different values in the different texts. For the i^{th} text, let p_i , λ_{1i} and λ_{2i} denote the values of the parameters. A hierarchical model allows us to make the reasonable assumption that the different values of p_i are not unrelated, but come from some (unknown) probability distribution. Similarly for the λ_{1i} and the λ_{2i}. These distributions form the second layer of the hierarchy and are often the distributions of most interest. These distributions provide overall information about the term's characteristics across

different texts. Moreover, the distribution of the p_i values need not be considered separately from the distributions of the λ_{1i} and the λ_{2i} values. Rather, they may all be given a joint distribution, adding to the richness of the model. Also, this scheme extends in an obvious way when there are sub-sub-populations, or sub-sub-sub-populations, etc.

A hierarchical model can thus be constructed with individual documents as the subpopulation and a population that consists of the entire dataset. In such a setting the term's characteristic in every document can be noted. A document where the term occurs only once will have a different set of parameters as compared to one where the term occurs several times and even different parameters where the term does not occur at all. Such a model can differentiate whether a term occurs with bursty or scattered characteristics in a document. Hierarchical models might thus help in differentiating between term characteristics across documents, like *topical* or *non-topical* behaviour.

In another setting each sub-population might consist of text of different genres (or topic) and the population consists of all the documents across all genres (or topics). Here the difference in parameters of the second layer of the hierarchy to the parameters of each of the sub-populations would provide a measure of difference between each genre (or topic). This measure can then further be used for the purpose of genre (or topic) classification. A new text document can then be modeled based on the hierarchical framework. And similarity in parameters of this document with each of the categories will help in deciding which genre (or topic) the document belongs to.

This concluding chapter summarizes the main contributions of the thesis. The thesis tackles the issue of term burstiness in text and provides large scale experimental evidence of term burstiness on a dataset. A model of term burstiness is proposed in the thesis based on a term's re-occurrence pattern as it is reflected in gaps between term occurrences. Predictions made by the model parameters are either superior or on par with methods based on frequency counts alone in capturing different

term behaviours. The current research has opened up several avenues for further research. These areas where the term burstiness model might be applied are then discussed. The model proposed in the thesis has certain limitations. These limitations are discussed and ways of overcoming these limitations by means of hierarchical modeling are proposed.

Appendix A

Example Documents

Defiant Saddam refuses to plead
Iraq's deposed leader Saddam Hussein has refused to enter a plea after detailed charges were formally presented at his trial in Baghdad.
The chief judge read out specific charges against him relating to the killings of Shia Muslims in 1982.
"This is no way to treat the president of Iraq," Saddam Hussein said when asked to plead guilty or not.
After Saddam and seven co-defendants heard the charges against them, the defense starting presenting its case. Under the Iraqi legal system, the court first hears the prosecution evidence and then the judges decided on the specific charges to be brought.
I am the president of Iraq according to the will of the Iraqis and I am still the president up to this moment Saddam Hussein. The charges read out by Chief Judge Raouf Abdel Rahman relate to the defendants' alleged roles in the crackdown on the town of Dujail in 1982 after a failed assassination attempt on Saddam Hussein.
Saddam Hussein was accused of ordering:
* The illegal arrest of 399 people
* The torture of women and children
* The destruction of farmland
* The murder of nine people in the early days of the crackdown
* The murder of 148 people in the later phase of the crackdown
Saddam Hussein, who if found guilty could face the death penalty, refused to enter a plea.
"I can't just say yes or no to this. You read all this for the sake of public consumption, and I can't answer it in brief," he said. "You are before Saddam Hussein, president of Iraq. I am the president of Iraq according to the will of the Iraqis and I am still the president up to this moment."
The judge ordered the court to record that Saddam Hussein had denied the charges and then read out charges against the other defendants. The first of these was Barzan al-Tikriti, the former head of the intelligence service, who was charged with the same crimes as his half-brother, Saddam Hussein.

Table A.1: Example of concept burstiness in an article from a BBC news where the underlined terms refer to the same concept.

Bush Stumbles On Pearl Harbor, But Crowd Applauds Him
Republican **George** Bush stunned an American Legion convention Wednesday by mistakenly saying Sept. 7 marked the anniversary of the attack on Pearl Harbor that led America into World War II. The gaffe which he corrected after prompting from the audience overshadowed a campaign day in which the vice president also stumped for Jewish votes at a B'nai B'rith convention in Baltimore, Md.
Bush said he opposes creation of an independent Palestinian state in the Middle East, but also cautioned Israel against annexation of the occupied territories or exercising permanent control by military occupation. In another matter signaling a departure from the Reagan administration, Bush told Knight-Ridder newspapers Wednesday he will soon outline a proposal calling for "some adjustment" in the minimum wage.
The administration opposes raising the minimum wage, which is now $3.35 an hour. But campaign sources told Knight-Ridder that Bush was expected to link a raise in the minimum wage to a sub-minimum wage "training differential" for teen-agers and other new hires. Congress has opposed such a differential in the past.
"We ought to maintain a minimum wage differential for training people," Bush said. Bush has been criticized by his opponents as insensitive to the needs of working people. The vice president received encouraging news from the latest ABC News-Washington Post poll which found Bush leading Dukakis by an eight-point margin.
The survey of 1,104 likely voters found that 51 percent favored Bush while 43 percent prefer Dukakis. The poll, conducted Aug. 31 to Sept. 6, had a margin of error of 3.5 percent. On the question of a Palestinian state, Bush sharply criticized Democratic rival Michael Dukakis, although he didn't mention him by name. "Anyone who has trouble making up his mind on this issue, or who proposes to leave it open, just doesn't understand the dangers to Israel and to the United States; just doesn't understand the very real threats that continue to exist," Bush said.
Bush opened the day in Louisville with a speech before about 6,000 veterans at their 70th annual meeting. Departing from his prepared remarks, Bush said, "I wonder how many Americans remember today is Pearl Harbor Day. Forty-seven years ago to this very day we were hit and hit hard at Pearl Harbor and we were not ready.
"In a Bush administration that lesson would not be forgotten," said Bush, who was a Navy flyer decorated for combat missions during the war. "It would guide my defense and foreign policy." The words were barely out of his mouth before the crowd turned uneasy. A buzz of mumurs rose among the audience as legionnaires began whispering about it. "I thought nobody would forget that date," one veteran commented to another. A minute after his mistake, Bush stopped dead in his speech, alerted by the stir in the audience and people waving at him over his error.
"Did I say Sept. 7th? Sorry about that," Bush said, adding quickly that the correct date of the attack was Dec. 7, 1941. The audience applauded his correction and generally gave him a warm, polite reception. Bush told reporters later, "I just got messed up. I wanted to work Pearl Harbor in and just got carried away." It was the second rocky day in a row for Bush's campaign, following a booing, jeering reception from shipyard workers in Portland, Ore. on Tuesday.
Sheila Tate, Bush's press secretary, said the vice president had ad libbed his Pearl Harbor remark, perhaps because it was 44 years ago this week that Bush, a decorated Navy flyer in the war, was shot down on over the Pacific. The gaffe marred Bush's attempt to highlight differences with Dukakis. He accused Dukakis of trying to "cancel and delay our strategic modernization with what amounts to an undeclared unilateral freeze."
Bush noted that Dukakis opposes the MX and Midgetman missiles, the Strategic Defense System and has called for cancellation of two aircraft carrier task forces. Bush said Dukakis has opposed the Stealth bomber in the past "although this may be changing." "Let me be clear," Bush said. "I do not question his patriotism. But patriotism is not the issue. The issue is how best to deter war, to keep the peace, to fulfill our country's special responsibility as leader of the free world."
Bush said, "I think America wants tough, tested, experienced leadership .... I am proud of having served my country in combat. I believe I have the fiber and the experience to lead this country." Later, in Baltimore, Bush underscored the closeness of U.S.-Israeli relations. "No threat, no stone thrown, is strong enough to divide us. No wedge will be driven between us," Bush pledged.
In warning against Israeli annexation of the occupied territories, Bush said, "There has got to be another way, and that's what the peace process is all about." Bush said he opposes creation of a Palestinian state on grounds it would be a threat to Israel as well as to Jordan, which he said is crucial to any lasting peace settlement.
Moreover, he said, "it would be contrary to American interests." On the flight from Louisville to Baltimore, Bush said that anti-abortion protestors who disrupted Dukakis' speech Tuesday in a Chicago suburb "in my view went too far."
Bush, who opposes abortion in most instances, said, "I would ask any supporter of mine to resist carrying their right to demonstrate to an extreme." Explaining his gaffe, Bush explained: "I just got messed up. I wanted to work Pearl Harbor in and just got carried away, and then I looked up and saw incredulity on the face of one particular guy down to my left and I thought, 'Whoops...My Heavens! I've done it!' So I was glad to correct it."

