American University in Cairo

AUC Knowledge Fountain

Theses and Dissertations

2-1-2015

Hyperlink-extended pseudo relevance feedback for improved microblog retrieval

Tarek Elganainy

Follow this and additional works at: https://fount.aucegypt.edu/etds

Recommended Citation

APA Citation

Elganainy, T. (2015).*Hyperlink-extended pseudo relevance feedback for improved microblog retrieval* [Master’s thesis, the American University in Cairo]. AUC Knowledge Fountain.
https://fount.aucegypt.edu/etds/1210

MLA Citation

Elganainy, Tarek. *Hyperlink-extended pseudo relevance feedback for improved microblog retrieval*. 2015. American University in Cairo, Master's thesis. *AUC Knowledge Fountain*.
https://fount.aucegypt.edu/etds/1210

This Thesis is brought to you for free and open access by AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact mark.muehlhaeusler@aucegypt.edu.

THE AMERICAN UNIVERSITY IN CAIRO

SCHOOL OF SCIENCES AND ENGINEERING

Hyperlink-Extended Pseudo Relevance Feedback for Improved Microblog Retrieval

A thesis submitted to

Department of Computer Science and Engineering

In partial fulfillment of the requirements for the degree of

Masters in Computer Science

By: Tarek Elganainy

Supervisor: Prof. Dr. Ahmed Rafea

Summer 2014

ACKNOWLEGMENTS

All the experiments conducted in this work are based on TREC (Text REtrieval Conference) 2011, 2012, and 2013 Microblog track data (trec.nist.gov/data/tweets/). We would like to thank TREC for offering the tweets ids, tools to crawl the dataset and the evaluation sets (www.github.com/lintool/twitter-tools/wiki/Tweets2011-Collection).

I would like to thank Dr. Jimmy Lin (www.umiacs.umd.edu/~jimmylin/), for his continuous help and support for the TREC Microblog Track APIs.

Special thanks goes to Dr. Ahmed Rafea for his great help, support, and patience which without I couldn’t reach this phase.

A very special acknowledgment goes to my fiancé, who stood beside me and helped me a lot during working on this thesis. Thanks so much for your love and care, I wouldn’t have reached this state without your support.

Last but not least, I thank my mom, dad, and brothers. My parents have sacrificed their lives for my brothers and me and provided unconditional love and care. I love them so much.

ABSTRACT

Microblog retrieval has received much attention in recent years due to the wide spread of social microblogging platforms such as Twitter. The main motive behind microblog retrieval is to serve users searching a big collection of microblogs a list of relevant documents (microblogs) matching their search needs. What makes microblog retrieval different from normal web retrieval is the short length of the user queries and the documents that you search in, which leads to a big vocabulary mismatch problem. Many research studies investigated different approaches for microblog retrieval. Query expansion is one of the approaches that showed stable performance for improving microblog retrieval effectiveness. Query expansion is used mainly to overcome the vocabulary mismatch problem between user queries and short relevant documents. In our work, we investigate existing query expansion method (Pseudo Relevance Feedback - PRF) comprehensively, and propose an extension using the information from hyperlinks attached to the top relevant documents.

Our experimental results on TREC microblog data showed that Pseudo Relevance Feedback (PRF) alone could outperform many retrieval approaches if configured properly. We showed that combining the expansion terms with the original query by a weight, not to dilute the effect of the original query, could lead to superior results. The weighted combine of the expansion terms is different than what is commonly used in the literature by appending the expansion terms to the original query without weighting. We experimented using different weighting schemes, and empirically found that assigning a small weight for the expansion terms 0.2, and 0.8 for the original query performs the best for the three evaluation sets 2011, 2012, and 2013. We applied the previous weighting scheme to the most reported PRF

configuration used in the literature and measured the retrieval performance. The P@30 performance achieved using our weighting scheme was 0.485, 0.4136, and 0.4811 compared to 0.4585, 0.3548, and 0.3861 without applying weighting for the three evaluation sets 2011, 2012 and 2013 respectively. The MAP performance achieved using our weighting scheme was 0.4386, 0.2845, and 0.3262 compared to 0.3592, 0.2074, and 0.2256 without applying weighting for the three evaluation sets 2011, 2012 and 2013 respectively.

Results also showed that utilizing hyperlinked documents attached to the top relevant tweets in query expansion improves the results over traditional PRF. By utilizing hyperlinked documents in the query expansion our best runs achieved 0.5000, 0.4339, and 0.5546 P@30 compared to 0.4864, 0.4203, and 0.5322 when applying traditional PRF, and 0.4587, 0.3044, and 0.3584 MAP when applying traditional PRF compared to 0.4405, 0.2850, and 0.3492 when utilizing the hyperlinked document contents (using web page titles, and meta-descriptions) for the three evaluation sets 2011, 2012 and 2013 respectively.

We explored different types of information extracted from the hyperlinked documents; we show that using the document titles and meta-descriptions helps in improving the retrieval performance the most. On the other hand, using the meta-keywords degraded the retrieval performance. For the test set released in 2013, using our hyperlinked-extended approach achieved the best improvement over the PRF baseline, 0.5546 P@30 compared to 0.5322 and 0.3584 MAP compared to 0.3492. For the test sets released in 2011 and 2012 we got less improvements over PRF, 0.5000, 0.4339 P@30 compared to 0.4864, 0.4203, and 0.4587, 0.3044 MAP compared to 0.4405, 0.2850. We showed that this behavior was due to the age of the collection, where a lot of hyperlinked documents were taken down or moved and we couldn't get their information.

Our best results achieved using hyperlink-extended PRF achieved statistically significant improvements over the traditional PRF for the test sets released in 2011, and 2013 using paired t-test with p-value < 0.05. Moreover, our proposed approach outperformed the best results reported at TREC microblog track for the years 2011, and 2013, which applied more sophisticated algorithms. Our proposed approach achieved 0.5000, 0.5546 P@30 compared to 0.4551, 0.5528 achieved by the best runs in TREC, and 0.4587, 0.3584 MAP compared to 0.3350, 0.3524 for the evaluation sets of 2011 and 2013 respectively.

The main contributions of our work can be listed as follows:

1. Providing a comprehensive study for the usage of traditional PRF with microblog retrieval using various configurations.
2. Introducing a hyperlink-based PRF approach for microblog retrieval by utilizing hyperlinks embedded in initially retrieved tweets, which showed a significant improvement to retrieval effectiveness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IR	Information Retrieval
PRF	Pseudo Relevance Feedback
HPRF	Hyperlink-extended Pseudo Relevance Feedback
URL	Uniform Resource Locator
TREC	Text REtrieval Conference
LM	Language Modeling
P@30	Precision at position 30
MAP	Mean Average Precision
TFIDF	Term Frequency Inverse Document Frequency

Chapter 1. INTRODUCTION

Microblogs are relatively a new type of social networks that enable users to share their thoughts, communicate with friends and read real-time news. We can define microblogging as "*a form of blogging that lets you write brief text updates (usually less than 200 characters) about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web.*" [1]. There are many microblogging services out in the market like Twitter (www.twitter.com), Tumblr (www.tumblr.com), Facebook (www.facebook.com) and Google+ (plus.google.com). One popular microblogging service is Twitter.

On Twitter, a user can share his thoughts in a short text named tweet consisting of maximum 140 characters. Twitter has more than 200 million users, and around 340 million tweets published per day[1]. There are many ways to post your tweet, you can write it down on your PC browser, using your mobile Internet-enabled device or using the SMS. The increasing popularity of Twitter made the researchers ask Why and How we use twitter in [2], [3]. Given the nature of tweets as they are real-time and crowd sourced, makes users start thinking of using it as a platform to provide them with information.

TREC (Text REtrieval Conference - trec.nist.gov), one of the most famous conferences for text retrieval, started a track in 2011 for Microblog retrieval. The TREC conference has several different tracks such as web track, blog track, video track and etc. For each track it preforms a sort of competition between the track participants. In order to do so it provides all track participants with the same datasets and evaluation sets. It also specifies the evaluation metrics, which the participants should use. The participants are required to submit the results of their methods using

[1] https://business.twitter.com/basics/what-is-twitter/ [Accessed 26 January 2013]

the specified evaluation metrics. The high ranks of the competition are the participant who achieves the best performance. The Microblog track, which started in 2011, provided the participants with tweet collections and their corresponding evaluation sets. The aim of the Microblog track is to get the best retrieval performance. The main evaluation metric used for this purpose is Precision of the first 30 retrieved results. The Microblog track was successful in its first year, and continued for the years 2012, 2013, and they are planning to continue it for the next year 2014. The Microblog track datasets and evaluation sets represent a benchmark that people can compare their approaches against. A lot of papers other than TREC reports where published based on the Microblog track data. In our work, all the experiments conducted are based on TREC datasets and evaluation sets. We compare our results to the best-submitted runs in TREC for the 3 years that the Microblog track was active.

Our best results achieved using the hyperlinked-extended PRF approach we proposed outperformed the best results achieved in the Microblog track for the years 2011 and 2013 for both evaluation metrics P@30 and MAP and achieved comparable results for the year 2012. We show that we can outperform state of the art retrieval approaches that are more complex and depending either on manual training data or uncontrollable third party components. Approaches that depend on manual training data are like learning to rank, which is a machine learning approach to re-rank the search results based on training data consisting of user queries and their corresponding relevant search results. Other approaches depends on uncontrollable third party components like search engines, where it is hard to reproduce the results if the search engine changed their algorithms.

1.1. Motivation

The special nature of microblogs, which are short documents (with a maximum of 140 characters in case of Twitter), introduced different search scenarios and retrieval challenges when compared to other search tasks, such as web search [4], [5]. The major challenge in microblog retrieval lies in the severe vocabulary mismatch between the user query and short relevant documents. This means that when one searching for a query term, a lot of relevant tweets might not be retrieved, since their short tweet text does not contain this specific term but other relevant terms the user did not mention. There are two main approaches, which present an effective solution to this problem, namely query expansion and document expansion.

For query expansion, the basic idea is to try to expand the user query with relevant terms. In this case, if the user searches for a query term, by adding all the other similar and relevant term to this query, there is a higher chance of retrieving all the relevant tweets, which are about the same subject. In order to enrich the context and get more relevant terms that the user forgot to mention in his query or don't know about it, the relevant terms are extracted from relevant tweets or using external sources like WordNet (http://wordnet.princeton.edu/), or web search. Query expansion has been used in [6], [7], [8], [9], [10], [11], [12].

On contrary, in document expansion the process focuses on expanding the tweets that the user searches in. In other words instead of expanding the text of the query, the texts of the tweets in the dataset are expanded. Document expansion has been used in [13], [14], [15], [16]. One of the most common sources of expanding the tweets is using the titles of the hyperlinked documents attached to them.

Most of the reported work on query expansion mainly applied pseudo relevance feedback (PRF) to find the relevant terms that should be added to the original query. When using PRF, after an initial search using only the original query, the top retrieved documents (tweets) are assumed to be relevant to the user query. Then, we select a set of top occurring terms and add them to the query after removing the original query terms and stop words. Finally, the expanded query is used for a second search aimed at better vocabulary matching and better retrieval performance. The numbers of feedback terms and the number of documents are important parameters in the PRF process and can highly affect the performance. However, in previous studies these parameters were typically selected subjectively. As a result there is still an absence of recommended PRF configuration in the literature. Having a recommended PRF configuration is useful in achieving optimal performance for microblog retrieval. In addition, some reported work focused on extracting terms from top retrieved documents without considering the hyperlinks embedded in many of the retrieved tweets. This represents additional unused content of microblog documents, since in most microblogs the main content is actually the embedded hyperlink [7], [8], [9], [10], [11].

Most of work reported that utilized hyperlinks was to improve retrieval effectiveness by applying document expansion or using them as features for learning to rank algorithms [13], [17], [15], [16].

1.2. Thesis Statement

The objective of this research is to study the impact of query expansion on microblog retrieval performance. Firstly, we do a comprehensive study on the traditional Pseudo Relevance Feedback (PRF). We analyze the impact of PRF under different configurations and discuss how it impacts the retrieval performance.

Secondly, we utilize the contents of hyperlinked documents attached to the top relevant search results and study the impact of using different types of information extracted form these hyperlinks on the retrieval performance.

1.3. Our Approach

We introduce a novel "hyperlink-extended PRF" query expansion approach for microblogs, to improve retrieval effectiveness. The retrieval effectiveness is improved by utilizing hyperlinks attached to the top relevant documents in the PRF process. Our approach helps in overcoming the problem of vocabulary mismatch between user queries and short relevant documents.

We use information from hyperlinks attached to the top relevant documents such as: page title, meta-description, and meta-keywords. Using information from hyperlinks attached to the top relevant documents, we enrich the original query with terms that are relevant to the query but were not available, neither in the original query nor in the top relevant documents. We then preform a second round search using the enriched expanded query. This will lead to retrieving new relevant documents in the second round search and will improve the retrieval effectiveness.

Overview of our proposed approach:

Step 1: Initially, We comprehensively analyze the standard PRF by preforming extensive experiments. We preform a set of experiment using different configurations to identify the effect of different parameters and to discover to what extent PRF can improve the retrieval effectiveness for microblog retrieval.

Step 2: Secondly, having the top relevant tweets retrieved in the previous step, we crawl the pages of the hyperlinks attached to them. From the pages of the attached hyperlinks we extract different information such as: titles/meta-description/meta-keywords. Based on the information extracted from the web pages, we choose relevant terms that can be used for expanding the query.

Step 3: Later, we use the terms extracted from step 2 to augment the terms extracted using the standard PRF in step 1. We use all of the extracted terms together to expand the original query and form the final expanded query. We call the described process the "hyperlink-extended PRF - HPRF" method.

Step 4: Finally, we use the expanded query in a second round retrieval to search the tweets collection again and get the final list of relevant tweets.

We apply our study on TREC microblog track datasets from 2011, 2012, and 2013. In Chapter 4, we will discuss our approach in more details, and we will show the experimental results in Chapter 5.

1.4. Organization of the Thesis

In chapter 2, we will discuss the theoretical background of our work; first we start from the term scoring functions. The Term Frequency Inverse Document Frequency (TFIDF) term scoring is described, followed by another up to date term scoring function (Okapi BM25). Next, we discuss each of the evaluation metrics we use to evaluate our approach. These evaluation metrics are Precision at position K and Mean Average Precision (MAP). We also discuss some other important background as follows: Porter Stemmer, which is the most commonly used stemming algorithm in language processing applications; Relevance Feedback specifically Pseudo Relevance Feedback (PRF); The baseline retrieval model we use (Language Modeling) from Lucene (http://lucene.apache.org/) "a famous open source search engine"; and the statistical significance paired t-test which we used to show the significance improvements achieved using our approach over the baseline traditional PRF approach.

In chapter 3, we will be surveying the previous work done on improving retrieval performance in microblogs. The survey includes previous works based on using query expansion and document Expansion, utilizing hyperlinks in Learning to Rank algorithms, and other tweet specific features that were used to improve the retrieval performance in microblogs.

In chapter 4, we explain our proposed approach in depth, and the system architecture. We elaborate more on all the modules we used or built in our retrieval system.

In chapter 5, we show and discuss our experimental results for the proposed approach. Finally, in chapter 6, we conclude and discuss our future work directions.

Chapter 2. THEORITICAL BACKGROUND

In this chapter, we will be presenting the theoretical background for the algorithms and tools we will be experimenting with in our proposed approach.

In the first section, we will discuss different ways for query and document representation and the basic stemming algorithm we apply before the retrieval process. Firstly, we will present the well know Term Frequency-Inverse Document Frequency (TFIDF) term scoring scheme, that we used throughout our work for expansion terms scoring. Secondly, we will explain a more up to date term scoring scheme, the Okapi BM25. Finally, we will show how Porter stemmer works, which we will use to stem all the terms we use in the expansion process to avoid term redundancies.

In the second section, we will explain the Pseudo Relevance Feedback (PRF) approach in deep, which is the traditional way of expanding user queries. In the third section, we will show how we calculate query and document similarity using the baseline retrieval model (Language Modeling).

Finally, in the forth section, we will elaborate on one of the evaluation metrics we use, Precision at position K, which calculates the ratio of the number of relevant documents retrieved over the total number of documents retrieved at position K. In the forth section, we will elaborate on another evaluation metric we used, the Mean Average Precision (MAP), which calculates the mean of the precision at all the different positions where you encounter a relevant document in the list of relevant search results. Then, we will illustrate the paired t-test that we will use to evaluate how our proposed approach is significantly better than the traditional PRF in the results section.

2.1. Document Representation

2.1.1. Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TFIDF) is a very well know term weighting scheme used in Information Retrieval (IR) and Text Mining [18]. TFIDF is a statistical measure of the importance of a certain term in a document. The TFIDF score of a term increases if it's not used widely in the corpus and appeared many times in the document, and vise versa.

TFIDF consists of two parts, TF, which is directly proportional with number of occurrences of a term in a given document, and IDF, which is inversely proportional with the number of occurrences of a term in the whole corpus. TFIDF is usually calculated as the product of both TF and IDF to balance the importance of the term both in the document and the whole corpus. For example, if a term occurs a lot in a document like stop words, the TF will be high, while it will occur a lot in the whole corpus and the value of the IDF will be low, and the total value of TFIDF will be low. On the other hand, when a term occurs less in the whole corpus the IDF value will be high, which indicates the importance of the term in the corpus, and the TFIDF value will be initially high and will increase when the value of the TF increases.

The basic way to calculate Term Frequency (TF) is by getting the ratio of number of occurrences of a term in a document to the total number of terms in the same document as follows:

$$TF(term\ t\ in\ document\ d) \\ = (\#\ of\ occurences\ of\ t\ in\ d)/(total\ \#\ of\ terms\ in\ d)$$

The basic version of Inverse Document Frequency (IDF) is calculated as the logarithm of the ratio between the total number of documents in the corpus to the number of documents in the corpus containing the term as follows:

$$IDF(term\ t\ in\ corpus\ c) = \log \frac{total\ number\ of\ documents\ in\ c}{number\ of\ documents\ containing\ t\ in\ c}$$

The IDF value ranges from 0 when all the documents contain the term of interest and $\log(total\ number\ of\ documents\ in\ c)$ when the term occurs in only one document. To avoid the division by zero problem when the term is never seen in the corpus, we add one to both the nominator and denominator as follows:

$$IDF(term\ t\ in\ corpus\ c) = \log \frac{1 + (total\ number\ of\ documents\ in\ c)}{1 + (number\ of\ documents\ containing\ t\ in\ c)}$$

There are many variations of how to calculate the TFIDF score of a term (t) in a document (d); the very basic one can be calculated by multiplying both the TF and the IDF as following:

$$TFIDF_{t,d} = TF_{t,d} * IDF_t$$

TFIDF sometimes is used to filter stop-words in fields like Information Retrieval or Text Classification.

2.1.2. Okapi BM25

Another up to date term scoring function widely used in Information Retrieval is Okapi BM25, which is introduced by Robertson et al. [19]. Okapi BM25 a.k.a as BM25 was introduced as part of the Okapi information retrieval system developed at London's city university in the 1980's – 1990's. BM25 is based on the probabilistic retrieval model introduced by Robertson et al. [19], where BM stands for "Best

Match". Generally, BM25 is considered the state of the art TFIDF like term scoring function.

Like TFIDF, BM25 assumes a bag of words representation of the document, where the document is modeled as a set of words regardless of the document structure and the inter-relationships between terms. BM25 consists of two components as TFIDF; the first component represents the term frequency TF, which is directly proportional with the number of occurrences of a term in a document, and inverse document frequency IDF, which is inversely proportional with the number of occurrences of term in the whole corpus.

The TF component is calculated as following:

$$TF(term\ t\ in\ a\ document\ d) = \frac{f(t,d) * (k+1)}{f(t,d) + k * (1 - b + b * \left(\frac{|d|}{avgdl}\right))}$$

Where $f(t,d)$ is the number of occurences of a term t in a document d; $|d|$ is the length of the document d; $avgdl$ is the average document length in the corpus; and k, b are tuning parameters usually choosen as 2 and 0.75 respectively.

The IDF component is calculated as following:

$$IDF(term\ t\ in\ corpus\ c) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5}$$

Where N is the total number of documents in the whole corpus; and $n(t)$ is the total number of documents in the corpus containing the term t. One drawback for the previously mentioned IDF formula is for terms occurring in more than half of the documents in the corpus, the IDF component will be negative. This undesirable behavior can be solved by giving a floor of 0 to terms widely used in the corpus to filter them, as they are not distinguishing terms.

The BM25 score of a term is then calculated as the product of its TF and IDF component as following:

$$TFIDF_{t,d} = TF_{t,d} * IDF_t$$

2.1.3. Porter Stemmer

Stemming has been studied extensively in Linguistics and Computer Science. In Information Retrieval, stemming is a process of getting the root of the inflected/derived words. Stemming is not supposed to get the morphological root of the word, and in most cases it does not, it is sufficient if stemming can map similar words to the same stem. Stemming removes the common suffixes of inflated words such as, -ED, -ING, -ION, and -IONS. Table 1 shows an example for applying stemming on some inflections of the word “Connect”:

Table 1 Stemming Examples

Connected
Connecting
Connection
Connections
Stem = “Connect”

Porter [20] introduced a reliable stemming algorithm that can work on simple and compound suffixes to extract a word stem. One of the famous Porter Stemmer implementations is Snowball (http://snowball.tartarus.org/). In our work, we use Snowball implementation of Porter Stemmer to stem original query terms and the expansion terms extracted using different expansion methods. After applying stemming on the original and the expansion query terms, we remove the original query terms from the expansion terms. As a result, we ensure no redundancies between the original query terms and the expansion terms.

2.2. Pseudo Relevance Feedback

In Information Retrieval, Relevance Feedback is a commonly used method for query expansion to improve the retrieval performance. The idea behind Relevance Feedback is try to analyze the initial list of search results retrieved using the user original query and learn relevant terms that the user missed in his original query. The user may miss some terms either by forgetting them or not knowing them from the first place. For example, if the user is searching for “Egypt” at the time of revolution, he may miss the terms “Jan25” or “Tahrir”. The relevance feedback works as following:

1. Use the user original search query to search the document collection and get initial list of search results.
2. Assess the initial list of search results as relevant/non-relevant.
3. Use the relevant search results to learn new keywords that were not present in the original user query that may lead to improved retrieval performance.
4. Expand the original user query with the terms extracted from the previous step.
5. Do a second round retrieval with the expanded query.

There are three types of relevance feedback based on the way of assessing the initial list of search results:

1. Explicit Relevance Feedback: Assessing the initial list of search results is done manually, where the user is asked to choose the relevant/non-relevant search results and then resubmits the query after expansion.

2. Implicit Relevance Feedback: Assessing the initial list of search results is done implicitly by analyzing the user behavior such as, the pages he visit, and the time he spend in each page.
3. Blind Relevance Feedback: a.k.a Pseudo Relevance Feedback assess the initial list of search results by assuming the top retrieved search results relevant.

Pseudo Relevance Feedback (PRF), also known as Local Feedback or Blind Relevance Feedback, is a very widely used method for query expansion based on relevance feedback as discussed before [21], [22]. In PRF, the process of choosing relevant documents is automated by assuming the top retrieved results as relevant. PRF offers the user improved retrieval performance without any manual intervention from his side.

The basic idea of PRF is to extract the top occurring terms from the list of most relevant documents to enrich the original user query as following:

1. First, the user query is used to get the list of most relevant documents from the search collection.
2. Then, the terms that have occurred the most in the list of most relevant documents are extracted and filtered from the original query terms.
3. The terms extracted from the previous step are then sorted by a ranking function according to their importance, for example: using TFIFD term scoring function, and the top ranked terms are used to form the list of expansion term.
4. Afterwards, the list of expansion terms is combined with the original query to form the new expanded query.

5. Finally, the expanded query is used for second round retrieval.

After applying PRF, some new documents that were missed in the fist retrieval round may now be retrieved. These newly retrieved documents were missed in the first retrieval round since they did not contain the original query keywords. In the second retrieval round, however, the newly added terms by PRF may help retrieving these new documents. So, if the newly introduced terms are truly relevant to the query, they can help in retrieving relevant documents that were missed in the first retrieval round. However, it should be noted that the performance of PRF is highly dependent on the quality of the expansion terms. If the terms added to the query are truly relevant terms that will lead to better retrieval performance. On the other hand, if the added terms are non-relevant noisy terms, this may harm the retrieval performance by diluting the original query.

2.3. Query and Document similarity (Language Modeling)

Language Modeling has been used effectively in many fields like Statistical Machine Translation, Speech Recognition, Part of Speech Tagging and Information Retrieval (IR). For IR, Language Modeling is used as a formal technique to model documents over the heuristic TFIDF representation. A language model is built for each document to estimate the probability that it generated a user query. The probability a document generated a query is called the Query Likelihood Model.

Given a user query, the query likelihood probability is calculated for all the documents in the corpus. Then, all documents are ranked descending with the probabilities that they generated the query. Finally, the ranked list of documents is shown to the user with the most relevant ones in the top of the list.

The query likelihood is calculated by assuming the query to be observed as a random sample from a document as following [23]:

Given a query $Q = w_1, w_2, w_3, \dots, w_k$ where w_i is the query word at position i, the likelihood of query Q to be generated from a document language model $\hat{\theta}_D$ is

$$P(Q|\hat{\theta}_D) = \prod_{i=1}^{k} P(w_i|\hat{\theta}_D)$$

The previous way of estimating query likelihood suffers the data sparsity problem. In other words, when a document does not contain a query term, the likelihood of this document will be equal to zero. To solve the data sparsity problem we employ smoothing techniques to give probability value for unseen terms within the document [24].

2.4. Evaluation Metrics

2.4.1. Precision at position K (P@K)

Precision at position K is an effective measure widely used in the field of information retrieval to assess the performance of retrieval algorithms. The main idea of precision is to measure how much the system is precise when deciding a document is relevant. A system is more precise when the number of documents it retrieves as relevant is truly relevant. In information retrieval the precision for a retrieval system is calculated at a certain position K in the list of search results, to compare with other systems. Generally, K is chosen heuristically as the number of search results that the user maybe interested in.

The precision at position K is usually referenced as P@K. We can calculate P@K as the ratio of the number of truly relevant documents retrieved at position K to the total number of retrieved document (K) as follows:

$$P@K = \frac{|D_r \cap D|}{|D|}$$

Where D_r is the truly relevant documents retrieved at position K; D is the full list of relevant search results retrieved at position K; and $P@K$ represents how precise is the system in deciding the top K search results relevant. The value of $P@K$ ranges between 0, where the system is totally non-precise and always decides non-relevant documents as relevant and vice versa and 1, where the system is totally precise and never decides that a non-relevant document is relevant. In other words, P@K will be equal to 1 if the list of top relevant search results at position K is all truly relevant. On the other hand, P@K will be equal to 0 if the list of top relevant search results at position K is all truly non-relevant. For TREC Microblog Track K is typically selected to be 30.

2.4.2. Mean Average Precision (MAP)

Another widely used evaluation metric for retrieval effectiveness is the Mean Average Precision (MAP). MAP builds over Precision at position K discussed in the previous section, so it's better to read section 3 before proceeding in this section. MAP is used to measure how the system is effective by showing the user the truly relevant documents condensed in the beginning of the list of relevant search results.

The value of MAP is directly proportional not only to the number of the truly relevant search results retrieved, but their position in the list of retrieved search results. MAP tries to imitate the user need to see the truly relevant search results in the first search pages. The Mean Average Precision (MAP) is calculated by breaking it down into its three main components, Precision at position K, Average Precision, and the Mean Average Precision as following:

1. ***P:*** The precision at different K positions for each query:

   Given a list of relevant search results, at each position K where you encounter a relevant search results, calculate P@K.

2. ***AP:*** The average of the precision values calculated in the previous step for each query in the test set:

   Average of the list of P@K values calculated in the previous step for each query, as following:

   $$AP = \frac{\sum P@K}{N}$$

   Where $\sum P@K$ is the summation of $P@K$ for different values of K whenever you encounter a relevant document in the search results; and N is the total number of truly relevant documents retrieved in the list of relevant search results.

3. ***MAP:*** The Mean of the Average Precisions for all the queries in the test set:

   MAP is calculated as the mean of the Average Precisions (AP) calculated in the previous step for all the search queries in the test set as follows:

   $$MAP = \sum_{i=1}^{M} AP_i$$

   Where M is the total number of queries in the test set; and AP_i is the average precision of the query at position i.

MAP is always calculated for a set of user queries associated with a set of search results with a predefined size. For TREC Microblog track, the typical size for the full list of search results, is ten thousands search results.

2.4.3. Paired t-test

Paired t-test is a statistical significance test used to ensure that a certain behavior did not happen by chance and is statistically consistent [25]. The paired t-test operates on two input samples where each value in the first sample has a natural associated value in the second sample. The two samples are commonly test scores before and after applying an intervention. A paired t-test analyzes the differences between the two input samples, taking into consideration the distribution of the values within each sample. The output from the paired t-test is a single value known as t-value. To calculate the t-value, the mean and the standard deviation of the differences between the two input samples are calculated.

t-value is calculated as following:

$$t - value = \frac{Mean - null\_Hypothesis}{Standard\ Deviation / \sqrt{N}}$$

Where the $null\_Hypothesis$ in our case is equal to 0 by assuming there is no difference between the two samples; N is the number of instances in any of the input samples; and the denominator is commonly referred to as the standard error

t-value is then converted to p-value which is the probability that the two input samples came from the same group. The relationship between the t-value and p-value is as following:

- t-value is inversely proportional to p-value.
- p-value is always positive, while t-value can be positive or negative.
- p-value of a negative t-value is as same as the equivalent positive one.

There are different thresholds for the p-value to measure how much strong is the assumption that the two input samples are from different groups, as following:

- $p < 0.01$, very strong
- $0.01 < p < 0.05$, strong
- $0.05 < p < 0.1$, low
- $p > 0.1$, not significant

In our work we use a software package (http://commons.apache.org/proper/commons-math/) to calculate the p-values and we measure the significance based on $p\text{-value} < 0.05$.

Chapter 3. RELATED WORK

In this chapter, we will discuss how our work is related to the previous work done in literature. As our work is based on TREC Microblog Track datasets, most of the related work discussed in this chapter will be based on the same datasets. TREC microblog track was introduced in 2011 due to the increased interest in microblog retrieval. Ad-hoc search task for microblogs were studied over the past three years using two tweets collections and three query sets [15], [16], [14]. The datasets include two tweet collections (2011 and 2012) and three query sets (2011, 2012, and 2013). TREC Microblog Track 2011 dataset has been extensively studied in the literature including TREC reports. For 2012 and 2013 datasets, most of the work done is only reported in TREC.

Different approaches were investigated for microblog retrieval [15], [16], [14] to overcome the special nature of microblog documents [4], [5]. One of the main challenges in microblog retrieval is term mismatch problem between short queries and short relevant documents. Researches tackled the term mismatch problem in Microblogs either by doing document or query expansion. Another line of research in improving retrieval performance in microblogs focused on using twitter specific features in re-ranking the search results, specifically using the hyperlinked documents attached to the tweets.

In section one, we will discuss most of the work done on document expansion for improving the retrieval effectiveness in microblogs, especially the work done using the TREC Microblog track datasets. The main idea behind document expansion is to enrich short documents that the user search in with relevant terms that may help in matching corresponding short user queries. We will show the importance of document expansion in microblogs and how the problem of vocabulary mismatch between short user queries and short documents is much severe in microblogs compared to normal web search.

As our work is mainly founded on query expansion, in section two, we discuss the previous work done in query expansion for microblogs in details. Query expansion, specifically pseudo relevance feedback (PRF) has been widely used as a strong baseline for microblog retrieval. We will show the techniques presented in previous work and how they used to choose the parameters in the expansion process, to motivate our study. Finally, we show that there was no comprehensive study for different configurations for the query expansion process in microblogs; in addition, to the best of our knowledge previous studies did not focus on utilizing the hyperlinked document attached to the top relevant documents in the query expansion process.

In section three, we show that some twitter specific features were used as important features in re-ranking search results by applying learning to rank algorithms. One of the main features used in learning to rank algorithms to indicate the relevance of a tweet, is the existence of hyperlinks, which motivates our work in considering the hyperlinks embedded in top relevant search results a good source for relevant terms extraction. Moreover, we discuss other twitter specific features that helped improving retrieval effectiveness in microblog retrieval.

Finally, in section four, we summarize all the previous work done in literature and discuss their drawbacks and how we can achieve comparable or even better retrieval performance using much more simpler technique that we propose in our work.

3.1. Document Expansion

Having a large corpus of short documents, where each document contains few words, applying traditional Information Retrieval (IR) models and techniques will be difficult [13]. The first issue you face is the vocabulary mismatch problem, given the short length of the document; the probability that the few query terms fail in matching the short document is high. Secondly, the main IR models rely on term scoring functions such as: TFIDF, and BM25 where the term frequency in the document plays a big role in deciding if a document is relevant or not. In microblogs, where the documents are very short, 140 character in the case of twitter, most of the terms occur only once in a document, which makes it very hard to estimate the language model for a document.

