ETA – Final Project Report
Guangda Zhu (gz6xw@virginia.edu)
DS 5001
April 30 2020

Introduction

As news swarm outs from different sources, the impact of these different sources on the news that are reported becomes mysterious. Could there be any patterns hidden behind the content of these news? Have these patterns changed throughout the past years? This project aims to use Text Analytics models and tools to investigate on the differences in news from various sources and check the impacts of time on the news reported.

Data Description

The data used in this project comes from the newzy file on Box folder. It collects the news content from various sources such as Google News, PowerLine, Guardian, etc. Each of the entry includes a doc_id, sources, title, content, the date it was published, and the url on which it is posted. The sources of these news can be liberal, conservative, or somewhere in the middle. To conduct this study, we select the two sources that have the full content, which are PowerLine and Daily Kos. We randomly select 1000 news from each source and form the final corpus that is analyzed.

Powerline, started by three lawyers, has shown a leaning towards the conservative party, whereas Daily Kos is mainly focused on center-left liberal politics news. The differences in the political stances of these two sources could potentially lead to different point of focus in their news, and thus attracting different audience.

Preprocessing

The data seems to come with some missing values. To clean this up for the study, we remove the rows where doc_content is missing. The content of PowerLine news all start with the author's name in the parenthesis, while Daily Kos news does not include such information. To process the content so that news from both sources are including same level of information, we first remove the author name from PowerLine news. We further split the date column into year and month to look into whether the time in the year would affect the news reported.

Tables

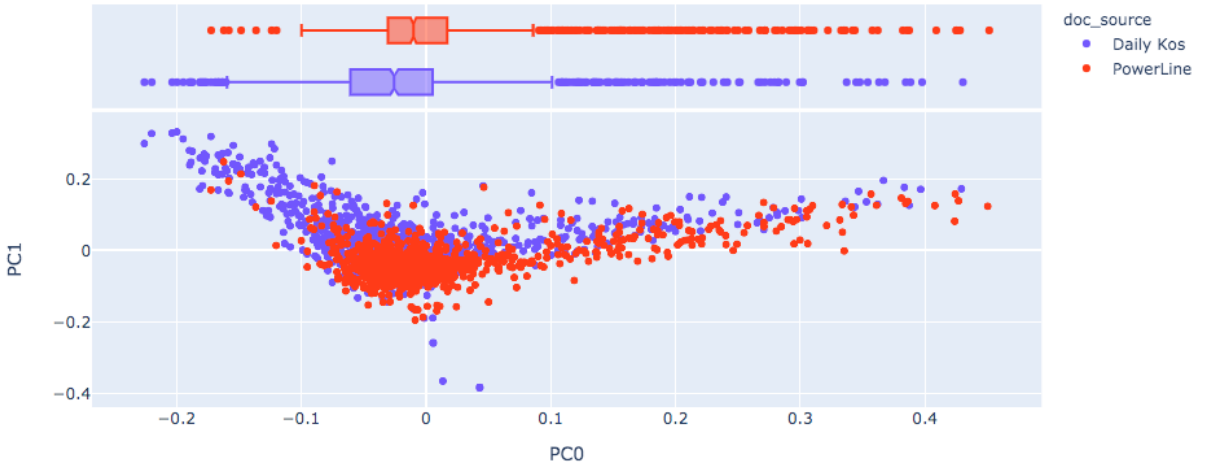
The LIBRARY table consists of information on the news level, which includes the doc_id, title, year, month and the source. We further split this into the sentence level for DOC table, and token level with part of speech information for TOKEN table. We then convert the TOKEN table into VOCAB table by counting the frequency and assign Boolean value for differentiating the terms that are numbers from the rest. We also annotate the VOCAB table with stop words and stemmer information.