Table A.2: First Example of term burstiness from the AP dataset for the term bush.

Bush's Campaign Strategy: 'I'm Not Going to Mess Up'
George Bush, relaxing on Air Force Two as it raced across California, summed up his campaign strategy in just six words: "I'm not going to mess up." Once, twice, three times, four times he said it over and over within a period of a minute. It was as if he had been reciting the same sentence to himself for months, rehearsing a simple formula for winning the White House. "I can't look back and I can't look forward beyond Nov. 8, either unifocus," Bush said.
By all appearances, the man given to gaffes has succeeded in not messing up. A loser in Iowa's first-in-the-nation caucuses, he rallied to bring the primary season to a surprisingly early close. Down by 17 points in the polls in late July, he climbed back up, methodically following a game plan that eventually gave him a wide lead over his rival. He didn't loose his focus, either. Simply put, it was to paint Dukakis as an incorrigible liberal, an inexperienced governor whose leadership could not be trusted.
"Do not gamble on another liberal Democrat coming out of nowhere," Bush warned audiences. Emphasizing that Republicans have brought peace and prosperity, Bush said Dukakis wasn't up to dealing with the Soviet Union and he was sure to raise taxes. Bush concentrated his time and resources on nine battleground states: California, Texas, Ohio, Illinois, New Jersey, Michigan, Pennsylvania, Wisconsin and Missouri.
In the process, he surprised a lot of people. Once considered weak and awkward on the stump, he instead was tough and aggressive, willing to wage what was widely viewed as a negative campaign. He wasn't a wimp. He seemed confident and poised. In televised debates with Dukakis, Bush came across as steady and relaxed a far cry from the nervous, hyperactive performance he turned in four years earlier in a debate with then Democratic vice presidential nominee Geraldine Ferraro.
Of course, there were occasional foul-ups. His biggest miscue was saying that Sept. 7 instead of Dec. 7 was the anniversary of the Japanese attack on Pearl Harbor. Mixing up his words on another occasion, he said, "I hope I stand for anti-bigotry, anti-Semitism, anti-racism." He quickly had aides assure people he was not an anti-Semite.
George Herbert Walker Bush, scion of the late Connecticut senator and New York financier Prescott Bush and Dorothy Walker Bush, was born to a life of privilege in Greenwich, Conn. Bush amassed a fortune in his own right in the oil fields of Texas, and then followed his father's lead and built up an impressive resume of public service: member of Congress, ambassador to the United Nations, chairman of the Republican Party, director of the Central Intelligence Ageny, vice president. Still, Bush had not won an election in his own right since 1968.
He lost badly in the first event of the 1988 presidential campaign, the Iowa caucuses, but then regrouped and came back in New Hampshire. Then, after losses in South Dakota and Minnesota, Bush jumped ahead of the pack and virtually locked up the GOP presidential nomination with 16 primary wins on Super Tuesday. Lingering questions about Bush's role in the Iran-Contra affair threatened to derail his presiential aspirations, as did a perception that he was not his own man. Even his selection of a running mate created controversy, and Bush was forced to defend Dan Quayle throughout the fall campaign.
In the end, though, it didn't seem to matter. Bush's aides said the turning point was the Republican convention in August, when there was, in effect, a transfer of leadership and Bush stepped out of President Reagan's shadows. Bush played two roles: sweet and sour, good cop and bad cop. He slashed deeply at Dukakis. At the same time, he portrayed himself as a family man interested in family values and "a gentler, kinder nation."
Like Ronald Reagan's campaign before his, Bush's was not loaded with bold, new initiatives. He did not veered from the course of the last eight years. "When you have to change horses in midstream, doesn't it make sense to switch to the one who's going in the same way?," Bush asked. There were few specifics about how a Bush administration would work.
Bush's answer to the huge budget deficit? A "flexible freeze" that provided no clues about where he would cut spending. His campaign was loaded with "feel good" themes balloons, children, sun-drenched wheat fields similar to those Reagan used. Indeed, his commercials were made by the same team that produced Reagan's successful "Morning in America" advertisements in the 1984 presidential race.
And if Bush's speeches sometimes sounded like Reagan's, no wonder. The vice president hired former White House speechwriter Peggy Noonan, one of Reagan's top wordsmiths, for his own campaign. Bush's commercials hammered Dukakis as soft on crime, weak on defense and on the wrong side of environmental issues, such as the cleanup of the Boston Harbor echoing themes that Bush sounded on the campaign trail.
Pollster Lou Harris said, "The Bush commercials have had an enormous impact. Really more than the debates, more than anything else, they have determined the set of the election up until now." Ben Wattenberg, an American Enterprise Institute senior fellow, said, "They went after Dukakis on all the L-word (liberal) stuff, which they had some nice symbols for, prison furloughs and Pledge of Allegiance and ACLU and stuff like that."
Republican political consultant David Keene said, "Bush seized on issues of little substance (such as) the Pledge of Allegiance, prison furloughs ... that created a caricature Dukakis couldn't escape from."
In this way, Keene said, Bush changed the public perception of Dukakis. Democrats accused Bush of fanning racial tensions by emphasizing the case of Willie Horton, a black man who raped a white woman after escaping from a Massachusetts prison while on furlough. As governor of Massachusetts, Dukakis had defended granting furloughs to murderers.
"There isn't any racism," Bush said. Bush also hit Dukakis for being in favor of gun control and abortion.
By focusing on the hot-button issues, Bush hoped to win over the swing voters the Democrats who voted for Reagan in 1980 and 1984 but might go home in 1988. As the race drew to a close, polls showed Bush significantly ahead of Dukakis. "Guy asked me yesterday, 'Are you overconfident?' There's no sense of overconfidence. It's more, you know, keep on moving," Bush said. "I'm going to run like I'm 10 points back all across this country."