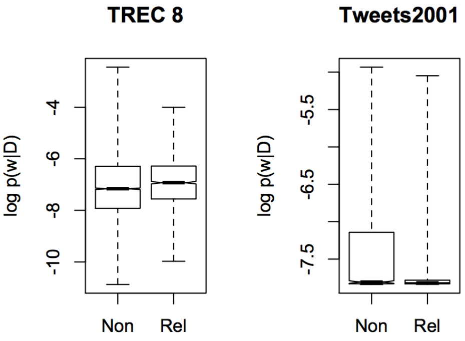
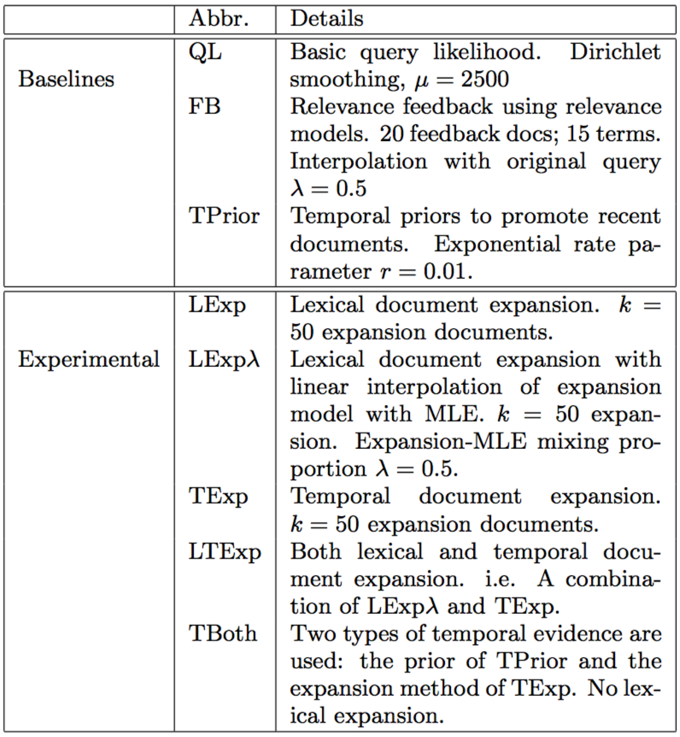


Figure 1 log-Probabilities of Query Terms in Relevant, and Non-Relevant Documents in Two Corpora [13]

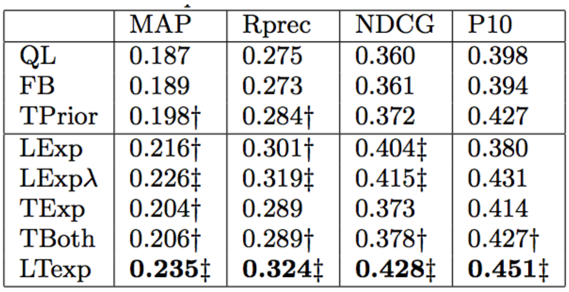
Efron et al. [13] compared a news article dataset represented in TREC 8 data to a tweet dataset represented in TREC 2011 microblog data. In Figure 1 they show the distributions of the log-probabilities of query terms for the first 100 retrieved documents using a simple language model. From Figure 1 we can realize two problems, firstly, tweets unlike longer news articles leads to strongly peaked distribution for query terms for the top retrieved results. The tweets that contain the query words contain it once most of the time, and tweets are almost all equal in length. Secondly, we can notice that for TREC 8 data, the mean and the median log-probabilities of a query term in a relevant document is higher than in non-relevant document, unlike the case for the tweets.

They introduced a massive document expansion approach to enrich tweets with additional terms from the top retrieved results using each document as a pseudo query. Their approach relies on assuming that each tweet talks about a single topic. Each tweet in the tweets corpus is expanded with relevant terms to help getting better retrieval performance. They submit each tweet in the corpus as a pseudo query to the search engine, and retrieve the most relevant documents to be used later for expanding the document itself. The expansion process is done based on two evidences, the lexical evidence, and the temporal evidence.



	Abbr.	Details
Baselines	QL	Basic query likelihood. Dirichlet smoothing, $\mu = 2500$
	FB	Relevance feedback using relevance models. 20 feedback docs; 15 terms. Interpolation with original query $\lambda = 0.5$
	TPrior	Temporal priors to promote recent documents. Exponential rate parameter $r = 0.01$.
Experimental	LExp	Lexical document expansion. $k =$ 50 expansion documents.
	LExpλ	Lexical document expansion with linear interpolation of expansion model with MLE. $k = 50$ expansion. Expansion-MLE mixing proportion $\lambda = 0.5$.
	TExp	Temporal document expansion. $k = 50$ expansion documents.
	LTExp	Both lexical and temporal document expansion. i.e. A combination of LExpλ and TExp.
	TBoth	Two types of temporal evidence are used: the prior of TPrior and the expansion method of TExp. No lexical expansion.

Figure 2 Efron et al. Baseline and Experimental Retrieval Names and Parameters [13]



	MAP	Rprec	NDCG	P10
QL	0.187	0.275	0.360	0.398
FB	0.189	0.273	0.361	0.394
TPrior	0.198†	0.284†	0.372	0.427
LExp	0.216†	0.301†	0.404‡	0.380
LExpλ	0.226‡	0.319‡	0.415‡	0.431
TExp	0.204†	0.289	0.373	0.414
TBoth	0.206†	0.289†	0.378†	0.427†
LTexp	**0.235**‡	**0.324**‡	**0.428**‡	**0.451**‡

Figure 3 Efron et al. results on TREC 2011 microblog track data [13]

Figure 2, and Figure 3 show their experimental results, where † symbol indicates results achieving statistical significance improvements over baseline, using paired t-test with p-value < 0.05 and ‡ symbol indicate p-value < 0.01. Their approach managed to improve retrieval effectiveness significantly, specially when utilizing both the lexical and the temporal evidence. By comparing our best MAP results achieved in Table 22, we show that we achieved a lot better results using a less complicated approach.

Other approaches for document expansion were reported in TREC microblog track that mostly showed improvement to the retrieval effectiveness [15], [16]. One of the most applied approaches for document expansion was expanding tweets containing hyperlinks with corresponding titles of the hyperlinked documents [26], [27].

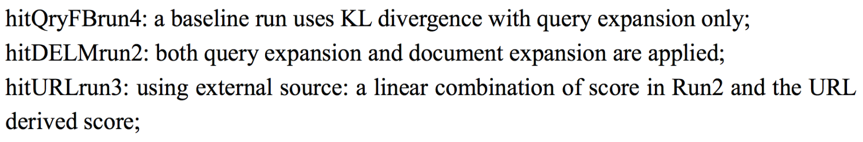
Han et al. [26] applied Document Expansion Language Model (DELM) to improve how short documents are represented in the search corpus. In DELM, each document language model is smoothed using its k nearest neighbors, where the influence of the neighbor documents is controlled by their cosine similarity with the original document. Moreover, they did analysis on the importance of hyperlinked documents attached to the tweets. They showed that a tweet containing a hyperlink is more likely to contain substantially important information.

They built a linear model to combine the relevance score of the tweet itself and the relevance score based on the contents of hyperlinked document attached to it as shown in Figure 4.

$$\mathrm{Score}(Q,D) = (1-\lambda) * \mathrm{score\_D}(Q,D) + \lambda * \mathrm{Score\_Url}(Q,D) * \delta$$

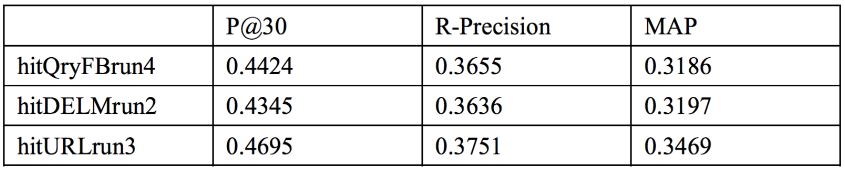
Figure 4 Han et al. scoring scheme proposed for document expansion [26]

Where score_D(Q,D) is the relevance score of the tweet text calculated using the query likelihood model; Score_Url(Q,D) is the relevance score of the web-page content to the query and calculated the same way as previous; λ is a parameter to control the effect of the hyperlink document content on the retrieval process, they set it empirically to 0.8; and δ is the zoom ratio which makes score_D(Q,D), and Score_Url(Q,D) comparable, which is calculated as the ratio of their average scores.



hitQryFBrun4: a baseline run uses KL divergence with query expansion only;
hitDELMrun2: both query expansion and document expansion are applied;
hitURLrun3: using external source: a linear combination of score in Run2 and the URL derived score;

Figure 5 Han et al. Baseline and Experimental Retrieval Names and Parameters [26]



	P@30	R-Precision	MAP
hitQryFBrun4	0.4424	0.3655	0.3186
hitDELMrun2	0.4345	0.3636	0.3197
hitURLrun3	0.4695	0.3751	0.3469

Figure 6 Han et al. results on TREC 2012 microblog track data [26]

As shown in Figure 5, and Figure 6 their approach achieved big improvements over the baseline, especially when utilizing the contents of the hyperlinked documents attached to the tweets in hitURLrun3. We achieve comparable results to their reported results, with less overhead, as in their case they crawl the whole contents of the hyperlinked documents unlike our approach where we crawl only their meta-data. Another reason for our results not to perform better than their reported number, is the age of the collection when we performed the experiments, where the number of broken hyperlinked documents in the collection was higher as shown in Table 4.

Duc et al. [27] analyzed the relevance judgments of the TREC 2011 microblog track to realize the importance of the hyperlinked documents attached to tweets. They show that "around 94% of the highly relevant tweets and 80% of all relevant tweets, as opposed to 53% in the non-relevant tweets" [27]. They suggested that crawling the whole contents of the hyperlinked documents is not practical, in addition to, that most of the hyperlinks point to graphics or multimedia. As a result, they crawled only the hyperlinked documents titles to use them for document expansion. After crawling the hyperlinked documents titles, they simply append them to their corresponding tweets.

Their approach achieved improvements over the baseline, nevertheless, our approach outperform their reported numbers regarding the best run achieved using the hyperlinked documents titles in document expansion. They achieved 0.3323 P@30 for 2012 evaluation set, while our best run achieved 0.4339 P@30 for the same evaluation set.

Although document expansion potentially lead to improved retrieval effectiveness for microblog search, its computational cost is high, since using each tweet to search the tweet collection as in [13] or accessing embedded hyperlinks to extract page titles, and contents [15], [16] for all tweets in a collection is seen impractical for the current large tweets streams.

3.2. Query Expansion

An alternative to overcome the vocabulary mismatch problem is query expansion [28]. Several studies showed the effectiveness of using query expansion to improve the performance of microblog retrieval [6], [7], [29], [9], [10], [11], [12].

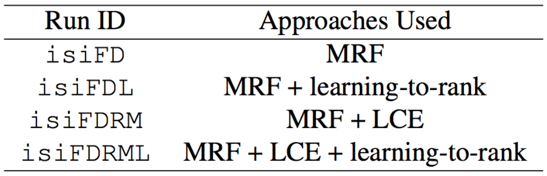
Traditional PRF, which selects some terms from initially top retrieved documents, was reported by many participants in TREC microblog track to improve retrieval effectiveness [15], [16], [14]. The reported work for PRF selected specific numbers for documents and terms heuristically for the feedback process without a comprehensive study to find the best configuration.

In this section, we discuss the previous work done on query expansion leveraging different types of evidences. First, we will show the use of the internal evidence by utilizing information extracted from the tweets corpus itself in the query expansion process, especially the top relevant tweets. Secondly, we will show different types of external evidences used in enriching short user queries, like web search or existing open source corpora. Finally, we will present a line of research focused on the real time nature of microblogs and how the change in time can affect the relevance of the terms used in the expansion process.

3.2.1. *Expansion using Internal Evidence*

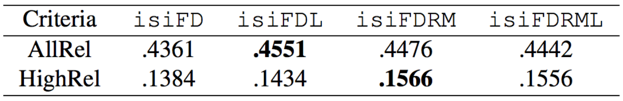
Metzler et al. [30] analyzed the ad-hoc search task in TREC microblog track to define the challenges they need to address. They define the challenges as following: the very short length of the documents, the highly varied document quality, the language identification issues, temporally biased queries, retrieval metrics, and lack of training data. They made use of best practices in ranking techniques like term, phrase, and proximity based text matching. They did text matching using Markov Random field model (MRF), pseudo relevance feedback using Latent Concept Expansion (LCE), and feature-based learning to rank model.

As shown in Figure 7 and Figure 8 all their runs achieved improvements over the baseline. isiFDL that utilize learning to rank and isiFDRM that utilizes query expansion using latent concept expansion reported the best performances for all relevance and high relevance query sets respectively on TREC 2011 microblog track datasets.



Run ID	Approaches Used
isiFD	MRF
isiFDL	MRF + learning-to-rank
isiFDRM	MRF + LCE
isiFDRML	MRF + LCE + learning-to-rank

Figure 7 Metzler et al. Baseline and Experimental Retrieval Names and Parameters [30]



Criteria	isiFD	isiFDL	isiFDRM	isiFDRML
AllRel	.4361	**.4551**	.4476	.4442
HighRel	.1384	.1434	**.1566**	.1556

Figure 8 Metzler et al. results on TREC 2011 microblog track data [30]

Li et al. [31] proposed two methods for query expansion: Word Activation Force Algorithm (WAF) and Electric Resistance Network. The WAF is based on the assumption that there exists a force in the document that makes human brain activate associates of a word, like ‘hospital’ activates ‘doctor’ or ‘nurse’. The electric resistance network performs on the WAF network to expand the tweet with relevant terms. They confirmed the effectiveness of query expansion in improving microblog retrieval performance. Their best runs achieved 0.4388 P@30 and 0.4000 MAP on TREC 2011 microblog data.

Roegiest et al. [32] decided to build a baseline for the first year of the microblog track using existing methodologies and then improve upon it. They used the Wumpus search engine, developed at University of Waterloo to do some basic experimentation. They tried applying different basic Information Retrieval methods, where they achieved the best performance when applying pseudo relevance feedback (PRF) using internal evidence and with the help of and external evidence, the GOV2 corpus from TREC Terabyte Track as a language model. They applied PRF based on the KL-divergence, and Okapi models, on the top 15 retrieved documents and choosing the best 8 terms for the expansion. From their results, we can notice that the best-performing method is query expansion, especially when using a tweet-based language model, where they achieved 3.45 P@30 on the TREC 2011 datasets.

Hong et al. [33] proposed three techniques to get better search results. Firstly, they used hashtags as an additional type of information for the query, and they grouped any two consecutive terms in the query and added them to the final query. Secondly, they tried query expansion using pseudo relevance feedback (PRF). Finally, they applied affinity propagation method, a non-parametric clustering algorithm to group tweets according to their similarities. They applied PRF by assuming the top 10 search results as relevant and selecting 10 terms as feedback terms. They found out that applying query expansion using PRF is very effective in improving the retrieval performance. Moreover, affinity propagation can achieve comparable results to PRF if combined with other techniques. Their best reported numbers using query expansion where 0.403 P@30 on the TREC 2011 microblog track data.

Karimi et al. [34] experimented using different preprocessing and query expansion methods. They observed that hashtags could act as an explicit marker for the topic of the tweet. As a result, they used hashtags as a simple form of pseudo relevance feedback (PRF). They do an initial round retrieval to get a set of relevant search results and then do a second round retrieval after adding hashtags extracted from the initial list of search results to the original query. Additionally, they applied traditional PRF by assuming the top 10 search results relevant and using the whole tweet text to extract expansion terms not only the hashtags. Finally, they applied a twitter specific named entity recognizer on the top relevant search results and used the extracted named entities as the only expansion terms in a typical PRF process as mentioned before. Their best runs achieved were by considering the whole tweet text in the PRF process. They achieved 0.3639 P@30, 0.3108 MAP and 0.1445 P@30, 0.1537 MAP on TREC 2011 and TREC 2012 microblog datasets respectively.

Aboulnaga et al. [35] explored the use of Frequent Itemsets Mining (FIM) in discovering text patterns from tweet streams and use it later for query expansion. FIM has been widely used to mine data streams, as it is computationally simple and can be parallelized in some of its steps. They use the BM25 model as the baseline retrieval model in their work. First, they collect the frequent itemsets and index them as normal tweets. Next, they use the TREC topics to search the corpus of frequent itemsets and select the most relevant itemsets to be used later for query expansion. They apply query expansion using 10 expansion terms and by giving the original and the expansion terms the same weight to avoid concept drift. Their approach showed promising results on TREC 2011 dataset, where they achieved 0.4525 P@30 and 0.3764 MAP for their best run. On they other hand, their approach did not improve the performance much for TREC 2012 dataset, where they achieved 0.3819 P@30 and 0.2467 MAP for their best run.

3.2.2. *Expansion using External Evidence*

The main source of information used to expand user queries is to analyze the same corpus used for retrieval in favor of extracting relevant terms not present in the original query to improve the retrieval performance. Another approach widely used is to utilize external resources a.k.a external evidence to collect relevant terms to the user search query to help overcome the vocabulary mismatch problem between short user queries and short search documents. Multiple external sources were studied in the literature suck as: web search results in [7], [36], [37], [38], [39], Wikipedia pages in [40], [41], and WordNet (http://wordnet.princeton.edu/) in [42].

Some work utilized web search as an external evidence for query expansion in microblogs [7], [36], [37], [38], [39]. Saad El Din et al. [38] searched Google in the time frame of the tweets collection with the original query to get the first search result title to be used later for query expansion. The web page title extracted in the previous step is used to expand the original user query after removing the website name. The web page title is appended directly to the original query with no term weighting involved. Additionally, they tried incorporating traditional pseudo relevance feedback (PRF) with their proposed web-based query expansion, where they achieved their best results. For the PRF configuration, they considered the first 50 search relevant to extract 10 expansion terms. Their approach showed very promising improvements by enriching the original query with terms from web search results. Their best run achieved 0.365 P@30, 0.2548 MAP on TREC 2012 microblog track data.

El-Ganainy et al. [7] used Google API to retrieve web search results matching microblogs query at the same time period of the collection. For each Google search result, the title and the snippets of the web page is extracted and used later in the query expansion process. The web pages titles and snippets extracted in the previous step are used to extract expansion terms in a traditional PRF manner. The number of feedback documents they used for the PRF process was 50 to extract 12 feedback terms. For the web-based PRF, the number of web search results used was the top 3 search results. They tried different combinations of using only expansion terms generated from the traditional PRF process or by combining them with terms extracted from the web-based PRF. The best results they achieved were using the combination of the terms extracted from traditional PRF and the web-based PRF. The expansion terms are combined with the original query using a weight to avoid concept drift; they assumed a weight of 0.2 for the expansion terms and 0.8 for the original user query. They show that query expansion alone can lead to superior results either using the traditional PRF or by including terms from external sources like web search results, where they achieved 0.4849 P@30, 0.3030 MAP and 0.5356 P@30, 0.3444 MAP on TREC 2013 microblog track dataset using PRF, and PRF combined with web-based PRF respectively.

Louvan et al. [37] did some experiments incorporating different scoring functions, query reformulation, and query expansion. They applied traditional pseudo relevance feedback (PRF) by assuming the top search results relevant, in addition to, utilizing web-snippets returned from web search. Moreover, in their customized scoring function they utilized other features and methods that can improve the retrieval performance such as: retweet count, phrase query identification and proximity search. They utilized web search results by using snippets of web results returned from Google search engine. For proximity search, they used Lucene proximity operator with 10 words distance. For phrase identification and part of speech tagging (POS), they used Stanford POS tagger[1]. For the retweet value, they consider the tweets having higher retweet value more important and can be more helpful in the query expansion process. They achieved their best results by applying query expansion utilizing web search result, and including the re-tweet value in the customized scoring function. Their best-submitted runs achieved 0.414 P@30 on TREC 2011 microblog track dataset.

Zhu et al. [39] applied query expansion by utilizing internal and external evidences. They apply query expansion based on internal evidence by applying

[1] http://nlp.stanford.edu/software/tagger.shtml

pseudo relevance feedback (PRF) on the initial set or retrieved search results. For the external evidence, they utilize the web page titles of the three search results retrieved using Google web search engine. They combined terms extracted from the traditional PRF and search results from Google. Their best runs achieved 0.2384 P@30, 0.2093 MAP on TREC 2012 microblog track datasets.

Other studies emphasized the importance of different external sources to be used for enriching queries other than web search results, such as, Wikipedia pages, and WordNet [40], [41], [42].

Small et al. [40] utilized Wikipedia pages in the query expansion process. For each query, stop words are removed. Then, each term in the query is used to get the corresponding Wikipedia page. In case a term matches a disambiguation Wikipedia page, they use it for expansion, instead of the disambiguated pages. For each Wikipedia page after removing the unnecessary tags, they extract the top four occurring term. Finally, across all the lists formed from different Wikipedia pages, the top four occurring terms across the entire lists are used for expanding the original query. Their best run achieved 0.1808 P@30 on the high relevance query set in TREC 2012 microblog track datasets.

Wu et al. [41] applied query expansion by detecting concepts in the search queries and use their corresponding Wikipedia pages to extract expansion term. For each query they detect the concepts it entitles. The concepts may be part of the query or the whole query. After detecting the concepts, they use them to get their corresponding Wikipedia pages. For each Wikipedia page, they calculate its language model based on the term frequencies of the concept terms in the page itself. The expansion terms are chosen as the top 20 terms extracted for each concept. The term weights of the expansion terms are determined based on the estimated language model for the corresponding Wikipedia page. They showed that their approach for query expansion could improve the retrieval performance for microblogs; where they achieved 0.1161 MAP on the high relevance query set in TREC 2012 microblog track datasets.

Zhang et al. [42] used WordNet to extract expansion terms related to the query terms. WordNet is a " large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept"[1]. For all the query terms they get their synonyms from WordNet depending on the part of speech and use it for expanding the original query. Additionally, they apply query reformulation based on electric resistance network. They showed that using WordNet as an external source for query expansion can help improving the retrieval performance; where they achieved 0.2028 P@30, 0.1555 MAP on the high relevance query set in TREC 2012 microblog track datasets.

3.2.3. *Expansion using Temporal Evidence*

Microblogs topics are so time dependent by nature. What people are talking about today on twitter may be different than what they are taking about tomorrow, even on hourly basis it may change. Topics change over time may lead to different judgment of terms to be relevant or irrelevant regarding using them for query expansion. For example: if someone is searching for the word "Egypt" at the time of the revolution in 2011, relevant terms that may be added to the query are "Tahrir", "Jan25", on the other hand, if some one is searching for the same word in 2014 at the time of the presidential elections, the names of the presidential candidates will be more relevant. In addressing this problem, a line of research focused on applying

[1] http://wordnet.princeton.edu/

temporal-based query expansion for microblog retrieval [6], [9], [10], [12], [43], [44], [45].

Choi et al. [6] discuss the impact of time information on relevance in microblog retrieval for queries that are sensitive to events and trends. They show that recent work on time-based retrieval in microblogs has promising results. They proposed selecting time period based on user behavior (e.g. retweets) to extract relevant tweets to be used for the expansion process. Then, they apply pseudo relevance feedback (PRF) on the peak times determined in the previous step. Moreover, they used a timely based retrieval model as the baseline retrieval model. They show that their approach is effective compared to other approaches, where their best results achieved 0.5429 P@30, 0.3226 MAP on TREC 2011 microblog track datasets.

Metzler et al. [10] showed that using temporal co-occurrence of terms is much more effective than using the traditional term co-occurrence for choosing expansion terms. Similarly, Massoudi et al. [9] developed a language retrieval model tailored for microblogs, taking in consideration textual and microblog specific characteristics. They use some quality indicators that emphasis the relevance of a tweet based on some microblog specific features such as: emoticons, post length, shouting, capitalization, and the existence of hyperlinks. In addition, they propose a dynamic query expansion for microblogs. They assumed that selecting terms temporally closer to the query time are more effective in expansion. Their proposed system achieved good improvements over the baseline, which confirms its effectiveness. For their best runs they achieved 0.4820 MAP on TREC 2011 microblog track datasets.

In a more recent work, Miyanishi et al. [12] confirms the importance of pseudo relevance feedback (PRF) in improving retrieval performance in microblogs. They state that one main drawback for PRF is that the initial list of search results used to extract the expansion terms may contain many non-relevant documents that will affect the retrieval performance negatively. They tried to address PRF weaknesses by proposing a two-stage relevance feedback model using manual tweet selection and incorporating lexical and temporal evidence to the model. They showed that manually selecting one relevant tweet could significantly improve retrieval effectiveness; where their best runs achieved 0.5354 P@30, 0.5384 AP, and 0.4910 P@30, 0.3584 AP on TREC 2011, and 2012 microblog track datasets respectively. Their best runs achieved very high results that outperformed TREC best systems and our best run, but they can’t be put in comparison as they involve human intervention in the relevance feedback process.

Ferguson et al. [43] investigated the term weighting schemes for ranking tweets. For the baseline they used the Okapi BM25 model for tweets retrieval. Also, they incorporated standard query expansion by applying pseudo relevance feedback (PRF). Moreover, they introduced temporal reweighting of tweets based on the temporal distribution of the relevant ones. They down weight tweets occurring far from the center of the tweets assumed relevant in the PRF process. Their proposed temporal-based query expansion approach downgraded the performance of the baseline they used. Their baseline achieved 0.4211 P@30, 0.2109 MAP on TREC 2011 microblog track data.

Gao et al. [44] built a system based on the Okapi BM25 retrieval model. They applied a peak detection algorithm before doing pseudo relevance feedback (PRF) on the most relevant tweets for each search query. They start by building a time histogram for the tweets by grouping them hourly based on the 24 hours in the day; they call their basic unit of grouping a “bin”. Then, their algorithm starts detecting the peak when it encounters a significant increase in the bin count over the historical mean of bin counts, and stops when the rate goes back to the same value when it started, or when it encounters a new significant increase. Moreover, they automatically combined the relevance assessments of multiple retrieval techniques, where they combined the output from using only the BM25 model, and the BM25 after applying traditional PRF, and the BM25 after applying temporal based PRF. The results show that finding tweeting peeks and incorporating them while applying relevance feedback improves the retrieval performance. Their temporal based PRF achieved improvements over the baseline, where it achieved 0.171 P@30, 0.150 MAP while the baseline achieved 0.150 P@30, 0.133 MAP. But their best run submitted in terms of P@30 was using only the traditional PRF, where it achieved 0.182 P@30, 0.154 MAP. While their best run submitted in terms of MAP was the combination between the three relevance models, where they achieved 0.178 P@30, 0.157 MAP. All the reported numbers are conducted on TREC 2012 microblog track data.

Willis et al. [45] investigated different approaches using the temporal information in microblogs to improve the retrieval performance. They submitted four runs, where three of them used the temporal information and the fourth, which is the baseline, did not. Their baseline depends on the Markov Random Field (MRF) model proposed by Metzler and Croft [23], and applying pseudo relevance feedback (PRF) using Indri[1] search engine. Then, they applied a temporal based retrieval model by giving tweets occurring in relevant times a prior over other tweets. Moreover they investigated two approaches for using temporal evidence in query expansion. Firstly,

[1] http://www.lemurproject.org/indri/

they applied the recency-based query expansion, where they favor recent relevant tweets to be used for relevance feedback over older ones. Secondly, they applied relevance feedback on tweets posted in time periods having a high concentration of top relevant results. Finally, they combine the previous two temporal based query expansion approaches. They show the effectiveness of their combined temporal based query expansion, where they achieved their best results using it. Their best-submitted run achieved 0.204 P@30, 0.173 AP on the high relevance query set in TREC 2012 microblog track datasets.

3.3. Learning to Rank

Hyperlinks embedded in tweets has been used in documents expansion to improve the retrieval performance in microblogs as we discussed in section 1. Most of the work done in literature focused on expanding the tweets text with the corresponding web-page title crawled using the hyperlink embedded in it. In addition to using embedded hyperlinks in tweets for document expansion, hyperlinks have been used widely as features in learning to rank algorithms for improving microblog retrieval [17], [15], [16]. In this section we will give a glimpse on the work done on using hyperlinks in learning to rank algorithms as a good feature in determining the relevance of the tweet. Learning to rank is discussed in this section to motivate the importance of hyperlinks embedded in tweet. So, we will not discuss deeply about the use of learning to rank algorithms in microblog retrieval in this section, as it's not the main focus of our work.

McCreadie et al. [17] investigated how the contents of embedded hyperlinks in tweets can help estimating the tweet's relevance. They experiment with three approaches to show the importance of hyperlinks embedded in the tweets for improving microblog retrieval performance.

Firstly, They tried the straightforward solution, where they used the hyperlinked documents contents to extend the tweets as a virtual document. They appended the hyperlinked documents contents to the corresponding tweets so they can enrich the context of each tweet containing hyperlinks. The score for each query Q, and tweet t is calculated as following:

$$score_{hyperlink}(Q, t, d, d_l) = score(Q, t, d + d_l)$$

Figure 9 McCreadie et al. Virtual Document scoring function [17]

Where d is the tweet content without the hyperlink, and d_1 is the content of the hyperlink (the virtual document). One drawback for this weighting scheme, that the size of the hyperlinked document content is order of magnitude bigger than the tweets, and this will bias poor retrieval models that does not normalize the relevance score based on it's length.

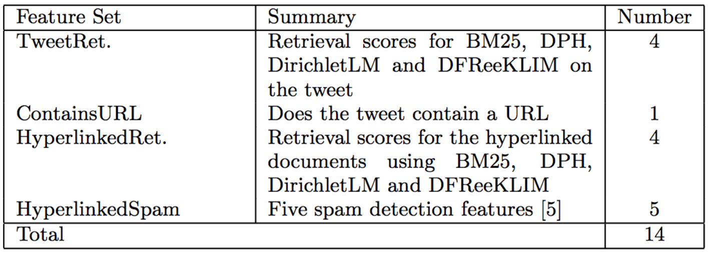
Secondly, they tried to address the drawbacks of the first approach. One solution is, to consider the tweet and the contents of the hyperlinked document two fields of the same document, and use field weighting support in some search engines to weight them. So, for each tweet a new document is created d_f, where it consists of two fields, d, and d_1. Then the score for each query Q, and tweet t is calculated as following:

$$score_{hyperlink}(Q, t, d, d_l) = score_f(Q, t, d_f, \mathbf{C}, \mathbf{W})$$

Figure 10 McCreadie et al. Field-Based Weighting scoring function [17]

Where C is a vector of field normalization parameters, for example: in the case of using Okapi BM25, we can add b (term frequency non-linearity), and D is a vector of the weights assigned to each field.

Finally, they decided to use the information of the hyperlinks in machine learning based re-ranking algorithm "learning to rank". Where the algorithm takes a set of features representing if the tweet can be relevant or non-relevant based on their existence/non-existence or the number of their occurrences, and training data consisting of tweets and their relevance assessment if they are relevant/non-relevant. The algorithm then tries to learn some weights for the features to determine based on them when a new unseen tweet comes if it's relevant or not. Figure 11 shows the set of the features they used to build their model.



Feature Set	Summary	Number
TweetRet.	Retrieval scores for BM25, DPH, DirichletLM and DFReeKLIM on the tweet	4
ContainsURL	Does the tweet contain a URL	1
HyperlinkedRet.	Retrieval scores for the hyperlinked documents using BM25, DPH, DirichletLM and DFReeKLIM	4
HyperlinkedSpam	Five spam detection features [5]	5
Total		14

Figure 11 McCreadie et al. Learning to Rank feature sets, descriptions and the number of features per set [17]

From their results they show that using the contents of the hyperlinked documents in tweets can improve the retrieval effectiveness over than using only the tweets text, and even just utilizing their existence is useful in determining the importance of a tweet. Their best runs achieved were 0.4252 P@30, 0.3810 MAP and 0.2091 P@30, 0.2112 MAP on TREC 2011, and 2012 microblog track datasets respectively.

3.4. Summary

Aforementioned work proposed various approaches to improve the retrieval effectiveness for microblog search. Although these approaches are highly advanced and usually led to improvements, they are either:

- Computationally costly (e.g. document expansion), where you have to process all the documents in the corpus before the indexing time.
- Dependent on third-party uncontrolled components (e.g. using web-search-based expansion from Google), where third-party components are considered black boxes that you can't build scientific proof on it. Moreover the algorithms they apply may change from one day to another.
- Require user's manual intervention, or require the presence of training data (e.g. learning to rank), which are not available freely. Additionally, there is no existing benchmark regarding the training data that people can compare their algorithms based on it.

Less attention was directed toward the utilization of tweets embedded hyperlinks in query expansion, which is seen to be more efficient and less complicated than other approaches, especially if optimized correctly.

Chapter 4. PROPOSED APPROACH

In this chapter we will present our proposed approach and describe the detailed architecture of the system we used to do the experimentation. In section 1, we will present our system architecture, how the data flow through the system, and a brief description of each module we used or built.

In section 2, we will elaborate on how we search the tweets collection offered by TREC, discuss the basics of the Lucene search engine used in our system as the baseline retrieval model. Moreover, we will present the APIs exposed by TREC to access the tweets collections. In section 3, we discuss how we select the top relevant documents (tweets) to be used in the PRF expansion process. In section 4, we present the preprocessing steps we apply on the tweets text before using them in the expansion process to remove noise and undesirable terms. In section 5, we discuss the important information we collect for the hyperlinked documents attached to the most relevant documents such as: the web pages titles, meta-description, and meta-keywords. In section 6, we show the steps we follow to extract the expansion terms that we will use it later in the query expansion process. In section 7, we discuss how the expansion terms are ranked using different term scoring functions to extract the most important terms that we believe will help in improving the retrieval performance. In section 8, we propose a way to control the effect of the expansion terms on the original query by giving a weight for both of them so we will not dilute the short original user queries with longer expansion terms.

Finally, in section 9 we show how we evaluate our proposed approach based on the evaluation scripts offered by TREC using different evaluation metrics.

4.1. System Architecture

We propose an end-to-end system for microblog retrieval leveraging query expansion using the traditional PRF with different configurations, in addition to; utilizing the contents of the hyperlinked documents attached to the top relevant search results.

As shown in Figure 12, our system operates as follows:

1. Search Tweet collections:

   Submit the user query to the search engine.

2. Select Top Relevant Tweets:

   Retrieve a list of relevant search results. Depending on the configuration parameters select the top n_d relevant search results to be used in the query expansion process.

3. Extract Expansion Terms:

   Extract the text of the top relevant search results after removing the original query words and stop words to be used later for expansion. Then, extract URLs of hyperlinked documents attached to top relevant search results. Finally, crawl page titles, meta-description, and meta-keywords of URLs extracted in the previous step.

4. Select Expansion Terms:

   Depending on the configuration of the expansion process combine the top n_t terms extracted from previous step with terms extracted from the tweets text to be used for forming the list of expansion terms.

5. Combine Selected Terms with the Original Query:

   Use the list of expansion terms to enrich the original query. The expansion terms are appended to the original user query with a weight so they won't dilute the effect of the original query given its short length. Finally, do a second round retrieval using the expanded query to serve the user a final list of search results.

