AN APPROACH FOR THE VISUALIZATION OF TOPIC EMERGENCE IN POPULAR LANGUAGE

An Undergraduate Research Scholars Thesis

by

ANDREW LARAMORE, KAVEET LAXMIDAS, KEVIN MEDEIROS, AND LAUREN VOIGT

Submitted to the Undergraduate Research Scholars program
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Research Advisor: Dr. Bruce Gooch

May 2016

Major: Computer Science

TABLE OF CONTENTS

ABSTRACT

An Approach for the Visualization of Topic Emergence in Popular Language

Andrew Laramore, Kaveet Laxmidas, Kevin Medeiros, and Lauren Voigt
Department of Computer Science & Engineering
Texas A&M University

Research Advisor: Dr. Bruce Gooch
Department of Computer Science & Engineering

The collision of cultures introduces change: from the manifestation of ideas in physical realms to less tangible avenues of discussion, cultural fusion presents interesting and particular trends that this research team aims to explore. In prior experiments in text analysis and interactive storytelling, preliminary observations on this phenomena were made. These lead to a fundamental question on the feasibility of visualizing cultural change drawing entirely from the written material of different scopes. More specifically, the team considers the influence of culture on language, interest patterns, and prevalence in two opposing pools of literature and written material, observing and graphing the relationship in a manner supplemental to the understanding of cultural behavior. The presented methodology details the cultural collision between the ”Computer Hacker” and mainstream media cultures, yet serves to exemplify an innovation on the visualization of colliding cultures in general.

NOMENCLATURE

COCA	Corpus of Contemporary American English
Crawler	A software tool for the traversal and storage of information, particularly information found on the open web
MySQL	An open-source relational database management system
n-gram	A sequence of *n* words in specific order
RDBMS	Relational Database Management System

CHAPTER I

INTRODUCTION

Inspiration and aim

People and communities have always passed accumulated knowledge on to future generations through stories, often words of mouth almost infamous for their tendency to be swayed and morphed by current events, personal experience, and cultural context. For this research group, storytelling became a topic of interest as a means of providing interactive feedback tool for students. Following research in language patterns and processing it became readily apparent that storytelling, in the context of time, contains particular markers that can be exploited to illustrate larger observations in culture. Thus the team aims to explore whether these markers (more specifically trends in language construction and diction) contain substantial coordinated information to serve as a basis for visualizing cultural changes. In line with the interests of the group are ”Hacker Culture” and popular media culture, whose abundant documentation in literature and written material are crucial to the understanding of common language bases in cultural trends. Currently much research exists in the realm of natural language processing and cultural trends, yet few prior research efforts achieve intuitive presentation of story-derived data. As such, it is a primary effort of this research team to invent methodology for visualizing cultural collision in a manner that provides valuable insights into the integration of existing cultures and ultimately the emergence of new cultural trends.

CHAPTER II

METHODS

Data collection and storage

Due to the inherent dichotomy in the data (the hacker and popular media lexicons being separate collections), a method for handling and storing segmented language data was devised. Following standard language research convention, sourced data was categorized using the *n*-gram model; phrases were binned based on the number of constituent words (1-grams, 2-grams, etc) and paired with frequency data to establish a preliminary relationship between words and their relative frequency. *N*-gram data acquired from the Corpus of Contemporary American English (COCA) contained historical frequencies for the top one million 2, 3, 4, and 5-grams and served as the primary dataset for phrase prominence as a result of time. To store this data a relational database was designed and implemented using the MySQL relational database management system (RDBMS). Each order of *n*-gram received its own table within the parent database, with structure as illustrated below:

Field Name	word_0	word_1	word_2	freq
Data Type	VARCHAR(50)	VARCHAR(50)	VARCHAR(50)	FLOAT

Fig. II.1. Standardized storage format for *n*-gram data

This structure was extended to the available data (tables for *n*-grams of length 1-5) and the sourced data was imported into the appropriate tables. To represent the ”hacker culture” element of the data, the publicly available Jargon File (http://www.catb.org/jargon/) was parsed and imported into an identical table structure. With the RDBMS in place and the source data organized, manipulation and analysis of the data was then possible.

Qwerying the Google Ngram database

In addition to querying the available COCA data, tools to compare dictionary files against the Google Ngrams dataset were also developed. Using the BigQuery API, queries possible with the self-hosted COCA data were thus available with the Ngram data via this approach. Identical approaches were taken in the storage of this information.

Public data crawling

For supplemental data outside of the two *n*-gram sources, a web crawler and variations were created. The database crawler reads in a dictionary of terms (a plain-text file extracted from the various jargon dictionaries) and compares all of the dictionary terms to against available data. For matching entries between the two datasets, the crawler saves the corresponding years and allows the creation of a timeline of the terms use. The quote crawler works similarly, but rather than line-by-line, it reads quote-by-quote. When crawling the famous movie quotes from IMDB, there was a peak frequency of the number of hacker jargon found in the time frame from 1984-1996. This is the time frame for the emergence of personal computing and common internet use. The data found from crawling a computer science dictionary had similar results to the hacker jargon dictionary. This data paves the way for a novel way to present the data intuitively, as follows in the next section.

