Natural Language Processing of Corporate SEC Filings for Political Influence Detection

Joey Hejna[1 2*], under supervision of Andrew B. Hall[4]

Abstract
Despite recent developments in technology, little is known about the means or ways in which corporations interact with American government. Andrew Hall, Associate Professor at Stanford University, suspects that the majority of capital corporations spend on the political sphere is outside of lobbying. In this paper, we take a numerical approach to analyzing corporate SEC filings to try and draw correlations between the language a firm uses and changes in government regulatory policy over time.

Keywords
Corporate Policital Influence — SEC Filings — Natural Language Processing

[1] *Research Intern, Political Science Department, Stanford University, Stanford, California, U.S.A.*
[2] *Los Altos High School Class of 2017*
[3] *Stanford University Summer Session 2016, Stanford University, Stanford, California, U.S.A.*
[4] *Assistant Professor of Political Science, Stanford University, Stanford, California, U.S.A.*
*__Corresponding author__: joey.hejna@gmail.com

Introduction

Large corporations with large capital resources may exert political influence through channels other than typical lobbying. At Stanford University's Political Science Department, a team of graduate students and undergraduates led by Associate Professor Andrew B. Hall is researching the relationship between corporations and legislation through the analysis of large text based datasets. Their goal is to examine temporal correlations between the language used in SEC Corporate filings and legislative changes affecting corporations. Identifying linguistic trends may reveal part of a larger narrative of corporate tendencies and whether specific legislation alters industry behavior, ultimately increasing public transparency. This research effort demands significant data mining and text processing techniques to analyze the 1.3 Terabytes of 10-K and 10-Q filings obtained from the U.S. Securities and Exchange Commission. I worked as the group's primary programmer, decomposing the filings into language datasets and creating a data visualization tool. This paper describes the systems I developed and prototyped which will be used by Stanford's research group for their multi-year project and hopefully later made publicly available for other academic researchers.

1. Parsing SEC Files

The basic approach was to format, parse, and then split the text files from corporate filings into n-grams, units of varying numbers of words, in order to make statistical correlations against time, industry, or firm. Initial analysis of the text files was completed through Python's Natural Language Toolkit. Extraneous information, including tables, HTML code, and stop words (like "the", "a", and "and") in addition to sections of the filings irrelevant to our analysis were eliminated and all words were stemmed. Key information, such as firm location and industry, were parsed using regular expressions. The end goal was to minimize file size to decrease computation time. Python's multiprocessing library was used for multi-threading to speed up the process and map functions were utilized to ensure our list operations were backed in C code.

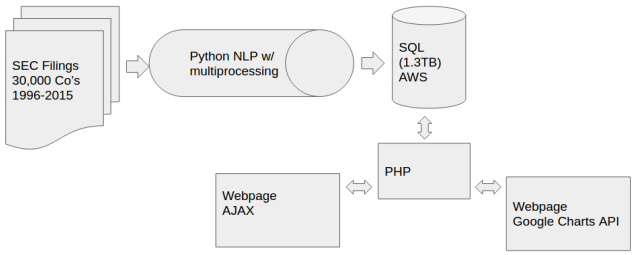

Figure 2. Architecture for Visualizing Word Tokens

2. Bigram Dataset

I created a sparse-matrix style dataset of bigram (word pair) frequencies for graduate students working in Andrew Hall's lab hoping to determine if certain keywords can predict a firm's political affiliation (right or left) and thus their donation tendencies. I first took the extracted information for each given firm and collapsed all of their filings into a single profile containing only the bigrams that the firm had used in greater than 10% and less than 90% of their filings. The final list of predictor bigrams was derived through a second elimination step, taking only bigrams that existed in more than 15% and