Methods – PCA

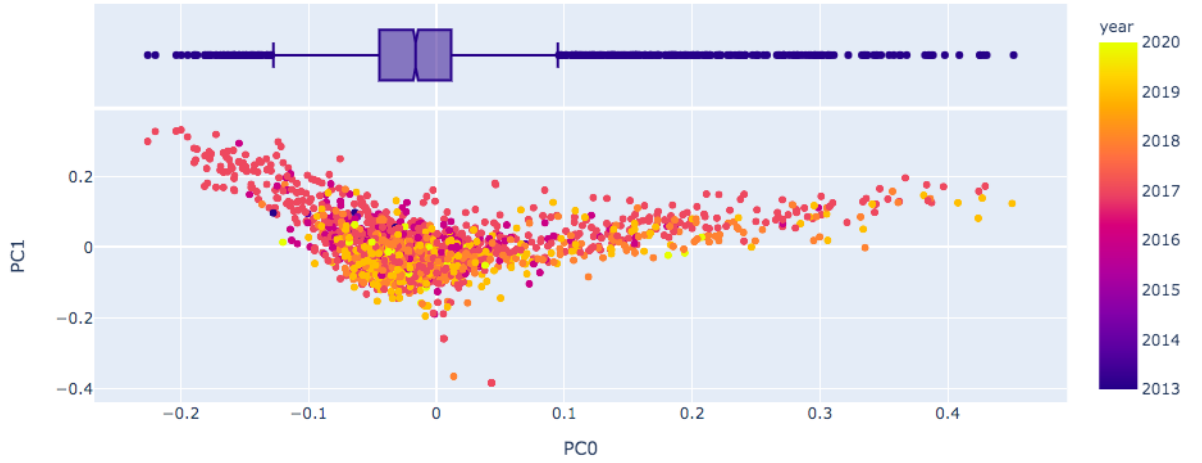
The PCA model is first ran on the corpus to check for the meaningful components. For comparison reason, I picked the first two components and check out the words that are of positive and negative relationship with them. For PC0, the words such as health care, medicaid, tax are of positive relation, whereas fbi, Russia, investigation are negative. For PC1, the positive words are: twitter, mattbors, omar, etc, and the negative words include senate, house, comey.

```
docs PC0+ health obamacare bill care insurance republicans medicaid tax percent you
docs PC0- fbi comey investigation mueller russian russia intelligence flynn nunes trump
docs PC1+ mattbors follow twitter facebook she her omar my me jensorensen
docs PC1- republicans obamacare bill senate health care comey house republican medicaid
```

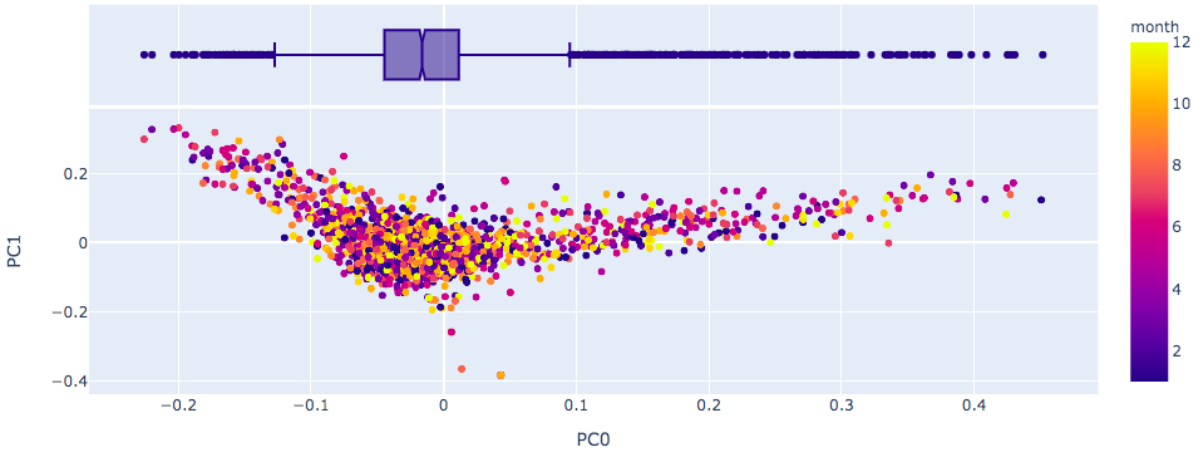
To further split these by the sources, we make the following plot, from which we can see that both Daily Kos and Powerline seem to cover the same region regarding the principal components. The data points from the below scatterplot overlap in many area.