Table A.3: Second Example of term burstiness from the AP dataset for the term bush.

US Contributed To Settlements With Challenger Families Who Couldn't Sue
The U.S. government contributed 40 percent to settlements for two Challenger astronauts who worked for the government even though **federal** law protects the government against claims like theirs. A key legal reason: The government was concerned that Morton Thiokol would end up passing claims against it along to the **federal** government. Thiokol made the defective rockets that were blamed for the Challenger explosion, but government officials had also spurned warnings by Thiokol engineers that a launch on Jan. 28, 1986, might be unsafe because of cold temperatures at Cape Canaveral. Relatives of all seven astronauts were free to sue Morton Thiokol for damages, but only the families of the two non-government employees high school teacher Christa McAuliffe and Hughes Aircraft employee Gregory Jarvis were allowed to sue the **federal** government. Documents released by the government this week to settle a lawsuit filed under the Freedom of Information Act by The Associated Press and six other news organizations show that the government and Morton Thiokol paid $7,735,000 in cash and annuities to settle with the four spouses and six children of Dick Scobee, Ellison Onizuka, Jarvis and McAuliffe. Scobee was a civilian astronaut with the National Aeronautics and Space Administration. Onizuka was an Air Force lieutenant colonel. The government has refused to contribute to settlements reached by Morton Thiokol with the survivors of civilian NASA astronauts Ronald McNair and Judith Resnik. And last month, government lawyers opposing a lawsuit filed by the family of the other military officer aboard, Navy Cmdr. Michael Smith, got the government dropped as a defendant because he was a military officer. The government documents showed that Morton Thiokol paid 60 percent and the government 40 percent of the settlement with the four families. In a letter to the news organizations that sued, Justice Department lawyer Joanne S. Marchetta stipulated that the "single percentage split ... is applicable to all four settlements in the aggregate." The government did not require Morton Thiokol to bear a larger share of the settlement with the Scobee and Onizuka relatives, even though those families could sue only the company. In a brief filed in the FOI case, the government suggested why it contributed at all to the Scobee and Onizuka settlements. "It was entirely possible under the facts presented by the Challenger accident that the government would be brought in as a third party to any claims by the families brought against government contractors like MTI," the brief said. That reference was to the objections by Morton Thiokol engineers to launching the shuttle in extreme cold objections that were withdrawn under government pressure. Ronald Krist, a Houston lawyer who represented the McNairs, Jarvis' father and Resnik's mother, noted that a 1983 Supreme Court decision allows a contractor who pays damages to a **federal** employee's survivors to try to recover part of those damages from the government if **federal** employees shared the blame. So far, however, Morton Thiokol has made no move to do this. "That law is on the books and that would give the government the theoretical incentive to settle" with the families of government workers, Krist said. Thus, by entering into joint settlements with Thiokol and the families, the government extinguished the potential third-party claim of Thiokol against the government in these four cases and avoided trials that would have replayed the painful government errors that led to the fateful decision to launch in cold weather.

Table A.4: Example of term burstiness from the AP dataset for the term government and **federal**.