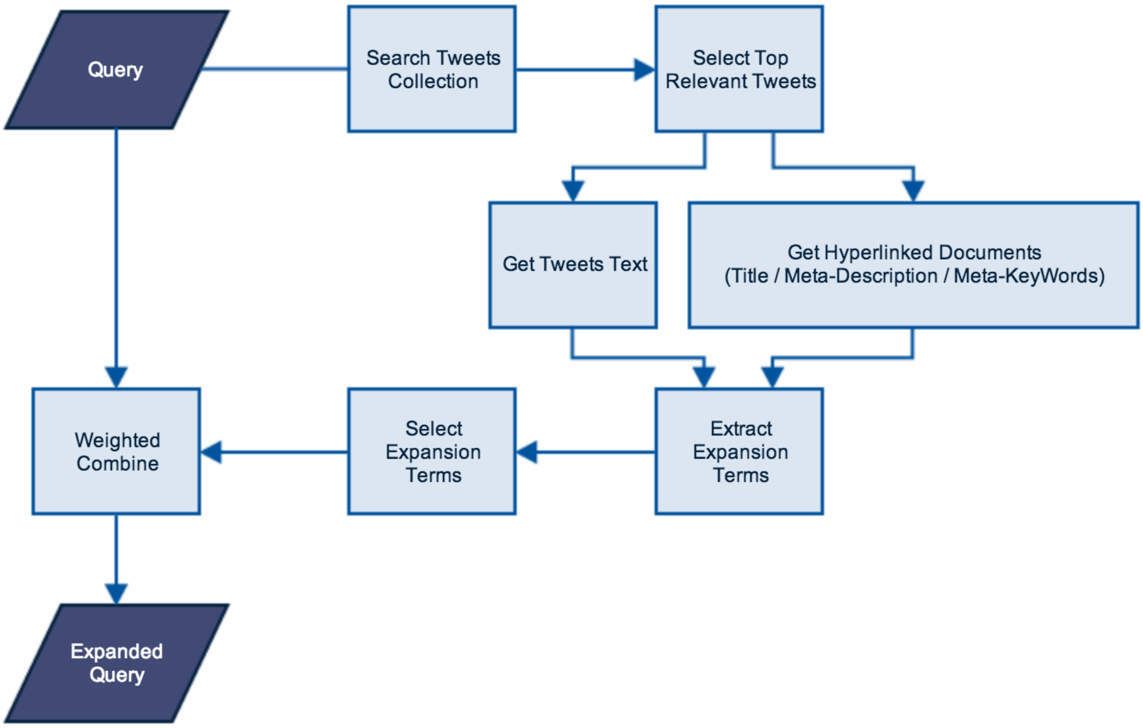


Figure 12 System Overview

In the previously mentioned process, we either used existing modules or build our own modules to process the data. Here is a list of the names of the modules we used:

- Search Tweets Collection
- Select Top Relevant Tweets
- Get Tweets Text
- Get Hyperlinked Documents
- Extract Expansion Terms
- Select Expansion Terms
- Weighted Combine

- Evaluation

In the next sections we will describe the main modules used in details.

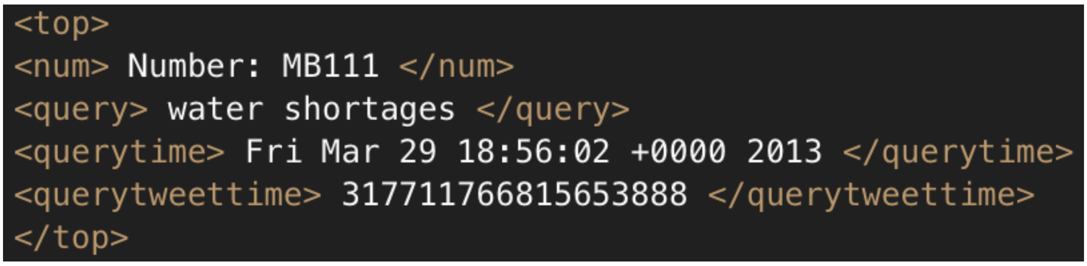
4.2. Search Tweets Collection

The search process starts by having the information need itself "User Queries". User queries generally consists of a small set of terms represent a certain information need for the user, who is looking for relevant search results in the big collection of tweets available in the system. TREC introduced three sets of user queries over three different years 2011, 2012, and 2013. The naming convention used by TREC for the user queries is "topics". Table 2 shows statistics for the number of topics offered by TREC for each year:

Table 2 TREC topics statistics

Year	# of Topics
2011	50
2012	60
2013	60

TREC topics are available in XML format as shown in Figure 13:



```
<top>
<num> Number: MB111 </num>
<query> water shortages </query>
<querytime> Fri Mar 29 18:56:02 +0000 2013 </querytime>
<querytweettime> 317711766815653888 </querytweettime>
</top>
```

Figure 13 Sample of TREC topics

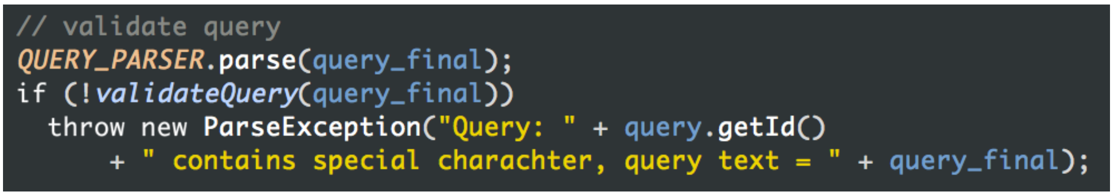
The details of the tags in the XML files is as following:

- <top> represent the full details of each topic.
- <num> represent the topic number, which is a serial number used later in the evaluation to map each topic with it's relevant documents.

- <query> the query text itself.
- <querytime> represent the time the query was issued, as all the tweets posted after the query time are not considered relevant.
- <querytweettime> represent the ID of the latest tweet posted before the query was issued, as all the tweets posted after the query time are considered irrelevant.

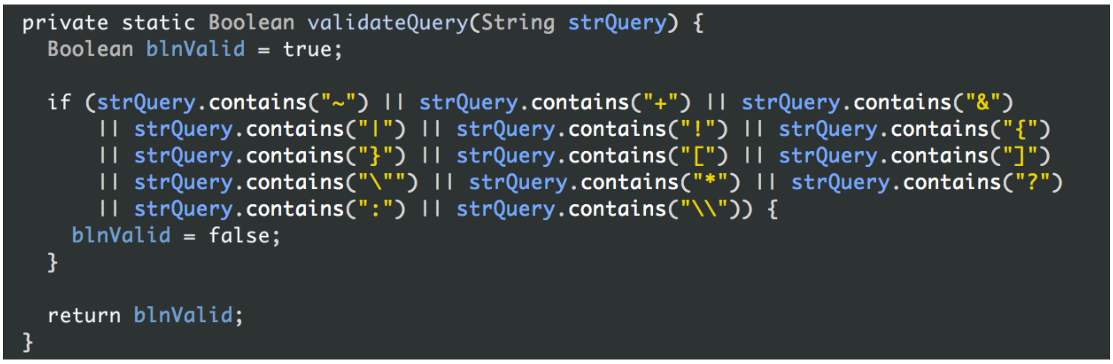
The XML files for all the three set of topics are parsed to extract the information needed for doing the search process such as: <query>, <querytime>, and <querytweettime>. Other information is extracted for the evaluation purposes such as <num>, which maps each query with its relevant search results.

As a validation step, before the queries are submitted to the search engine, we do parse them using Lucene query analyzer to make sure they don't contain any special character that may lead to failure in the retrieval process as shown in Figure 14 and Figure 15.



```
// validate query
QUERY_PARSER.parse(query_final);
if (!validateQuery(query_final))
  throw new ParseException("Query: " + query.getId()
      + " contains special charachter, query text = " + query_final);
```

Figure 14 Query Analyzer API call

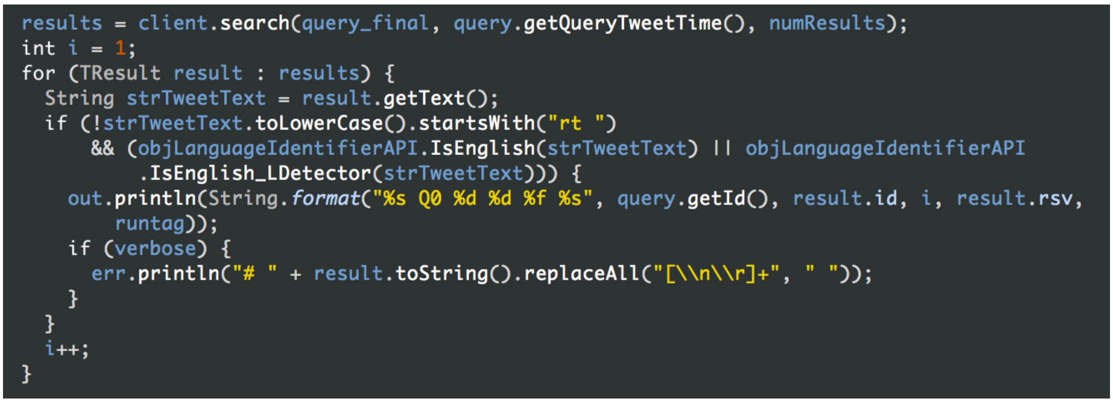


```
private static Boolean validateQuery(String strQuery) {
  Boolean blnValid = true;

  if (strQuery.contains("~") || strQuery.contains("+") || strQuery.contains("&")
      || strQuery.contains("|") || strQuery.contains("!") || strQuery.contains("{")
      || strQuery.contains("}") || strQuery.contains("[") || strQuery.contains("]")
      || strQuery.contains("\"") || strQuery.contains("*") || strQuery.contains("?")
      || strQuery.contains(":") || strQuery.contains("\\")) {
    blnValid = false;
  }

  return blnValid;
}
```

Figure 15 Check on Lucene special characters in the search queries

Then, all the queries extracted are submitted to the search engine to retrieve the best matching search results. The search engine and the tweets collections itself are on servers maintained by TREC, where they expose some APIs to do the search process. Figure 16 show how the search API works, the *client.search* method has three parameters as following:

- query_final: which is the final user query you are searching with, this maybe the original user query, or the final query after applying query expansion.
- query.QueryTweetTime(): which represents the latest tweet ID posted before the user submits his query, this information is provided in the < querytweettime> tag provided in the topics files.
- numResults: which is the number of results requested by the user from the search engine, typically we choose ten thousands results that we apply MAP on.



```
results = client.search(query_final, query.getQueryTweetTime(), numResults);
int i = 1;
for (TResult result : results) {
  String strTweetText = result.getText();
  if (!strTweetText.toLowerCase().startsWith("rt ")
      && (objLanguageIdentifierAPI.IsEnglish(strTweetText) || objLanguageIdentifierAPI
          .IsEnglish_LDetector(strTweetText))) {
    out.println(String.format("%s Q0 %d %d %f %s", query.getId(), result.id, i, result.rsv,
        runtag));
    if (verbose) {
      err.println("# " + result.toString().replaceAll("[\\n\\r]+", " "));
    }
  }
  i++;
}
```

Figure 16 TREC Search API sample call

The TREC microblog trach guidelines states that retweets and non-English tweets are considered non-relevant. So, as we can see in the code snippet in Figure 16, the retweets are filtered, by checking on the tweet text if it contains the pattern “rt ” in the beginning. Also, non-English tweets are filtered using two open source language identifiers to make sure if the tweet is English or not.

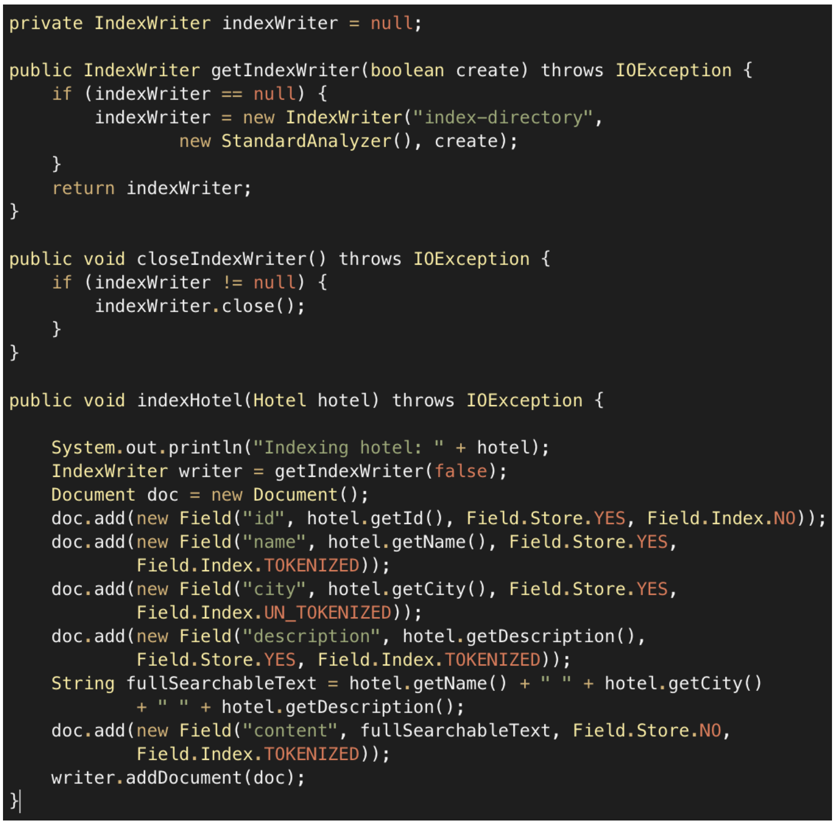
4.2.1. *Lucene Search Engine*

All the search API calls are served by the Lucene search engine. Lucene as in (http://lucene.apache.org/core/) is “a high-performance, fully-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform”. The baseline retrieval model offered by Lucene is based on the state of the art negative KL-divergence language modeling approach with Dirichlet prior smoothing parameter μ set to 2,500. Full-text search in Lucene takes place on two steps as following [46]:

1. Creating an index for the whole documents in the corpus.
2. Parse the queries submitted by the user to serve him a list of relevant search results out of the index built in the previous step.

Firstly, to create an index, you have to parse all the corpus you have, partition it to small chunks like: title, body, etc… “which is not the case for tweets”. Then, you have to instantiate a *Document*, and fill all the fields with the extracted data from the previous step. Finally, the *Document* is written in the *Index* using *IndexWriter*. Figure 17 shows a code snippet for how to use Lucene Java API to create a *Document* object, add *Field*s to it, and then write it in the *Index* using *IndexWriter*.

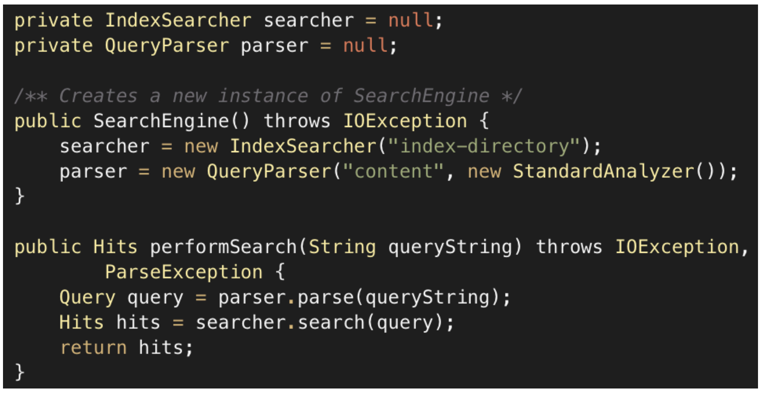
Secondly, after creating the index, users starts searching it using their queries. To support full-text search using Lucene Java API you need two classes: *QueryParser* and *IndexSearcher*. *QueryParser* is used to parse the query string submitted by the user and instantiate a *Query* object. The *Query* object is then used to search the index through *IndexSearcher.search()*.



```java
private IndexWriter indexWriter = null;

public IndexWriter getIndexWriter(boolean create) throws IOException {
    if (indexWriter == null) {
        indexWriter = new IndexWriter("index-directory",
                new StandardAnalyzer(), create);
    }
    return indexWriter;
}

public void closeIndexWriter() throws IOException {
    if (indexWriter != null) {
        indexWriter.close();
    }
}

public void indexHotel(Hotel hotel) throws IOException {

    System.out.println("Indexing hotel: " + hotel);
    IndexWriter writer = getIndexWriter(false);
    Document doc = new Document();
    doc.add(new Field("id", hotel.getId(), Field.Store.YES, Field.Index.NO));
    doc.add(new Field("name", hotel.getName(), Field.Store.YES,
            Field.Index.TOKENIZED));
    doc.add(new Field("city", hotel.getCity(), Field.Store.YES,
            Field.Index.UN_TOKENIZED));
    doc.add(new Field("description", hotel.getDescription(),
            Field.Store.YES, Field.Index.TOKENIZED));
    String fullSearchableText = hotel.getName() + " " + hotel.getCity()
            + " " + hotel.getDescription();
    doc.add(new Field("content", fullSearchableText, Field.Store.NO,
            Field.Index.TOKENIZED));
    writer.addDocument(doc);
}
```

Figure 17 Lucene indexing example (code snippet) [46]

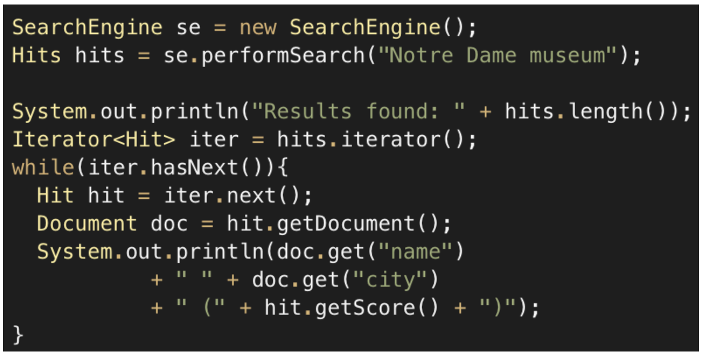
Figure 18 shows a code snippet of how the *Query* object, and *IndexSearcher.search()* are used to search the index built in the previous step and return a list of relevant search results in *Hits* object. The constructor of SearchEngine class firstly instantiate an object from *IndexSearcher* using the index created in the previous step. Then, it instantiates a *QueryParser* object to be used later to parse user queries. *QueryParser* constructor takes two paramaters, the first determines the field it is going to parse, and the second determines the *Analyzer* that will be used while parsing the query string. Next, *performSearch* method is called using the query string to first parse the query, then search the index, and finally return a list of relevant search results represented in *Hits* object.



```
private IndexSearcher searcher = null;
private QueryParser parser = null;

/** Creates a new instance of SearchEngine */
public SearchEngine() throws IOException {
    searcher = new IndexSearcher("index-directory");
    parser = new QueryParser("content", new StandardAnalyzer());
}

public Hits performSearch(String queryString) throws IOException,
        ParseException {
    Query query = parser.parse(queryString);
    Hits hits = searcher.search(query);
    return hits;
}
```

Figure 18 Lucene search example (code snippet) [46]

Finally, as shown in Figure 19, we iterate on the *Hits* object to print the details of each item in the list of relevant search results returned by Lucene *IndexSearcher*.



```
SearchEngine se = new SearchEngine();
Hits hits = se.performSearch("Notre Dame museum");

System.out.println("Results found: " + hits.length());
Iterator<Hit> iter = hits.iterator();
while(iter.hasNext()){
  Hit hit = iter.next();
  Document doc = hit.getDocument();
  System.out.println(doc.get("name")
           + " " + doc.get("city")
           + " (" + hit.getScore() + ")");
}
```

Figure 19 Lucene parse results example (code snippet) [46]

4.3. Select Top Relevant Tweets

After using the original search query to retrieve the initial list of search results, comes the second step in the relevance feedback process, where we select the top relevant tweets. In most of the studies that applied Pseudo Relevance Feedback on Microblogs [6], [7], [8], [9], [10], [11], [12] the number of relevant tweets used was based on heuristics, and there was no comprehensive study on how the number of the feedback documents can strongly affect the retrieval performance.

In our study, we comprehensively experiment with different values for the number of feedback documents (tweets) and measure their effect on the retrieval performance in Microblogs. Here is a list of the different values we use for the number of feedback documents based on the commonly used values in literature:

Table 3 Different values for number of feedback documents used

Number of feedback documents used
10
20
50
100

4.4. Get Tweets Text

For all the top relevant documents (tweets) we extract their text to be used for expansion terms extraction. As the tweet text extracted will be used later in extracting the expansion terms, we want to ensure that the expansion terms added is not junk and does not include redundant information from the original query. As a result, we extract the text of each tweet after applying some preprocessing steps as following:

- Apply Porter Stemmer on the original query terms and the expansion terms.
- Remove original query terms from expansion terms after stemming.

- Remove hyperlinks from expansion terms.
- Remove any HTML/CSS/JavaScript codes from the expansion terms.
- Eliminate any special character from the expansion terms, except # and @ as they are for special use in Twitter.
- Remove stop words from the expansion terms.
- Discard any non-English term in the expansion terms, as all our experiments are on English data sets, and non-English tweets are considered irrelevant.

4.5. Get Hyperlinked Documents

For all the top relevant documents (tweets) selected in the relevance feedback process, we parse their text and extract the embedded hyperlinks if they exist. Then, the hyperlinks extracted are used to crawl some important information related to the web-pages they point to that may help adding relevant term to the expansion process that did not occur in the original user query or the top relevant tweets itself. We extract three major pieces of information for each hyperlink:

1. Page title: generally try to summarize the page target in a very short length, and most of the time contains the website name.
2. Meta-Description: a very rich short document that get the most important information out of the page, usually used by search engines for indexing.
3. Meta-Keywords: an old way to help search engines index the web pages by assigning some tags (keywords) to each web page.

Titles and meta-description of hyperlinked documents may include unneeded text. For example, titles usually contain delimiters like '–' or '|' before/after page domain name, e.g., "... | CNN.com" and "... – YouTube". We clean these fields through the following steps:

- Split page titles on delimiters and discard the shorter substring, which is assumed to be the domain name.
- Detect error page titles, such as "404, page not found!" and consider them broken hyperlinks.
- Remove special characters, URLs, and snippet of HTML/JavaScript/CSS codes.

This process helps in discarding terms that are potentially harmful if used in query expansion.

4.6. Extract Expansion Terms

A hyperlink in a tweet is more than a link to related content as in webpages, but actually it is considered a link to the main focus of the tweet. In fact, sometimes tweet's text itself is totally irrelevant, and the main content lies in the embedded hyperlink, e.g. "This is really amazing, you have to check htwins.net/scale2".

By analyzing the TREC microblog dataset over the past three years, we found more than 70% of relevant tweets contain hyperlinks. This motivates utilizing the hyperlinked documents content in an efficient way for query expansion.

The content of hyperlinked documents in the initial set of top retrieved tweets is extracted and integrated into the PRF process. Titles of hyperlinked pages usually act like heading of the document's content, which can enrich the vocabulary in the PRF process. Similarly, meta-description of a webpage usually summarizes its content, which can be useful to further enrich the expansion terms. Furthermore, meta-keywords in webpages could help introduce relevant keywords to the contents of the web page.

We apply hyperlinked documents content extraction on five different levels:

- Tweets level (**PRF**): which represents the traditional PRF, where terms are extracted from the initial set of retrieved tweets while neglecting embedded hyperlinks.
- Hyperlinked document titles level (**HPRF-1**): where the page titles of the hyperlinked documents in feedback tweets are extracted and integrated to tweets for term extraction in the PRF process.
- Hyperlinked documents meta-description level (**HPRF-2**): the titles and meta-description of hyperlinked documents are extracted and integrated to tweets for term extraction.
- Hyperlinked documents meta-keywords level (**HPRF-3**): the titles, meta-description, and meta-keywords of hyperlinked documents are extracted and integrated to tweets for term extraction.

4.7. Select Expansion Terms

TFIDF and BM25 are used for ranking the top terms to be used for query expansion. We calculate the score for a term x as follows:

$$score(x) = [tf_t(x) + A * tf_{h_t}(x) + B * tf_{h_d}(x) + C * tf_{h_k}(x)] \cdot log\frac{N}{df(x)} \quad (1)$$

Where $tf_t(x)$ is the term frequency of term x in the top n_d initially retrieved tweet documents used in the PRF process. $tf_{h_t}(x)$ is the term frequency of term x in the titles of hyperlinks in the top n_d tweets. $tf_{h_d}(x)$ is the term frequency of term x in the meta-description of hyperlinks in the top n_d tweets. $tf_{h_k}(x)$ is the term frequency of term x in the meta-keywords of hyperlinks in the top n_d tweets. A, B, and C are binary functions that equal to 0 or 1 according to the content level of hyperlinked documents used in the expansion process. $df(x)$ Is document frequency of term x in the collection. N is the total number of documents in the collection.

Terms extracted from the top n_d initially retrieved documents are ranked according to equation 1, and top n_t terms with the highest score are used to formulate Q_E for the expansion process.

4.8. Weighted Combine

To balance the effect of the expansion terms on the final list or results we do give them weight not to dilute the original query effect. Weighted geometrical mean is used to calculate the final score of retrieval for a given Q query according to equation 2:

$$P(Q|d) = \sqrt{P(Q_0|d)^{1-\alpha} \cdot P(Q_E|d)^{\alpha}} \quad (2)$$

Where Q_0 is the original query. Q_E is the set of extracted expansion terms. $P\ (Q|d)$ is the probability of query Q to be relevant to document d. α is the weight given to expansion terms compared to original query (when $\alpha = 0$ no expansion is applied).

The weighted combine is an out of the box feature implemented in Lucene query language. Lucene reference the weighted combine as boosting. In Lucene you can apply boosting for a term or a phrase. Boosting allows you to control the importance of a term in a query, for example if you are searching for the query:

```
jakarta apache
```

And you want to give the term "jakarta" a boost to be important as 4 times as the term "apache", you use the caret "^" symbol to assign the boost, like following:

jakarta^4 apache

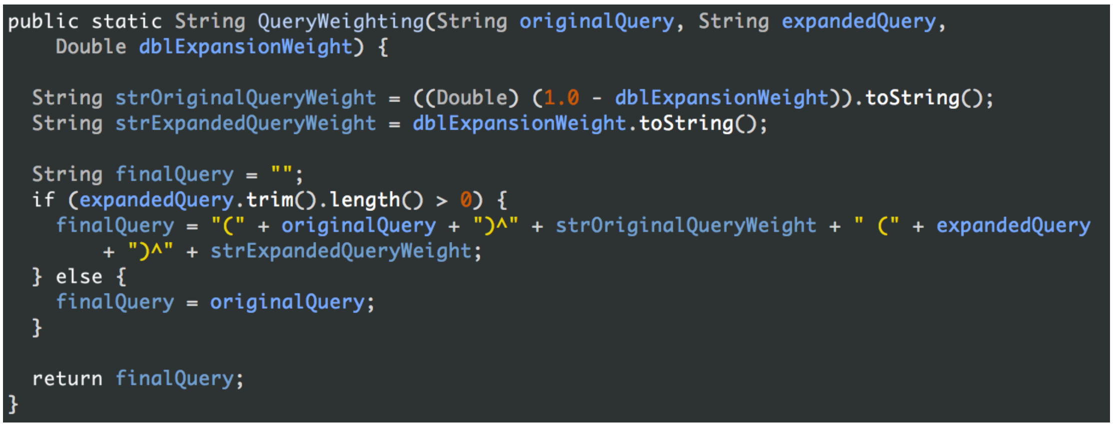
Boosting can also be applied on term phrases as following:

(jakarta apache)^4 (Apache Lucene)

The default value for the boost is 1 for all the terms, and the boost values should always be positive. Also, you can use boost values less than one such as: 0.2, which is our case.

We show how we implement the query weighting in Figure 20. The QueryWeighting function takes three parameters as following:

- originalQuery: the original user query available in the topics file.
- expandedQuery: the expansion terms space delimited.
- dblExpansionWeight: the weight given for the expansion terms, which varies from 0.0, where no expansion happens to 1.0, where the original query is neglected and all the weight of the query goes to the expansion terms.



```
public static String QueryWeighting(String originalQuery, String expandedQuery,
    Double dblExpansionWeight) {

  String strOriginalQueryWeight = ((Double) (1.0 - dblExpansionWeight)).toString();
  String strExpandedQueryWeight = dblExpansionWeight.toString();

  String finalQuery = "";
  if (expandedQuery.trim().length() > 0) {
    finalQuery = "(" + originalQuery + ")^" + strOriginalQueryWeight + " (" + expandedQuery
        + ")^" + strExpandedQueryWeight;
  } else {
    finalQuery = originalQuery;
  }

  return finalQuery;
}
```

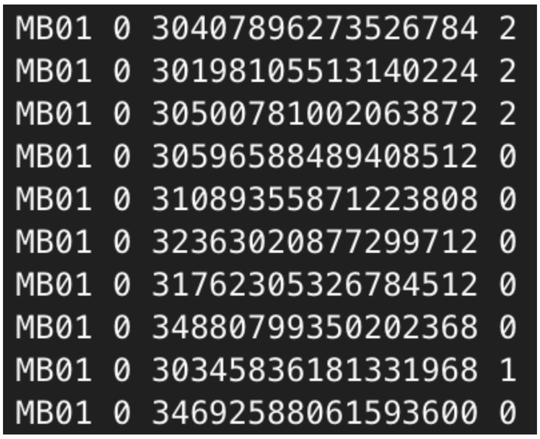
Figure 20 Code snippet for how the expansion terms weighting takes place

4.9. Evaluation

After submitting the query to the search engine and retrieving the list of search results, we want to evaluate the performance of the submitted query using the two evaluation metrics we use: P@30, and MAP. We do evaluate the results after searching the tweets collection with the original user query, and the result will stands for the baseline performance, in addition to, evaluating the list of search results after applying our different methods for query expansion to measure their effectiveness.