Moving on to see how the time factor affects these components. From the graph it seems like the more recent years such as 2020 and 2019 have smaller value along PC1, as compared to 2016. The year of 2016 is also more spread out along PC0 as compared to 2020.

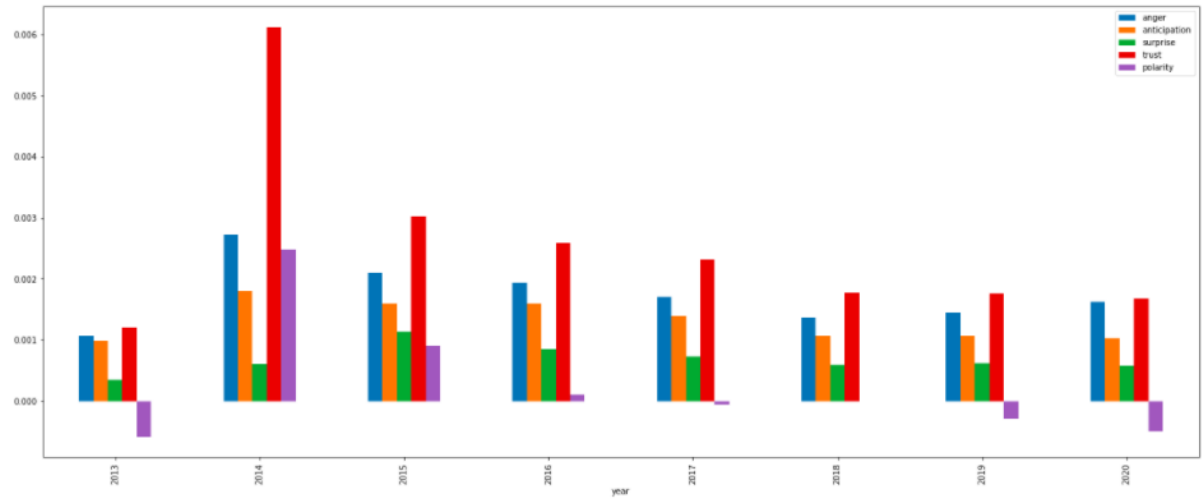


If we split these into month, points from different months seem to mingle together pretty well and no clear patterns stand out from this plot.