Woman Found Dead After SWAT Team Storms House
A woman opened fire in an elementary school, killing one child and critically wounding five others Friday. She was found dead when a SWAT team stormed a nearby house where she had wounded a seventh person. "I've just received information that the suspect has been found and she is dead," Police Chief Herbert Timm said nearly nine hours after the shooting began at Hubbard Woods Elementary School. "The situation is over."
She shot herself once in the head with a .32-caliber pistol, Timm said. "She indicated (in a telephone call to her parents) that she had shot some people and felt very sorry about that," he said. The woman, identified as Lori Dann, 30, was carrying three handguns around 10:45 a.m. when she opened fire inside the one-story, red-brick school, killing a child and wounding five others, police said.
"There was blood all over the classroom and desks knocked over" as the children panicked during the 15 minutes the shooting lasted, said Glencoe Patrolman John Cegliesk. "Kids were lying around in very serious condition," in the school, Timm said. "Kids were hiding under the desk as well as well as they could."
She then ran to a nearby house and wounded one of its occupants in a struggle. Hostage negotiators arrived moments later and Ms. Dann spoke by telephone to her parents, brought to the house in hopes of talking her out. Asked why police had not attempted to storm the house earlier, Timm replied, "We've had seven people shot already. We're not going in there without taking every precaution."
Timm said he decided to send the SWAT team into the house after repeated efforts by police and Ms. Dann's parents failed to contact her. The team was accompanied by FBI agents with sound detectors probing for movement in the house. Timm said police had yet to establish a motive for the rampage. Ms. Dann allegedly set a fire earlier Friday, eight blocks away at the home of a couple where she worked as a housekeeper in this affluent North Shore community.
She reportedly was upset, having been told she was losing her job because the family was moving, police said. Timm, reconstructing a day of violence at an evening news conference, said Ms. Dann set two fires and carried a gasoline can to the door of a day-care center before embarking on the shooting spree. Before going to the school Friday, police said, Ms. Dann went to the home of Padraig and Marion Rushe on Forest Glen to take two of their children, 4 and 6 years old, to the zoo.
Instead of going to the zoo, however, Timm said Ms. Dann took the children to Ravinia School in neighboring Highland Park, where she set off one of several incendiary devices she carried in her car Friday. Authorities found one of the devices burning there, but extinguished it before it could cause any damage. Timm said Ms. Dann drove next to a day-care center, where she went to the front door carrying a can of gasoline. "But she was met at the front door by a staff member who asked what she was doing and she took off," said Timm. He said Ms. Dann returned to the Rushe home, talked with Mrs. Rushe and the children in a family room in the basement, then went upstairs and set off another incendiary device. While the fire raged upstairs, Mrs. Rushe managed to push her two children through a small window in the basement, then crawl out herself. Ms. Dann, meanwhile, made her way about eight blocks to the school, police said. Timm said several other incendiary devices were found in Ms. Dann's car in the neighborhood near the house where the standoff took place. The FBI said it had "an investigative interest" in Ms. Dann and Madison, Wis., police Lt. William Sprague said she had been arrested there March 14 on a charge of retail theft. He gave no further details. Keith Wilson told the Wisconsin State Journal from his home in Los Angeles that he had roomed across the hall from Ms. Dann when they attended the University of Wisconsin. He described her as "just an absolutely bizarre lady" and said "sometimes she'd wander the halls and try to open doors." Timm, the police chief, said when the woman walked into the school around 10:45 a.m., her first stop was a boys' bathroom, where she shot a youngster. "She left him and ran into one classroom, telling the teacher there that a boy had been wounded," said Timm. "Then she entered a second classroom, announced she had a gun and opened fire." Police recovered a .357-caliber Magnum revolver from the school, for which Timm said Ms. Dann had been issued a permit. "How did a woman with that kind of background get licensed to carry a gun?" Timm said. Joe Sumner, director of Winnetka police operations, said the woman then ran through the woods from the school and entered the home, where four members of a family and their maid were present. "Apparently, she just walked in and confronted the homeowners," Sumner said. "The mother came running out and said, 'There's a woman in my home with a gun.' "Then we heard a shot. The son came out, holding a gun, and fell over in the driveway," Sumner said. The son, 20-year-old Philip Andrew, apparently struggled with the suspect, grabbed one of her weapons, and was shot in the chest, Sumner said. The father, maid and grandfather fled by a side door, leaving her alone inside. Wilmette police Sgt. Michael Geier said the woman was believed to have set fire to the home where she worked while a mother and her children were in the basement doing laundry. "This woman set a fire in the house, trapping the people in the basement," he said. "They escaped, got out, apparently without serious injury." Eight-year-old Nick Corwin died of his wounds at Highland Park Hospital, said spokesman Mark Newtwon. Another 8-year-old boy and Andrew were in critical condition at the hospital, said spokeswoman Sue Masaracchia. The four children at Evanston Hospital were reported in critical but stable condition with gunshot wounds following surgery, said spokeswoman Mary Ash. They were identified as Mark Tebourek, 8, Robert Trossman, 6, Lindsay Fisher, 8, and Kathryn Ann Miller, 7, said another spokeswoman, Cheryl Soohoo. At the school, parents clustered in groups, consoling each other. Village Manager Robert Buechner said children were released in small groups throughout the afternoon to ensure their safety. Ann Arnold, who lives across the street from one of the wounded children, said her own 7-year-old son made sure the doors were locked after he came home. "He's still scared something else is going to happen," Mrs. Arnold said.

Table A.5: Example of term burstiness from the AP dataset for the term said.

farm credit administration
Special Meeting
summary:
Notice is hereby given pursuant to the Government in the Sunshine Act (5 U.S.C. 552b(e)(3)), that the special meeting (53 FR 94, January 4, 1988) of the Farm Credit Administration Board (Board) scheduled for January 8, 1988 was cancelled due to inclement weather conditions. The matters scheduled to be considered at that meeting were addressed at the regular meeting held on January 12, 1988.
for further information contact: David A. Hill, Secretary to the Farm Credit Administration Board, 1501 Farm Credit Drive, McLean, Virginia 22102-5090, (703) 883-4003.
address: Farm Credit Administration, 1501 Farm Credit Drive, McLean, Virginia 22102-5090.
Dated: January 14, 1988.
David A. Hill,
Secretary, Farm Credit Administration Board.
FR Doc. 88-1018 Filed 1-14-88; 8:45 am
BILLING CODE 6705-01-M

Table A.6: Example document from the FR dataset with only one occurrence of the term of.

DEPARTMENT OF THE INTERIOR
Bureau of Land Management
Craig, Colorado Advisory Council Meeting
Time and Date: February 17, 1988, at 10:00 a.m.
Place: Little Snake Resource Area, 1280 Industrial Avenue, Craig, Colorado.
Matters To Be Considered:
1. Status of Little Snake Resource Management Plan Protests
2. High Desert 300 Race Monitoring Results
3. 1988 Motorcross EA
4. Potential Land Exchange in Piceance Basin
5. Status of Oilshale Tract Ca, Cb, and Wolf Ridge Corporation's Nahcolite EIS
6. Election of Officers
Contact Person For More Information: Mary Pressley, Craig District Office, 455 Emerson Street, Craig, Colorado 81625-1129, Phone: (303) 824-8261.
Dated: December 10, 1987.
Mary Pressley,
Acting Associate District Manager.
FR Doc. 87-30179 Filed 12-31-87; 8:45 am
BILLING CODE 4310-JB-M

Table A.7: Example document from the FR dataset with only one occurrence of the term the.