For the three different query sets we have, TREC offered the ground truth evaluation files they obtained manually. As shown in Figure 21, the format of the evaluation files is as following:

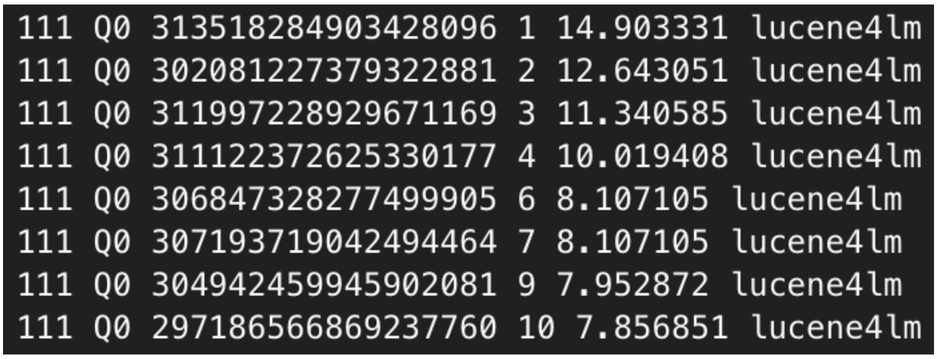
- Topic ID
- Dummy column used in other tracks
- Tweet ID
- Relevance score:
    - 0 => non-relevant, used in the cases you want to train a learning to rank algorithm and in need of negative data.
    - 1 => relevant.
    - 2 => highly relevant.



```
MB01 0 30407896273526784 2
MB01 0 30198105513140224 2
MB01 0 30500781002063872 2
MB01 0 30596588489408512 0
MB01 0 31089355871223808 0
MB01 0 32363020877299712 0
MB01 0 31762305326784512 0
MB01 0 34880799350202368 0
MB01 0 30345836181331968 1
MB01 0 34692588061593600 0
```

Figure 21 Snippet from the TREC Microblog track evaluation files

Having the ground truth evaluation files is the first step in the evaluation. The second step in the evaluation is to obtain the list of search results retuned from the search API after submitting the query. An example for the search results file returned from the search API is shown in Figure 22.



```
111 Q0 313518284903428096 1 14.903331 lucene4lm
111 Q0 302081227379322881 2 12.643051 lucene4lm
111 Q0 311997228929671169 3 11.340585 lucene4lm
111 Q0 311122372625330177 4 10.019408 lucene4lm
111 Q0 306847328277499905 6 8.107105 lucene4lm
111 Q0 307193719042494464 7 8.107105 lucene4lm
111 Q0 304942459945902081 9 7.952872 lucene4lm
111 Q0 297186566869237760 10 7.856851 lucene4lm
```

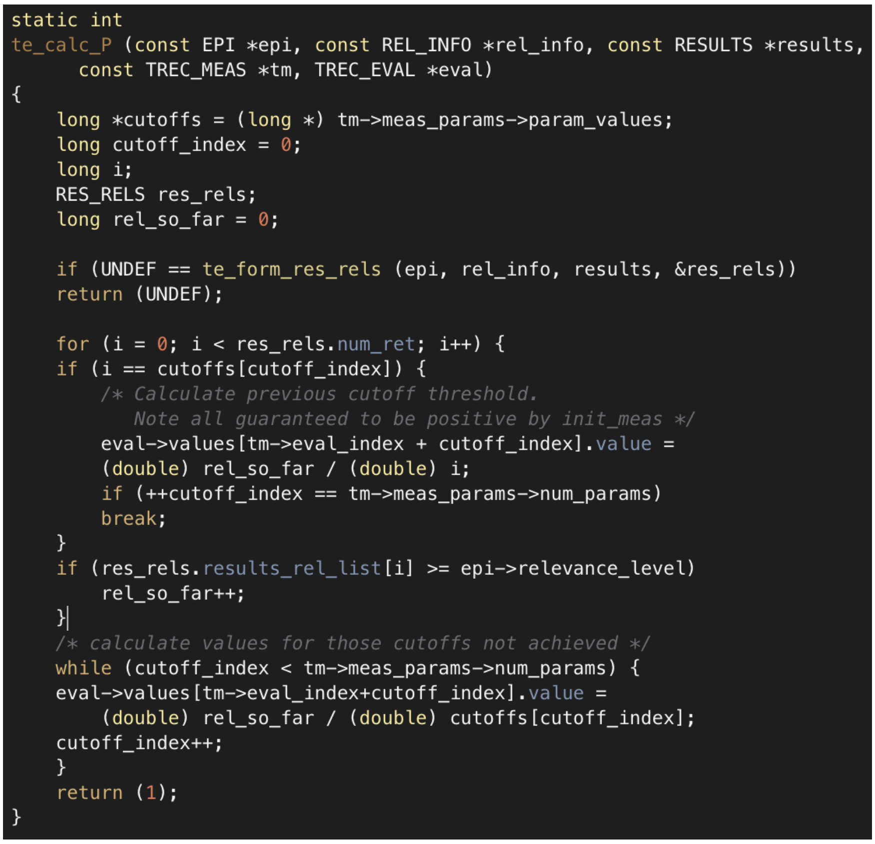
Figure 22 Sample for the search results file returned from the search API

The format of the search results returned from the search API is as following:

- Topic number
- Dummy column used in other TREC tracks
- Tweet ID
- Position in the search results list
- Relevance score

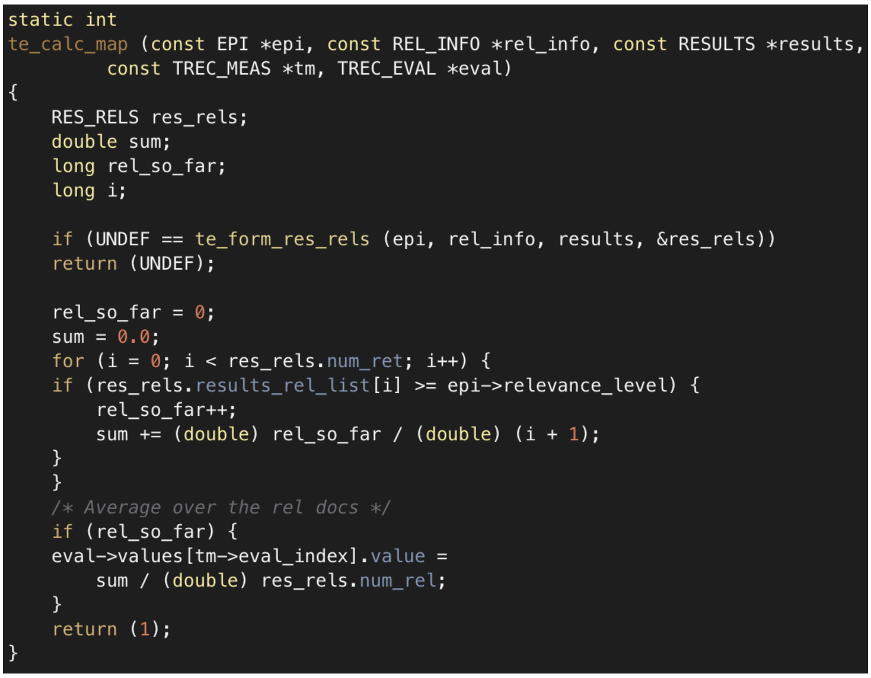
- Search engine name

Finally, the ground truth evaluation files and the list of search results returned from the search API are used to calculate the effectiveness of the approach used as discussed in the theoretical background chapter (pages: 16, 17). Figure 23 and Figure 24 show code snippets for how we calculate precision at position K and MAP respectively using TREC evaluation scripts.



```
static int
te_calc_P (const EPI *epi, const REL_INFO *rel_info, const RESULTS *results,
      const TREC_MEAS *tm, TREC_EVAL *eval)
{
    long *cutoffs = (long *) tm->meas_params->param_values;
    long cutoff_index = 0;
    long i;
    RES_RELS res_rels;
    long rel_so_far = 0;

    if (UNDEF == te_form_res_rels (epi, rel_info, results, &res_rels))
    return (UNDEF);

    for (i = 0; i < res_rels.num_ret; i++) {
    if (i == cutoffs[cutoff_index]) {
        /* Calculate previous cutoff threshold.
           Note all guaranteed to be positive by init_meas */
        eval->values[tm->eval_index + cutoff_index].value =
        (double) rel_so_far / (double) i;
        if (++cutoff_index == tm->meas_params->num_params)
        break;
    }
    if (res_rels.results_rel_list[i] >= epi->relevance_level)
        rel_so_far++;
    }
    /* calculate values for those cutoffs not achieved */
    while (cutoff_index < tm->meas_params->num_params) {
    eval->values[tm->eval_index+cutoff_index].value =
        (double) rel_so_far / (double) cutoffs[cutoff_index];
    cutoff_index++;
    }
    return (1);
}
```

Figure 23 Snippet of TREC Precision at position K evaluation script



```
static int
te_calc_map (const EPI *epi, const REL_INFO *rel_info, const RESULTS *results,
         const TREC_MEAS *tm, TREC_EVAL *eval)
{
    RES_RELS res_rels;
    double sum;
    long rel_so_far;
    long i;

    if (UNDEF == te_form_res_rels (epi, rel_info, results, &res_rels))
    return (UNDEF);

    rel_so_far = 0;
    sum = 0.0;
    for (i = 0; i < res_rels.num_ret; i++) {
    if (res_rels.results_rel_list[i] >= epi->relevance_level) {
        rel_so_far++;
        sum += (double) rel_so_far / (double) (i + 1);
    }
    }
    /* Average over the rel docs */
    if (rel_so_far) {
    eval->values[tm->eval_index].value =
        sum / (double) res_rels.num_rel;
    }
    return (1);
}
```

Figure 24 Snippet of TREC MAP evaluation script

Chapter 5. EXPERIMENTS AND RESULTS

In this chapter, we will include a description for all the experiment settings we did and discuss their results. Based on the evaluation set we have from TREC, we will evaluate our experiments for the three evaluation sets available (2011, 2012, and 2013), using P@30 and MAP for the first ten thousand results.

In section one, we will describe the two tweet collections offered by TREC for the two years 2011, and 2013, in addition to, we will describe the associated three test sets released by TREC for the three consecutive years 2011, 2012, and 2013. For each tweet collection / test set, we will mention the statistics of the dataset, the time period it was collected in, and general analysis. In section two, we will present the performance numbers achieved using the baseline run, without doing any query expansion.

In section three, we will discuss the parameter tuning process we do on the expansion weight. We will comprehensively experiment using different weights for the expansion terms, either by using 1:1 setup reported in the literature or using different values varying from 0 where no expansion happens, to 0.9 which is the highest expansion weight while keeping an effect for the original query.

In section four, we will present all the experiments we did on query expansion and report the performance improvements we achieved measured in P@30 and MAP for the top ten thousands search results. We start by applying the traditional Pseudo Relevance Feedback (PRF) using different configurations for the number of feedback documents and terms. Next, we apply our proposed hyperlink-extended pseudo relevance feedback approach (HRPF) on three levels, web page titles, meta-description, and meta-keywords.

Finally, in section five, we summarize all the experiments we did and show the best runs achieved using different approaches for different configurations. We show that our HPRF proposed approach achieves the best result when utilizing the web pages titles and meta-description and using the BM25 weighting scheme.

5.1. Test Collection

We used the TREC microblog track datasets from 2011, 2012, and 2013 tracks in our experimentation. The datasets include two tweet collections and three query sets. The TREC 2011 microblog collection contains 16 million tweets representing a sample stream between January 23 and February 7, 2011. Two sets of 50 and 60 topics were released in 2011 and 2012 respectively against this collection along with relevance judgments done manually. A much larger collection of 243 million tweets was released in 2013 as a sample stream between February 1 and March 31, 2013 along with 60 topics and relevance judgments done manually.

We applied a baseline run by searching the tweets collections with the original queries Q_0. All hyperlinks appearing in the top 100 results for each query were extracted. Title, meta-description, and meta-keywords of the corresponding hyperlinked document (webpage) were extracted. We found that on average 60-65% of the results contained embedded hyperlinks, which confirms the importance of utilizing these hyperlinks effectively. Table 4 presents the number of extracted hyperlinks from the top ranked 100 results for each query of each year, and shows the number of extracted page titles and meta-description for these hyperlinks. As shown in Table 4, the percentage of broken links for 2011 and 2012 were much larger than 2013. This is expected because of the age of the 2011 tweet collection.

Table 4 Number of extracted hyperlinks and the number of extracted content of them

dataset	*hyperlinks*	*extracted titles/desc.*	*broken%*
2011	3164	2116 / 1612	33.0%
2012	3781	2360 / 1760	37.6%
2013	3694	2920 / 2099	20.9%

5.2. Baseline

We started by searching the two tweets collections using the original three query sets without applying any query expansion. Table 5 presents the baseline runs of searching the collections without applying query expansion (i.e. $\alpha = 0$ in equation 2). This baseline is just for comparison with examined PRF methods. From the results we can see that the query set for 2012 is performing the worst, which will be the same behavior after applying our approach, nevertheless, we achieve significant improvements over the baseline for the three evaluation sets.

Table 5 Baseline results for TREC Microblog test sets

	2011	**2012**	**2013**
P@30	0.4238	0.3565	0.4500
MAP	0.3882	0.2275	0.2524

5.3. Parameter Tuning

Initially, we applied PRF while setting n_d to 50 documents and n_t to 10 terms based on the most reported setup in TREC microblog track [15], [16] to find the optimal expansion weight α. We tested several values of α ranging from 0.0, where no expansion where applied to 0.9, which is the highest weight for the expansion terms without discarding the original query effect. Then, we monitored the retrieval effectiveness measured by P@30 and MAP. From the results shown in Table 6 and Table 7, we can conclude that the value of α that achieved the best performance for the three topic sets was 0.2 for both P@30 and MAP. This is different than literature that typically used 1:1 weighting between original and expanded query terms. We set $\alpha = 0.2$ for the rest of our experiments.

Table 6 PRF P@30 performance for different weighting schemes

P@30										
α	***0.1***	***0.2***	***0.3***	***0.4***	***0.5***	***0.6***	***0.7***	***0.8***	***0.9***	***1:1***
2011	0.4755	0.485	0.4818	0.4796	0.4592	0.4252	0.3946	0.3673	0.3408	0.4585
2012	0.3994	0.4136	0.4102	0.3938	0.3554	0.3158	0.2791	0.2565	0.2384	0.3548
2013	0.475	0.4811	0.4467	0.4244	0.3856	0.3461	0.3167	0.2844	0.255	0.3861

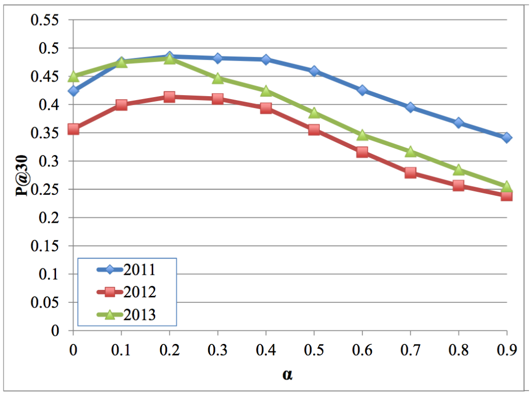


Figure 25 PRF P@30 performance for different weighting schemes

Table 7 PRF MAP performance for different weighting schemes

MAP										
α	***0.1***	***0.2***	***0.3***	***0.4***	***0.5***	***0.6***	***0.7***	***0.8***	***0.9***	***1:1***
2011	0.4385	0.4386	0.4261	0.4086	0.3627	0.3185	0.2747	0.2439	0.2164	0.3592
2012	0.282	0.2845	0.2798	0.2575	0.2169	0.1768	0.1465	0.1241	0.1088	0.2074
2013	0.3226	0.3262	0.3134	0.2895	0.2432	0.1889	0.1527	0.1244	0.1046	0.2256

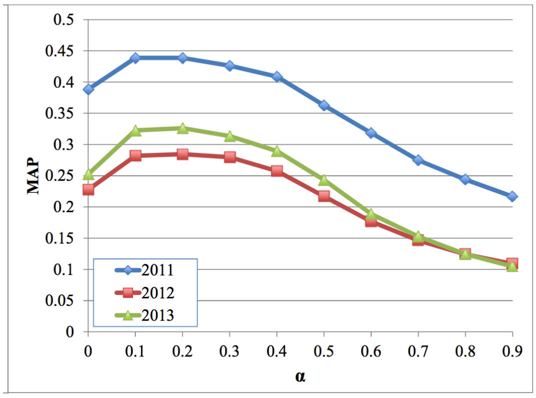


Figure 26 PRF MAP performance for different weighting schemes

5.4. Experiments

Once we had α set, we ran different configurations for PRF, HPRF-1, HPRF-2, and HPRF-3 using different values of expansion document and terms. n_d values tested were {10, 20, 50, and 100}, and $n_t$values tested were {5, 10, 15, 20, and 30}. Moreover, we try different term weighting schemes for selecting the top n_t term, such as TFIDF and Okapi BM25. We aim from this extensive experimentation to study the effect of different configurations of relevance feedback on the retrieval performance in microblogs. In addition, try to get the best-recommended setup for PRF in microblog retrieval rather than heuristically selecting numbers that may be suboptimal.

For the different configurations, we compared the performance of applying hyperlink-extended PRF to traditional PRF via comparing P@30 and MAP, and applying statistical significance test using paired t-test with p-value < 0.05.

5.4.1. *Traditional PRF*

Objective:

The objective of this set of experiments is to study the effect of traditional Pseudo Relevance Feedback (PRF) on the performance of microblog retrieval. In addition to, measuring to what extent the relevance feedback configuration regarding the number of feedback documents and terms can affect the retrieval performance.

Methodology:

In this set of experiments we apply the traditional PRF under different configurations regarding the number of feedback documents {10, 20, 50, and 100} and the number of feedback terms {5, 10, 15, 20, and 30}. The basic steps of applying PRF is as following (refer to Chapter 2 for details, page 13):

- Search the tweets collection using the original user query.
- Consider the top n_d documents relevant.
- Select the top n_t term from the top relevant documents.
- Use the top selected terms in the previous step to expand the original user query.
- Search the tweets collection with the expanded query aiming better retrieval performance.

Results and Discussions:

Table 8 and Table 9 reports full results of PRF using tweets content only with different number of feedback documents (n_d) and terms (n_t), while underlining runs that achieved the best scores. Table 8 reports the full results by selecting the top n_t terms using the TFIDF weighting scheme, while Table 9 reports the full results by selecting the top n_t terms using the Okapi BM25 weighting scheme. Almost all the configurations led to big improvements over the baseline when compared by P@30 and MAP, which confirms the effectiveness of PRF in microblog retrieval.

Test sets 2011 and 2013 showed the best performance achieved when using only 10 documents in the feedback process, while 2012 best performance was when using 50 documents in feedback. Best results achieved using PRF for TREC 2011 dataset outperformed reported results achieved by other approaches on the same test collection, which was studied extensively in the literature, including: document expansion [13], temporal-based query expansion [6], [11], web-search-based query expansion [36], [7], and learning-to-rank [11], [15]. The main reason for this result may be that previous studies did not give much attention to using small number of feedback document in the PRF process, which we showed to achieve the best results for 2011 and 2013 test sets.

Table 8 PRF (TFIDF) results for different number of feedback document and terms for TREC microblog dataset when α=0.2

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4878	**0.4939**	0.4837	0.4830	0.4898	0.3672	0.3763	0.3977	0.3972	0.3977	0.4983	0.5083	0.5156	0.5067	**0.5178**
	20	0.4796	0.4850	0.4918	0.4789	0.4769	0.3870	0.3927	0.3932	0.3977	0.4113	0.4906	0.4972	0.5028	0.4928	0.4922
	50	0.4769	0.4810	0.485	0.4837	0.4741	0.3881	0.4102	**0.4147**	0.413	0.4136	0.4944	0.4706	0.4789	0.4822	0.485
	100	0.4701	0.4769	0.4789	0.4816	0.4810	0.3938	0.3847	0.3972	0.4006	0.4073	0.4883	0.4944	0.4878	0.4756	0.4689
MAP	**10**	0.4387	**0.4452**	0.4386	0.4398	0.4401	0.2632	0.2651	0.2720	0.2710	0.2721	0.3165	0.3259	0.3374	0.3368	**0.3421**
	20	0.4363	0.4411	0.4382	0.4347	0.4314	0.2678	0.2729	0.2765	0.2796	0.2825	0.3226	0.3291	0.3325	0.3235	0.3254
	50	0.4353	0.4363	0.4387	0.4359	0.4341	0.2748	0.2826	0.2848	0.2855	**0.2925**	0.3229	0.3147	0.3251	0.3233	0.3236
	100	0.4210	0.4267	0.4318	0.4311	0.4276	0.2725	0.2746	0.2803	0.2842	0.2842	0.3153	0.3226	0.3186	0.3141	0.3094

Table 9 PRF (BM25) results for different number of feedback document and terms for TREC microblog dataset when α=0.2

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4707	0.4803	0.4810	**0.4864**	0.4857	0.3576	0.3797	0.3915	0.4006	0.4017	0.5106	0.5239	0.5156	0.5278	**0.5322**
	20	0.4694	0.4837	0.4844	0.4741	0.4776	0.3915	0.3949	0.3977	0.4096	0.4124	0.5094	0.5133	0.5111	0.5094	0.5078
	50	0.4728	0.4810	0.4707	0.4694	0.4694	0.4119	**0.4203**	0.4186	0.4158	0.4164	0.4894	0.4856	0.4906	0.4839	0.4878
	100	0.4639	0.4653	0.4660	0.4701	0.4612	0.4130	0.4068	0.4045	0.4011	0.4000	0.4822	0.475	0.4744	0.4772	0.4783
MAP	**10**	0.4248	0.4373	**0.4422**	0.4405	0.4392	0.2578	0.2677	0.2707	0.2717	0.2766	0.3251	0.3368	0.3388	0.3443	**0.3492**
	20	0.4218	0.4337	0.4381	0.4320	0.4270	0.2690	0.2737	0.2760	0.2835	0.2842	0.3196	0.3268	0.3347	0.3333	0.3394
	50	0.4245	0.4298	0.4318	0.4289	0.4223	0.2801	0.2850	**0.2951**	0.2917	0.2922	0.3128	0.3164	0.3231	0.3267	0.3267
	100	0.4093	0.4220	0.4173	0.4158	0.4101	0.2784	0.2781	0.2840	0.2855	0.2890	0.3116	0.3110	0.3128	0.3143	0.3161

5.4.2. *Hyperlink-extended PRF*

5.4.2.1. Hyperlinked-extended PRF (titles)

Objective:

The objective of this set of experiments is to study the effect of our proposed hyperlink-extended pseudo relevance feedback (HPRF) on the performance of microblog retrieval by utilizing the web page titles of the hyperlinks embedded in the top relevant documents. In addition to, measuring to what extent the relevance feedback configuration regarding the number of feedback documents and terms can affect the retrieval performance.

Methodology:

In this set of experiments we apply our proposed HPRF approach by utilizing the web page titles of the hyperlinks embedded in the top relevant documents "HPRF-1". We apply HPRF-1 under different configurations regarding the number of feedback documents {10, 20, 50, and 100} and the number of feedback terms {5, 10, 15, 20, and 30}. The steps to apply HPRF-1 is as following:

- Search the tweets collection using the original user query.
- Consider the top n_d documents relevant.
- Extract the web-page titles of the hyperlinks attached to the top relevant documents (tweets) and append them to their corresponding tweets.
- Select the top n_t term from the top relevant documents.
- Use the top selected terms in the previous step to expand the original user query.
- Search the tweets collection with the expanded query aiming better retrieval performance.

Results and Discussions:

Extending text used for PRF by hyperlinked documents titles led to improvements in the retrieval effectiveness, which indicates its importance in the expansion process.

We noticed that the best runs in HPRF-1 performed a lot better than PRF for 2013 datasets compared to 2011, and 2012. One interpretation for the fair performance on 2011, 2012 evaluation sets is the percentage of broken hyperlinks in the feedback tweets, see Table 4. Doing further analysis on the three evaluation sets, we show that the performance of HPRF-1 is proportional to the number of broken hyperlinks in the top retrieved tweets as in Figure 27. In Figure 27, we show the effect of broken hyperlinks on the average P@30 gain from HPRF-1 over PRF for all the evaluation sets, in case of using 5 relevant feedback tweets.

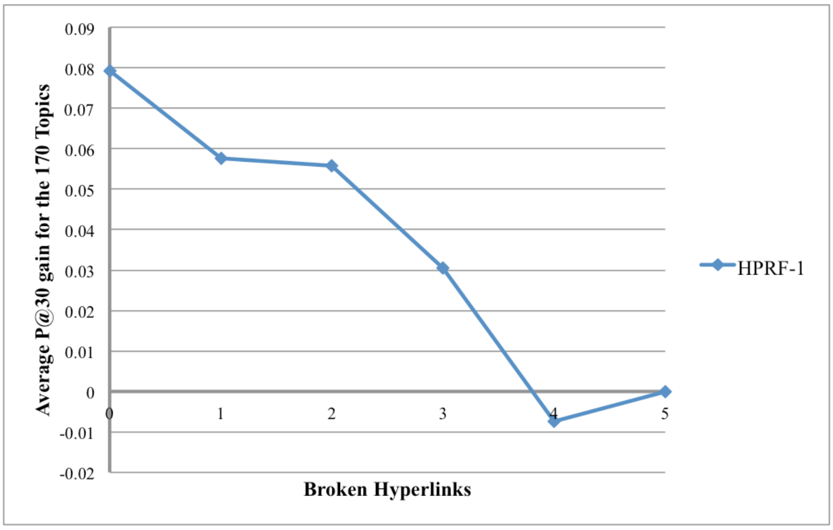


Figure 27 Using Hyperlinked Documents titles is not effective when the number of broken Hyperlinks in the top retrieved results increases

Table 10 and *Table 12* reports full results of different configurations for HPRF-1 and highlights the runs that achieved statistically significant improvement over the corresponding configuration in PRF. *Table 10* reports the full results by selecting the top n_t terms using the TFIDF weighting scheme, while *Table 12* reports the full results by selecting the top n_t terms using the Okapi BM25 weighting scheme. Comparing results in *Table 10* to that in *Table 8*, about 80% of the different configurations of PRF were improved using HPRF-1.

Significance test:

We applied statistical significance test using paired t-test with p-value < 0.05 to check the runs where HPRF-1 is performing significantly better than PRF. As shown in *Table 11* the t-test results, where runs having p-value < 0.05 are considered significantly better than PRF using the TFIDF weighting scheme. From the results we can see that significant improvements were achieved for some of the PRF configurations of TREC 2012, and 2013 test sets. *Table 13* shows the t-test results, where runs having p-value < 0.05 are considered significantly better than PRF using the Okapi BM25 weighting scheme. From the results we can see that significant improvements were achieved for some of the PRF configurations of TREC 2011, 2012, and 2013 test sets.

From the significance test results we can notice that using Okapi BM25 term weighting scheme leads to more consistent behavior across the three test sets when compared to TFIDF.

Table 10 HPRF-1 (TFIDF) results for different number of feedback document and terms for TREC microblog dataset when α=0.2. Gray cells indicates statistically significant improvement over corresponding PRF configuration

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4864	0.4946	0.4952	**0.498**	0.4966	0.3842	0.3915	0.3972	0.4073	0.413	0.5339	**0.5483**	0.5378	0.5267	0.5283
	20	0.4823	0.4905	0.4905	0.4871	0.4816	0.3887	0.3994	0.4164	0.4237	0.4215	0.51	0.53	0.5261	0.5194	0.505
	50	0.4694	0.4796	0.4762	0.481	0.4728	0.4085	0.4192	**0.4249**	0.422	0.4226	0.4933	0.5	0.4972	0.495	0.4933
	100	0.4782	0.4755	0.4769	0.4741	0.4687	0.4017	0.4068	0.4113	0.4102	0.4	0.4978	0.4817	0.4733	0.4767	0.4906
MAP	**10**	0.4381	**0.4558**	0.4525	0.4525	0.4504	0.2703	0.274	0.2781	0.2795	0.2822	0.3314	0.3434	0.3491	0.3471	**0.3498**
	20	0.4376	0.4458	0.4499	0.4405	0.4284	0.2747	0.2836	0.2898	0.2908	0.2937	0.3279	0.3428	0.345	0.344	0.3433
	50	0.4279	0.434	0.4305	0.4373	0.4252	0.2827	0.2949	0.2962	**0.2994**	0.2987	0.3264	0.3277	0.3316	0.3358	0.3359
	100	0.4284	0.4278	0.421	0.4198	0.4164	0.2783	0.2859	0.2901	0.2924	0.2912	0.3159	0.3179	0.3193	0.3188	0.3255

Table 11 HPRF-1 (TFIDF) and PRF statistical significance test results using paired t-test with p-value < 0.05

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.9019	0.9412	0.2170	0.1114	0.1922	0.0790	0.1343	0.9481	0.3198	0.1255	0.0257	0.0329	0.1394	0.0295	0.1525
	20	0.7704	0.6401	0.8848	0.3656	0.6314	0.8163	0.4990	0.1076	0.0207	0.2978	0.1885	0.0946	0.2239	0.1436	0.3012
	50	0.2965	0.7960	0.2946	0.6673	0.8361	0.0577	0.4063	0.2260	0.2086	0.2709	0.9398	0.0523	0.1804	0.4146	0.5423
	100	0.1473	0.8338	0.7015	0.2198	0.1007	0.3089	0.0218	0.1442	0.2214	0.3431	0.6019	0.5815	0.5077	0.9492	0.1608
MAP	**10**	0.9278	0.2584	0.1031	0.0570	0.0372	0.0938	0.0302	0.1349	0.0788	0.0301	0.0227	0.0093	0.0767	0.1182	0.1538
	20	0.8813	0.6023	0.1751	0.3443	0.5797	0.2590	0.1080	0.0170	0.0195	0.0514	0.3999	0.0487	0.1204	0.0043	0.0062
	50	0.1235	0.7668	0.3683	0.7850	0.2165	0.1510	0.0363	0.0576	0.0206	0.2455	0.6285	0.0470	0.3220	0.0782	0.1109
	100	0.3463	0.8704	0.1738	0.1802	0.2060	0.2488	0.0215	0.0895	0.1040	0.1762	0.9500	0.6330	0.9423	0.5535	0.0698

Table 12 HPRF-1 (BM25) results for different number of feedback document and terms for TREC microblog dataset when α=0.2. Gray cells indicates statistically significant improvement over corresponding PRF configuration

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4701	**0.4946**	0.4912	0.4932	0.4945	0.3825	0.3859	0.3910	0.4068	0.4107	0.5344	**0.5394**	0.5339	0.5328	0.5306
	20	0.4687	0.4823	0.4850	0.4884	0.4844	0.3915	0.4006	0.4011	0.4119	0.4198	0.5167	0.5294	0.5139	0.5083	0.5133
	50	0.4714	0.4680	0.4646	0.4714	0.4633	0.4000	0.4147	0.4203	0.4209	**0.4237**	0.4894	0.4956	0.4961	0.4939	0.4922
	100	0.4626	0.4599	0.4599	0.4551	0.4599	0.3989	0.3989	0.4045	0.4079	0.4062	0.4817	0.4711	0.4828	0.4817	0.4717
MAP	**10**	0.4318	**0.4510**	0.4487	0.4495	0.4493	0.2656	0.2718	0.2739	0.2796	0.2829	0.3306	0.3396	0.3457	0.3452	**0.3492**
	20	0.4245	0.4345	0.4419	0.4349	0.4279	0.2730	0.2819	0.2827	0.2882	0.2941	0.3262	0.3356	0.3355	0.3379	0.3441
	50	0.4251	0.4290	0.4298	0.4285	0.4214	0.2776	0.2848	0.2931	0.2962	**0.2974**	0.3109	0.3197	0.3250	0.3262	0.3285
	100	0.4021	0.4157	0.4135	0.4155	0.4123	0.2704	0.2789	0.2831	0.2865	0.2865	0.3127	0.3072	0.3165	0.3129	0.3123

Table 13 HPRF-1 (BM25) and PRF statistical significance test results using paired t-test with p-value < 0.05

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.8715	0.1298	0.1168	0.1424	0.0142	0.0662	0.3935	0.9135	0.3839	0.1388	0.0492	0.0336	0.0173	0.0033	0.6648
	20	0.8928	0.8933	0.9083	0.0077	0.0958	0.9997	0.5542	0.7638	0.7620	0.0963	0.2069	0.2844	0.8361	0.8923	0.0447
	50	0.6993	0.0476	0.4108	0.6507	0.3347	0.1152	0.4340	0.7918	0.3493	0.2560	0.9998	0.1536	0.6168	0.2804	0.5696
	100	0.8367	0.2527	0.3402	0.0261	0.8703	0.1328	0.3085	0.9990	0.3408	0.3516	0.9598	0.6885	0.4762	0.7035	0.5044
MAP	**10**	0.3366	0.0464	0.1640	0.0394	0.0018	0.0848	0.1898	0.1774	0.0491	0.0710	0.3691	0.4822	0.0496	0.6953	0.9961
	20	0.6860	0.8975	0.3557	0.5910	0.7761	0.1330	0.0481	0.1424	0.2214	0.0048	0.0465	0.1416	0.8563	0.1865	0.0027
	50	0.9010	0.8826	0.7176	0.9361	0.7760	0.5967	0.9498	0.5189	0.2058	0.1932	0.6942	0.3001	0.6961	0.9326	0.7153
	100	0.3116	0.4424	0.5366	0.9763	0.6551	0.1081	0.8506	0.7614	0.7809	0.3899	0.8215	0.4803	0.4601	0.7718	0.5116

5.4.2.2. Hyperlinked-extended PRF (titles + meta-descriptions)

Objective:

The objective of this set of experiments is to study the effect of our proposed hyperlink-extended pseudo relevance feedback (HPRF) on the performance of microblog retrieval by utilizing the web page titles and meta-descriptions of the hyperlinks embedded in the top relevant documents. In addition to, measuring to what extent the relevance feedback configuration regarding the number of feedback documents and terms can affect the retrieval performance.

Methodology:

In this set of experiments we apply our proposed HPRF approach by utilizing the web page titles and meta-descriptions of the hyperlinks embedded in the top relevant documents “HPRF-2”. We apply HPRF-2 under different configurations regarding the number of feedback documents {10, 20, 50, and 100} and the number of feedback terms {5, 10, 15, 20, and 30}. The steps to apply HPRF-2 is as following:

- Search the tweets collection using the original user query.
- Consider the top n_d documents relevant.
- Extract the web-page titles and meta-descriptions of the hyperlinks attached to the top relevant documents (tweets) and append them to their corresponding tweets.
- Select the top n_t term from the top relevant documents.
- Use the top selected terms in the previous step to expand the original user query.
- Search the tweets collection with the expanded query aiming better retrieval performance.

Results and Discussions:

Extending text used for PRF by hyperlinked documents content led to improvements in the retrieval effectiveness. We noticed that HPRF-1 led to less significant improvements than HPRF-2, which indicates that using the meta-description of hyperlinked webpage documents in PRF further improve the results.

Table 14 and Table 16 reports full results of different configurations of HPRF-2 and highlights the runs that achieved statistically significant improvement over the corresponding configuration in PRF. Table 14 reports the full results by selecting the top n_t terms using the TFIDF weighting scheme, while Table 16 reports the full results by selecting the top n_t terms using the Okapi BM25 weighting scheme. Comparing results in Table 14 to that in Table 8, over 90% of the different configurations of PRF were improved using HPRF-2. The noticed better performance for 2013 test set over 2012 and 2011 may be a result of the percentage of broken hyperlinks of 2012 and 2011 was much higher than 2013 leading to less content to be added for query expansion, see Table 4 and Figure 27.

Significance test:

We applied statistical significance test using paired t-test with p-value < 0.05 to check the runs where HPRF-2 is performing significantly better than PRF. Table 15 shows the t-test results, where runs having p-value < 0.05 are considered significantly better than PRF using the TFIDF weighting scheme. From the results significant improvements were achieved for most of the PRF configurations of TREC 2013 test set, and some of 2012 test set. Table 17 shows the t-test results, where runs having p-value < 0.05 are considered significantly better than PRF using the Okapi BM25 weighting scheme. From the results significant improvements were achieved for most of the PRF configurations of TREC 2013 test set, and some of 2011, and 2012 test sets.

From the significance test results we can notice that using Okapi BM25 term weighting scheme leads to more consistent behavior across the three test sets when compared to TFIDF.