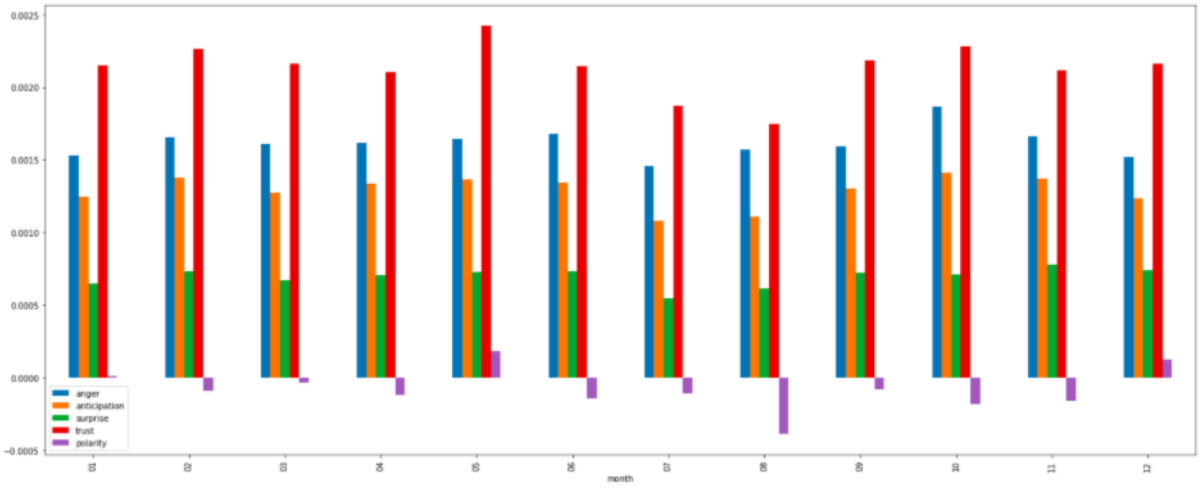


Modes – Sentiment Analysis

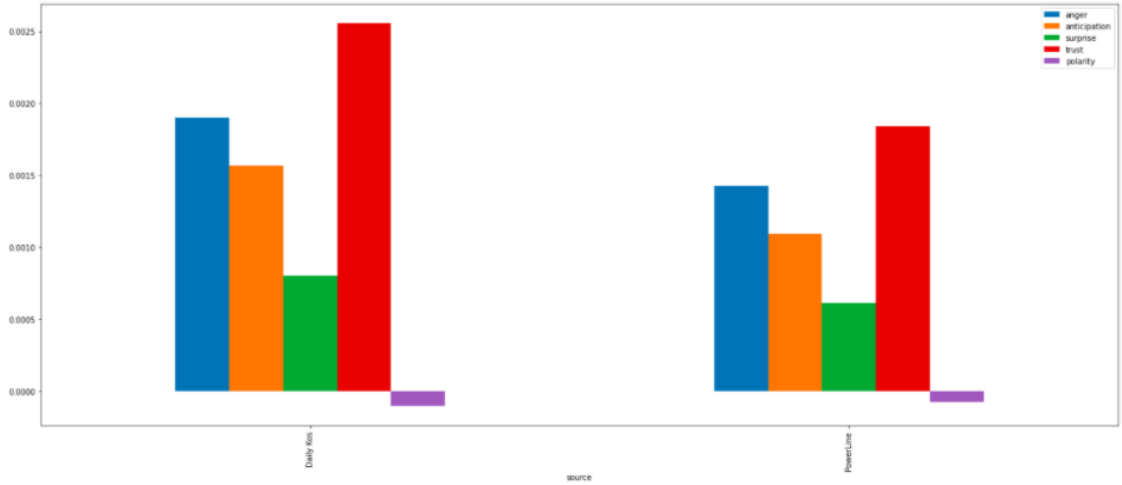
Out of the emotions we learned in class, I pick the four that seem to have more extreme feelings, which are anger, anticipation, surprise, trust and the corresponding polarity. The follow graph shows how these emotions evolve throughout the year. While the polarity goes up and down in the past decade, the four emotions seem to remain in a relative order, with trust being the largest value, followed by anger, anticipation, then surprise. The trust level strikes in 2014, which could be caused by some historical event that happened around the time.



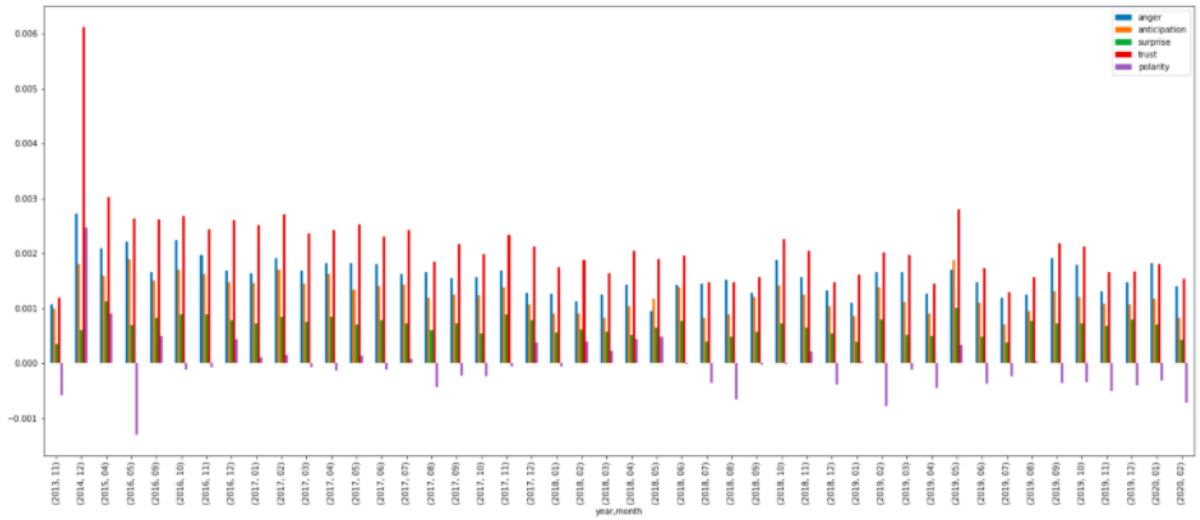
If we group by the month and take the average across different years, we can see that the emotions shown do not fluctuate so much. However, if we look closely, it seems to follow a curve that goes up around May, down around August, and then go back up near October. One hypothesis is that the emotions fluctuate in a half-year cycle, where it goes up every five month and then drop down. This would need to be further testified by more research, but logically speaking, it makes sense for emotions to fluctuate in a certain pattern slightly as time goes.