NEW TO THIS ADVISORY: WASHINGTON-dated AM-Markets-Crisis
Gameplan, The Bush administration conferred closely with foreign allies, the Federal Reserve Board and Wall Street officials through the weekend in mapping a strategy for averting a "Black Monday" crisis that, in the end, never developed, sources said. Latest version moved as f0342. NEW YORK-dated AM-Markets-Explainer, a question-and-answer explanation of Monday's events in the financial markets. Latest version moved as f0353.
NEW YORK Stock prices gyrate in the heaviest morning of trading in Wall Street history, shoved around by speculation in stock-index futures in Chicago. Slug AM-Markets Rdp. Latest version moved as f0348.
NEW YORK With a worldwide audience watching, the stock market refused to reenact the crash of 1987 Monday. Instead, it zigzagged wildly in a record-breaking barrage of trading that left the analysts and statisticians who follow it gasping to keep up. Slug AM-Market Analysis. Latest version moved as f0294.
UNDATED European stock markets tumble in the wake of Wall Street's Friday the 13th selloff, ignoring the moderate selloff earlier in Tokyo and girding for further turmoil. Slug AM-Markets-Foreign. Latest version moved as f0269.
WASHINGTON President Bush says he's not worried by the stock market's gyrations the Monday after the big drop, while Federal Reserve Chairman Alan Greenspan reassures jittery bankers that he is coordinating closely with the administration and foreign governments to keep any fallout from harming the U.S. economy. Slug AM-Markets-Government. Latest version moved as f0281.
With:
AM-Markets Glance. Moved as f0190.
AM-Markets-Glossary. Moved as f0170.
AM-Markets-Dow-Quarter Hour. Moved as f0293.
AM-Markets-Dow Point Gains. Moved as f0292.
AM-Markets-NYSE Volume List. Moved as f0296.
NEW YORK-dated AM-Markets-Bonds, The U.S. Treasury bond market falls sharply as investors cash in on Friday's big price runup that had stemmed from a move to safer fixed-income investments from a plummeting stock market. High-yield "junk" bonds also continue to decline in early activity. Latest version moved as f0315.
BOSTON-dated AM-Market-Mutual Funds, Mutual fund investors flood their funds with phone calls and many shift their money out of stocks in reaction to the turmoil on Wall Street. Latest version moved as f0206.
CHICAGO-dated AM-Markets-Index Futures, Stock index futures rebound from an opening selloff as nervous traders return to the pits following Friday's plunge in prices. Latest version moved as f0345.
NEW YORK-dated AM-Markets-Circuit Breakers, How the trading restrictions aimed at cooling off any severe market selloffs are supposed to work and how they actually did in the wake of the Friday the 13th plunge. Latest version moved as f0254.
NEW YORK-dated, AM-Markets-Takeovers, Takeover-related issues that were hit hard in Friday's selloff take further lumps as speculators register doubts about the future of such deals. Latest version moved as f0332.
LONDON-dated AM-Markets-Europe, Share prices nosedive across Europe, disrupting trading in several financial centers and prompting authorities to step in to try to contain the damage. Latest version moved as f0288.
TOKYO-dated AM-Markets-Japan, to be updated with Tokyo Stock Exchange trading that begins 8 p.m. EDT. Latest version moved as f0204.
NEW YORK Investors on Main Street hold their breath while watching the wild gyrations on Wall Street, but there are no early signs of frenzy. "Most brokers are virtually frozen into immobility because the market's moved so fast," a Chicago investor says. Meanwhile, a Texas trader notes, "I don't see any panic at all."
Slug AM-Markets-Main Street. Latest version moved as f0196.
NEW YORK Out-of-town tourists, New Yorkers with a few hours to kill and a dozen scraggly self-proclaimed anarchists shouting "sell today, jump tomorrow" ring the cavernous New York Stock Exchange building in southern Manhattan, where trading spasms raised fears of another Black Monday panic.
Slug AM-Markets-Scene. Latest version moved as f0311.
FORT LAUDERDALE, Fla. Juggling calls from nervous clients while watching the market continue to fall, stockbroker John Rodstrom Jr. faced a morning reminiscent of Black Monday two years ago. But the broker at Kidder Peabody; Co. does not believe the market is looking at a similar crisis.
Slug AM-Markets-Stockbroker. Latest version moved as f0274.
AP Business News.

Table A.8: Example document from the AP dataset with frequent occurrences of the term as.

Structural defects in GaAs related to excess As were characterized and their behavior upon heat treatments studied. The observed defects included precipitates and dislocations. Results showed most of the precipitates in As-rich GaAs to the rhombohedral arsenic. Two exceptions were observed in an In-doped LEC (liguid encapsulated Czochralski) GaAs, which were As-rich but could not be further identified. Some of the observed As precipitates showed a simple orientation relationship with the matrix which yields structural coherence between As precipitates and GaAs matrix. Other As precipitates showed less coherent orientation. The dislocation loops in As-rich GaAs consisted a faulted loop with Shockley type Burgers vector and a perfect loop associated with an extra l_brace111r_brace plane. It was proposed that these loops were formed **as** a result of dual condensation of both excess As interstitials and Ga vacancies, followed by generation and movement of Shockley partial dislocations. These precipitates and dislocation loops disappear after annealing, indicating a solvus temperature between 600–700degreeC. The EL2 concentration increased **as** the defects dissolved, showing the defects to be the source of the excess As required to form EL2. The implication is that the As interstitial and Ga vacancies coexist in GaAs at high temperatures, which indicates that these point defects are responsible for the formation of arsenic antisites by direct combination. During the cooling period, they freeze into the matrix **as** point defects during a rapid cooling and condense **as** dislocation loops and precipitates during very slow cooling, in the dislocation-free region of the crystals. Around dislocations, the excess As precipitates heterogeneously even during rapid cooling. 217 refs.

Table A.9: Example document from the DOE dataset where the Arsenic, As is confused and hence conflated to the function word **as**.

Appendix B

WinBUGS modeling code

```
model
{
  for(i in 1:N)
  {
    for (j in (k[i]+1) :  k[i+1])
    {
      w[j] ~ dweib(1, mu[i]) I(cen[j], )
    }
     mu[i] ← lambda[M[i]]
     M[i] ~ dcat(P[] )
  }
   P[1:2] ~ ddirch(alpha[])
   alpha[1] ← 1
   alpha[2] ← 1
   theta ~ dnorm(0.0,1.0E-6) I(0, )
   lambda[2] ← lambda[1] + theta
   lambda[1] ~ dnorm(0.0, 1.0E-6) I(0, )
}

# INITIAL PARAMETER VALUES
 list( lambda = c(0.1, NA), theta = 0.1)
```

Table B.1: WinBUGS code (first version, used in this thesis) for the term re-occurrence model used for modeling term burstiness.

```
model
{
  for (j in 1 :  1)
  {
     w[j] ~ dweib(1, mu[j]) I(cen[j], )
     mu[j] <- lambda[M[j]]
     M[j] ~ dcat(P[])
  }

  for (j in 2 :  T)
  {
     w[j] ~ dweib(1, mu[j]) I(cen[j], )
     mu[j] <- lambda[Mod[j]]
     Mod[j] <- equals(cen[j], 0)*(1 - equals(cen[j-1], 0))
         + (1 - equals(cen[j], 0) )*M[j]
     M[j] ~ dcat(P[])
  }

   P[1:2] ~ ddirch(alpha[]);
   alpha[1] <- 1;
   alpha[2] <- 1;
   lambda[1] ~ dnorm(0.0, 1.0E-6) I(0, )
   limit2 <- 1 - lambda[1]
   theta ~ dnorm(0.0, 1.0E-6) I(0, limit2)
   lambda[2] <- lambda[1] + theta
}

# INITIAL PARAMETER VALUES
 list( lambda = c(0.1, NA), theta = 0.1)
```

Table B.2: WinBUGS code (improved) for the term re-occurrence model used for modeling term burstiness.

