NLP Analysis on Obama Speech

Modern Data Analytics[G0Z39a]

Prof. De Spiegeleer Jan

Chaojie Huang (r0816192)

08/26/2022

KU LEUVEN

1 Project Process

1.1 Web Scraping

Firstly, BeatutifulSoup is a powerful web scraping framework in python, which is the method I used in this project to download PDF files. Then we use PDFplumber framework to extract text from PDFs. Next in text preprocessing, I used regular expression to remove speech headers and footers. What's more, I corrected missing values and selected speeches that were given during Obama's tenure.

1.2 Data Processing

In the process of data cleaning, I took six steps, including removing stopwords, removing punctuations, lowering words, tokenizing, lemmatizing and extracting nouns.

1.3 Topic Modelling

1.3.1 Top2Vec

Top2Vec is an algorithm for topic modeling. It automatically detects topics present in text and generates jointly embedded topic, document and word vectors. By using package top2vec, I built this model that automatically detected the number of topics was 4. This technique was just a preliminary analysis of topic modeling to give me insights into the best number of topics in following K-means Cluster and LDA method

1.3.2 K-means Cluster

K-means belongs to Unsupervised Learning. It aimed at grouping similar data points together and discover underlying patterns. It identifies k number of centroids, and then allocates every data point to the nearest cluster. Package Cluster was used in this method. By using elbow method to find the optimal number of clusters, I picked 9 topics. However, the most common words in cluster 3, 5, 6 are the same, which means the result of this topic model is not very credible.

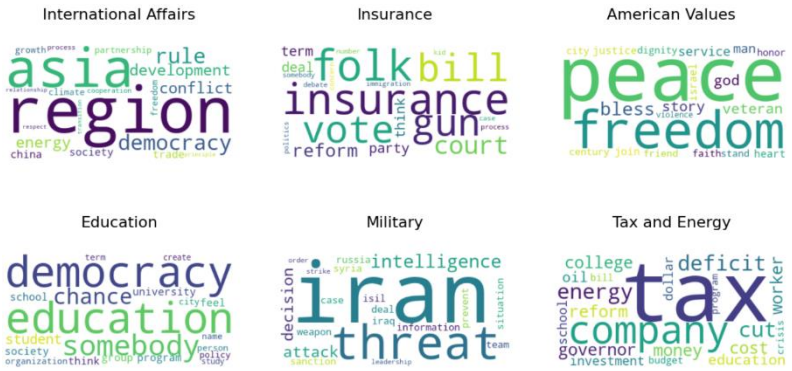
1.3.3 LDA- Latent Dirichlet Allocation

(LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. The package mostly used for LDA is 'gensim' To

find the optimal topic number, I calculated coherence score and then picked 6 topics. The topic visualization by package pyLDAvis shows that it's a good topic model.

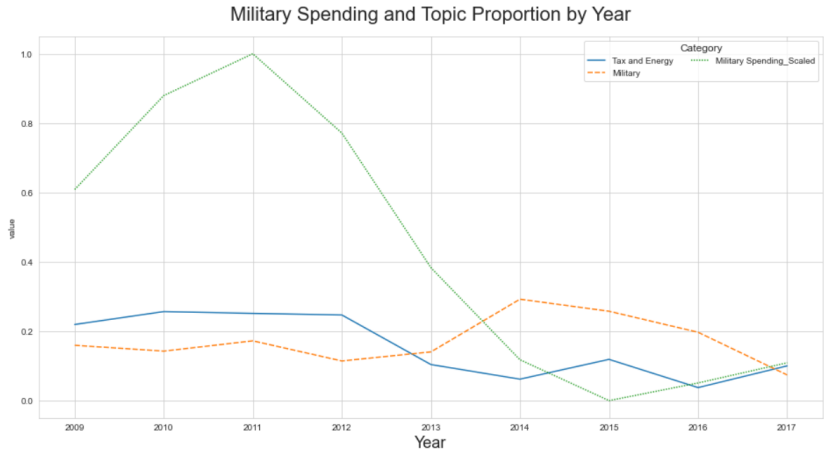
2. Results

2.1 Word Cloud of topics



2.2 External Data

According to topics shown, I added external variables which are military spending, uninsured rate, tax revenue, energy consumption and Obama support rate. I compared the topic proportion and scaled external data. The following figure shows some examples of trends between the proportion of speech topics and extra variables like US military spending.



3. Conclusion

In my analysis, Obama's speeches can be mainly divided into six topics. And during his tenure, Obama gave a lot of speeches on American values. In addition, there are similar trends between topics and external variables based on speeches given by Obama.