CS372 Project Interim Report: How Does a President Talk Their Economy into Success?

Youngin Lee 20170473_Youngin Lee
Haksun Son 20170340_Haksun Son
Sejun An 20160351_Sejun An
Akkanit Pornpattanadul 20180815_Akkanit Pornpattanadul

1 Introduction

Presidential speeches are an important indicator of a government's policies and what they think is important and should be focused on. However, it is hard to quantify the effort a president is willing to put into a certain topic just by listening to each speech.

This project aims to build a model that can quantify the emphasis on a certain topic, then affirm that this emphasis is carried onto real life. Specifically we aim to examine the emphasis on the American economy within presidential speeches, then compare the output of the model to real-life economic indicators such as growth rate, GDP and trade. Economics is a good field to observe, as there are many indicators that measure the growth.

2 Problem Description

Given a set of presidential speeches, categorized by the year they were given, we aim to build a model that can extract a vector quantifying the emphasis on economy in those speeches. We validate our model by examining the goodness of fit between our output and various economic indicators in real life, such as GDP growth and unemployment. After the validation of the model, we can check the parameters of our model to confirm a positive or negative relationship between the emphasis and economic growth. Finally, we can conduct separate validations for the Democratic Party and the Republican Party to explore whether there is a difference in the emphasis on economy between the two parties.

3 Technical Approaches

3.1 Quantifying the Emphasis on the Economy

Our main approach to quantifying the emphasis on economics in our corpus will be investigating the frequency and distribution of words related to the economy. The quantifying model works in three steps; finding groups of keywords related to the economy, searching and counting the relative frequency of those keyword groups, and modeling the distribution of keyword groups.

Finding keywords: To find keywords that are relevant to the economy, we first start with a list of seed words. Using WordNet, we find the synset that each words belong to, then find all hyponyms, direct and indirect, of this synset. We can collect the lemmas for all of the synsets, which will result in a list of searchable keywords. We also add the lemmas harvested from the sister synsets of our original seed word and words that are frequent in the definitions of the relevant synsets. Finally, for each lemma, we must consider the different forms of our keywords. We do this by adding any of the keywords' related forms, namely the derivationally related forms and pertainyms defined in WordNet. It is important that we group our keywords by the seed word they originated from, as we can perform a more fine-grained analysis of different aspects of the economy; for example, seed words such as 'work', 'price' and 'trade' can be indicators of employment, inflation and international trade.

Counting frequency: After finding keywords to search for, we can search our documents for the frequency of these words. After we search for our keywords, we can aggregate the counts separately for each group of keywords. Because we want to compare our results between multiple documents with differing lengths, we use relative frequencies instead of absolute. The frequency for each keyword group becomes a dimension of our output vector.

Modeling distribution: In order to conduct a more sophisticated analysis, we can look at the distribution of the groups of keywords. We divide the document into a set number of parts, such as three for introduction, body and conclusion, and count the frequencies separately for these parts. The frequency counts for each part of the document becomes a dimension of our output vector.

3.2 Verfiying Our Quantifying Model

After we construct a model to quantify the emphasis, we can validate this number by performing linear regression against the economic indicators in the second part of our dataset. To increase the accuracy of the linear regression, we can use various solvers such as L-BFGS and SAG. We can evaluate our results by the goodness of fit between our metric and economic indicators, measured by the p-values and the R^2 values. In addition to the measures of goodness of fit, we can also examine the parameters of the model, to see whether the emphasis correlates with economic growth in a positive or negative way; in other words, we can see whether the emphasis on economy is a result of the president's policy priorities or a result of an economic depression posing a more pressing problem.

3.3 Comparing the Republican Party and the Democratic Party

We also plan to conduct an experiment to compare the Republican Party and the Democratic Party in terms of the emphasis on economics in their speeches. We plan to use linear regression separately on the two parties and observe any differences from the original models. From these attempts we can look for differences in the distribution, as in the mean and the variance, in the two parties.

4 Preliminary Results

We conducted our preliminary results using the Inaugural corpus in nltk, and the GDP growth rate for the US economy in the inaugural year of presidents. In our final work we plan to expand our corpus to presidential speeches divided into years.

4.1 Deciding seed words

Finding the appropriate keywords is crucial in order to obtain meaningful information on the emphasis of economy. Our selection of seed words are one of the most influential factors in this crucial step; therefore, we investigated whether the number of seed words really influences the experimental results and whether the seed words should be related to economic terms.

We experimented with eleven seed words, where eight words had a very close relationship with the economy (economy, industry, work, trade, worker, interest, money, market). The remaining three words were more ambiguous in its relationship to the economy (opportunity, success, government). Starting with terms that are closely related to the economy, the mean square error was measured by adding seed words one by one. The results were as shown in the figure below.

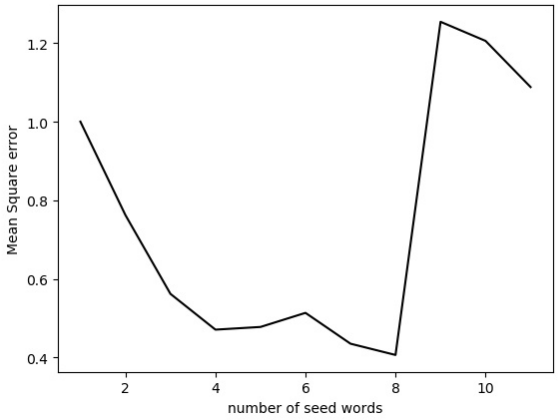

Figure 1: Number of seed words and mean square error

For the first eight seed words, which were relevant to the economy, the error tend to decrease as more seed words where added. On the other hand, when terms unrelated to the economy where added, the error increased significantly.

From this result we can infer that using appropriate terms related to the economy is important, as well as using multiple seed words to cover all aspects related to the economy.