Appendix C

English stop word list

a about above across after afterwards again against all almost alone along already also although always
am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere
are around as at
back be became because become becomes becoming been before beforehand behind being below beside
besides between beyond bill both bottom but by
call can cannot cant co computer con could couldnt cry
de describe detail do done down due during
each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything every-
where except
few fifteen fify fill find fire first five for former formerly forty found four from front full further
get give go
had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how
however hundred
i ie if in inc indeed interest into is it its itself
keep last latter latterly least less ltd
made many may me meanwhile might mill mine more moreover most mostly move much must my myself
name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere
of off often on once one only onto or other others otherwise our ours ourselves out over own
part per perhaps please put rather re
same see seem seemed seeming seems serious several she should show side since sincere six sixty so some
somehow someone something sometime sometimes somewhere still such system
take ten than that the their them themselves then thence there thereafter thereby therefore therein
thereupon these they thick thin third this those though three through throughout thru thus to together
too top toward towards twelve twenty two
un under until up upon us very via
was we well were what whatever when whence whenever where whereafter whereas whereby wherein
whereupon wherever whether which while whither who whoever whole whom whose why will with within
without would
yet you your yours yourself yourselves

Table C.1: Standard English stop word list obtained from the *University of Glasgow* website.

References

[Aar99] Scott. Aaronson. Stylometric clustering: a comparison of data driven and syntactic features. Available on: `http://www.cs.berkeley.edu/ aaronson/sc.doc`, 1999.

[Ada01] Lada A. Adamic. Zipf, power-laws, and pareto - a ranking tutorial. Available online from: `http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html`, 2001.

[Aiz03] Akiko Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing Management*, 39(1):45–65, 2003.

[AKFS03] S. Argamon, M. Koppel, J. Fine, and A. R. Shimony. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346, 2003.

[Arg06] Atelach Alemu Argaw. Burstiness patterns in query analysis. GSLT Machine Learning II, Department of Computer and Systems Sciences, Stockholm University/KTH, May 10, 2006.

[Baa01] H. Baayen. *Word frequency distributions.* Kluwer Academic Publishers., 2001.

[BB98] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proc. 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th Conf. on Computational Linguistics (COLING)*, pages 79–85, San Francisco, California, 1998.

[BDM+02] Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. Shallow methods for named entity coreference resolution. In *Proceedings of TALN*, 2002.

[Bre97] M. R. Brent. Toward a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research*, 26:363–375, 1997.

[BS74] Abraham Bookstein and Don R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Sceince*, 25(5):312–318, 1974.

[Bur80] Quentin L. Burrell. A Simple Stochastic Model for Library Loans. *Journal of Documentation*, 36:115–132, 1980.

[BYRN99] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press / Addison-Wesley, 1999.

[Cav02a] Gabriela Cavagliá. Measuring corpus homogeneity using a range of measures for inter-document distance. In *Third International Conference on Language Resources and Evaluation.*, pages 426–431, 2002.

[Cav02b] Gabriela Cavagliá. Measuring the homogeneity of different varieties of language. In *Proceedings of the* 5^{th} *National Colloquium for Computational Linguistics in the UK*, pages 37–44, 2002.

[CC96] Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[CDAR97] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB*

'97: Proceedings of the 23rd International Conference on Very Large Data Bases, pages 446–455, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[CG95a] K. Church and W. Gale. Inverse document frequency (idf): A measure of deviation from poisson. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130, 1995.

[CG95b] K. Church and W. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[Cha01] C. Chaski. Empirical evaluation of language-based author identification techniques. *Forensic Linguistics: International Journal of Speech, Language and Law*, 8(1):1–64, 2001.

[Chu95] Kenneth Ward Church. One term or two? In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 310–318, New York, NY, USA, 1995. ACM Press.

[Chu00] K. Church. Empirical estimates of adaptation: The chance of two noriega's is closer to p/2 than p^2. In *COLING*, pages 173–179, 2000.

[CK01] Gabriela Cavagliá and Adam Kilgarriff. Corpora from the Web. In *Proceedings of the* 4^{th} *National Colloquium for Computational Linguistics in the UK.*, 2001.

[Col03] David Collett. *Modelling Survival Data in Medical Research.* Chapman and Hall/CRC Second edition, April 2003.

[CS03] Ross Clement and David Sharp. Ngram and bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, 18(4):423–447, 2003.

[Dai96] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts, 1996.

[DKLP03] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123, 2003.

[DP96] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Proceedings of the 13th International Conference on Machine Learning*, pages 105–112, 1996.

[DP97] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[DRSG04a] Anne De Roeck, Avik Sarkar, and Paul H. Garthwaite. Defeating the homogeneity assumption. In Gerald Purnelle, Cedrick Fairo, and Anne Dister, editors, *Proceedings of 7th International Conference on the Statistical Analysis of Textual Data (JADT)*, pages 282–294, De Louvain, Belgium, 2004. UCL Presses Universitaries.

[DRSG04b] Anne De Roeck, Avik Sarkar, and Paul H Garthwaite. Frequent term distribution measures for dataset profiling. In Maria Teresa Lino, Maria Francisca Xavier, Fatima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th International conference of Language Resources and Evaluation (LREC)*, pages 1647–1650, Paris, France, 2004. European Language Resources Association (ELRA).

[DRSG05] Anne De Roeck, Avik Sarkar, and Paul H. Garthwaite. Even very frequent function

words do not distribute homogeneously. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *The International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 429–436, Shoumen, Bulgaria, 2005. INCOMA Ltd.

[Dun93] Ted E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[Edm69] H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, 1969.