Table 14 HPRF-2 (TFIDF) results for different number of feedback document and terms for TREC microblog dataset when α=0.2. Gray cells indicates statistically significant improvement over corresponding PRF configuration

	n_t	**2011**					**2012**					**2013**				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4741	0.4932	0.4918	0.4946	**0.4959**	0.3836	0.3944	0.4045	0.4169	0.4164	0.5444	0.5489	**0.5544**	0.5433	0.5317
	20	0.4735	0.4884	0.4925	0.4884	0.4918	0.3893	0.3966	0.4028	0.4028	0.4254	0.51	0.5356	0.5411	0.5372	0.5322
	50	0.4741	0.4789	0.4796	0.4789	0.4755	0.4056	0.4056	0.4175	0.4271	**0.4311**	0.505	0.5033	0.5094	0.4967	0.5022
	100	0.4755	0.4714	0.4728	0.4707	0.4687	0.4107	0.4028	0.4113	0.4085	0.4203	0.4761	0.4944	0.485	0.4828	0.4744
MAP	**10**	0.4385	0.4461	0.4524	**0.4575**	0.4524	0.2705	0.2819	0.2832	0.2842	0.2856	0.3368	0.3503	0.3555	0.3534	**0.357**
	20	0.4365	0.447	0.4455	0.4415	0.4384	0.2783	0.2815	0.2862	0.291	0.297	0.3296	0.3478	0.3546	0.3543	0.3542
	50	0.4288	0.4353	0.4308	0.4318	0.4265	0.2829	0.292	0.2962	**0.3034**	0.3016	0.3299	0.3313	0.339	0.336	0.3399
	100	0.4341	0.4324	0.4321	0.4225	0.4196	0.2802	0.2806	0.2905	0.2907	0.2913	0.3209	0.334	0.3322	0.3293	0.3212

Table 15 HPRF-2 (TFIDF) and PRF statistical significance test results using paired t-test with p-value < 0.05

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.3533	0.9555	0.4690	0.3031	0.4371	0.1137	0.1378	0.6398	0.1756	0.1400	0.0089	0.0319	0.0155	0.0062	0.0700
	20	0.6278	0.7548	0.9430	0.3392	0.1897	0.7897	0.6408	0.3146	0.4260	0.1907	0.1973	0.0464	0.0290	0.0184	0.0161
	50	0.6047	0.7101	0.5072	0.4249	0.8428	0.1080	0.6643	0.7985	0.2358	0.1040	0.5141	0.0436	0.0502	0.2980	0.1912
	100	0.3063	0.2623	0.3402	0.1595	0.1009	0.1228	0.1085	0.1429	0.3059	0.0963	0.5239	0.9998	0.9014	0.7160	0.6927
MAP	**10**	0.9810	0.9076	0.1817	0.0761	0.0545	0.1027	0.0133	0.1010	0.0597	0.0498	0.0054	0.0033	0.0176	0.0346	0.0275
	20	0.9823	0.5396	0.4301	0.4574	0.3388	0.1134	0.2059	0.0413	0.0251	0.0151	0.2757	0.0277	0.0046	0.0002	0.0007
	50	0.3086	0.9183	0.3671	0.6427	0.4538	0.0706	0.1137	0.0892	0.0066	0.1218	0.4063	0.0371	0.0494	0.0729	0.0335
	100	0.2311	0.5088	0.9802	0.3374	0.4142	0.0969	0.2385	0.0568	0.1492	0.1307	0.5402	0.2886	0.1983	0.1382	0.1620

Table 16 HPRF-2 (BM25) results for different number of feedback document and terms for TREC microblog dataset when α=0.2. Gray cells indicates statistically significant improvement over corresponding PRF configuration

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4680	0.4912	**0.5000**	0.4993	0.4980	0.3842	0.3938	0.4045	0.4136	0.4175	0.5467	**0.5546**	0.5456	0.5417	0.5367
	20	0.4558	0.4823	0.4850	0.4844	0.4871	0.3893	0.3994	0.4096	0.4153	0.4209	0.5050	0.5389	0.5317	0.5289	0.5267
	50	0.4605	0.4707	0.4721	0.4673	0.4680	0.4062	0.4090	0.4107	0.4277	**0.4339**	0.4956	0.4878	0.4939	0.4978	0.5000
	100	0.4592	0.4524	0.4599	0.4592	0.4592	0.3904	0.4034	0.4090	0.4085	0.4051	0.4672	0.4728	0.4700	0.4806	0.4794
MAP	**10**	0.4310	0.4393	**0.4587**	0.4566	0.4497	0.2655	0.2780	0.2814	0.2832	0.2864	0.3350	0.3421	0.3511	0.3544	**0.3584**
	20	0.4188	0.4391	0.4403	0.4383	0.4367	0.2734	0.2830	0.2898	0.2931	0.2948	0.3189	0.3417	0.3454	0.3464	0.3519
	50	0.4116	0.4246	0.4276	0.4325	0.4205	0.2794	0.2842	0.2908	0.2982	**0.3044**	0.3134	0.3151	0.3207	0.3283	0.3333
	100	0.4009	0.4106	0.4125	0.4147	0.4114	0.2672	0.2752	0.2812	0.2813	0.2891	0.3065	0.3065	0.3073	0.3117	0.3169

Table 17 HPRF-2 (BM25) and PRF statistical significance test results using paired t-test with p-value < 0.05

	n_t	2011					2012					2013				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.6966	0.1662	0.0253	0.1234	0.1038	0.0426	0.1498	0.2342	0.2274	0.1219	0.0444	0.0349	0.0272	0.0354	0.5096
	20	0.0449	0.8772	0.9152	0.1083	0.2267	0.7788	0.5938	0.2368	0.5335	0.2221	0.4810	0.1202	0.1610	0.1480	0.0990
	50	0.0484	0.2820	0.8911	0.7595	0.8655	0.3855	0.2051	0.3404	0.2830	0.1172	0.5275	0.7775	0.6719	0.0437	0.1188
	100	0.4824	0.0394	0.4605	0.1029	0.8086	0.0235	0.6739	0.5739	0.4519	0.5219	0.2830	0.8204	0.6187	0.7119	0.8764
MAP	**10**	0.3838	0.8469	0.0490	0.0459	0.0440	0.0793	0.0802	0.0696	0.0276	0.1295	0.1608	0.0311	0.0432	0.0333	0.0228
	20	0.6826	0.5265	0.7585	0.2338	0.1995	0.2858	0.0694	0.0017	0.0421	0.0231	0.8553	0.0265	0.0452	0.0211	0.0331
	50	0.1178	0.6340	0.7240	0.6318	0.7238	0.8686	0.8395	0.2549	0.1753	0.1096	0.9099	0.7928	0.5915	0.7930	0.1951
	100	0.3000	0.0462	0.5685	0.9000	0.8652	0.0163	0.5237	0.4808	0.4031	0.9830	0.3284	0.4657	0.3349	0.6276	0.8985

5.4.2.3. Hyperlinked-extended PRF (titles + meta-descriptions + meta-keywords)

Objective:

The objective of this set of experiments is to study the effect of our proposed hyperlink-extended pseudo relevance feedback (HPRF) on the performance of microblog retrieval by utilizing the web page titles, meta-descriptions, and meta-keywords of the hyperlinks embedded in the top relevant documents. In addition to, measuring to what extent the relevance feedback configuration regarding the number of feedback documents and terms can affect the retrieval performance.

Methodology:

In this set of experiments we apply our proposed HPRF approach by utilizing the web page titles, meta-descriptions and meta-keywords of the hyperlinks embedded in the top relevant documents "HPRF-3". We apply HPRF-3 under different configurations regarding the number of feedback documents {10, 20, 50, and 100} and the number of feedback terms {5, 10, 15, 20, and 30}. The steps to apply HPRF-3 is as following:

- Search the tweets collection using the original user query.
- Consider the top n_d documents relevant.
- Extract the web-page titles, meta-descriptions, and meta-keywords of the hyperlinks attached to the top relevant documents (tweets) and append them to their corresponding tweets.
- Select the top n_t term from the top relevant documents.
- Use the top selected terms in the previous step to expand the original user query.
- Search the tweets collection with the expanded query aiming better retrieval performance.

Results and Discussions:

Extending HPRF-2 with meta-keywords extracted from hyperlinked documents to relevant tweets slightly degraded the retrieval performance.

Table 18 and Table 20 reports full results of different configurations of HPRF-3 and highlights the runs that achieved statistically significant improvement over the corresponding configuration in PRF. Table 18 reports the full results by selecting the top n_t terms using the TFIDF weighting scheme, while Table 20 reports the full results by selecting the top n_t terms using the Okapi BM25 weighting scheme. Comparing results in Table 18 to that in Table 8, over 81% of the different configurations of PRF were improved using HPRF-3. In addition, significant improvements were achieved for most of the PRF configurations of TREC 2013 test set, and some of 2012 test set. The noticed better performance for 2013 test set over 2012 and 2011 may be a result of the percentage of broken hyperlinks of 2012 and 2011 was much higher than 2013 leading to less content to be added for query expansion, see Table 4 and Figure 27.

Significance test:

We applied statistical significance test using paired t-test with p-value < 0.05 to check the runs where HPRF-3 is performing significantly better than PRF. Table 19 and Table 21 shows the t-test results, where runs having p-value < 0.05 are considered significantly better than PRF using 2 different term weighting schemes, TFIDF and Okapi BM25 respectively. Compared to HPRF-2, HPRF-3 got a bit better significance results over PRF for 2012 evaluation set, but did slightly worse on 2013 evaluation set.

From the significance test results we can notice that using Okapi BM25 term weighting scheme leads to more consistent behavior across the three test sets when compared to TFIDF.

Table 18 HPRF-3 (TFIDF) results for different number of feedback document and terms for TREC microblog dataset when α=0.2. Gray cells indicates statistically significant improvement over corresponding PRF configuration

	n_t	**2011**					**2012**					**2013**				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4728	0.4878	0.4884	**0.4925**	0.4884	0.391	0.3983	0.4062	0.413	0.4113	0.5322	0.5483	0.5517	**0.553**	0.5383
	20	0.4741	0.4871	0.4857	0.483	0.4857	0.3944	0.3966	0.4045	0.4034	0.4203	0.51	0.5261	0.5172	0.5244	0.5222
	50	0.4782	0.483	0.4871	0.4816	0.4762	0.4079	0.4124	0.4147	0.4282	**0.4294**	0.5	0.505	0.4978	0.4983	0.4989
	100	0.4728	0.4741	0.4735	0.4714	0.4769	0.4051	0.4113	0.4068	0.413	0.4073	0.4728	0.4741	0.4735	0.4714	0.4769
MAP	**10**	0.4357	0.4508	0.4513	0.4566	**0.4573**	0.2757	0.2828	0.2871	0.2875	0.2853	0.3318	0.3515	**0.3566**	0.3513	0.3597
	20	0.4376	0.4482	0.4447	0.4404	0.4408	0.2785	0.2837	0.2877	0.2891	0.2964	0.3277	0.3468	0.3495	0.3526	0.3549
	50	0.4287	0.4382	0.4347	0.4344	0.424	0.2818	0.289	0.2939	**0.3031**	0.3024	0.3321	0.3356	0.3339	0.3384	0.3369
	100	0.4349	0.4297	0.4286	0.4235	0.4172	0.2807	0.2862	0.2901	0.2907	0.2853	0.4349	0.4297	0.4286	0.4235	0.4172

Table 19 HPRF-3 (TFIDF) and PRF statistical significance test results using paired t-test with p-value < 0.05

	n_t	**2011**					**2012**					**2013**				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.2544	0.5723	0.6780	0.4085	0.8869	0.0472	0.0788	0.5675	0.2728	0.3289	0.0056	0.0411	0.0302	0.0025	0.1466
	20	0.6443	0.8392	0.5926	0.7006	0.4040	0.4537	0.6409	0.2030	0.4029	0.4077	0.1833	0.0721	0.2364	0.0288	0.0288
	50	0.7670	0.6843	0.8087	0.8139	0.8003	0.0826	0.8534	0.9982	0.1638	0.1100	0.7717	0.0405	0.1546	0.2778	0.2983
	100	0.5841	0.6558	0.4345	0.1896	0.6173	0.2419	0.0411	0.3820	0.1263	0.9998	0.4206	0.9399	0.8804	0.8494	0.7975
MAP	**10**	0.7454	0.4909	0.1781	0.1022	0.0725	0.0247	0.0120	0.0772	0.0645	0.1034	0.0009	0.0025	0.0097	0.0022	0.0136
	20	0.9052	0.4692	0.5410	0.5386	0.2466	0.1134	0.1039	0.0149	0.0453	0.0206	0.3680	0.0231	0.0090	0.0003	0.0002
	50	0.3282	0.8359	0.7212	0.8971	0.3708	0.1387	0.2100	0.1143	0.0069	0.1115	0.2943	0.0157	0.2382	0.0861	0.1113
	100	0.1940	0.7645	0.7671	0.4810	0.3258	0.0565	0.0323	0.1225	0.2158	0.8220	0.7645	0.4350	0.1762	0.1923	0.1745

Table 20 HPRF-3 (BM25) results for different number of feedback document and terms for TREC microblog dataset when α=0.2. Gray cells indicates statistically significant improvement over corresponding PRF configuration

	n_t	**2011**					**2012**					**2013**				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.4667	0.4816	0.4898	**0.4939**	0.4925	0.3859	0.3977	0.4051	0.4079	0.4158	0.5406	0.5467	**0.5478**	0.5467	0.5367
	20	0.4558	0.4850	0.4735	0.4782	0.4837	0.3876	0.4011	0.4051	0.4136	0.4181	0.5067	0.5117	0.5244	0.5272	0.5217
	50	0.4619	0.4707	0.4762	0.4762	0.4755	0.3972	0.4113	0.4113	0.4237	**0.4311**	0.4956	0.4961	0.4933	0.5000	0.4983
	100	0.4551	0.4585	0.4673	0.4633	0.4653	0.3859	0.3972	0.4034	0.4124	0.4096	0.4661	0.4600	0.4622	0.4711	0.4622
MAP	**10**	0.4281	0.4381	0.4520	**0.4559**	0.4504	0.2695	0.2807	0.2836	0.2855	0.2859	0.3375	0.3441	0.3517	0.3554	**0.3570**
	20	0.4183	0.4409	0.4387	0.4403	0.4432	0.2714	0.2818	0.2887	0.2941	0.2928	0.3240	0.3356	0.3425	0.3479	0.3485
	50	0.4106	0.4208	0.4284	0.4309	0.4231	0.2755	0.2847	0.2887	0.2976	**0.3010**	0.3184	0.3196	0.3246	0.3246	0.3281
	100	0.3994	0.4043	0.4111	0.4162	0.4146	0.2663	0.2721	0.2764	0.2835	0.2877	0.3080	0.3057	0.3023	0.3085	0.3079

Table 21 HPRF-3 (BM25) and PRF statistical significance test results using paired t-test with p-value < 0.05

	n_t	**2011**					**2012**					**2013**				
	n_d	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**	**5**	**10**	**15**	**20**	**30**
P@30	**10**	0.5581	0.8850	0.3718	0.3566	0.4938	0.0507	0.1352	0.2425	0.5967	0.2112	0.0402	0.0379	0.0149	0.0435	0.6024
	20	0.0463	0.8886	0.2316	0.5920	0.4639	0.6518	0.4867	0.4873	0.6723	0.4888	0.7042	0.8735	0.2658	0.0456	0.1552
	50	0.1211	0.2717	0.5049	0.4067	0.5346	0.0964	0.3100	0.3929	0.3860	0.0757	0.5426	0.2704	0.7532	0.0313	0.2079
	100	0.2331	0.3898	0.8638	0.3668	0.6968	0.0113	0.2987	0.9092	0.2449	0.2360	0.2416	0.1319	0.1975	0.5713	0.1420
MAP	**10**	0.6510	0.9394	0.288	0.1003	0.1957	0.0773	0.1063	0.0777	0.0640	0.1721	0.0382	0.1961	0.021	0.0245	0.0462
	20	0.6382	0.3992	0.9431	0.2714	0.0393	0.5546	0.0892	0.0036	0.0150	0.0651	0.2463	0.1357	0.1408	0.0116	0.1275
	50	0.0346	0.3390	0.7653	0.8590	0.9388	0.4001	0.9518	0.0675	0.2141	0.1205	0.4227	0.5691	0.7907	0.7400	0.8083
	100	0.2848	0.0438	0.4340	0.9643	0.6289	0.0244	0.1987	0.2003	0.7061	0.7640	0.5107	0.4303	0.0794	0.3241	0.2312

5.5. Summary

Table 22 summarizes the best runs achieved by each query expansion approach for each test set measured by P@30 and MAP, and compares it to baseline and traditional PRF. As shown, the more content used in the PRF process, the better the results, except for using the meta-keywords. Using meta-keywords in the expansion process slightly degraded the retrieval performance. One interpretation for this behavior is that meta-keywords are sort of deprecated technique of adding relevant words for the web pages to help the search engine index the page; most of the web pages currently use the meta-description instead of meta-keywords.

HPRF-1 achieved improvements over PRF, but the number of statistically significant runs was not that much, while HPRF-2 achieved further improvements, which made the number of the statistically significantly runs over PRF higher for both P@30 and MAP. In fact, the best achieved result in Table 22 for 2011, and 2013 test sets outperformed the best reported automatic run in microblog track, which applied query expansion and results re-ranking based on large number of features [14], [47].

Moreover, from the results we can notice that using Okapi BM25 as the term weighting scheme achieved the best results. In addition, it leads to more consistent behavior over TFIDF regarding the number of results achieving significance improvements using our proposed approach over the traditional PRF.

Our results for hyperlink-extended PRF show that configuring PRF properly and utilizing content of hyperlinked-document in tweets effectively produce an efficient retrieval system that outperforms other sophisticated techniques such as learning-to-rank and document expansion. This shows the importance of properly tuning and extending query expansion. Nonetheless, our extended HPRF can be combined with other methods to achieve even better results.

Table 22 Best runs achieved by each query expansion method.

** and + indicate statistical significant improvement over baseline and PRF respectively*

	P@30			**MAP**		
	2011	**2012**	**2013**	**2011**	**2012**	**2013**
Baseline	0.4238	0.3565	0.4500	0.3882	0.2275	0.2524
PRF (TFIDF)	0.4939*	0.4147*	0.5178*	0.4452*	0.2925*	0.3421*
PRF (BM25)	0.4864*	0.4203*	0.5322*	0.4405*	0.2850*	0.3492*
HPRF-1 (TFIDF)	0.4980*	0.4249*	0.5483*	0.4558*	0.2994*	0.3498*
HPRF-1 (BM25)	0.4946*	0.4237*	0.5394*$^{+}$	0.4510*$^{+}$	0.2974*	0.3492*
HPRF-2 (TFIDF)	0.4959*	0.4311*	0.5544*$^{+}$	0.4575*	0.3034*	0.3570*$^{+}$
HPRF-2 (BM25)	**0.5000***$^{+}$	**0.4339***	**0.5546***$^{+}$	**0.4587***$^{+}$	**0.3044***	**0.3584***$^{+}$
HPRF-3 (TFIDF)	0.4925*	0.4294*	0.5530*	0.4573*	0.3031*	0.3566*
HPRF-3 (BM25)	0.4939*	0.4311*	0.5478*	0.4559*	0.3044*	0.3517*
TREC Best System	0.4551	0.4424	0.5528	0.3350	0.3186	0.3524

Chapter 6. CONCLUSION AND FUTURE WORK

Microblogs has been a hot topic for research in the recent years due to the wide spread of social microblogging platforms such as Facebook, Google+, and Twitter. In our work, we focus on Twitter, which is a fast growing microblog platform. Twitter has more than 200 million users, and around 340 million tweets published per day. In Twitter, a user shares his thoughts, opinions, and news in a short text named “Tweet” in less than 140 characters. The increased popularity and use of microblog services made the users start using it as a platform to satisfy their information need [2], [3]. As a result, a big attention has been directed to improve microblog retrieval to satisfy the growing information need of the microblog services users.

Due to the popularity of Twitter as a microblogging platform, TREC (Text REtrieval Conference) one of the big conferences in the field of Information Retrieval started a track (a competition) in the year 2011 to get the best microblog retrieval performance among all the participants. The Microblog track was successful and continued for the next two years (2012 and 2013), and they are still planning to continue it for the year 2014. TREC Microblog track provides datasets and evaluation sets that are considered a benchmark for microblog retrieval. A lot of research has been done on the datasets offered by the TREC microblog track to investigate the effectiveness of different retrieval approaches on the performance of microblog retrieval. The best performance achieved till now in TREC microblog track is less than 0.56 on 1.0 scale, which opens a big room for research to enhance the retrieval performance. All our experimentations are based on TREC datasets.

One of the major problems in Microblog retrieval is the severe vocabulary mismatch between short user queries, and the short documents we are searching in.

Various approaches have been investigated in the literature for better query document matching in microblogs to improve the retrieval performance. Two main approaches have been investigated in the previous work to overcome the vocabulary mismatch problem in microblog retrieval, namely query expansion and document expansion. For document expansion, the aim is to extend all the documents (tweets) in the tweets collection we are searching in with relevant terms that may help match short user queries. Another effective approach to overcome the vocabulary mismatch problem is query expansion. In query expansion, the main idea is to extend short user queries with terms that are relevant to the user information need and may lead to retrieve relevant documents that couldn't be retrieved using the original user query.

The objective of our work is to study the effect of query expansion on microblog retrieval performance. First, we do a comprehensive study on the traditional Pseudo Relevance Feedback (PRF), by analyzing its impact under different configurations and how changing the configuration impacts the retrieval performance. Then, we utilize the contents of hyperlinked documents attached to the top relevant search results and study the impact of using different types of information extracted form these hyperlinks on the retrieval performance.

TREC offered three different evaluation sets (2011, 2012, and 2013) against two tweets collections (2011, and 2013). In the year 2013, TREC made a public API based on the well-know open source Lucene web search engine to access the two tweets collection and do search using the up-to-date KL-divergence language modeling retrieval model. We use TREC API to do a first round retrieval by retrieving the most matching tweets to each query in the evaluation set, then extract the expansion terms using different techniques. Finally, we make a weighted combination of the expansion terms and the original user query to avoid diluting the original short query with the long expansion terms. We do the weighted combination using Lucene query language.

We did a set of experiments in favor of studying the effect of different configurations on the performance of the traditional PRF. In addition, we did another three sets of experiments to study the impact of our proposed hyperlink-extended pseudo relevance feedback (HPRF) by using different types of information from hyperlinks attached to top relevant documents. We utilize three different types of information extracted from hyperlinks in our HPRF method, the web page titles, meta-descriptions, and meta-keywords. For each HPRF set of experiments, we compare our results to the baseline retrieval model, and after applying traditional PRF. Moreover, we do a statistical significance test to show how our HPRF approach is consistently performing better than existing approaches.

Our experimental results on the three datasets of TREC microblog track confirmed the effectiveness of PRF with various kinds of configurations. We found that using less number of feedback documents (n_d=10) in the PRF process was more effective than using larger numbers for two of our test sets. It was shown that an effective configuration of PRF parameters could lead to superior results over baseline and outperform other effective approaches.

Moreover, we examined two different term weighting schemes namely TFIDF, and a more up to date weighting scheme Okapi BM25. We show that BM25 can lead to slightly better results regarding the best performing configuration. In addition, we show that regarding the number of runs that our proposed HPRF approach achieves significant improvements over PRF, BM25 performs more consistent than using TFIDF on the three evaluation sets. Results also suggest that the optimal weighting scheme between the original user query and the expansion terms is 4:1 ($\alpha = 0.2$) not as the typical weighting 1:1 used in literature.

Finally, results for our introduced HPRF method showed that extending feedback documents with additional content from hyperlinked-documents leads to improved results. We found that including the titles and meta-description of hyperlinked-documents to the feedback documents can lead to significant improvement over traditional PRF. On the other hand, we show that using meta-keywords slightly degrade the retrieval performance. Our achieved results (using query expansion only) outperformed the best-submitted runs in TREC microblog track for the years 2011, and 2013. Our best runs achieved 0.5000, and 0.5546 P@30 compared to 0.4551, and 0.5528 for the best run submitted on TREC, and 0.4587, and 0.3584 MAP compared to 0.3350, and 0.3524 for the best run submitted on TREC for the two evaluation sets 2011, and 2013 respectively.

Future Work

We achieved superior retrieval performance using only query expansion compared to other studies that applied query expansion on microblogs before. Moreover, our approach outperforms most of the more complicated systems implemented in the literature or gives comparable results with less complications, less computational cost and less dependency on other third party components. The good retrieval performance achieved using our proposed approach solely motivates potential further improvements when other retrieval techniques get integrated. This represents our direct future work.

Further improvements can applied to the HPRF method we proposed by utilizing different types of information extracted from the hyperlinked documents. For example, using the Google page rank for the hyperlinks attached to the top relevant documents as a measure for the importance of those hyperlinks / documents, in addition to, the page rank of the domain name of each hyperlink. In addition, we plan to investigate the failure point of the expansion technique to work on avoiding it and improve the HPRF method.

Other term weighting schemes can be explored, like Robertson Term Selection value [48]. Moreover, as we showed in the experimental results, the retrieval performance is so sensitive to the configuration of the relevance feedback process, regarding the number of documents and terms used. Dynamic selection for the number for terms and documents used in the expansion process can be explored to select the best configuration for each query on it's own without any manual intervention.

REFERENCES

[1] (2014, April) Wikipedia. [Online, Accessed 26 January 2013]. http://en.wikipedia.org/wiki/Micro-blogging

[2] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng, "WhyWe Twitter: Understanding Microblogging Usage and Communities," in *Joint 9th WEBKDD and 1st SNA-KDDWorkshop*, San Jose, California , USA, 2007.

[3] Dejin Zhao and Mary Beth Rosson, "How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work," in *GROUP'04*, Sanibel Island, Florida, USA, 2009.

[4] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi, "Searching microblogs: coping with sparsity and document quality," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 183-188, 2011.

[5] Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris, "# TwitterSearch: a comparison of microblog search and web search," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 35-44, 2011.

[6] Jaeho Choi and W Bruce Croft, "Temporal models for microblogs," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2491-2494, 2012.

[7] Tarek El-Ganainy, Zhongyu Wei, Walid Magdy, and Wei Gao, "QCRI at TREC 2013 Microblog Track," in *Proceedings of The Twenty-Second Text REtrieval Conference*, 2013.

[8] Feng Liang, Runwei Qiang, and Jianwu Yang, "Exploiting real-time information retrieval in the microblogosphere," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pp. 267-276, 2012.

[9] Kamran Massoudi, Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp, "Incorporating query expansion and quality indicators in searching microblog posts," in *Advances in Information Retrieval.*: Springer, pp. 362-367, 2011.

[10] Donald Metzler, Congxing Cai, and Eduard Hovy, "Structured event retrieval over microblog archives," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 646-655, 2012.

[11] Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara, "Combining recency and topic-dependent temporal variation for microblog search," in *Advances in Information Retrieval.*: Springer, pp. 331-343, 2013.

[12] Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara, "Improving pseudo-relevance feedback via tweet selection," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 439-448, 2013.

[13] Miles Efron, Peter Organisciak, and Katrina Fenlon, "Improving retrieval of short texts

through document expansion," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 911-920, 2012.

[14] Jimmy Lin and Miles Efron, "Overview of the TREC2013 microblog track," in *Proceedings of the Twenty-Second Text REtrieval Conference*, 2013.

[15] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff, "Overview of the TREC-2011 Microblog Track," in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[16] Ian Soboroff, Ounis Iadh, Craig Macdonald, and Jimmy Lin, "Overview of the TREC-2012 microblog track," in *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*, 2012.

[17] Richard McCreadie and Craig Macdonald, "Relevance in microblogs: enhancing tweet retrieval using hyperlinked documents," in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pp. 189-196, 2013.

[18] Gerard Salton, Anita Wong, and Chung-Shu Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.

[19] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford, "Okapi at TREC-3," *NIST SPECIAL PUBLICATION SP*, pp. 109-109, 1995.

[20] Martin F Porter, "An algorithm for suffix stripping," *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130-137, 1980.

[21] Chris Buckley, Gerard Salton, and James Allan, "Automatic retrieval with locality information using SMART," in *Proceedings of the First Text REtrieval Conference TREC-1*, pp. 59-72, 1993.

[22] Efthimis N. Efthimiadis, "Query Expansion," *Annual Review of Information Systems and Technology*, vol. 31, pp. 121-187, 1996.

[23] Donald Metzler and W. Bruce Croft, "Combining the Language Model and Inference Network Approaches to Retrieval," Amherst, MA, Preprint submitted to Information Processing & Management 2004.

[24] Chengxiang Zhai and John Lafferty, "A Study of Smoothing Methods for Language Models applied to Ad Hoc Information Retrieval," in *SIGIR*, New Orleans, Louisiana, USA, 2001.

[25] Statistics Tutorial. [Online, Accessed 26 January 2013]. http://www.gla.ac.uk/sums/users/jdbmcdonald/PrePost_TTest/pairedt1.html

[26] Zhongyuan Han et al., "Hit at trec 2012 microblog track," in *Proceedings of Text REtrieval Conference*, 2012.

[27] Thong Hoang Van Duc, Thomas Demeester, Johannes Deleu, Piet Demeester, and Chris Develder, "UGent participation in the Microblog Track 2012," in *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*, pp. 1-5, 2012.

[28] Chris Buckley, Gerard Salton, and James Allan, "The effect of adding relevance information in a relevance feedback environment," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 292-300, 1994.

[29] Feng Liang, Runwei Qiang, and Jianwu Yang, "Exploiting real-time information retrieval in the microblogosphere," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pp. 267-276, 2012.

[30] Donald Metzler and Congxing Cai, "USC/ISI at TREC 2011: Microblog Track," 2011.

[31] Yan Li et al., "PRIS at TREC2011 Micro-blog Track," 2011.

[32] Adam Roegiest and Gordon V. Cormack, "University of Waterloo at TREC 2011: Microblog Track," 2011.

[33] Dzung Hong, Qifang Wang, Dan Zhang, and Luo Si, "Query Expansion and Message-passing Algorithms for TREC Microblog Track," 2011.

[34] Sarvnaz Karimi, Jie Yin, and Paul Thomas, "Searching and Filtering Tweets: CSIRO at the TREC 2012 Microblog Track," , 2012.

[35] Younos Aboulnaga and Charles L Clarke, "Frequent Itemset Mining for Query Expansion in Microblog Ad-hoc Search," in *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*, 2012.

[36] Ayan Bandyopadhyay, Kripabandhu Ghosh, Prasenjit Majumder, and Mandar Mitra, "Query expansion for microblog retrieval," *International Journal of Web Science*, vol. 1, no. 4, pp. 368-380, 2012.

[37] Samuel Louvan et al., "University of Indonesia at TREC 2011 Microblog Track," 2011.

[38] Ahmed Saad El Din and Walid Magdy, "Web-based Pseudo Relevance Feedback for Microblog Retrieval," in *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*, 2012.

[39] Bolong Zhu et al., "ICTNET at microblog track trec 2012," in *Proceeding of the Twenty-First Text REtrieval Conference*, Gaithersburg, 2012.

[40] Sharon Small, Karl Appel, Lauren Mathews, and Darren Lim, "Siena's Twitter Information Retrieval System: The 2012 Microblog Track," in *Proceedings of the Twenty-First Text REtrieval Conference*, 2012.

[41] Hao Wu and Hui Fang, "Concept Detection and Using Concept in Ad-hoc of Microblog Search," in *Proceedings of the Twenty-First Text REtrieval Conference*, 2012.

[42] Jiayue Zhang et al., "PRIS at 2012 Microblog Track," in *Proceedings of the Twenty-First Text REtrieval Conference*, 2012.

[43] Paul Ferguson, Neil O'Hare, James Lanagan, and Alan F. Smeaton, "CLARITY at the TREC 2011 Microblog Track," 2011.

[44] Wei Gao, Zhongyu Wei, and Kam-Fai Wong, "Microblog Search and Filtering with Time Sensitive Feedback and Thresholding based on BM25," in *Proceedings of the*

Twenty-First Text REtrieval Conference (TREC 2012), 2012.

[45] Craig Willis, Richard Medlin, and Jaime Arguello, "Incorporating Temporal Information in Microblog Retrieval," in *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*.

[46] John Ferguson Smart. (2006, April) Web Mining Lab in UCLA. [Online]. http://oak.cs.ucla.edu/cs144/projects/lucene/

[47] Siming Zhu, Zhe Gao, Yajing Yuan, Hui Wang, and Guang Chen, "PRIS at TREC 2013 Microblog Track," in *Proceedings of The Twenty-Second Text REtrieval Conference*, 2013.

[48] Stephen E Robertson, "On term selection for query expansion," *Journal of documentation*, vol. 46, no. 4, pp. 359-364, 1990.