To compare the emotions across the two sources, we can see that overall Daily Kos, the liberal news sources, reveals more emotions shown in the news, as compared to the conservative Powerline, for all the categories.

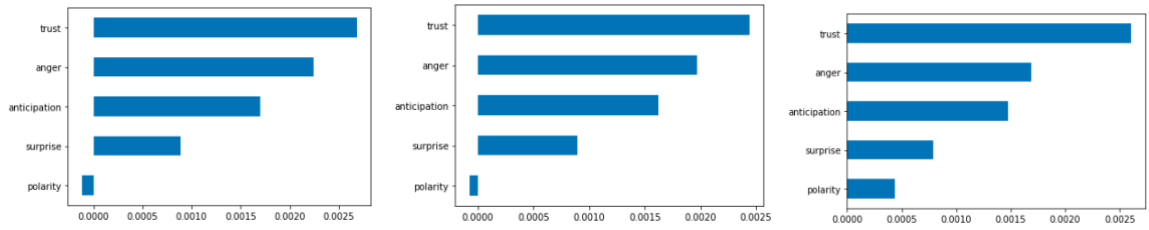


When I split the time into year and month to see how emotions have evolved in the past years, we can see that seemingly the emotions do not change drastically.

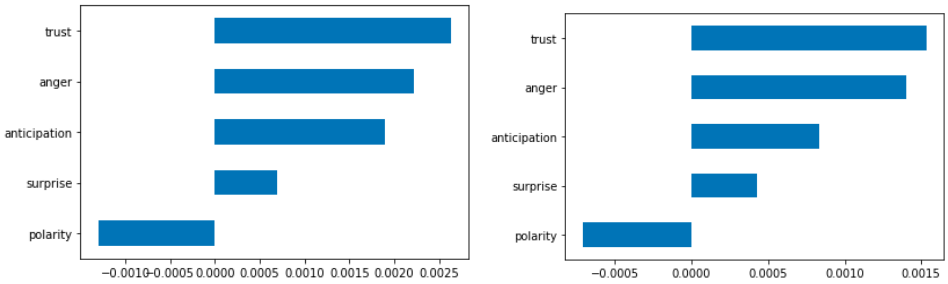


The last election happened on Nov,8 2016. To compare the differences in the situation, I check the sentiment from the months around Nov 2016, including October, November and December. To also see how the sentiment is built up across the country before elections, I select the month of 2020-02 and 2016-05. Ideally the comparison should be done for the same month that are further ahead of the election, however, due the limitation in the data set, these two months are chosen to represent the sentiment far ahead of election during that year.