[Elk05] Charles Elkan. Deriving TF-IDF as a Fisher kernel. In *In the Proceedings of the twelfth edition of the Symposium on String Processing and Information Retrieval (SPIRE)*, pages 296–301, 2005.

[EP98] Jim Entwisle and David Powers. The present use of statistics in the evaluation of NLP parsers. In Jill Burstein and Claudia Leacock, editors, *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Language Learning*, pages 215–224. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[Fin93] S. Finch. *Finding Structure in Language.* PhD thesis, University of Edinburgh, 1993.

[FK82] W. Francis and H. Kucera. *Frequency analysis of English usage.* John Wiley and Sons, Houghton Mifflin Company (Boston MA), 1982.

[FNSO04] T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura. Identification of bursts in a document stream. In *In First International Workshop on Knowledge Discovery in Data Streams (in conjunction with ECML/PKDD 2004)*, 2004.

[Fra97] Alexander Franz. Independence assumptions considered harmful. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 182–189, 1997.

[GA04] C.H. Gooi and J. Allan. Cross-document coreference on a large scale corpus. In *Proc. Human Language Technology/North American chapter of Association for Computational Linguistics annual meeting (HLT/NAACL)*, pages 9–16, Boston, USA, May 2004.

[GCSR95] A. Gelman, J. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, London, UK, 1995.

[GDR01] A Goweder and Anne De Roeck. Assessment of a significant Arabic corpus. In *In proceedings of the* 39^{th} *ACL Workshop on Arabic Language Processing*, 2001.

[Gil90] Larry Gillick. Probabilistic Models for Topic Spotting. In *Workshop Notebook for the WHISPER Meeting at MIT Lincoln Laboratory, July 25–26*, pages 206–211, 1990.

[Gil92] W. R. Gilks. Derivative-free Adaptive Rejection Sampling for Gibbs Sampling. *Bayesian Statistics*, 4:641–649, 1992.

[Har75a] S.P. Harter. A probabilistic approach to automatic keyword indexing, part 1: On the distribution of speciality words in a technical literature. *Journal of the American Society for Information Science*, 26:197–206, 1975.

[Har75b] S.P. Harter. A probabilistic approach to automatic keyword indexing, part 2: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, 26:280–289, 1975.

[Har92a] Donna Harman. The DARPA TIPSTER Project. *SIGIR Forum*, 26(2):26–28, 1992.

[Har92b] Donna Harman. Relevance feedback and other query modification techniques. *Information retrieval: data structures and algorithms*, pages 241–263, 1992.

[Hea94] Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[HH76] M.A.K. Halliday and Ruqaiya Hasan. *Cohesion in English.*, volume 9. English Language Series, Longman, London, 1976.

[HJ82] K Hofland and S Johansson. *Word frequencies in British and American English.* The Norwegian Computing Centre for the Humanities, 1982.

[Hoe91] Michael Hoey. *Patterns of Lexis in Text.* Oxford University Press, Oxford, 1991.

[Jan03] Martin Jansche. Parametric models of linguistic count data. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 288–295, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[JK69] N. L. Johnson and S. Kotz. *Discrete Distributions.* John Wiley and Sons, Houghton Mifflin Company (Boston MA), 1969.

[JM00] D. Jurafsky and J.H. Martin. *Speech and Language Processing.* Prentice Hall., New Jersey., 2000.

[JMRS91] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss. A dynamic language model for speech recognition. In *Proceedings of a workshop on Speech and natural language*, pages 293–295, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[Joa98] Thorsten Joachims. Text categorization with suport vector machines: Learning with

many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 137–142, London, UK, 1998. Springer-Verlag.

[Kat96] Slava M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–60, 1996.

[Kil96a] Adam Kilgarriff. Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved lob-brown comparison. In *Proceedings of ALLC-ACH Conference.*, Bergen, Norway., 1996.

[Kil96b] Adam Kilgarriff. Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition.*, Sussex, UK, 1996.

[Kil97] Adam Kilgarriff. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of ACL-SIGDAT Workshop on very large corpora*, Hong Kong, 1997.

[Kil01] Adam Kilgarriff. Comparing corpora. *Journal of Corpus Linguistics*, 6(1):1–37, 2001.

[Kil05] Adam Kilgarriff. Language is never, ever, ever, random. *Corpus Linguistics and Lingustic Theory*, 1(2):263–275, 2005.

[Kle02] Jon Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, New York, NY, USA, 2002. ACM Press.

[KM90] R. Kuhn and R. De Mori. A cache-based natural language model for speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(6):570–583, 1990.

[KNRT03] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[KPS04] Youngjoong Ko, Jinwoo Park, and Jungyun Seo. Improving text categorization using the importance of sentences. *Information Processing and Management*, 40(1):65–79, 2004.

[KR98] Adam Kilgarriff and Tony Rose. Measures for corpus similarity and homogeneity. In *Proceedings of* 3^{rd} *Conference on Empirical Methods in Natural Language Processing*, pages 46–52, Granada, Spain., 1998.

[Kwo96] K. L. Kwok. A new method of weighting query terms for ad-hoc retrieval. In *SIGIR*, pages 187–195, 1996.

[Lew98] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 4–15, London, UK, 1998. Springer-Verlag.

[LH97] Chin-Yew Lin and Eduard Hovy. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[Low99a] Stephen A. Lowe. The Beta-Binomial Mixture Model and its Application to TDT Tracking and Detection. In *Proceedings of the DARPA Broadcast News Workshop*, pages 127–132, 1999.

[Low99b] Stephen A. Lowe. The beta-binomial mixture model for word frequencies in documents with applications to information retrieval. In *In Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, pages 2443–2446, 1999.

[Luh58] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):155–164, 1958.

[Mac03] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[Man54] B. Mandelbrot. Structure formelle des textes et communication. *Word*, 10:1–27, 1954.

[Man83] B. B. Mandelbrot. *The fractal geometry of nature.* W. H. Freeman., New York, 1983.

[Mar61] M. E. Maron. Automatic indexing: An experimental inquiry. *Journal of ACM*, 8(3):404–417, 1961.

[MB04] Ketan Mane and Katy Borner. Mapping topics and topic bursts in PNAS. *PNAS*, 101:5287–5290, 2004.

[Min03] Tom Minka. Estimating a dirichlet distribution. Available on: `http://research.microsoft.com/ minka/papers/dirichlet/`, 2003.

[Mit97] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[Mit99] Ruslan Mitkov. Anaphora Resolution: The State of the Art. Working paper, University of Wolverhampton, 1999.