CHAPTER III

RESULTS

In comparing the collision of terms between the COCA, IMDB, and Google databases, it is strictly evident that each follows a similar trend in emergence. With the experimental input data, peaks of computer and hacker jargon appear in the 1985-2000 range consistently in each of the independent databases. Graphed data is fed from the previously constructed *n*-gram frequency tables to construct a graph of phrase collisions against time.

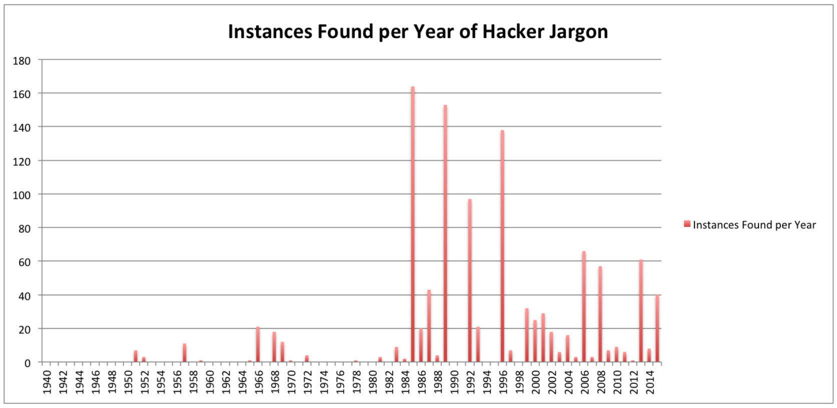

Fig. III.1. Demonstrating the crawler using the Jargon dictionary file against online sources

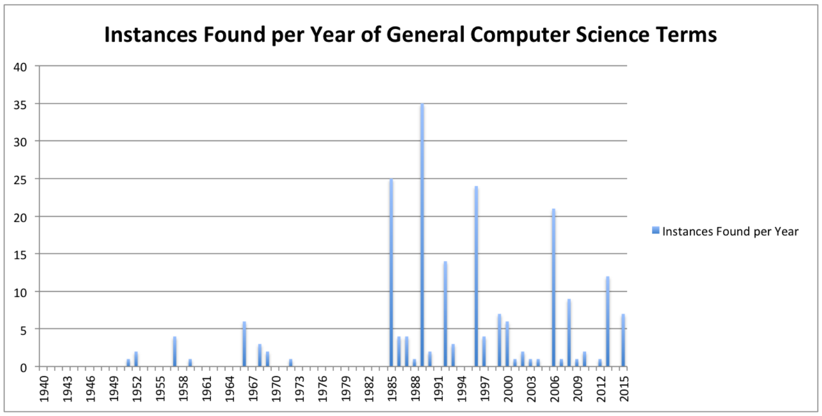

Fig. III.2. Demonstrating the crawler using the Computer Science terminology dictionary file against online sources

CHAPTER IV

CONCLUSION

Through this experimentation is becomes evident that topic emergence in popular media is reliably and consistently demonstrated in movie, text, and online sources. We find that although these domains are notably distinct in medium, they each reflect the current topic trends in popular culture. As such, the compounding of this emergence data (particularly through the n-gram approach) provides a technique of visualization that properly gauges and quantifies the strength of a topic at a given point in time. The *n*-gram storage technique provides a versatile scaffolding upon which to generate visuals: its ability to record and reproduce time-coordinated frequency data for phrase contexts rather than words is elemental in the approach to accurately illustrate topic popularity. Populating the internally-held data store proves to be quite trivial. Phrase data gathered from the three experimental sources is transformed gracefully between the original and destination data stores, a conversion process that is inexpensive in compute time and efficient as a means of representing entire libraries of knowledge. As such, the method is easily extensible to additional language data sources and provides the framework for visualization of any similar inputs.

REFERENCES

[1] James Abello, Peter Broadwell, and Timothy R. Tangherlini. Computational folkloristics. *Commun. ACM*, 55(7):60–70, July 2012.

[2] Erik Brunvand. The heroic hacker: Legends of the computer age. *Brunvand, JH, The truth never stands in the way of a good story*, pages 170–198, 2000.

[3] Pdraig Mac Carron and Ralph Kenna. Universal properties of mythological networks. *EPL (Europhysics Letters)*, 99(2):28002, 2012.

[4] Sara Graça da Silva and Jamshid J. Tehrani. Comparative phylogenetic analyses uncover the ancient roots of indo-european folktales. *Royal Society Open Science*, 3(1), 2016.

[5] Iwe Everhardus Christiaan Muiser and Mariet Theune. Visualizing the dutch folktale database.

[6] Saul Schleimer, Daniel S Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85. ACM, 2003.

[7] Timothy R Tangherlini. The folklore macroscope. *Western Folklore*, 72(1):7–27, 2013.

[8] Scott Weingart and Jeana Jorgensen. Computational analysis of the body in european fairy tales. *Literary and Linguistic Computing*, 28(3):404–416, 2013.