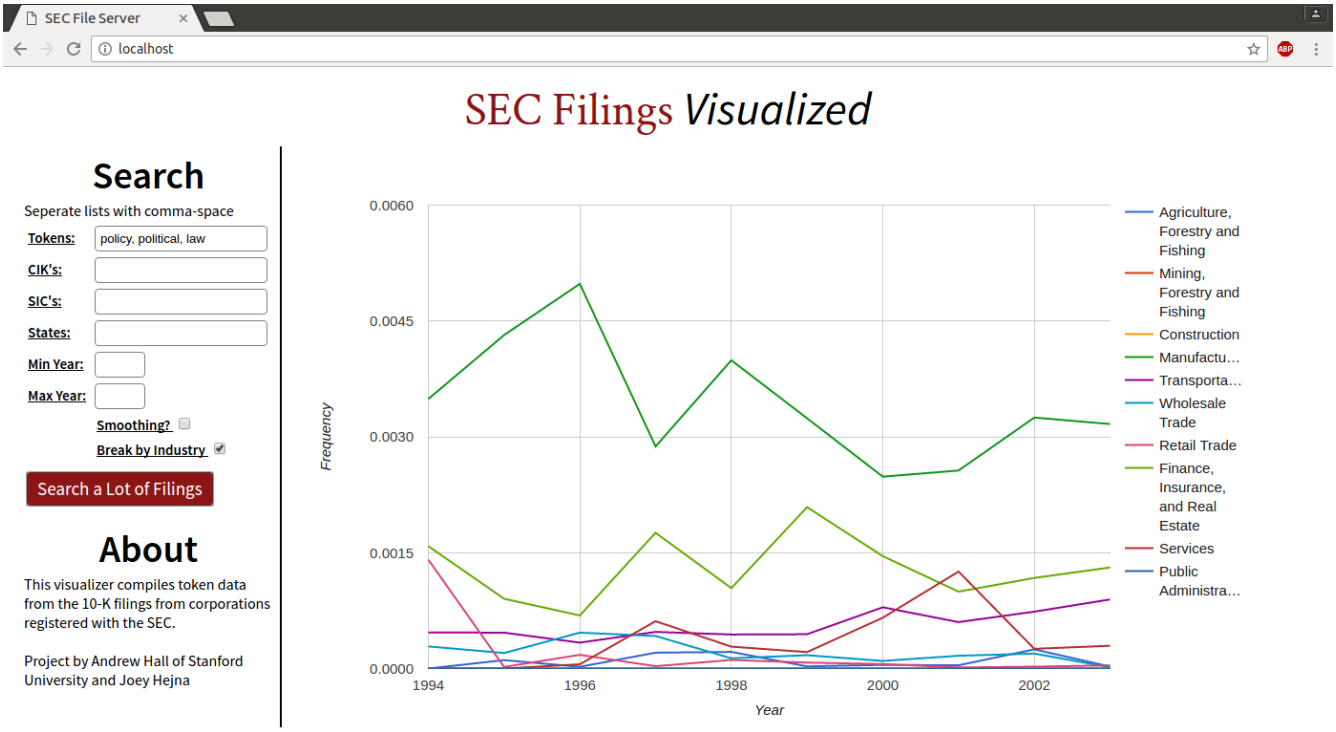

Figure 1. Web-Based Research Interface for Visualization of Word Tokens

Firm	Industry	SIC	CIK	ZIP	# Bigrams	secretari general	debt instrument
Technitrol Inc.	Electric Light & Wiring Equip	3640	96763	19053	103756	0	18
Transtechnology Corp	Cutlery, Handtools & General Hardware	3420	99359	7938	173881	12	12

Table 1. Table of Bi-Grams by Company (The dataset would continue to the right.)

less than 85% of the bigram "profiles" for each firm. A second pass over the files resulted in frequency counts for each bigram by firm.

3. Web Visualizer

Modeling the architecture of Google's N-grams, an online tool that allows users to query word frequencies in the entire Google Books database, I created a prototype web visualizer for word frequencies in SEC filings. Taking the preprocessed files analyzed in Python, I created a multi-table SQL schema. Tables were created for sections of the alphabet in order to decrease query time. Each table was indexed appropriately and only had entries for a firm's CIK identifier, year, token, and the token's count. Frequencies are later normalized by dividing a token's count by the number of tokens in a given year's filings. Based on the CIK identifier, extended information on the firm could be crossed checked with a different table containing firm metadata. I spent hours optimizing the search time and indexes. The database, which was integrated with AWS, interacted with a JavaScript based web page that utilized asynchronous requests (AJAX) in order to allow the user to customize queries. Users can query based on industry, time, and location.

4. Conclusion

In this project, we are attempting to uncover temporal correlations between language in SEC corporate filings and legislative changes that affect corporations. Correlations may be in either direction: i.e. frequent use of language in corporate filings may preceed a legislative change, or corporate filings may contain new language in response to legislation. Graphing the frequency of word tokens and pairs over time gives researchers the ability to see trends across companies, industries and events.

Acknowledgments

I would like to thank Assistant Processing Andrew Hall and his research team for giving me the opportunity to contribute to their research project.