APPENDICIES

Appendix A: List of stop-words used

Table 23 Stop-words list

he'd
etc
they'll
ours
its
yourselves
until
becoming
over
although
formerly
interest
thereby
she
something
along
serious
alone
these
else
nowhere
once
how
under
became
he
theirs
always
few
further
system
he's
bottom
herself
own
twelve
each
wherever
itself
we
somewhere
same
enough
has
must
did not
who
amount
after
nevertheless
couldn't
who's
would
whereby
any
had
be
what's
however
get
whose
so
fify
behind
much
and
whereas
that
co
i'd
often
than
against
doing
whom
thence
several
seeming
due
i'm
towards
whether
elsewhere
beyond
whence
mill
of
are
shouldn't
can't
among
describe
empty
on
only
her
yourself
everyone
that's
itself
move
therein
thereafter
others
we're
sixty
or
done
did
again
wasn't
without
third
many
not
he'll
nor
haven't
anyhow
cant
anyway
almost
were
please
more
toward
his
inc
we'd
there's
when
someone
three
myself
none
everywhere
throughout
onto
except
isn't
such
hers
we've
whereafter
here
himself
herself
thick
whole
thin
even
latterly
this
perhaps
ever
call
other
have
becomes
one

go
computer
she's
i've
give
before
next
made
hasnt
namely
six
de
hereafter
when's
she'd
side
could
whenever
do
whither
mostly
may
noone
we'll
whatever
a
back
us
front
won't
un
seem
cannot
fire
up
i
eg
either
what
amoungst
does not
nothing
they'd
having
down
part

ltd
does
hereby
shan't
can
bill
about
former
well
through
re
fifteen
name
above
four
too
all
yours
top
five
thus
sincere
moreover
she'll
at
as
you
still
cry
forty
therefore
neither
never
which
anything
see
i'll
take
am
anyone
whoever
there
an
off
our
why

now
them
then
will
ought
myself
some
why's
upon
indeed
might
put
eleven
most
across
they've
seems
Himself
latter
where's
rather
me
mine
aren't
hence
don't
it's
already
should
my
wouldn't
whereupon
per
how's
beforehand
mustn't
sometime
within
thereupon
ten
but
you're
afterwards
last
meanwhile
sometimes

from
let's
while
was
because
another
during
full
if
seemed
ie
below
they're
wherein
weren't
find
between
less
with
those
is
it
ourselves
besides
your
fill
into
the
in
around
two
beside
twenty
themselves
their
also
first
couldnt
found

here's
yet
keep
to
con
detail
thru
both
anywhere
least
become
you'd
you'll
by

very
out
they
nobody
via
somehow
for
hereupon
you've
no
nine
otherwise
everything
hundred

herein
hadn't
being
amongst
show
since
him
where
every
together
though
eight
been
hasn't

Appendix B: PRF best run expansion examples

Table 24 Expansion terms extracted for the best PRF runs with respect to P@30

Topic ID	***Original***	***PRF***
MB01	bbc world service staff cuts	650 outlin foreign quarter languag major offic new close job
MB02	2022 fifa soccer	cup qatar world blatter sepp winter presid chang stage summer
MB03	haiti aristide return	haitian polit duvali jeremiah wright rev allow america door media
MB04	mexico drug war	clinton hillari judici essenti reform flag violenc consen broadli sketchi
MB05	nist computer security	cloud guidanc tackl public new guidelin wiki virtual standard govern
MB06	nsa	secur relationship 2m showtim gizmodo directli expo analyst booti deliv
MB07	pakistan diplomat arrest murder	lahor reuter court charg paki american detent kill extend held
MB08	phone hacking british politicians	prime minist gordon brown voicemail wrote fear summ polic tabloid
MB09	toyota recall	vehicl million 17 nearli 15m fuel said corp leak motor
MB10	egyptian protesters attack museum	looter mummi destroi thousand protect loot shield surround youth form
MB11	kubica crash	renault robert pace formula test set fastest valencia intent signal
MB12	assange nobel peace nomination	prize laureat win wikileak yunu imprison osama alongsid muhammad bishop
MB13	oprah winfrey half-sister	secret famili reveal announc ha knew exist big seemingli prompt
MB14	release of the rite	hopkin anthoni horror box offic film ap weymouth exorcist wham
MB15	thorpe return in 2012 olympics	london venu game shape teessid paralymp nimrod 547 580 goalcom
MB16	release of known and unknown	fear fly rsn cemetari goodwin precip rumsfeld 219 hi peopl
MB17	white stripes breakup	northern know light cotl gbm black stijl great thump icki
MB18	william and kate fax save the date	middleton princ royal wed ferguson marriag diana sarah bbc invit
MB19	cuomo budget cuts	spend medicaid gov andrew propos york ny new 1329 lakesuccessni
MB20	taco bell filling lawsuit	beef meat claim laist freudian starch inaccur thei oxymoron bogu
MB21	emanuel residency court rulings	rahm chicago mayor appeal ballot meet requir isnt break illinoi
MB22	healthcare law unconstitutional	judg florida declar thi reform feder obama becam presid rule
MB23	amtrak train service	derail new station york nogo peddl cascad chp

		loma kap
MB24	super bowl seats	stadium cowboi befor fan unsaf readi becaus deni dalla sent
MB25	tsa airport screening	privat program shut door utter chronicl opt ditch halt patriot
MB26	us unemployment	benefit firsttim claim number file american fall eas youth rate
MB27	reduce energy consumption	build expenditur help hvac 901 wai 2030 divert envelop curtain
MB28	detroit auto show	intern polic green scamp car shenzhen balmi precinct greensboro cameo
MB29	global warming and weather	climat cold caus denier chang thei tlga circumst colder gore
MB30	keith olbermann new job	msnbc current countdown tv home host becom flamm commen smal
MB31	special olympics athletes	compet game winter michigan stadium celebr state spenc athen thiev
MB32	state of the union and jobs	address presid 2011 celebrit cnncom video devot penn deli react
MB33	dog whisperer cesar millan techniques	train spaniel cocker sanaa caesar filipino 225 dilla watch passeng
MB34	msnbc rachel maddow	report shrill murbarak internet shakeup hanniti character thei yippi pundit
MB35	sargent shriver tributes	wa sarg mourner potomac ap bono optimist u2 globe recal
MB36	moscow airport bombing	suicid kill busiest injur russia blast 31 mondai peopl domodedovo
MB37	giffords recovery	gabriel rep doctor road week rocki sprint come dai fro
MB38	protests in jordan	thousand amman opposit demand step pm support albania algeria modest
MB39	egyptian curfew	impos egypt cairo deploi state 7am 6pm televis militari accord
MB40	beck attacks piven	glenn fox franc anchorman threat taunt defi death franci thi
MB41	obama birth certificate	limbaugh releas hawaii demand secret stai coverup guv abercrombi shield
MB42	holland iran envoy recall	1979 analysi crisi egypt malaria iranian dutch checkout awar rais
MB43	kucinich olive pit lawsuit	denni settl sue suit congression cafeteria ohio sandwich toothi usatodaycom
MB44	white house spokesman replaced	carnei biden secretari claim press jai comment rcmp madelein dalei
MB45	political campaigns and social media	pew network american 22 2010 market bowl onlin wwwcmstothemaxcom super
MB46	bottega veneta	bag fragranc snakeskin rodart rtw teal pricei turquois 2011 payn
MB47	organic farming requirements	violent path produc children fresh indonesia offer street learn rodal
MB48	egyptian evacuation	begin state plan egypt thei 286 katrina stark amid

		afp
MB49	carbon monoxide law	poison detector plugin famili amazoncom comet epa backup blown displai
MB51	british government cuts	cameron lockerbi spend signific tax thei rule advis exert uk major bomber thi afp previou
MB52	bedbug epidemic	bite let dont ar rid warn abat thei pesticid goodnight hotel insect fight citi got
MB53	river boat cruises	nile midnight vike thi valentin 12th feb europ wa dai cancel sat march travel citi
MB54	the daily	ipad new newspap murdoch launch corp app rupert unveil magazin digit azeroth free public paper
MB55	berries and weight loss	acai diet plan healthi adi supplement imp lose reduct cleans fruit new popular promot colon
MB56	hugo chavez	venezuelan boss presid venezuela thi sai offic oliv enemi caraca honest golf stone 99 2011
MB57	chicago blizzard	weather forecast cam thi warn snow watch channel brace web 2011 break pictur histor lake
MB58	fda approval of drugs	contrav weightloss deni loss weight wont ha declin orexigen diet astrazeneca preterm agenc thi oral
MB59	glen beck	coco thi glenn fox piven franc insan watch obama trend answer question ngag funni just
MB60	fishing guidebooks	thi nack blue red question cyberstalk nick commonli travel lure modif aquarium istanbul tropic cuba
MB61	hu jintao visit to the united states	china presid obama chines dinner ar thi meet washington pre relat power diplomat bureau beij
MB62	starbucks trenta cup	bottl wine entir hold new size bigger stomach coffe caffein rt venti huge drink ha
MB63	bieber and stewart trading places	jon kristen justin new thi daili video martha 911 rocki bodi actress twilight loui hous
MB64	red light cameras	redlight studi depot traffic fatal ar crash updat cut oceansid citi intersect speed driver save
MB65	michelle obama's obesity campaign	ladi childhood rate new weight crescent role militari atlanta insur respons major loss rise plai
MB66	journalists' treatment in egypt	attack beaten arrest detain foreign ar protest target new egyptian media guardian jazeera world cairo
MB67	boston celtics championship	laker nba vs thi marqui kendrick bruis pierc wa final daniel paul ar game angel
MB68	charlie sheen rehab	enter actor check goe half men voluntarili ha home intox hiatu caller tv said 911
MB69	high taxes	ar pai corpor incom rate thi increas thei return wai state ha bui unemploy alreadi
MB70	farmers markets opinions	thi somervil winter todai vendor ferri new pig argu check bother breath smell nobodi build
MB71	australian open djokovic vs. murray	novak andi final tenni ferrer feder 2011 win david semifin live men titl stream semi
MB72	kardashians opinions	kim khloe kourtnei new 115000 watch thi fashion

		style sister award khlo pictu face humphri
MB73	iran nuclear program	talk weapon power collaps ar lead offici fail intern new dev adva dprk attempt nanotechnolog
MB74	credit card debt	relief consolid settlement elimin option help settl wai solut best pai bad payment combin want
MB75	aguilera super bowl fail	christina anthem nation sing flub xlv fumbl botch 2011 line submit defend repeat lyric upload
MB76	celebrity dui violations	attornei seattl lawyer pressli guilti new jaim antitrust indepth counti arrest wa charg thei plead
MB77	ncis	watch glee nntn troubl angel 33 episod dai e13 tmnya obssess whoonga angkat2 ziva la
MB78	mcdonalds food	ronald fast hostag wa fastfood held hungri price liber think campaign armi im cost rais
MB79	saleh yemen overthrow	presid abdullah ali dictat thei govern leader egyptian protest egypt mubarak yemeni seek term wa
MB80	chipotle raid	flotilla isra priyanka legal probe thi tax priv incom israel aid hsu aspca katrina gaza
MB81	smartphone success	android nokia mobil new app blackberri comscor ha os monthli overal pass unveil chipmak market
MB82	illegal immigrant laws	new arizona state missouri suprem adopt ralli termin court rule right nebraska enforc raid face
MB83	stuxnet worm effects	iran chernobyl comput malwar expert claim sai boomerang secur investig duck reuter new warn confick
MB84	athlete concussions	teen savard rais multipl risk health crosbi bruin ar sidelin think ha thi helmet reduc
MB85	best buy improve sales	depart excus close thi estat home market suster real tip editor forecast skill guest cheap
MB86	joanna yeates murder	jo vincent tabak charg man newsom remand accus court william butler suspect 1939 polic sadden
MB87	chicken recipes	cutlet fri thi easi salad grill fettuccin delit fresh casserol food homemad curri new roast
MB88	kings' speech awards	guild sag win produc director best boardwalk empir nomin academi hooper firth reign oscar thi
MB89	supreme court cases	illinoi rahm emanuel rule ballot appeal thei issu adopt ha high hear decis chicago missouri
MB90	anti-bullying	bulli teen school bellow srilanka anti program launch ar parent video h8 psa legisl thi
MB91	michelle obama fashion	look ladi militari oprah state famili obam honor dress flap mr design winfrei power wear
MB92	stock market tutorial	dtn invest new trade price tip 2011 rise video ap todai report amid high dow
MB93	fashion week in nyc	thi new fashiolista york design notch cover friend 9th upcom polo princess ani collect start
MB94	horse race betting	monei tip balai best tipster make fantast free earn win sport lose todai anita thi
MB95	facebook privacy	german set deal congress user data ar mark 10 extrem new everi know zuckerberg need
MB96	sundance attendees	sxsw meetup fellow mashabl connect 2011 thi 20

		march registr 11 creativ forward come look
MB97	college student aid	financi grant approv loan barefoot tuition scholarship hightech appli 2311 educ new quick babysitt avg
MB98	australian floods	18fashiontoaid qld appeal launch fashion design 19 tax australia prime minist propos gillard flee julia
MB99	superbowl commercials	ar watch best wa dure thei better thi time funni justin pepsi dorito funniest quarter
MB100	republican national committee	hous chairman democrat budget thi senat polici debt ryan paul ar presid obama parti thei
MB101	natalie portman in black swan	wa oscar actriz nomine actress watch veri deserv kuni award star mila mejor amaz movi
MB102	school lunches	time eat im child todai obes haha advantag box pack dont foodz high bore againi
MB103	tea party caucus	senat rubio scalia thi republican speak billionair antonin marco 2012 congression question join caller meet
MB104	texting and driving	dont shouldnt thei phone danger cell argu think seth thi talk know ye work sai
MB105	the avengers	sevenfold captain america cast seiz rev thi hayden dear thor toxic compil dai marvel knight
MB106	steve jobs' health	appl care gate blockquot address destroy rhode compos creator healthcar alli eric consum insur tech
MB107	somalian piracy	pirat prais thi like look thei somali sai forehead dont doe boost lool polit ar
MB108	identity theft protection	cost infograph doe type crime secur center kind tip social varieti cyber scam resourc concern
MB109	gasland	oscar nomin fox screen josh industri frack seen environment thi environ documentari nod doc watch
MB110	economic trade sanctions	belaru iranian eu outlook iran polici expert toughen bilater dollar binari develop widen turkish currenc
MB111	water shortages	theme asharam @maudebarlow #buranamanoholihai groundwat @globalvoic defunct bapu scarciti face plummet @priyankachopra zua acut bangalor cove rango ha activist britain farmer thn iran spoke govt devic singapor domin lectur rage
MB112	florida derby 2013	hors race @aaronrayk #previewpredict ivl lousiana @gulfstreampark gulfstream ani virtuou #odd #horserac orb #ironi jockei ballot 100000 merit cougar louisiana drawn alejandro #sport prep kentucki sampl casino predict 64 rout
MB113	kal penn	actor #lovenewgirl @westernu @kalpenn tv eurotrip #greatmindsthinkalik malh sunil @newgirlonfox #ldnont kumar usc yaar raffl xma #twin olympu summari harold ha fascin western

		pirat tmrw fallen geniu escap cam india
MB114	detroit efm undemocratic	financi support convict council appoint #citycouncil @1mikesteel #letitrip citi #backchannel @mhpshow confus #nerdland lmaoooooooooo kwame snyder #mi #detroit loom freewai verdict baptist takeov coon whyi activist pledg coincid minist recogn
MB115	memories of mr. rogers	neighborhood birthdai 35 handknit fact wa tbqh 85th neighbourhood nsfw devour sanctuari hi happi guidanc todai dow circu worn pleas nod sweetest sweater childhood unfortun b4 85 wouldv steve just
MB116	chinese computer attacks	hacker suspect post washington #securityguard report 100plu cyberattack wapo sophist persist cyber militari target york blog month ti new time ar
MB117	marshmallow peeps dioramas	librari contest chick fragonard @ydrcom make easter teikoku #giggl extravaganza 2013 #contest rotten dessert annual 700 bunni dc thi paint skip center scene alma studio egg winner public uk creat
MB118	israel and turkey reconcile	ankara restor ap relat apolog erdogan diplomat jerusalem minist ti prime suggest @algemein quick obama surpris rej frontpag zionism agre brothel latvia rapproch hungari supp broker kil compen switzerland austria
MB119	colony collapse disorder	pesticid bee @huffpostgreen research rise death point backup robot earthtechl wnc plan huffpost ccd unusu rescu larg term happen right
MB120	argentina's inflation	censur supermarket freez price imf data twomonth nation report becom #imf impos bearish overrun #econom inaccur blatantli @theeconomist goldman indec pluck spiral curb bloomberg soar halt hyper percent comp econom
MB121	future of moocs	higher shape educ linkedin flexibilti @hastac @gerrycanavan scoopit #eli2013 edudem upenn convers #mooc tuft #yam rebuk spook dropout mook nuke napoleon aprendiendo moscow hybrid academ structur evolut mock graphic experienc
MB122	unsuccessful kickstarter applicants	job #hatethemal @sherlockmr @indiaknight endeavour bcu rabak workplac ha nottingham time feedback tgh #thank #fuck useless unfortun sia tk sedih mcdonald contact pope receiv review email cheer futur ugh luck
MB123	solar flare	biplan geomagnet intens @smh erupt cme unleash coal disrupt hippi fic longlast incom spit halo 5th wassup massiv blunt warn wave storm activ chip wind earth sun bc radio dure
MB124	celebrity dui	attornei counti #troubl king #celebr seattl defens mirandakerr nht bulin temecula #sober nimen

		britton @stlouisblu xiang hous bellevu hin sox reliv nvr pitcher bloom closest behavior shi lawyer phoenix orlando
MB125	oscars snub affleck	director ben award argo best win dga guild hi 65th #argo #gossip triumph winni #celebr sweep new ceremoni categori nod despit shrug jare speech exclu wee thi entertain honor ridicul
MB126	pitbull rapper	lohan lindsai lawsuit ajasa lose sarkodi dismiss rb dont africa #grammi battl wayn court terribl suit fail pop worst becom sing smh isnt yall rememb world everyon best shit realli
MB127	hagel nomination filibustered	gop senat republican 7127 cruis vote repub @cnnbrk stall wage #tcot schedul ship clear tuesdai instead cover read break end wai wa
MB128	buying clothes online	store shop contain amadiu neex stuff #magaluf erri sheer directori trendi psych accessori phoenix nowadai korean topic cheap guid xoxo embarrass secur articl im fashion junior carri design safe tip
MB129	angry birds cartoon	rovio seri toon march premier game bow deliv anim episod releas month 16 tv fan helsinki #indian #suryarai @suryarai bundl #game maker reuter consum introduc debut quest appear entertain launch
MB130	lawyer jokes	@bigboyv86 highfiv cement prosecutor skunk copper need regul wire penni loll wa sarcast besti invent partner neck hall aha ly loud offic road truth ahh present swear wasnt send differ
MB131	trash the dress	shoot bride cibolo boudoir garner 151 aniston receipt intuit trashi spiel demonstr hawaii #photographi iam coolest concept ar dont nana pub love director #oscar fab submit repeat wet carri gorgeou
MB132	asteroid hits russia	meteor metor fanpic videozapi prelud soviet rk meteorit discoveri 7pm unreal giant appear channel earth random ago stupid minut 14 die happen someth tonight video alwai realli feel look todai
MB133	cruise ship safety	drill fiv result death gone wrong crew kill member lifeboat canari di #travel emerg abc island fell dure new
MB134	the middle tv show	malcolm fridg empti thing slow kaburutz nutan haha internet cut power season annoi #faceoff ott #notfair best sickest fairli thei rapid randomli ot indian africa target batteri ea east cancel
MB135	big dog terminator robot	militari darpa build bootleg mow concret dynam recov musik slip boston bsa cloud brain web block throw ic googl youtub futur lama jd pick goe main act isnt walk real
MB136	gone girl reviews	book post breaker wild @conquerdg new #goodbook spring #shakeitup #barbado #marri

		korin sociopath 1939 gillian marci #dad sistar19 yor flynn #drink harmoni tokyo forum kpop civil ps3 sunshin request itll
MB137	cause of the super bowl blackout	wacki flurri xlvii light quarter produc beyonc dure sundai went cbss disarrai #wsj freakout ligh confus ticker twitter cbc abnorm momentum youngest thrown cnn document gop broadcast breaker ot polici
MB138	new york city soda ban blocked	bloomberg vow reuter @stfudustin judg mayor @badkidandrew fight newsnew caprici arbitrari sugari mayo bbc larg court sale drink
MB139	artists against fracking	minidoc yoko ono gimm @nygovcuomo yorker truth dont health protect york ha mother @simplynonna join @lavenderblue27 @youtub tell @yokoono pennsylvania occupi new lennon @sharethi common sean america present sing dream
MB140	richard iii burial dispute	cathol minster academ led research york reburi epetit anglican odon 20000 @fuckitsdustin leicest sai imo reckon christina depart sho @badkidandrew edg remain nearli cheap govern given ala opinion receiv poor
MB141	mila kunis in oz movie	becaus new wizard wanna #favoriteactress @andrewdolphin11 caus gabbana strictli dolc neil prep breast mysteri just press wednesdai gorgeou secret #oomf worth cuz fake fact wasnt busi date wear big week
MB142	iranian weapons to syria	maliki stop revolutionari kerri iraq iran command nuclear israel guard seek urg leader carri @lydisdeck #voanew sai overflight seckerri offici power @globalpost kill airstrik 35th censur shipment beirut clerk provinc
MB143	maracana stadium problems	threaten worker strike brazil fear grow fifa england cup refus emblemat valck readi #worldcup world eurosport fret jerom fiasco secretari #footbal indoor #sport construct brazilian clash insist properti dou despit
MB144	downton abbey actor turnover	anoth exit froggatt @msntv join season onlin lose hunki heartthrob @ew joann obrien reportedli maid recap goddamn spoiler elli forum miranda anna bate 2pm alert british shock hero latest tom
MB145	national parks sequestered	servic fewer ranger spring mean cut impact affect @natlparkservic @suryaray3 yosemit #surya avert @nprnew 100m #suryarai @suryarai @usatodai minim econom ev prepar forc public tout earli march isnt make
MB146	gmo labeling	groundbreak grassroot campaign washington launch state food consum store montvil wa confus @signon monsanto mandatori connecticut 2018 impli supplier ingredi illinoi pleas prop groceri

		declar initi ct patch contain solv
MB147	victoria's secret commercial	calvin kline onset thi panten twe kerr moistur barb got soap electron klein miranda ar reward weve model mark beer ton shoot alright worth 33 deal oo goe isnt car
MB148	cyprus bailout protests	ralli @ijreview journal youth independ america review eu #genel subm vote roil @ajel artemi nicosia jazeera troika amid accompani look chairman parliament presidenti controversi estim good palac 1500 packag andrea
MB149	making football safer	goodel nfl sport seahawk roger youth usa q13 #seahawk brute #seattl embassi look #tgdn #justsayin chapel #footbal border barri violent pleas nathan nc #tcot goo michel yahoo fox hill leagu
MB150	uk wine industry	#wine growth rais viticultur #vawin lasco accoun divers 11 expl growler #rva hous underscor #gadget perc output thei scholarship sweden dec grape #food virginia foundat economi concept adopt interact canadian
MB151	gun advocates are corrupt	nra @nicolejpearc control @havraha guncontrol @1phd @memomo agst flabbergast #gunskillpeopl #memphi sioux libertarian cbi #ow capitol maim lawmak enforc #p2 lobbi iowa strongest dozen violent rel flood gop govt liber
MB152	iceland fbi wikileaks	refus investig minist aid deni help kick thedailywhat nationa #ap #wikileak #technolog interior agent arriv london notic ap fridai #rt said try right thei
MB153	lighter bail for pistorius	african south oscar paralymp sprinter lawyer @youranonnew hear battl condit argu dai track star restrict charg plea arriv #new second
MB154	anti-aging resveratrol	ag coq10 spritz review slow formul red fountain ingredi youth wine effect lnv unfledg studi support endometriosi promis suffix juven champaign sinclair cognit substanc hormon unemploi pup globe oe grape
MB155	obama reaction to syrian chemical weapons	rebel toppl sai activi aleppo govern trade activist alass syria bashar presid media state fight #folyb killi moscow reportedli cla ministri russian russia foreign shadow israel defend fm themselv claim
MB156	bush's dog dies	barnei georg 12 rip pass dead sad becaus awai rustl caus @politico chenei condol shotgun wound regardless ws abt laura cam deni itun involv 43 ont #new web america announc
MB157	kardashian maternity style	kim stylist pregnanc fascin curv outfit explain formichetti cover complet black #bestweekev blogg littl peplum trou evolut time nicola pump defend pregnant fashion experi continu fix awkward favorit post photo

MB158	hush puppies meal	cardddd @99poni #hush ericson bearcat collegi hse #sub clog cincinnati ctfuu perci deaf mhm ankl hmu pig bastard youuu boot spirit slut mf mall shop men okai didnt said wish
MB159	circular economy initiatives	resourc @fastcoexist compani brandl creat #wastenot rapanui cloth epr agre holist thorough start examin crunch concept bullet tend silver benefit achiev china afraid truli wast kei report short run veri
MB160	social media as educational tool	techniqu strategi new todai strateg enhanc brief monitor essenti object #twitter corpor roi speaker abil engag intellig hashtag lover user brand offer product market award internet march differ isnt 14
MB161	3d printing for science	3dprint launch scientist pirat world bai newscientist offer gun ha new opensourc plan nanotechnolog embryon @gizmodo cornel man 4d #creativ #geek sculptur replica doodl @stfudustin stem printer fabric percent skull
MB162	dprk nuclear test	defens condemn conduct resolut ministri korean confirm @yourtowndal @officialdal @omarwaraich pass xinhua #dale @newspin denounc hous @zerohedg sanction sina successfulli new underground republ democrat tension certainli germani russia slam region
MB163	virtual currencies regulation	bitcoin alltim amazon coin treasuri insist govern announc rais fincen hackl high final administ #watch #tip margin guidanc summari new unnecessari comp revolut exchang kindl introduc major tuesdai note step
MB164	lindsey vonn sidelined	tiger wood date ar medallist #livewireathlet caroli airlift #livewir skier thrust sai intrigu fearless fro psycholog recov weekli olymp curiou #beauti associ led injuri ski suffer condit crash pregnant articl
MB165	acpt crossword tournament	ibadat @anggaraditya94 @xiyuanaweshum @paponmus uwesss requst subhi allah laaaaaaaaa mai ameen jul wechat magrib puzzl hav pend coke yesss prayer plz sent tp facebook lah tak hahaha need time
MB166	maryland casino table games	lotteri blackjack debut live result defens gqw fvw mjx april hindmost perryvil denouement imman arundel physician roulett dealt contemporari gae gravi imi poker newest hollywood progress attend launch crap prove
MB167	sequestration opinions	gop cut peo blame word mean someth onli automat ha choos avella layoff #sequestr dreamwork @politico depa rah 350 mount feder budget process fox known obama middl david tea anim

MB168	us behind chaevez cancer	rosi identifi breast firm surviv leader claim myeloid report special canc enzym leukemia mutat variat acut @bloombergnew 20000 newark mai genet #chavez beverag bloomberg drinker #u mason #venezuela percent reuter
MB169	honey boo boo girl scout cookies	june kept sale mama campaign shut word facebook sell ha onlin #honeybooboo gist aft fiasco 225 @tmz fresco trash badg sold wor bag remind pm report fuckin suck check boi
MB170	tony mendez	argo cia spy real #argo hispan latino aka @jdhawaii20 im screenplai memoir affleck autograph adapt wire meow geek injuri cast ty 60 base histori grow stori ben bro seriou super

Appendix C: HPRF-1 best run expansion examples

Table 25 Expansion terms extracted for the best HPRF-1 runs with respect to P@30

Topic ID	*Original*	*HPRF-1*
MB01	bbc world service staff cuts	outlin languag 650 caribbean statement quarter major shut lose new close job guyana shed brief qu irish foreign fund confirm
MB02	2022 fifa soccer	cup world qatar winter blatter sepp plan russia2018rusia stage presid eurosport summer sport digest chang plenti held pen discuss ahead
MB03	haiti aristide return	jeremiah haitian wright rev duvali polit want allow america door new media okin open pou passport democrat affair wa reuter
MB04	mexico drug war	clinton hillari judici essenti reform flag violenc risk heat legal consen broadli overhaul 2011 lethal patrol surg lawsuit secretari border
MB05	nist computer security	cloud guidanc tackl public new technolog virtual informationweek inform govern guidelin threat wiki standard address advic provid connect challeng program
MB06	nsa	secur relationship analyst global appl head watchdog googl encount gizmodo booti adult benefit buddi secret group date question sex fun
MB07	pakistan diplomat arrest murder	court detent extend held lahor charg reuter doubl paki american detain kill mount illeg pressur judg block releas order hand
MB08	phone hacking british politicians	scandal prime minist tabloid gordon summ brown amid voicemail dismiss editor sue wrote famou fear sourc polic mail phonehack everi
MB09	toyota recall	vehicl million 17 new fuel nearli global 15m car said corp leak abc involv motor wa reuter announc uk 245000
MB10	egyptian protesters attack museum	looter mummi destroi offici artifact crackdown loot shield sweep stolen revolt defend form cairo human possibl nation order mubarak sad
MB11	kubica crash	renault pace formula test set robert valencia new intent signal f1 surgeri involv seriou grandprix face campa r31 underlin eurosport
MB12	assange nobel peace nomination	prize laureat wikileak founder china winner visit famili schwab yunu imprison entrepreneurship alongsid muhammad elbaradei summit foundat scienc professor wife
MB13	oprah winfrey half-sister	secret famili reveal ha big shock announc sister share marque seemingli prompt sai reunion buzz stage file th surpris promis
MB14	release of the rite	hopkin anthoni box offic horror film ap weymouth exorcist usa review movi new
MB15	thorpe return in 2012 olympics	london venu shape nimrod scrap stadium plane fear latest new updat teessid goalcom gazett playbook 550 tottenham postpon 201 propos
MB16	release of known and unknown	precip rumsfeld 219 memoir nebraska breezi began donald reflect gd mp region cultur island central north airport entertain ar pop

MB17	white stripes breakup	zebra northern announc offici light rttnew black great realtim certifi documentari forex econom trailer broken gold whatev quit break busi
MB18	william and kate fax save the date	middleton princ royal ferguson wed sarah invit lookalik shortag marriag contributornetwork testino hit telegraphcouk epidem outcast greyson com new diana
MB19	cuomo budget cuts	spend medicaid lakesuccessni nanotechnolog new york sham ny assess governor revers expos gov andrew propos effort trick seek dirti past
MB20	taco bell filling lawsuit	beef meat expos ground claim opinion suit ignor laist consumerist mean realli forb shortli msnbc new motion fals lean guilti
MB21	emanuel residency court rulings	rahm chicago mayor ballot appeal meet requir run doe relat break chicagotribun huffingtonpostcom suntim post cour illinoi spy econom associ
MB22	healthcare law unconstitutional	judg feder rule declar florida void struck sai thi barack reform becam reuter surpris presid health obama mondai care good
MB23	amtrak train service	derail station york new chp loma collid vox encount penn car eagl strike victoria rare oil attack south outsid dead
MB24	super bowl seats	fan deni xlv stadium cowboi befor stairwel dalla unsaf fiasco readi becaus root 400 troubl latest sent wasnt fail home
MB25	tsa airport screening	privat program shut door opt ditch longer test hartsfield utter chronicl halt patriot new tf atl cnn travel bs updat
MB26	us unemployment	claim benefit fall firsttim number initi 000 youth weekli 42 tonga file american poverti tennesse eas concern rate skew unchang
MB27	reduce energy consumption	transport 75 account build hvac 901 specifi 2030 envelop effici emploi peak construct resourc consult reader equip written standard electr
MB28	detroit auto show	green cameo car chrysler hornet bigger wa intern polic make number 2010 stai everi shenzhen final mean precinct wai greensboro
MB29	global warming and weather	bizarr whale crop destroi updat right chang
MB30	keith olbermann new job	current msnbc tv home becom countdown regret host announc updat flamm commen smal alec baldwin pseudo todai tucson ha gore
MB31	special olympics athletes	winter compet michigan game celebr state thiev stadium row target tar central prepar smart gold energi heel host track drink
MB32	state of the union and jobs	2011 address presid video student obama ipad celebrit cnncom pennsylvania devot penn deli time react barack peak liber philli rat
MB33	dog whisperer cesar millan techniques	dilla watch rj trailer tag train spaniel tv world cocker check sanaa behavior medicin bite crash style anim group articl
MB34	msnbc rachel maddow	olbermann hoax internet idiot prayer union speech state market shakeup free character video spoof russo make favr sal rodger consult
MB35	sargent shriver tributes	bono u2 mourner sarg potomac wa ap optimist buri recal funer ideal chariti cape wife capet cofound breitbart tmcnet grandchildren

MB36	moscow airport bombing	suicid busiest domodedovo russia kill deadli terrorist injur blast 35 moment airpo hit wit deton slashdot video aftermath peopl 145
MB37	giffords recovery	road doctor rocki rep gabriel face long week sprint come dai sullivan fro complic despit injuri ahead brain continu li
MB38	protests in jordan	thousand demand step pm reform opposit support albania amman algeria modest erupt redempt uncomfort yemen belov stre activist chant aljazeera
MB39	egyptian curfew	impos unparallel deploi militari expert level cairo protest 0700 egypt 1800 unrest riot hosni author east middl local presid jan
MB40	beck attacks piven	glenn fox franc anchorman taunt threat defi death depict franci thi controversi fals fran scari target york act new hurt
MB41	obama birth certificate	secret stai limbaugh releas bookmark autom hawaii demand social coverup guv abercrombi shield governor etern privaci mysteri egg yahoo law
MB42	holland iran envoy recall	1979 analysi crisi egypt malaria iranian dutch checkout hosni awar row rais mubarak media social sahra oklahoman iaea wisner newscomau
MB43	kucinich olive pit lawsuit	denni settl congression sue cafeteria suit sandwich case ohio rep hi file toothi memeorandum su usatodaycom 150k 150000 pfft incid
MB44	white house spokesman replaced	madelein mccann biden secretari claim press sourc obama ap pick famili rcmp carnei clarenc mitchel complaint moscow brutal suspend terror
MB45	political campaigns and social media	pew network american 22 2010 market bowl onlin super twitter unleash weigh dii impact launch line featur ad pastrana tv
MB46	bottega veneta	fragranc bag rodart muse palm beyonc ell daughter beach fashion design snakeskin beauti chae face fan pewter rtw teal paltrow
MB47	organic farming requirements	path indonesia offer street learn kid job fig garcia violent root restaur children china fresh dinner welcom iask food rodal
MB48	egyptian evacuation	begin threat egypt ohio campu bomb forc colleg american state 286 stark amid afp globe victim flood aussi washington warn
MB49	carbon monoxide law	poison detector plugin amazoncom comet epa backup blown displai suffer propos alarm alert batteri repeat danger requir avoid effect digit
MB51	british government cuts	lockerbi cameron bomber tax spend advis major blast signific minist case sai handl rule exert
MB52	bedbug epidemic	twitter abat rid citi warn pest specialist suspend hotel crisi provinc ar help account vital
MB53	river boat cruises	travel midnight nile 12th feb sat vike thi valentin danub rhine wa demian cancel world
MB54	the daily	ipad new newspap corp unveil launch murdoch rupert app appl free store magazin public digit
MB55	berries and weight loss	acai diet adi plan healthi new reduct live health cleans lose supplement cardiovascular antioxid imp
MB56	hugo chavez	venezuelan venezuela presid boss power new enemi golf zinfandel 12 kecam year alyssa milano amerika