[MKE05] Rasmus E. Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness

using the dirichlet distribution. In *In the Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany*, 2005.

[MN98] Andrew McCallum and Kamal Nigam. A comparison of event models for Naive Bayes text classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.

[MS99] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts, 1999.

[Mun03] Robert Munro. A queuing-theory model of word frequency distributions. In *Proceedings of the First Australasian Language Technology Workshop (ALTW'03)*, 2003.

[MW84] Frederick Mosteller and David L. Wallace. *Applied Bayesian and Classical Inference. The Case of* The Federalist *Papers.* Springer-Verlag, New York, 2nd edition, 1984.

[MY03] G.S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *Proc. 7th Conf. on Natural Language Learning (CoNLL)*, pages 33–40, Edmonton, Canada, 2003.

[NSJ64] R.M. Needham and K. Sparck-Jones. Keywords and clumps. *Journal of Documentaton*, 20(1):5–15, 1964.

[Pie80] J. R. Pierce. *Introduction to Information Theory: Symbols, Signals, and Noise.* Dover Publications, New York, 1980.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[PNH06] Xuan-Hieu PHAN, Le-Minh NGUYEN, and Susumu HORIGUCH. Personal name resolution crossover documents by a semantics-based approach. *IEICE Transactions on Information and Systems*, E89-D(2):825–836, 2006.

[Pop02] Ioan-Iovitz Popescu. On the Lavalette's Nonlinear Zipf's law. Available online from: `http://alpha2.infim.ro/ ltpd/Zipf's_Law.html`, 2002.

[Pow96] David M. W. Powers. Learning and application of differential grammars. In *CoNLL97: ACL Workshop on Computational Natural Language Learning.* Association for Computational Linguistics, Madrid, 1996.

[Pow98] David M. W. Powers. Applications and explanations of Zipf's law. In Jill Burstein and Claudia Leacock, editors, *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Language Learning*, pages 151–160. Association for Computational Linguistics, Somerset, New Jersey, 1998.

[RG00] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *In proceedings of the Workshop on Comparing Corpora*, pages 1–6, 2000.

[RH92] Ronald Rosenfeld and Xuedong Huang. Improvements in stochastic language modelling. In *Fifth DARPA Workshop on Speech and Natural Language*, pages 107–111, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[RH97] Tony Rose and Nick Haddock. The effects of corpus size and homogeneity on language model quality. In *In proceedings of the ACL-SIGDAT Workshop on Very Large Corpora*, pages 178–191, 1997.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval.* Butterworth-Heinemann, Newton, MA, USA, 1979.

[RJ76] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[Rob96] Christian. P. Robert. Mixtures of distributions: inference and estimation. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 441–464, 1996.

[RSA97] Korin Richmond, Andrew Smith, and Einat Amitay. Detecting subject boundaries within text: A language independent statistical approach. In Claire Cardie and Ralph Weischedel, editors, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 47–54. Association for Computational Linguistics, Somerset, New Jersey, 1997.

[RSTK03] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pages 616–623. AAAI Press, 2003.

[RW94] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[Sal71] Gerard Salton. *The SMART Retrieval System–Experiments in Automatic Document Processing.* Prentice Hall, Englewood Cliffs, NJ, 1971.

[SB88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523, 1988.

[SB97] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Readings in information retrieval*, pages 355–364, 1997.

[Sch04] Karl-Michael Schneider. On word frequency information and negative evidence in Naive Bayes text classification. In José Luis Vicedo, Patricio Martínez-Barco, Rafael Muñoz, and Maximiliano Saiz Noeda, editors, *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004*, LNAI 3230, pages 474–485, Alicante, Spain, 2004. Springer Verlag.

[SDR04] Avik Sarkar and Anne De Roeck. An evaluation framework to judge the suitability of a non-english corpus for language engineering research. In Maria Teresa Lino, Maria Francisca Xavier, Fatima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th International conference of Language Resources and Evaluation (LREC)*, pages 1639–1642, Paris, France, 2004. European Language Resources Association (ELRA).

[SDRG05] Avik Sarkar, Anne De Roeck, and Paul H. Garthwaite. Term re-occurrence measures for analyzing style. In Shlomo Argamon, Jussi Karlgren, and James G. Shanahan, editors, *Proceedings of the SIGIR 2005 workshop on Stylistic Analysis Of Text For Information Access.*, 2005.

[Seb02] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[SGDR05] Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. A Bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the Ninth Conference on*

Computational Natural Language Learning (CoNLL-2005), pages 48–55, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[Sim55] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[SJ72] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[SL68] G. Salton and M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM*, 15(1):8–36, 1968.

[SP98] Robert Steele and David Powers. Evolution and evaluation of document retrieval queries. In David M. W. Powers, editor, *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning*, pages 163–164, Flinders University, Adelaide, Australia, 1998. ACL Association for Computational Linguistics.

[STBL03] D.J. Spiegelhalter, A. Thomas, N. G. Best, and D. Lunn. WinBUGS: Windows version of Bayesian inference Using Gibbs Sampling, version 1.4, 2003.

[TK03] Jaime Teevan and David R. Karger. Empirical development of an exponential probabilistic model for text retrieval: Using textual analysis to build a better model. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*, 2003.

[TRE] TREC. Text REtrieval Conference. Online: `http://trec.nist.gov/`.

[TST97] A.A. Tsonis, C. Schultz, and P.A. Tsonis. Zipf's law and the structure and evolution of languages. *Complexity*, 2(5):12–13, 1997.

[UC00] K. Umemura and K. Church. Empirical term weighting and expansion frequency. In *Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 117–123, 2000.

[WC93] Zheng Wang and Jon Crowcroft. Analysis of burstiness and jitter in real-time communications. In *SIGCOMM '93: Conference proceedings on Communications architectures, protocols and applications*, pages 13–19, New York, NY, USA, 1993. ACM Press.

[WS92] John Wilbur and Karl Sirotkin. The automatic identification of stop words. *Journal of Information Science*, 18(1):45–55, 1992.

[YL99] Y. Yang and X. Liu. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkley, August 1999.

[YP97] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[YW96] Yiming Yang and John Wilbur. Using corpus statistics to remove redundant words in text categorization. *Journal of the American Society of Information Science*, 47(5):357–369, 1996.

[Zip49] George K. Zipf. *Human Behaviour and the Principle of Least-Effort.* Addison-Wesley, Cambridge MA, 1949.