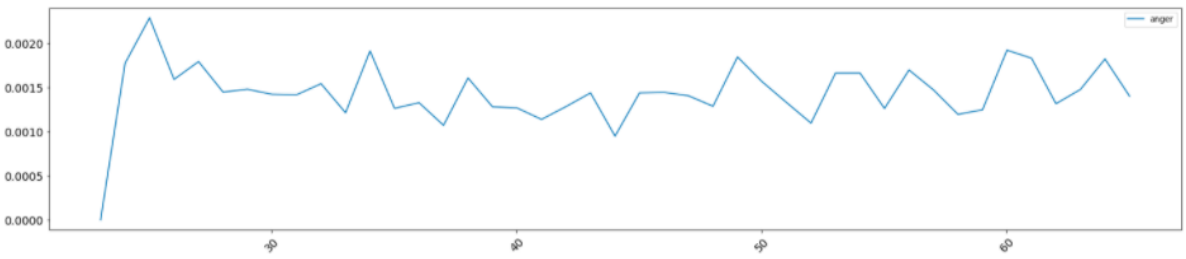
The following three plots show the emotions of October, November, and December of 2016. It seems like the five emotions did not change much and remained in the same order, with the highest being trust, then anger, anticipation, surprise. One thing that changes the most here is the polarity. We can see that from October and November, the polarity are both negative, whereas in December, it increases to around 0.0005 in positive. This could mean that there are more positiveness in the news compared to negativeness after the election.

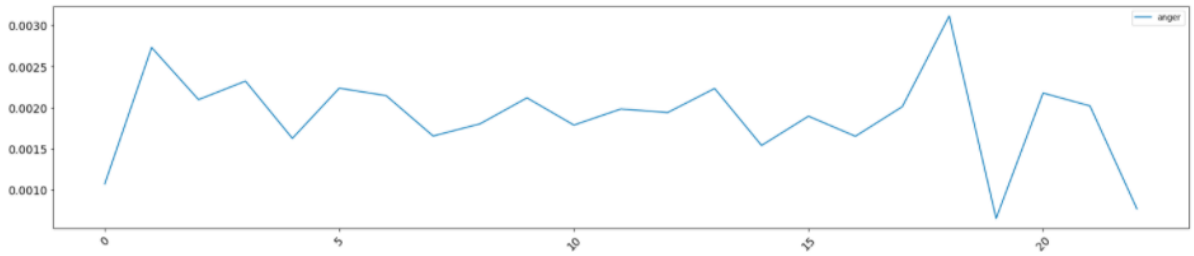


The following two graph are from May 2016 and February 2020. It seems like even for different month, people's emotions are rather similar before elections when its further away from election. After the election this year, it would be interesting to see if these emotions changes the same way as they did in 2016.



Below is how the emotion 'anger' changes across different years for news from PowerLine (top) and Daily Kos (bottom). While PowerLine has a more flat line, Daily Kos shows more of a fluctuating curve recently.



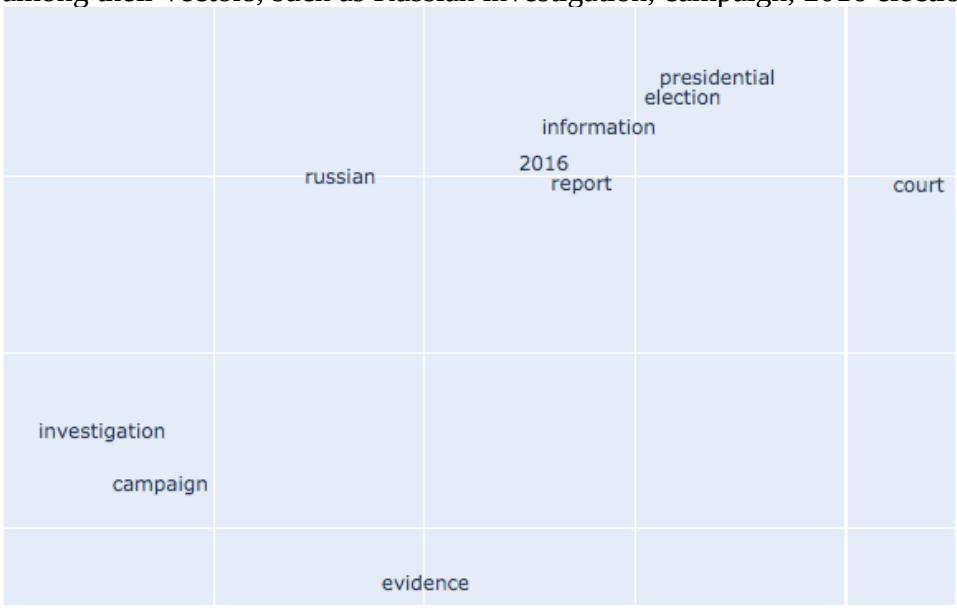


Models – Word Embedding

To further see the differences between the two sources, we purposely choose to make two embedding models, one for each source, and compare their results. We compare the results using the analogy with words ‘2016’, ‘election’, and ‘president’. The PowerLine model gives ('let', 0.9996554851531982) and ('policy', 0.9996533989906311), whereas the model for Daily Kos gives ‘2017’ and ‘said’.