MB57	chicago blizzard	forecast 2011 cam weather snow warn thi resourc inform watch websit channel web pictur sale
MB58	fda approval of drugs	contrav weight deni weightloss loss orexigen preterm diet declin wont ha won astrazeneca reduc agenc
MB59	glen beck	glenn piven franc fox godspe ridg pagan gala grate week scari target letter polit support
MB60	fishing guidebooks	aquarium travel thi question cyberstalk fsta cch guid linki commonli lagoon lure 2011 modif automot
MB61	hu jintao visit to the united states	china chines presid obama diplomat beij meet washington success thi american dinner tcl wasteland turbul
MB62	starbucks trenta cup	bottl wine entir hold new terrifi huge size gothamist squid wino stomach american rt laugh
MB63	bieber and stewart trading places	jon kristen justin new rocki bodi ian white board snow teen video ar martha 911
MB64	red light cameras	studi redlight fatal crash traffic depot cut updat oceansid save ar speed driver live citi
MB65	michelle obama's obesity campaign	ladi atlanta childhood weight rate role new crescent militari oprah come loss fat polici plai
MB66	journalists' treatment in egypt	attack beaten protest detain guardian foreign new arrest target egyptian clinton ordeal media alongsid condemn
MB67	boston celtics championship	nba marqui bruis kendrick daniel vs perkin spine laker collis ha ormond doc river final
MB68	charlie sheen rehab	check enter actor hiatu home tv goe exclus half month men intox hospit caller 911
MB69	high taxes	ginorm court defer return truste parcel offer bombai amnesti ballot dodg illinoi vodafon climat new
MB70	farmers markets opinions	winter somervil vendor thi park roadrunn homestead derri morph video molto adelaid recreat digest vimeo
MB71	australian open djokovic vs. murray	novak andi tenni final ferrer 2011 live david feder stream semifin roger titl ved ap
MB72	kardashians opinions	kim kourtnei khloe vulgar khlo 115000 humphri extravag new fashion unisex odom sister fragranc savag
MB73	iran nuclear program	talk collaps power nuke fail weapon expo progress concern israel languag wa featur new doubt
MB74	credit card debt	relief consolid settlement option elimin settl pai help wai solut best bad payment combin rid
MB75	aguilera super bowl fail	christina anthem nation xlv sing flub 2011 fumbl repeat line botch video rehears singer nail
MB76	celebrity dui violations	seattl guilti attornei lawyer antitrust pressli plead jaim bucki badger new tweetmem directori meme kristoff
MB77	ncis	regulatori effort e13 intonow help start redirect watch cage freedom episod season onlin free
MB78	mcdonalds food	ronald hostag held price campaign fastfood fast spoof rais liber armi cost certainti colonel chill
MB79	saleh yemen overthrow	presid abdullah ali step 2013 yemeni protest leader ouster seek readout rachman term reelect mlg
MB80	chipotle raid	flotilla isra priyanka legal gaza probe israel katrina aid panel regrett kaif chopra deadli immigr
MB81	smartphone success	android comscor nokia overal app infineon new profit pass io subscrib tablet os blackberri refund
MB82	illegal immigrant laws	missouri adopt termin court state rule new face right arizona ralli tough enforc fiscal raid

MB83	stuxnet worm effects	expert chernobyl boomerang duck secur warn iran russia comput sai claim new caus malwar reuter
MB84	athlete concussions	teen rais multipl crosbi risk health bruin savard vonn footbal conscious sidelin lawmak new suffer
MB85	best buy improve sales	depart excus close thi home forecast tip estat surpass skill market beat suster increas ar
MB86	joanna yeates murder	jo tabak newsom vincent accus remand charg man court sadden deepli neighbour funer chang smith
MB87	chicken recipes	cutlet fri salad easi fettuccin fresh thi homemad curri new roast anis sauc melang thai
MB88	kings' speech awards	guild sag director win produc hooper boardwalk empir reign oscar tom 2011 best dga firth
MB89	supreme court cases	illinoi emanuel rahm ballot rule appeal hear adopt missouri immigr illeg termin decis chicago high
MB90	anti-bullying	bulli teen school nj fckh8 anti new program creat parent video h8 psa help thu
MB91	michelle obama fashion	militari state ladi oprah flap famili design honor dress dinner power stylelist horyn flotu obam
MB92	stock market tutorial	dtn trade price tip chart invest 2011 new dow ap emini 2007 statoil todai concern
MB93	fashion week in nyc	fashiolista new york hotspot thi cover aritzia pari win frappuccino gainesvil menswear copenhagen fly trip
MB94	horse race betting	request problem link tip monei balai thi best sandown todai tipster profit make fantast burch
MB95	facebook privacy	set congress user german deal 10 everi know need data new mark roundup ar zuckerberg
MB96	sundance attendees	sxsw meetup fellow mashabl connect 2011 march onstar 11 20 creativ widow review retreat webinar
MB97	college student aid	financi loan certif cours appli scholarship univers onlin injur approv new barefoot babysitt abroad internship
MB98	australian floods	18fashiontoaid qld appeal launch fashion design 19 tax propos australia prime minist damag gillard victim
MB99	superbowl commercials	volkswagen forc automot blitz thi 2011 feb2 busi 350k ad chevi 000 fave favourit popular
MB100	republican national committee	hous chairman budget unit senat obama infrastructur convent transport upcom urg spend tour thi studi
MB101	natalie portman in black swan	star oscar nomine dga radiant celebr mcl opera boost prais pregnant royal director wa cinema
MB102	school lunches	child obes advantag parslei fingerprint pack berkelei suzann bakeri rose cake bui tonni eat 0230
MB103	tea party caucus	senat rubio scalia marco 2012 join address billionair question republican antonin meet contour ar converg
MB104	texting and driving	seth work cell shouldnt phone argu talk blog ye godin thi distract twitter dog facebool
MB105	the avengers	captain cast america sevenfold comic smulder prequel cobi spot toxic drawn tv hardest forex escap
MB106	steve jobs' health	appl destroy compos creator care product blockquot hipaa address tasteless rhode apologis schmidt gifford absenc
MB107	somalian piracy	prais thi pirat boost anim ndtvcom sale internet wotn bighead oxfam rampant maritim bittorr criticis
MB108	identity theft protection	cost infograph doe type crime secur kind tip social varieti cyber scam resourc concern warn
MB109	gasland	oscar nomin industri fox screen 1921 gassi shale

		barnett frack seam examinercom refuge filmmak coal
MB110	economic trade sanctions	belaru iranian outlook iran polici expert eu toughen bilater develop widen turkish currenc new appeal
MB111	water shortages	scarciti acut farmer iran rage global sh secur forc @maudebarlow
MB112	florida derby 2013	gulfstream orb race cougar casino odd #previewpredict hors park ivl
MB113	kal penn	yaar summari parti movi malh sunil uwo #ldnont raffl fascin
MB114	detroit efm undemocratic	financi appoint snyder loom baptist takeov support convict pledg minist
MB115	memories of mr. rogers	35 fact birthdai devour sanctuari happi circu neighborhood ipod download
MB116	chinese computer attacks	hacker suspect post washington cyberattack report persist militari york #securityguard
MB117	marshmallow peeps dioramas	librari contest make chick 2013 center alma studio winner public
MB118	israel and turkey reconcile	apolog rapproch obama diplomat raid coup begin talk supp broker
MB119	colony collapse disorder	pesticid bee research rise death point @huffpostgreen backup robot plan
MB120	argentina's inflation	censur imf supermarket freez price data nation econom bearish overrun
MB121	future of moocs	higher educ shape academ linkedin hq freedom @hastac @gerrycanavan scoopit
MB122	unsuccessful kickstarter applicants	
MB123	solar flare	earth erupt cme unleash disrupt spit halo massiv wave activ
MB124	celebrity dui	attornei counti britton #troubl bellevu king #celebr behavior seattl arrest
MB125	oscars snub affleck	ben director dga guild award argo win honor new present
MB126	pitbull rapper	lohan lindsai lawsuit lose battl court dismiss rb suit sing
MB127	hagel nomination filibustered	gop senat republican stall wage clear 7127 tatler vote wai
MB128	buying clothes online	store shop phoenix secur amadiu safe tip ikeji sheer directori
MB129	angry birds cartoon	rovio seri game march deliv premier toon anim tv fly
MB130	lawyer jokes	skunk regul present highfiv everi prosecutor need twitpic new time
MB131	trash the dress	hawaii lizel lotter state kobu fair ttd yolanda bride #photographi
MB132	asteroid hits russia	meteorit meteor videozapi prelud giant mi se
MB133	cruise ship safety	drill result death fiv kill crew gone wrong abc member
MB134	the middle tv show	kaburutz nutan annoi ott thing rapid fridg ot indian empti
MB135	big dog terminator robot	darpa militari build dynam boston video brain hors youtub goe
MB136	gone girl reviews	breaker 1939 gillian flynn spring wind wild longer new #goodbook
MB137	cause of the super bowl blackout	wacki light freakout twitter flurri abnorm xlvii momentum went gop
MB138	new york city soda ban blocked	judg bloomberg vow mayor newsnew fight mayo

		reuter bbc
MB139	artists against fracking	yoko ono minidoc pennsylvania ha mother gimm doc @nygovcuomo yorker
MB140	richard iii burial dispute	cathol ashdown funer john telegraph dr sai buri church reburi
MB141	mila kunis in oz movie	
MB142	iranian weapons to syria	kerri israel shipment revolutionari iraq iran command guard maliki urg
MB143	maracana stadium problems	threaten worker strike brazil fear grow england fret fiasco despit
MB144	downton abbey actor turnover	anoth join lose maid recap leav spoiler miranda froggatt chat
MB145	national parks sequestered	servic fewer ranger spring mean npr cut impact affect avert
MB146	gmo labeling	groundbreak grassroot campaign washington launch food state consum monsanto mandatori
MB147	victoria's secret commercial	onset kerr moistur miranda shoot twe model ag beauti tweet
MB148	cyprus bailout protests	ralli youth america @ijreview eu accompani vote look good journal
MB149	making football safer	goodel nfl seahawk shrink url youth paid usa sport roger
MB150	uk wine industry	#wine rais viticultur scholarship 11 grape foundat growth economi canadian
MB151	gun advocates are corrupt	guncontrol flabbergast #memphi sioux capitol shrink lawmak lobbi url iowa
MB152	iceland fbi wikileaks	investig refus aid deni minist help kick thedailywhat agent arriv
MB153	lighter bail for pistorius	south lawyer oscar africa lift ban argu travel restrict judg
MB154	anti-aging resveratrol	ag red ingredi wine effect coq10 endometriosi promis spritz review
MB155	obama reaction to syrian chemical weapons	assal toppl claim alass rebel bashar presid media state fight
MB156	bush's dog dies	barnei georg 12 itun rip pass dead hot awai rustl
MB157	kardashian maternity style	kim formichetti stylist explain nicola fashion peplum evolut pregnant experi
MB158	hush puppies meal	bearcat clog cincinnati #hush ericson collegi verkauf ankl boot spirit
MB159	circular economy initiatives	resourc compani creat rapanui cloth epr @fastcoexist start crunch bullet
MB160	social media as educational tool	techniqu strategi exclus infograph articl new enhanc brief post corpor
MB161	3d printing for science	3dprint launch embryon world stem fabric percent scientist skull pirat
MB162	dprk nuclear test	condemn resolut allafrica hous sanction pass conduct germani russia slam
MB163	virtual currencies regulation	bitcoin amazon coin govern rais hackl alltim announc treasuri shrink
MB164	lindsey vonn sidelined	tiger wood date fearless ar led injuri condit crash pregnant
MB165	acpt crossword tournament	puzzl nerdcor
MB166	maryland casino table games	blackjack debut april live roulett crap set arundel dealt come
MB167	sequestration opinions	gop cut peo blame word mean someth onli automat ha

MB168	us behind chaevez cancer	rosi identifi breast firm surviv leader claim myeloid report special
MB169	honey boo boo girl scout cookies	june kept sale mama campaign shut word facebook sell ha
MB170	tony mendez	argo real hispan cia spy im screenplai #argo memoir autograph

Appendix D: HPRF-2 best run expansion examples

Table 26 Expansion terms extracted for the best HPRF-2 runs with respect to P@30

Topic ID	*Original*	*HPRF-2*
MB01	bbc world service staff cuts	languag 650 outlin close job caribbean lose plan understood foreign statement seven quarter fund major loss program announc million offic shut radio new like end guyana alban todai shed brief
MB02	2022 fifa soccer	cup qatar world blatter sepp winter presid plan russia2018rusia held chang stage eurosport summer sport digest end plenti year pen discuss ahead ha deal wcup lausann aggreg work competitor maverick
MB03	haiti aristide return	jeremiah haitian wright rev duvali okin polit bertrand america media pou exil want new democrat wa allow presid non door viabil open bezwen hei guerr ye rele increasingli influenti ethnic
MB04	mexico drug war	clinton hillari reform judici essenti flag secretari state violenc outdat applaud legal ongo patrol lawsuit border crimin justic effort risk heat file agent profil trip view visit mondai consen punit
MB05	nist computer security	cloud technolog guidanc tackl new public issu standard guidelin informationweek institut nation manag virtual includ inform govern draft busi wednesdai set peer analysi threat wiki risk profession research address advic
MB06	nsa	secur relationship global watchdog appl analyst googl secret group com report date postgradu fun head naval mistress impli warfar intim inappropri cnet encount 1994 gizmodo experienc affair flirt academi tap
MB07	pakistan diplomat arrest murder	court held lahor pakistani kill detent charg consular american extend judg doubl thursdai dai order accus employe reuter said paki prosecutor offici hand detain mount illeg investig despit pressur allow
MB08	phone hacking british politicians	tabloid prime minist gordon sourc scandal brown wrote polic mail editor summ amid voicemail dismiss sue world cnn famou situat fear inform summer voic new sundai told phonehack close edmondson
MB09	toyota recall	vehicl million fuel nearli 17 new said corp leakag worldwid leak car concern involv motor wa global latest 15m announc uk manufactur 000 world salt lake abc reuter japan wednesdai
MB10	egyptian protesters attack museum	looter mummi destroi offici mubarak authoritarian artifact crackdown deploi nationwid loot curfew defi shield sweep stolen revolt hosni militari defend nearli govern form cairo presid human continu possibl saturdai nation
MB11	kubica crash	renault robert formula pace test set fastest intent signal new surgeri pre valencia lap face campaign time underlin post f1 potenti shoulder involv foot broken hospit fix seriou grandprix stai
MB12	assange nobel peace nomination	prize wikileak laureat founder protest china winner imprison visit vibrant alongsid famili elbaradei julian activist observ world russian movement arab right

		wife human jintao run hour schwab xia egypt yunu
MB13	oprah winfrey half-sister	secret famili reveal ha sister shock big sai announc knew share thei marque seemingli todai shook prompt patricia distribut new spill sundanc reunion adopt birth quickli bean auction buzz core
MB14	release of the rite	hopkin anthoni box horror offic film ap weymouth exorcist notabl bump oscar usa review weekend mani movi new
MB15	thorpe return in 2012 olympics	london nimrod scrap venu shape fear new stadium plane decis latest updat teessid game surveil aircraft goalcom prompt gazett playbook 550 tottenham postpon 201 gap propos benefit ham imag secur
MB16	release of known and unknown	cultur entertain pop celebr precip rumsfeld abcnew 219 memoir nebraska puls breezi began donald new reflect gd mp region island central north airport ar latest interview wind write grand ne
MB17	white stripes breakup	zebra rttnew northern black meg forex econom jack sai announc offici light busi myriad god great realtim preserv certifi know com new documentari analysi haha belong statement deliv trailer broken
MB18	william and kate fax save the date	middleton princ royal lookalik shortag ferguson wed sarah invit marriag contributornetwork testino hit telegraphcouk epidem outcast greyson bridal nationwid com new diana princess mario gossip process guest famou lesson speech
MB19	cuomo budget cuts	spend new nanotechnolog medicaid sham york governor andrew lakesuccessni radiu lead ny assess revers expos unveil gov propos effort trick seek dirti tuesdai past 1329 plan 10b alwai substanti layoff
MB20	taco bell filling lawsuit	beef meat laist shortli expos guilti pleasur ground claim opinion suit ignor came new donnel consumerist mean sharpton scarborough schultz hay realli maddow forb lawrenc msnbc motion matthew fals rachel
MB21	emanuel residency court rulings	rahm chicago mayor ballot appeal chicagotribun meet requir break run doe relat new huffingtonpostcom suntim post world video photo cour wsj illinoi com analysi spy headlin coverag econom associ financi
MB22	healthcare law unconstitutional	judg feder rule florida obama presid health mondai barack declar reform insur care pensacola vinson overhaul dealt void struck sai legisl oppos thi sweep primari signatur district mechan initi coverag
MB23	amtrak train service	derail station penn york collid rail new eagl car parlor chp loma outsid vox encount passeng stretch strike victoria rare oil afternoon attack south mondai dead sundai turn leav picturesqu
MB24	super bowl seats	fan stadium cowboi deni 400 stairwel xlv unsaf readi befor dalla becaus game sent sport sundai fail fiasco temporari section root wa troubl appar nfl latest ticket wasnt stori weren
MB25	tsa airport screening	privat program shut door standstil administr screener transport allow secur neutral opt brought replac wa said govern test month ditch new longer travel hartsfield just contractor utter chronicl instantli halt
MB26	us unemployment	benefit claim eas fall firsttim tennesse number initi 000 youth weekli 42 unchang tonga stage file govern

		rate expect american close poverti set percent decemb week concern wa time skew
MB27	reduce energy consumption	build specifi construct transport resourc consult written 75 profession industri engin premier account commerci review hvac 901 friend 2030 envelop instantli effici emploi peak reader equip standard expert electr trick
MB28	detroit auto show	chrysler green intern 2011 cameo car hornet bigger wa polic make time number 2010 stai everi shenzhen final mean precinct thei chock wasn wai greensboro safer sai thi visitor new
MB29	global warming and weather	bizarr whale crop destroi updat grantham right legendari jeremi weigh fund manag chang
MB30	keith olbermann new job	msnbc current tv countdown home host gore becom regret al announc act updat sinc flamm onlin commen smal said talk alec baldwin hi peopl pseudo todai tucson ha departur amid
MB31	special olympics athletes	winter compet game michigan state celebr stadium row neuro prepar energi track thiev drink new target tar central smart gold heel host bodi chanc sundai kfmb wyo vi hanniti 760
MB32	state of the union and jobs	presid address 2011 student obama video pennsylvania ipad highlight campu speech challeng univers view white celebrit parti hous barak cnncom pillar sai devot penn deli time react barack peak employ
MB33	dog whisperer cesar millan techniques	dilla rj trailer sanaa watch behavior medicin check crash anim tag train album link spaniel tv blog world cocker bandcamp yin sophia insight writer bite provid style group articl answer
MB34	msnbc rachel maddow	internet olbermann hoax prayer speech market stori free idiot union video make fake state abuzz parti shakeup character wishlist rach spoof merger russo caleb comcast favr pep nbc uber keith
MB35	sargent shriver tributes	bono u2 buri mourner sarg breitbart cape potomac wa wife ap optimist recal funer ideal chariti capet cofound tmcnet grandchildren frontman alwai remembr pio fond maryland pioneer cathol globe md
MB36	moscow airport bombing	suicid kill domodedovo busiest terrorist injur russia blast 35 peopl deadli explos video 31 moment airpo hit wit sitemap deton lacross slashdot aftermath 145 marketplac bomber provinci editori submiss column
MB37	giffords recovery	rep gabriel doctor road rocki sprint week come gunshot dai face wound long marathon ahead sai head brain rehabilit li sullivan fro complic despit injuri center continu buildup talk congressman
MB38	protests in jordan	thousand demand step opposit amman pm support reform unemploy inflat gather prime minist econom mass arab wave spread 3rd polit countri street fridai world albania leftist algeria modest islamist erupt
MB39	egyptian curfew	expert militari level protest impos unparallel egypt widespread unpreced deploy deploi escal regim cairo sai hold power question 0700 1800 instantli unrest riot hosni author connect east middl local presid
MB40	beck attacks piven	glenn fox franc threat anchorman taunt death defi new leftw fran wa depict york franci academ amid thi

		controversi veri fals led campaign scari target host ar view speak act
MB41	obama birth certificate	secret autom hawaii releas social stai law limbaugh bar info bookmark submit engag demand tool content secur polic network fast websit media save coverup guv facebook googl abercrombi twitter amend
MB42	holland iran envoy recall	1979 analysi egypt crisi dutch burial tehran malaria mubarak iranian checkout hosni awar row washington rais presid media social sahra oklahoman iaea wisner newscomau overse bahrami compli disown netherland ambassador
MB43	kucinich olive pit lawsuit	denni cafeteria settl congression sandwich sue tooth ohio su bit suit file hi congressman longworth repres suffer rep 2008 wrap hous crack case capitol fridai incid thi dental said split
MB44	white house spokesman replaced	madelein mccann clarenc famili biden mitchel secretari journalist attempt hack claim press sourc obama mobil ap pick report believ phone rcmp carnei hi voicemail complaint moscow brutal suspend terror access
MB45	political campaigns and social media	2010 network market bowl pew unleash super american 22 onlin pastrana bjorn twitter featur ad environ travi candid democrat weigh dii tv republican impact elect becam regular solut video launch
MB46	bottega veneta	rodart fragranc bag ell beauti muse palm fan beyonc galleri daughter beach fashion design art snakeskin chae face pewter rtw teal paltrow gwyneth turquois 2011 abstract sculptur tame payn knot
MB47	organic farming requirements	learn violent path fig children indonesia offer street iask kid job restaur help thei new fuller hotel food garcia creatur dine ws anthoni led concern root background skill produc china
MB48	egyptian evacuation	threat ohio campu bomb forc begin colleg egypt american state 286 stark northeast amid afp globe victim flood staff aussi washington accord warn earlier allow return cairo unit student mail
MB49	carbon monoxide law	poison detector plugin amazoncom comet epa backup blown displai suffer propos alarm alert batteri repeat danger requir avoid effect digit limit tuesdai begin dy goe state deal feb famili 50
MB51	british government cuts	lockerbi cameron bomber tax minist spend signific advis david prime exert rule deficit releas previou major megrahi scottish blast document auster case libya uk accord warn sai handl convict britain
MB52	bedbug epidemic	expert instantli connect import rid celebr favorit break friend new twitter hotel follow citi bug pest abat ar provinc warn specialist suspend council help fight infest crisi offspr invest battl
MB53	river boat cruises	travel midnight vike thi feb 12th wa nile sat demian world valentin bangladesh lena inform ar danub rhine dai cancel budapest david march love eiffel 2011 li hector dvd net
MB54	the daily	ipad new newspap murdoch corp launch unveil rupert store appl app connect public digit itun free magazin

		thi shropshir download updat saver antoin live week 26th event review design web
MB55	berries and weight loss	acai diet plan adi healthi ar new lose reduct supplement amazon health cleans thi promot live colon fruit number free cardiovascular current antioxid imp best expect inclin fast fact american
MB56	hugo chavez	venezuelan presid venezuela boss power new 12 year golf hi promis amerika mark threaten oliv enemi caraca pool breitbart zinfandel kecam stone 99 press bank krisi 2011 alyssa socialist yfrog
MB57	chicago blizzard	weather snow 2011 inform thi forecast ar cam storm power telli warn break com look resourc watch wriglei 1967 new topic video suburb thundersnow sourc arkansa relat issu gener upload
MB58	fda approval of drugs	contrav weight loss declin deni administr weightloss orexigen advisori committe ha preterm food recommend diet ignor obes wont reduc thi won tuesdai decid astrazeneca treat birth new risk depress agenc
MB59	glen beck	glenn piven fox ridg gala franc godspe pagan behalf grate theatr week buzz scari target letter polit ar london new support high mediait veri school doctrin imageri home disclaim nwo
MB60	fishing guidebooks	thi 2011 travel guid aquarium automot istanbul browser question cyberstalk fsta cch japan rai linki commonli lagoon thing lure servic modif write web scholarship tropic parad mosquito divorc 11 ask
MB61	hu jintao visit to the united states	china presid chines obama diplomat ar beij washington meet thi relat american com strateg success week dinner haiti sector power econom onlin israel tcl arab senior wasteland turbul world includ
MB62	starbucks trenta cup	bottl entir wine hold new ounc squid stomach terrifi size huge laugh gothamist coffe largest spare larger human room graphic averag wino nearli twitoast someth glass american rt ic drink
MB63	bieber and stewart trading places	jon kristen justin new ian daili thi rocki white snow bodi board baseman hedlund twilight opposit museum alec garrett teen video memori ar martha 911 kingston rod pitch sex actress
MB64	red light cameras	studi redlight fatal crash ar traffic depot save citi speed cut oceansid updat driver intersect new ticket live legislatur accid instal deadli opelika help lawmak road violat start prom percent
MB65	michelle obama's obesity campaign	ladi atlanta childhood weight new rate militari oprah stop role loss plan ar america travel crescent fat exercis come presid health nation polici plai lack insur food healthi global speech
MB66	journalists' treatment in egypt	beaten attack protest detain guardian foreign egyptian arrest new demonstr condemn record target clinton cairo violenc alongsid mubarak media shenker continu polic plaincloth outsid ordeal cnn peopl govern intimid ar
MB67	boston celtics championship	nba marqui bruis daniel kendrick perkin laker inform ormond ha electron spinal doc eject river coupon vs new spine collis showdown collect final fashion cord angel hope sport fine com

MB68	charlie sheen rehab	actor enter check voluntarili half men undisclos hiatu home exclus goe tv hi month hospit tmz sourc ha cb intox martin caller said 911 thi celebr product com fridai rep
MB69	high taxes	new defer truste parcel court bombai ballot free offer vodafon incom ginorm return price offic vote amnesti deficit food hear dodg illinoi climat percent tv ir growth stadium corpor school
MB70	farmers markets opinions	winter somervil thi vendor vimeo park produc convent oregon brave video food truro roadrunn homestead vega derri adam com ar local morph blogspot pop visit roll shut molto adelaid main
MB71	australian open djokovic vs. murray	novak andi final ferrer tenni david 2011 feder live roger ved stream semifin win thi ladbrok semi highlight finalen watch titl master anden hi ap singl men set cincinnati melbourn
MB72	kardashians opinions	kim kourtnei khloe teen khlo 115000 mom odom new vulgar fashion savag lamar pier 115 000 morgan grant watch humphri extravag michael weight unisex celebr sister peopl aniston fragranc immun
MB73	iran nuclear program	talk power collaps new nanotechnolog fail world aerospac nuke doubt thi dprk intern satellit weapon expo infowar rocket offici progress concern israel languag wa featur sanction tech iranian deputi ar
MB74	credit card debt	consolid relief settlement option elimin pai rid settl help bad decent wai solut best bankruptci fastest negoti payment citizen combin benefit want loan greatest free ar unsecur compani becom tip
MB75	aguilera super bowl fail	christina anthem nation flub xlv sing line repeat perform 2011 fumbl singer video lyric botch rehears lea belt sang apolog star michel nail kickoff start unbeliev submit com univers gaff
MB76	celebrity dui violations	seattl guilti lawyer attornei pressli plead jaim antitrust directori bucki badger new kristoff charg law search thei suspicion arrest googl dwi inform tweetmem angel kennedi meme soap privaci marin arizona
MB77	ncis	regulatori intonow effort e13 help start redirect guidanc fda watch cage televis chart engag approv firm freedom assist yahoo provid program expect episod cours easi season favorit fun onlin free
MB78	mcdonalds food	ronald hostag price held campaign rais fastfood fast spoof liber menu increas armi restaur cost prefer certainti rise colonel chill commod sale thi franchis kidnap ellen corp boost kfc ultim
MB79	saleh yemen overthrow	presid abdullah ali protest 2013 step yemeni term leader tunisia egypt ouster seek sanaa new wednesdai govern erupt countri announc brotherhood right hi readout rachman cair muslim thousand capit reelect
MB80	chipotle raid	flotilla isra gaza aid legal israel priyanka probe deadli regrett katrina report kaif blockad commiss chopra ship immigr panel incom grill attempt wa tax law intern hundr food mexican chain
MB81	smartphone success	android comscor nokia mobil infineon ha app profit io subscrib refund googl new tablet os monthli overal pass data chipmak ir blackberri quarter symbian tax

		unveil idc iphon statu boost
MB82	illegal immigrant laws	missouri adopt state termin arizona court rule new suprem right face tuesdai ralli nebraska tough realiti enforc guatemalan fiscal hundr number imprison overturn wa consent capitol center program ar thi
MB83	stuxnet worm effects	expert comput iran chernobyl boomerang russia secur malwar duck confick claim new warn investig purportedli viru success caus sai wa thi hacker wildli mani nato cultur reuter gibson haunt threat
MB84	athlete concussions	crosbi teen rais multipl risk health bruin footbal savard suffer sidelin new urg vonn effect sagna game parent fact conscious sidnei pennsylvania ha lawmak detect hockei standard solut arsen race
MB85	best buy improve sales	depart excus close suster thi grp ventur market entrepreneur home ar partner capit tip beat dark estat skill write gone forecast expect busi new ha time retail surpass real increas
MB86	joanna yeates murder	tabak vincent jo newsom accus remand charg man court neighbour new sadden landscap architect deepli bristol funer intrus uk chantilli chang custodi death artist market headlin kill mp seo smith
MB87	chicken recipes	cutlet easi fri salad fettuccin thi healthi delici carb melang bruschetta curri new sauc thai rice mozzarella pistachio soup meal entre caper mustard fresh ingredi homemad best roast anis bbq
MB88	kings' speech awards	guild sag director produc win hooper oscar boardwalk tom thi best 2011 empir firth reign film winner academi america crown dga announc wa claim honour colin nomin movi rule lead
MB89	supreme court cases	illinoi emanuel rahm ballot rule chicago mayor missouri adopt immigr appeal termin tuesdai ha hear order decis state illeg justic guatemalan high resid new agre wa right motion issu expedit
MB90	anti-bullying	bulli school teen nj new help psa parent fckh8 asburi mtv anti program radiu creat helpless video martial h8 pick puls buk thi ad kid district pepsi thu audit homo
MB91	michelle obama fashion	state militari ladi oprah design famili new dinner dress flap winfrei style wear critic honor renta union gown mr cut wore power signal stylelist break horyn flotu obam massiv campaign
MB92	stock market tutorial	dtn new trade chart instantli expert invest emini connect dow shiller import price celebr tip intradai favorit jone break 2011 bell ar nasdaq dai friend histor ap com share report
MB93	fashion week in nyc	fashiolista thi new york cover hotspot win cincinnati img design com list aritzia minkoff pari item frappuccino thei check friend gainesvil menswear copenhagen fly trip tumblr 2011 tip paco media
MB94	horse race betting	tip request problem link thi monei best racecours burch todai track cheltenham balai profit onlin result brighton win make reliabl lose 86 sandown highli lingfield trial odd tipster taunton fantast
MB95	facebook privacy	user set everi german data congress deal new 10 mark know aspect need zuckerberg unfortun account explain tune chang ceo roundup secur ar littl disabl hq interrog

		ebai dai congression
MB96	sundance attendees	sxsw meetup fellow mashabl connect 2011 march creativ austin tex 11 south 20 southwest confer annual technolog peopl entertain togeth share media idea widow come style webinar review onstar camp
MB97	college student aid	financi certif univers loan cours onlin inform program abroad internship tutor educ resourc languag leapfrog learn prepar ar busi facilit studi babysitt train appli new test middleburi 3221 text book
MB98	australian floods	18fashiontoaid qld appeal launch fashion tax design propos 19 prime minist gillard australia damag julia cost victim billion repair rate pm flee govern pai levi homebuy connolli reconstruct new coal
MB99	superbowl commercials	forc volkswagen automot 2011 blitz thi busi major watch feb2 350k ad thei thirti chevi 2014 darth 000 vader fave discov gm favourit cash popular size view glee channel second
MB100	republican national committee	hous chairman budget unit senat thei ar infrastructur thi ronald reagan transport obama washington vote ey convent spend websit rnc upcom cut superintend urg ban ryan new medicar leader paul
MB101	natalie portman in black swan	star oscar dga radiant nomine celebr opera boost royal director babi mcl award break movi hous prais pregnant wa 2011 cinema jp boom nomin dress sinchew sale alreadi 63rd ptz
MB102	school lunches	obes child advantag fingerprint pack parslei berkelei suzann sixth grader regularli bakeri bui leed michigan percent 000 mp brought rose children cake ar parent 29 studi reason save question schoolchildren
MB103	tea party caucus	senat rubio scalia republican marco join thursdai address 2012 meet question rand lawmak antonin presidenti plan conven conserv congress movement campaign new concord paul capitol read ar billionair iowa clue
MB104	texting and driving	cell seth argu phone talk ye shouldn work blindfold dog shouldnt death idiot don think blog lead els someon godin video twitlong yfrog twitter thi distract post new favorit facebool
MB105	the avengers	cast captain sevenfold america comic drawn smulder prequel cobi spot toxic thi tv hardest forex escap fm wa forum video hill receiv agent maria worst artist commun record com album
MB106	steve jobs' health	appl compos thi destroy product address hipaa creator portrait thei tasteless ceo medic new care apologis gifford absenc commenc analyst hi gabriel break blockquot chief leav ha brilliant cnn wa
MB107	somalian piracy	prais thi pirat boost anim ndtvcom sale internet wotn bighead oxfam rampant maritim bittorr criticis widen somali somalia envoi stanc famil drought cio nonprofit sai censor mumbai trigger ethic restrict
MB108	identity theft protection	cost infograph doe type crime secur kind tip social varieti cyber scam resourc concern warn credit center ring appl store number chanc informatio restitut schuster grownup frontlin increasingli entrepreneurship shaw

MB109	gasland	oscar nomin industri fox screen 1921 gassi shale barnett frack seam examinercom refuge filmmak coal ga glamour thi hbo drill climat environ dont documentari observ sydnei academi fals nod josh
MB110	economic trade sanctions	belaru iranian outlook iran polici expert eu toughen bilater develop widen turkish currenc new appeal usd firm region slap leader dollar effect data join cyclingnew marat link developm ecopi lukashenko
MB111	water shortages	scarciti sh forc acut hit farmer iran rage sever global secur @maudebarlow gaffei burn bbmp
MB112	florida derby 2013	cougar gulfstream race mad thunder casino 64 odd tournament #previewpredict hors treat park ivl virtuou
MB113	kal penn	yaar raffl summari enter place tv meet parti movi malh sunil uwo #ldnont ha want
MB114	detroit efm undemocratic	financi appoint emerg snyder manag loom baptist takeov support convict pledg minist area troubl mayor
MB115	memories of mr. rogers	35 birthdai fact happi devour sanctuari late circu neighborhood nod steve ipod download film fun
MB116	chinese computer attacks	hacker suspect post offlin access rang taken paper befor washington cyberattack report persist cyber militari
MB117	marshmallow peeps dioramas	librari contest public make center alma studio winner chick extravaganza 2013 art open favorit creation
MB118	israel and turkey reconcile	apolog obama raid rapproch diplomat pledg coup begin presid fridai embattl talk yellin naval supp
MB119	colony collapse disorder	pesticid bee research rise death point caus culprit hive @huffpostgreen rel percent scientist accord york
MB120	argentina's inflation	censur supermarket imf freez data econom price nation indec soar halt januari expuls bearish overrun
MB121	future of moocs	higher educ shape abuzz silicon ventur academ sustain sweep vallei academia freedom hype cours folk
MB122	unsuccessful kickstarter applicants	
MB123	solar flare	erupt earth sun cme massiv activ radio dure unleash disrupt long spit halo wave storm
MB124	celebrity dui	counti king britton bellevu attornei arrest #troubl drake 425 sox #celebr pitcher behavior seattl gossip
MB125	oscars snub affleck	ben director guild award dga argo 65th honor win present triumph new despit shrug hi
MB126	pitbull rapper	lohan lindsai lose lawsuit battl court uk defam dismiss rb actress gossip yahoo magazin scene
MB127	hagel nomination filibustered	senat gop republican defens clear pentagon vote stall secretari wage wai confirm rais 7127 tuesdai
MB128	buying clothes online	store tip shop phoenix contain guid secur inform amadiu safe purchas ikeji sure thi sheer
MB129	angry birds cartoon	rovio seri game march tv anim toon debut deliv premier weekend episod releas thi base
MB130	lawyer jokes	skunk regul present highfiv everi prosecutor need twitpic new time hall ly road differ dead
MB131	trash the dress	hawaii lizel lotter state kobu fair ttd yolanda bride #photographi coolest photographi wed choic shoot
MB132	asteroid hits russia	meteorit meteor prelud giant videozapi hd feb 14 valentin 15 hard 2013 dai mi se
MB133	cruise ship safety	drill result crew death canari member thomson fiv island gone wrong kill spain said lifeboat

MB134	the middle tv show	annoi fridg empti thing slow kaburutz nutan internet cut power ott rapid ot indian africa
MB135	big dog terminator robot	darpa militari build dynam boston new video real life brain web hors releas ic youtub
MB136	gone girl reviews	breaker wind 1939 gillian flynn spring itun connect wild selznick longer store cukor releas new
MB137	cause of the super bowl blackout	light wacki momentum releas outag polici senat oreo freakout remain energi market twitter night flurri
MB138	new york city soda ban blocked	bloomberg judg vow mayor restaur newsnew fight sugari pioneer struck appeal mayo reuter theater bbc
MB139	artists against fracking	yoko pennsylvania ono minidoc gastown #dontfrackni york mother cuomo filmmak ha governor documentari gimm doc
MB140	richard iii burial dispute	cathol funer ashdown academ anglican sai led research given john telegraph dr yorkshir wa buri
MB141	mila kunis in oz movie	
MB142	iranian weapons to syria	kerri revolutionari iraq iran command israel guard shipment sai kill maliki urg leader stop john
MB143	maracana stadium problems	brazil worker fear threaten refus strike grow cup fret england host mondai work despit grew
MB144	downton abbey actor turnover	anoth maid finneran season brien leav hunki siobhan crawlei join ladi reportedli lose tv new
MB145	national parks sequestered	servic fewer cut ranger yosemit spring sequestr becaus mean forc npr budget tourist impact season
MB146	gmo labeling	groundbreak grassroot washington food campaign state launch consum mandatori illinoi initi solv 37 latest wa
MB147	victoria's secret commercial	onset kerr moistur miranda shoot twe occas advantag pump brand model ag took beauti person
MB148	cyprus bailout protests	ralli restructur bank plan resign opposit commerci biggest break countri youth head america @ijreview vote
MB149	making football safer	goodel nfl roger seahawk shrink usa url youth paid leagu continu sport chapel hill look
MB150	uk wine industry	viticultur #wine rais scholarship grape foundat economi 8bn italian spanish account shop 11 growth canadian
MB151	gun advocates are corrupt	guncontrol flabbergast #memphi sioux capitol shrink lawmak lobbi url iowa dozen flood gop execut ralli
MB152	iceland fbi wikileaks	investig refus aid deni minist thedailywhat cooper kick agent help ago interior year said arriv
MB153	lighter bail for pistorius	south oscar restrict travel african judg track star charg lawyer eas lift oversea condit africa
MB154	anti-aging resveratrol	ag red ingredi wine endometriosi studi sinclair cognit effect grape coq10 miracl publish promis spritz
MB155	obama reaction to syrian chemical weapons	claim rebel assal toppl reportedli alass govern confirm bashar presid media state 25 fight kill
MB156	bush's dog dies	barnei georg 12 presid pass hot awai itun announc store rip app iphon dead sad
MB157	kardashian maternity style	kim formichetti stylist fashion nicola explain pregnant photo peplum continu evolut gossip carpet hollywood outfit
MB158	hush puppies meal	bearcat clog cincinnati amazon shoe #hush ericson collegi verkauf handbag qualifi ankl boot spirit ship
MB159	circular economy initiatives	resourc cloth start compani creat rapanui epr

		@fastcoexist achiev kei pioneer crunch veri bullet silver
MB160	social media as educational tool	techniqu strategi exclus market infograph enhanc new intellig articl product strateg todai 14 brief post
MB161	3d printing for science	3dprint embryon launch stem scientist cell 75 replac world ha new fabric percent skull pirat
MB162	dprk nuclear test	allafrica condemn sanction resolut hous korea conduct germani north denounc pyongyang obama pass new audienc
MB163	virtual currencies regulation	bitcoin amazon coin treasuri govern rais fincen insist hackl kindl introduc alltim announc shrink summari
MB164	lindsey vonn sidelined	tiger wood date fearless ar injuri crash thrust recov uniqu led suffer condit knee pregnant
MB165	acpt crossword tournament	puzzl nerdcor mangrov rufu vom toll von american foto
MB166	maryland casino table games	blackjack debut april live crap roulett set perryvil arundel dealt 122 come descend hollywood ac
MB167	sequestration opinions	gop cut peo blame word mean someth onli automat ha choos avella layoff #sequestr dreamwork
MB168	us behind chaevez cancer	rosi identifi breast firm surviv leader claim myeloid report special canc enzym leukemia mutat variat
MB169	honey boo boo girl scout cookies	june kept sale mama campaign shut word facebook sell ha onlin #honeybooboo gist aft 225
MB170	tony mendez	argo real hispan cia spy im screenplai #argo memoir autograph adapt wire meow geek ty

Appendix E: HPRF-3 best run expansion examples

Table 27 Expansion terms extracted for the best HPRF-3 runs with respect to P@30

Topic ID	***Original***	***HPRF-3***
MB01	bbc world service staff cuts	languag 650 outlin close job caribbean lose plan understood foreign statement seven quarter fund major loss program announc million offic
MB02	2022 fifa soccer	cup qatar world blatter sepp winter presid plan russia2018rusia held chang stage eurosport summer sport digest end plenti year pen
MB03	haiti aristide return	haitian duvali jeremiah wright rev polit new okin bertrand america media pou exil want passport pacif democrat econom wa allow
MB04	mexico drug war	clinton hillari reform judici essenti flag secretari patrol lawsuit state border legal violenc file agent outdat view applaud ongo crimin
MB05	nist computer security	cloud technolog public standard guidanc virtual institut tackl inform new nation issu guidelin cybersecur informationweek manag govern includ cyber draft
MB06	nsa	secur relationship global watchdog appl date analyst googl secret group com report site postgradu fun onlin head naval mistress impli
MB07	pakistan diplomat arrest murder	court held lahor pakistani kill detent charg consular american extend judg doubl thursdai unit dai afghanistan order state accus employe
MB08	phone hacking british politicians	tabloid prime minist gordon scandal sourc brown polic wrote mail world dismiss editor new summ amid voicemail lawsuit resign sue
MB09	toyota recall	vehicl million fuel nearli 17 new leak car japan said corp leakag worldwid salt lake concern involv motor wa global
MB10	egyptian protesters attack museum	looter mummi destroi mubarak hosni offici authoritarian egypt artifact crackdown deploi nationwid loot curfew defi shield moham sweep elbaradei stolen
MB11	kubica crash	renault formula robert pace test set surgeri fastest intent signal face valencia new eurosport pre lap f1 campaign time underlin
MB12	assange nobel peace nomination	prize wikileak laureat founder china protest visit egypt world winner imprison right vibrant alongsid famili elbaradei julian activist observ tunisia
MB13	oprah winfrey half-sister	secret famili reveal ha sister shock big writer smith kevin sai announc knew share state red thei marque seemingli todai
MB14	release of the rite	hopkin anthoni box offic horror film ap oscar weymouth exorcist fincher notabl nolan foster grit christoph academi bump fighter swan
MB15	thorpe return in 2012 olympics	london nimrod scrap venu shape fear new stadium plane decis latest updat teessid game surveil aircraft goalcom prompt gazett playbook
MB16	release of known and unknown	cultur entertain pop celebr interview new movi tv music precip rumsfeld abcnew 219 memoir nebraska

		puls breezi began donald reflect
MB17	white stripes breakup	rttnew zebra meg forex econom jack realtim busi northern black analysi stock entertain sai announc offici light com new market
MB18	william and kate fax save the date	middleton princ royal lookalik shortag ferguson wed sarah invit marriag new contributornetwork testino hit telegraphcouk epidem outcast greyson bridal nationwid
MB19	cuomo budget cuts	spend new nanotechnolog medicaid sham york governor andrew lakesuccessni radiu lead ny assess revers expos unveil gov propos effort trick
MB20	taco bell filling lawsuit	beef meat laist shortli new expos guilti pleasur ground claim opinion suit ignor came donnel consumerist mean sharpton scarborough schultz
MB21	emanuel residency court rulings	rahm chicago mayor ballot appeal break chicagotribun meet requir wsj run new doe headlin world video relat com busi onlin
MB22	healthcare law unconstitutional	judg feder rule florida obama presid health mondai barack declar reform insur care pensacola vinson overhaul dealt void struck sai
MB23	amtrak train service	derail station penn york eagl parlor car collid vox rail new victoria chp loma outsid encount passeng bald stretch strike
MB24	super bowl seats	stadium fan cowboi xlv deni sport 400 stairwel unsaf readi dalla befor jerri jone becaus nfl latest game sent articl
MB25	tsa airport screening	privat program shut door administr transport secur standstil screener allow neutral opt brought replac wa said govern test month ditch
MB26	us unemployment	claim benefit initi eas fall firsttim jobless tennesse number employ labor 000 youth economi weekli 42 unchang tonga stage file
MB27	reduce energy consumption	specifi build construct consult industri engin hvac commerci transport resourc equip written electr 75 profession premier account review power green
MB28	detroit auto show	chrysler green car intern 2011 cameo time stai fuel hornet bigger wa polic make number 2010 www everi naia shenzhen
MB29	global warming and weather	grantham bizarr whale jeremi crop destroi updat chang right commod wheat legendari climat weigh henri fund manag
MB30	keith olbermann new job	msnbc current tv countdown home host gore becom regret al announc act updat sinc flamm onlin commen smal said talk
MB31	special olympics athletes	winter compet game michigan state celebr kfmb stadium row neuro prepar energi 760 slater track bodi thiev drink vi new
MB32	state of the union and jobs	presid address 2011 obama ipad student video pennsylvania highlight campu speech challeng univers view white celebrit parti hous barak cnncom
MB33	dog whisperer cesar millan techniques	rj sanaa dilla crash trailer train watch behavior medicin check bite anim tag tip album link spaniel clicker tv blog
MB34	msnbc rachel maddow	hoax internet prayer speech olbermann market stori idiot union free fake rach video state caleb palin

		research make februari media
MB35	sargent shriver tributes	buri cape wife bono u2 mourner sarg breitbart potomac wa ap optimist recal funer ideal chariti capet cofound tmcnet grandchildren
MB36	moscow airport bombing	suicid domodedovo kill terrorist russia busiest explos injur blast 35 video peopl deadli 31 moment airpo hit wit sitemap deton
MB37	giffords recovery	gabriel rep doctor road rocki sprint gunshot week come rehabilit dai brain face wound long marathon ahead sullivan injuri sai
MB38	protests in jordan	thousand demand amman step reform opposit pm aljazeera support arab east middl new unemploy inflat pacif gather prime minist econom
MB39	egyptian curfew	protest expert egypt militari level cairo hosni impos unparallel sadat east middl presid widespread unpreced deploy deploi mubarak dunn escal
MB40	beck attacks piven	glenn fox franc threat anchorman taunt death defi new york polit leftw fran scari target wa veri depict media franci
MB41	obama birth certificate	autom secret social bookmark hawaii releas websit media stai facebook law limbaugh bar info twitter optim multipl monitor submit engag
MB42	holland iran envoy recall	1979 analysi egypt crisi dutch burial tehran malaria mubarak iranian checkout hosni awar row washington rais presid media social sahra
MB43	kucinich olive pit lawsuit	denni cafeteria settl congression sandwich suit polit sue su tooth ohio bit file hi congressman gawker longworth repres suffer rep
MB44	white house spokesman replaced	madelein mccann clarenc famili biden mitchel secretari journalist attempt hack claim press sourc obama mobil ap pick report believ phone
MB45	political campaigns and social media	network 2010 market pastrana bowl onlin pew unleash super american 22 travi elect solut bjorn mastercard prepaid twitter featur ad
MB46	bottega veneta	rodart fragranc bag ell muse palm galleri beauti fan beach paltrow gwyneth art beyonc daughter imag fashion design dopium snakeskin
MB47	organic farming requirements	fig iask learn violent path restaur children indonesia garcia offer street ws food kid job help china thei new fuller
MB48	egyptian evacuation	threat ohio campu bomb forc begin colleg egypt american state 286 stark northeast amid afp globe victim flood staff aussi
MB49	carbon monoxide law	poison detector plugin amazoncom comet epa backup blown displai suffer propos alarm alert batteri repeat danger requir avoid effect digit
MB51	british government cuts	lockerbi cameron bomber tax minist spend signific advis david prime exert rule deficit releas previou major uk megrahi scottish blast warn new document auster case libya accord herbal sai handl
MB52	bedbug epidemic	expert instantli rid connect new import break celebr favorit bug friend hotel twitter pest follow citi summit abat ar ny provinc warn crack bed specialist suspend council help fight fed

MB53	river boat cruises	travel midnight vike thi feb 12th bangladesh valentin world lena wa nile sat inform demian explor citi vacat china ar tip dvd net danub love bgv rhine dai cancel budapest
MB54	the daily	ipad new newspap murdoch rupert corp unveil store launch itun appl app connect digit free public saver media magazin thi shropshir review download debut updat antoin live liverpool blog comput
MB55	berries and weight loss	acai diet plan adi new health healthi ar sport featur lose reduct supplement amazon cleans thi free promot live colon fruit number cardiovascular current antioxid imp best expect inclin fast
MB56	hugo chavez	venezuelan presid venezuela new boss power golf 12 year caraca pool 99 alyssa hi milano promis amerika mark threaten problem oliv mesir enemi moriarti kijiji onlin breitbart zinfandel mcdowel kecam
MB57	chicago blizzard	weather snow 2011 storm forecast new inform telli break thi ar cam twitvid upload video thundersnow power warn winter com gener look resourc watch share wriglei 1967 topic busi suburb
MB58	fda approval of drugs	contrav weight loss administr declin deni weightloss orexigen food preterm advisori committe ha diet risk recommend ignor health obes depress new wont reduc birth thi won tuesdai decid astrazeneca agenc
MB59	glen beck	glenn piven fox franc ridg gala theatr scari target letter godspe london pagan behalf veri grate theater week buzz video polit ar new media support high mediait school doctrin imageri
MB60	fishing guidebooks	travel 2011 automot guid thi browser japan rai aquarium istanbul servic web question cyberstalk fsta niemann url cch california hide linki 03 lagoon commonli cloak thing redirect lure modif write
MB61	hu jintao visit to the united states	china presid obama chines washington diplomat ar beij relat meet thi com american power israel onlin inform intern strateg success week secret world tender unemploy dinner train servic haiti sector
MB62	starbucks trenta cup	bottl wine entir hold new ounc squid coffe stomach terrifi size huge laugh food gothamist largest spare graphic nightlif restaur eater larger human bar review room averag chef wino katya
MB63	bieber and stewart trading places	jon kristen justin new daili white ian snow hedlund bodi thi teen rocki alec garrett board kingston baseman actress twilight opposit huntsman museum sex sean switch video role hous memori
MB64	red light cameras	studi redlight traffic fatal crash ar citi new speed depot save driver cut oceansid updat ticket intersect accid live legislatur instal deadli opelika prom help gii hollywood insur theme local
MB65	michelle obama's obesity campaign	ladi atlanta childhood new weight health loss role plan rate militari oprah stop ar america travel crescent fat exercis come plai 2012 presid fashion signal cycl nation topnew ajc polici
MB66	journalists' treatment in egypt	attack beaten protest detain foreign egyptian guardian new demonstr media cairo arrest mubarak condemn record target clinton violenc alongsid shenker continu

		polic peopl plaincloth hosni newspap outsid ordeal cnn kurz
MB67	boston celtics championship	nba marqui bruis daniel kendrick perkin electron sport collect doc river new laker apparel car inform ormond com final ha spinal eject ebai sinc coupon vs spine onlin collis showdown
MB68	charlie sheen rehab	actor enter check half men voluntarili undisclos exclus hiatu tv home celebr goe hi month hospit tmz martin sourc com ha cb gossip intox new caller said 911 thi product
MB69	high taxes	bombai free new court vodafon defer truste parcel prepar ballot deficit offer offic state corpor incom coupon ginorm tv union softwar ir aarp pensacola school return turbotax price plc vote
MB70	farmers markets opinions	winter somervil local food park thi vendor produc vimeo truro san francisco new convent oregon cocktail citi brave video explor sushi vacat recommend badg roadrunn restaur homestead vega derri chicago
MB71	australian open djokovic vs. murray	novak andi tenni final ferrer 2011 david feder live roger ved master stream semifin win cincinnati thi ladbrok semi men highlight finalen watch titl anden atp hi ap singl set
MB72	kardashians opinions	kim kourtnei khloe new celebr gossip teen khlo entertain mom odom vulgar fashion savag lamar photo pier 115 000 115000 peopl morgan grant humphri extravag michael weight unisex sister aniston
MB73	iran nuclear program	talk new power collaps world nanotechnolog fail aerospac weapon intern nuke iranian doubt featur thi dprk satellit expo infowar rocket intellig offici progress concern israel korea languag break wa sanction
MB74	credit card debt	consolid relief settlement option elimin pai rid settl solut bankruptci help bad decent wai best fastest negoti payment citizen altern combin benefit want loan greatest free program ar unsecur compani
MB75	aguilera super bowl fail	christina anthem nation xlv flub sing line 2011 singer repeat perform lyric video fumbl botch lea michel star rehears belt sang apolog pittsburgh new nail entertain nfl kickoff univers bai
MB76	celebrity dui violations	seattl lawyer attornei guilti pressli antitrust directori jaim plead bucki law new googl badger dwi kristoff charg angel search arizona penalti brief malpractic crimin bush chart thei litig salt sec
MB77	ncis	regulatori intonow effort bioinformat e13 help start genom redirect guidanc fda watch cage televis chart engag approv firm freedom assist yahoo provid program expect episod cours easi season favorit fun
MB78	mcdonalds food	ronald hostag price held campaign restaur spoof commod rais fastfood fast menu increas chill new liber corp armi cost prefer certainti china rise strong colonel search citi sale thi franchis
MB79	saleh yemen overthrow	presid abdullah ali protest 2013 tunisia step yemeni egypt term leader ouster new state govern thousand seek sanaa wednesdai erupt unit countri announc brotherhood right hi readout rachman cair muslim

MB80	chipotle raid	flotilla isra gaza priyanka aid probe legal israel katrina deadli regrett blockad report chopra ship immigr kaif commiss incom panel hundr grill attempt wa new tax law intern home irishcentr
MB81	smartphone success	android comscor nokia googl mobil io infineon iphon ha app profit subscrib tablet refund new tax statu appl os monthli overal pass data idc chipmak ir blackberri quarter symbian oper
MB82	illegal immigrant laws	new missouri adopt state termin arizona court rule right suprem ralli nebraska face enforc tuesdai tough clarionledg realiti alinski monson guatemalan fiscal hundr number imprison legislatur overturn recipi wa consent
MB83	stuxnet worm effects	iran secur comput expert viru malwar new chernobyl boomerang russia duck confick claim warn investig purportedli success caus sai wa thi hacker wildli threat mani nato cultur iranian reuter gibson
MB84	athlete concussions	crosbi teen rais health multipl footbal risk sport new bruin savard suffer sagna sidnei pennsylvania sidelin game hockei standard urg bacari arsen vonn effect parent fact conscious delawar wenger ha
MB85	best buy improve sales	depart excus close suster thi home market grp estat tip ventur skill entrepreneur busi ar partner capit new beat dark real write gone forecast expect ha time decemb retail distress
MB86	joanna yeates murder	tabak vincent jo newsom accus remand charg new man uk court neighbour chantilli sadden landscap architect deepli bristol funer artist intrus headlin chang custodi latest death billboard woman market kill
MB87	chicken recipes	cutlet easi salad fri fettuccin healthi meal cook thi melang curri mozzarella pistachio sauc thai rice soup delici mustard new carb bruschetta feta entre honei caper onlin east fresh quick
MB88	kings' speech awards	guild sag director produc hooper win oscar boardwalk 2011 tom empir firth movi academi film america best thi winner reign colin new actor crown dga announc season wa claim apatow
MB89	supreme court cases	emanuel illinoi rahm ballot rule chicago mayor missouri adopt immigr appeal justic termin new tuesdai ha hear order decis state illeg guatemalan high resid elect agre wa right motion issu
MB90	anti-bullying	bulli school nj teen asburi new help mtv buk psa parent fckh8 kid monmouth stop anti press program radiu creat helpless cc park video martial h8 fight pick puls consolid
MB91	michelle obama fashion	state militari ladi new design oprah renta dinner famili style dress oscar flap winfrei union polit horyn flotu signal wear critic campaign break honor cathi video 2012 franca gown tea
MB92	stock market tutorial	dtn new chart trade invest emini instantli intradai price expert jone nasdaq bell connect dow shiller histor com import indic celebr tip share wordpress close financi dai favorit break michel
MB93	fashion week in nyc	fashiolista new york thi cincinnati cover com pari hotspot win img ticket voucher tip design sustain list

		2011 aritzia minkoff contest item frappuccino thei check friend gainesvil menswear copenhagen fly
MB94	horse race betting	tip request problem link monei thi best profit lingfield racecours result taunton new burch todai track cheltenham balai onlin lai brighton win make reliabl daili lose cherri 86 sandown highli
MB95	facebook privacy	set user everi german data congress deal new mark 10 know zuckerberg aspect need featur unfortun account explain secur tune chang ceo trend list roundup ar face littl disabl hq
MB96	sundance attendees	sxsw meetup mashabl austin fellow connect 2011 sxswi march creativ tex 11 south 20 southwest confer event annual technolog peopl entertain network togeth share media social idea webinar film music
MB97	college student aid	financi univers certif loan cours onlin abroad internship educ tutor program languag prepar inform studi train test text book new learn middleburi resourc busi leapfrog uic ar injur facilit babysitt
MB98	australian floods	18fashiontoaid qld appeal launch fashion tax design propos 19 prime minist australia gillard damag julia cost govern victim billion new connolli repair rate pm flee coal pai levi homebuy billi
MB99	superbowl commercials	volkswagen forc automot blitz busi 2011 bowl thi darth ad vader super cash major watch vote car feb2 350k bedava volkswagon tv thei passat video rk thirti chevi 2014 viral
MB100	republican national committee	hous budget unit chairman senat thei obama new vote ar upcom ey infrastructur washington thi convent polit ronald reagan spend transport medicar cut urg testimoni virginia democrat websit rnc corpor
MB101	natalie portman in black swan	star oscar celebr dga radiant nomine opera boost royal director babi mcl award break movi hous prais pregnant wa 2011 cinema jp boom nomin london dress sinchew sale alreadi 63rd
MB102	school lunches	obes child advantag pack fingerprint children leed nutrit bui parslei program health berkelei suzann sixth grader regularli bakeri michigan percent 000 medicin mp youth brought rose diet cake breakfast ar
MB103	tea party caucus	senat rubio marco scalia republican polit join 2012 rand presidenti thursdai antonin meet movement conven address paul conserv plan campaign question lawmak primari iowa demint read new congress clue maori
MB104	texting and driving	cell seth argu phone talk ye shouldn work blindfold dog shouldnt death idiot twitlong don think blog lead els video twitter femdom someon godin bondag nylon yfrog spank thi free
MB105	the avengers	captain comic cast america sevenfold drawn smulder forum agent prequel cobi spot vengeanc hayden toxic book thi com fred tv hardest forex escap metal fm wa hero video hill receiv
MB106	steve jobs' health	appl compos blockquot hipaa thi new destroy product gifford rupert address murdoch creator portrait gabriel thei tasteless ceo break medic cnn care apologis absenc commenc ipad analyst hi video chief

MB107	somalian piracy	prais thi pirat boost anim ndtvcom sale internet wotn bighead oxfam rampant maritim bittorr criticis widen somali somalia envoi stanc famil drought cio nonprofit sai censor mumbai trigger ethic restrict
MB108	identity theft protection	cost infograph doe type crime secur kind tip social varieti cyber scam resourc concern warn credit center ring appl store number chanc informatio restitut schuster grownup frontlin increasingli entrepreneurship shaw
MB109	gasland	oscar nomin industri fox screen 1921 gassi shale barnett frack seam examinercom refuge filmmak coal ga glamour thi hbo drill climat environ dont documentari observ sydnei academi fals nod josh
MB110	economic trade sanctions	belaru iranian outlook iran polici expert eu toughen bilater develop widen turkish currenc new appeal usd firm region slap leader dollar effect data join cyclingnew marat link developm ecopi lukashenko
MB111	water shortages	scarciti sh forc acut hit farmer iran rage sever global secur @maudebarlow gaffei burn bbmp groundwat @globalvoic defunct moment ani
MB112	florida derby 2013	cougar race gulfstream orb park mad kentucki thunder casino 64 odd tournament best tc13 #previewpredict hors treat ivl coffe virtuou
MB113	kal penn	yaar raffl summari enter place tv meet parti movi malh sunil uwo #ldnont ha want fascin council actor obama student
MB114	detroit efm undemocratic	financi appoint emerg snyder manag loom baptist takeov support convict pledg minist area troubl mayor #citycouncil legal fight kevyn kilpatrick
MB115	memories of mr. rogers	35 birthdai fact sanctuari late happi devour film circu neighborhood nod steve ipod releas download fun smile amaz befor 85th
MB116	chinese computer attacks	hacker suspect post offlin access rang technolog taken paper befor cyberattack report cyber militari mandiant washington persist york #securityguard 61398
MB117	marshmallow peeps dioramas	contest librari chick winner art 2013 twinki public make easter bunni sugar post center alma studio extravaganza wapo vii open
MB118	israel and turkey reconcile	obama apolog raid rapproch presid diplomat pledg coup jordan avigdor begin fridai lieberman shimon embattl ufuk erdogan tayyip netanyahu talk
MB119	colony collapse disorder	pesticid bee research rise death point caus culprit hive @huffpostgreen rel percent scientist green accord york commerci 40 lost 50
MB120	argentina's inflation	censur supermarket nation freez imf price data monetari econom soar halt report fund polit indec supermercado devolop market gold januari
MB121	future of moocs	higher educ shape academ abuzz freedom silicon ventur sustain sweep vallei academia hype cours folk big land deserv sign aprendiendo
MB122	unsuccessful kickstarter applicants	
MB123	solar flare	cme earth sun erupt eject activ disrupt radio mass massiv wave coron dure power unleash long spit nasa halo storm

MB124	celebrity dui	counti king britton bellevu 425 attornei drake sox arrest #troubl com drunk red #celebr pitcher new behavior seattl gossip washington
MB125	oscars snub affleck	ben director award guild argo dga honor interaksyon new 65th tv5 win present triumph onlin academi despit shrug best entertain
MB126	pitbull rapper	lohan lindsai lose lawsuit battl court review uk defam bizkit underworld korn nsync manson radiohead hendrix u2 muzik limp backstreet
MB127	hagel nomination filibustered	senat gop republican defens clear pentagon vote stall secretari wage cb wai confirm rais 7127 tuesdai new tatler anoth end
MB128	buying clothes online	store tip shop phoenix contain guid secur inform amadiu safe purchas ikeji sure thi sheer directori trendi look make psych
MB129	angry birds cartoon	rovio seri toon game tv march anim uncategor entertain premier debut deliv weekend episod releas 16 thi televis base short
MB130	lawyer jokes	skunk regul present highfiv everi prosecutor need twitpic new time hall ly road differ dead favorit thei
MB131	trash the dress	hawaii state fair bride lizel lotter kobu ttd yolanda groom carniv photoshoot dai #photographi coolest photographi session wed choic shoot
MB132	asteroid hits russia	meteorit meteor prelud giant meteoroid videozapi near hd feb 14 valentin 15 hard 2013 dai mi se
MB133	cruise ship safety	drill death result canari crew thomson island member fiv gone wrong kill spain said lifeboat dure di routin emerg abc
MB134	the middle tv show	annoi fridg empti thing slow kaburutz nutan internet cut power ott rapid ot indian africa target batteri ea east exam
MB135	big dog terminator robot	darpa militari build dynam new video boston real life brain web cryptozoolog hors releas ic youtub goe gait surv walk
MB136	gone girl reviews	1939 wind selznick cukor breaker itun gillian store flynn spring movi georg connect wild longer david shantytown releas yanke havilland
MB137	cause of the super bowl blackout	superdom outag oreo light wacki xlvii momentum releas power twitter polici senat freakout remain energi market footbal night flurri abnorm
MB138	new york city soda ban blocked	bloomberg judg vow mayor restaur michael newsnew fight sweenei sugari tingl burton pioneer milton struck doug appeal mayo reuter theater
MB139	artists against fracking	yoko ono pennsylvania lennon york minidoc gastown fox josh #dontfrackni sean mother cuomo filmmak tennesse ha new gasland video governor
MB140	richard iii burial dispute	cathol funer ashdown academ minster anglican john sai led research york yorkshir given westminst cathedr telegraph dr bentlei wa buri
MB141	mila kunis in oz movie	
MB142	iranian weapons to syria	iran israel revolutionari kerri command guard iraq maliki shipment kill lawmak sai urg leader stop attack respond john offici syrian
MB143	maracana stadium problems	brazil worker fear threaten refus strike grow england cup despit friendli fret host fiasco autom mondai work

		grew charl readi
MB144	downton abbey actor turnover	anoth tv maid finneran new season brien leav hunki siobhan crawlei consolid seri join ladi reportedli lose recap spoiler debt
MB145	national parks sequestered	servic cut ranger fewer sequestr yosemit spring becaus mean forc npr budget tourist impact season gear earli march recent figur
MB146	gmo labeling	food grassroot washington groundbreak campaign state launch consum monsanto 37 mandatori illinoi prop initi solv organ latest wa market nation
MB147	victoria's secret commercial	onset kerr moistur miranda shoot ag twe tweet defi occas routin advantag pump brand model took beauti person
MB148	cyprus bailout protests	ralli jazeera aljazeera bank restructur youth pacif plan resign asia opposit europ africa east commerci biggest middl break report countri
MB149	making football safer	goodel nfl roger seahawk shrink usa url youth paid leagu continu sport chapel hill look sai ap q13 #seahawk everyth
MB150	uk wine industry	viticultur #wine rais scholarship grape shop foundat economi 8bn italian spanish account 11 growth canadian critic washington canada impress glass
MB151	gun advocates are corrupt	guncontrol flabbergast #memphi sioux capitol shrink lawmak lobbi url iowa dozen flood gop execut ralli journal traffic polit entertain paid
MB152	iceland fbi wikileaks	investig refus aid deni minist interior thedailywhat cooper kick order year assang agent polic help jonasson ago clear classifi new
MB153	lighter bail for pistorius	south oscar restrict travel african judg track star charg lawyer eas lift oversea condit africa ban murder argu allow girlfriend
MB154	anti-aging resveratrol	ag red ingredi sinclair wine scienc endometriosi studi cognit effect new grape coq10 glaxosmithklin medicin miracl publish promis research spritz
MB155	obama reaction to syrian chemical weapons	claim rebel assal toppl reportedli alass govern confirm bashar presid media state 25 fight kill shadow israel sai peopl begin
MB156	bush's dog dies	barnei georg 12 hot itun store app presid iphon pass dead awai laura ipod ipad announc rip touch robillard sad
MB157	kardashian maternity style	kim formichetti fashion stylist nicola celebr carpet pregnant entertain explain red instyl photo peplum continu emmi evolut globe pregnanc gossip
MB158	hush puppies meal	bearcat clog cincinnati verkauf amazon net shoe #hush ericson collegi handbag qualifi ankl boot spirit 34 ship return store order
MB159	circular economy initiatives	resourc cloth start compani creat rapanui epr @fastcoexist veri achiev china afraid kei pioneer crunch bullet silver benefit program north
MB160	social media as educational tool	techniqu market strategi exclus intellig infograph enhanc new brief engag articl product strateg todai 14 post corpor speaker lover network
MB161	3d printing for science	3dprint embryon launch world stem scientist skull cell 75 replac gizmodo ha new fabric percent pirat rocket medicin bone engin

MB162	dprk nuclear test	condemn allafrica resolut korea hous sanction conduct north germani new obama pass slam africa denounc pyongyang successfulli barack democrat audienc
MB163	virtual currencies regulation	bitcoin amazon coin treasuri kindl govern rais fincen insist hackl introduc alltim announc shrink summari new url exchang final paid
MB164	lindsey vonn sidelined	tiger wood date injuri fearless nordegren ar ski knee crash world weight alpin showbiz thrust elin recov olymp surgeri uniqu
MB165	acpt crossword tournament	puzzl nerdcor mangrov rufu vom toll von american foto
MB166	maryland casino table games	blackjack debut april live roulett crap set hollywood perryvil expans arundel dealt 122 come descend gambl ac add king player
MB167	sequestration opinions	gop cut peo blame word mean someth onli automat ha choos avella layoff #sequestr dreamwork @politico depa 350 mount feder
MB168	us behind chaevez cancer	rosi identifi breast firm surviv leader claim myeloid report special canc enzym leukemia mutat variat acut @bloombergnew newark mai genet
MB169	honey boo boo girl scout cookies	june kept sale mama campaign shut word facebook sell ha onlin #honeybooboo gist aft 225 @tmz fresco badg wor pm
MB170	tony mendez	argo real hispan cia spy im screenplai #argo memoir autograph adapt wire meow geek ty aka base histori stori bro