This is quite interesting in that the powerline model is able to capture the role of president and link it to the word policy. On the other hand, the year 2016 was represented in the vector space and is matched to the year 2017.

Also, to take a look at the scatterplot of these vector representations, we can see that some of the interesting words that usually are referred together also have relatively small distances among their vectors, such as Russian investigation, campaign, 2016 election, evidence, etc.



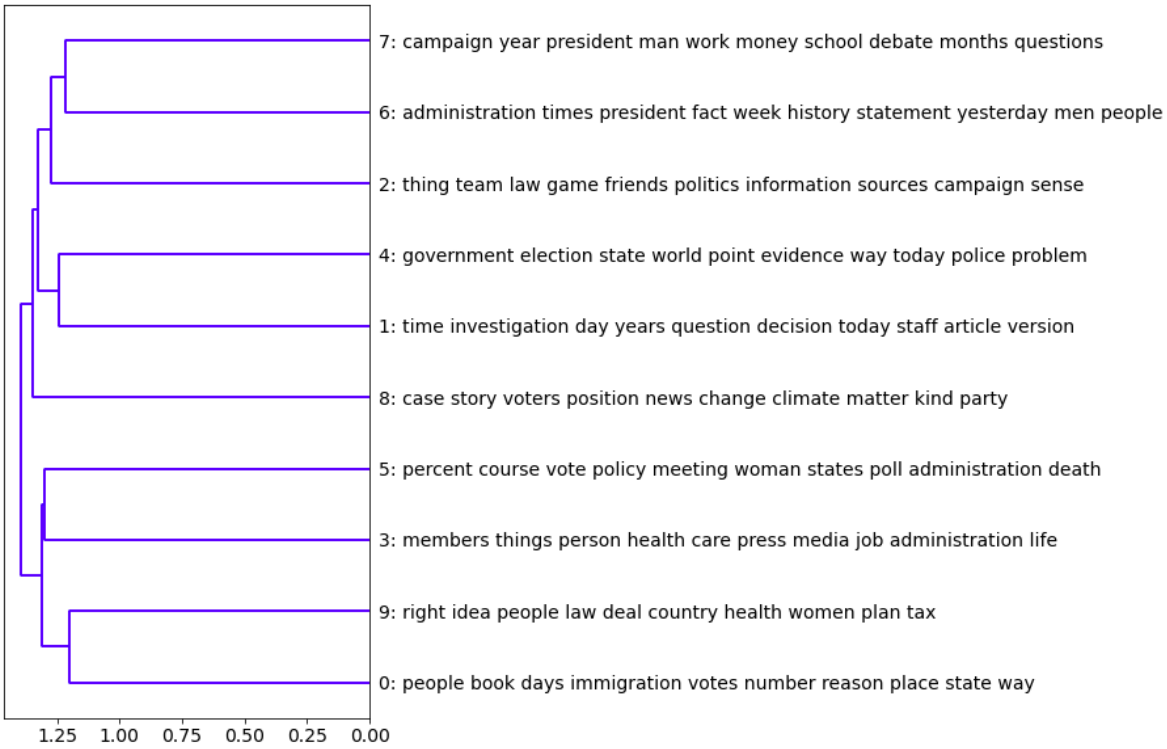
Models – LDA

For topic modeling, we use LDA model from Scikit-learn to generate ten topics from the corpus. The following table shows each of these topics and the popular words from each topic. As we can see, some of the topics talks about immigration, votes, investigation, campaigns, etc. the word administration appears in three topics, showing that it is quite a common theme among different topics in news. Some other words that are popular across topics are president, votes / voters, and people.

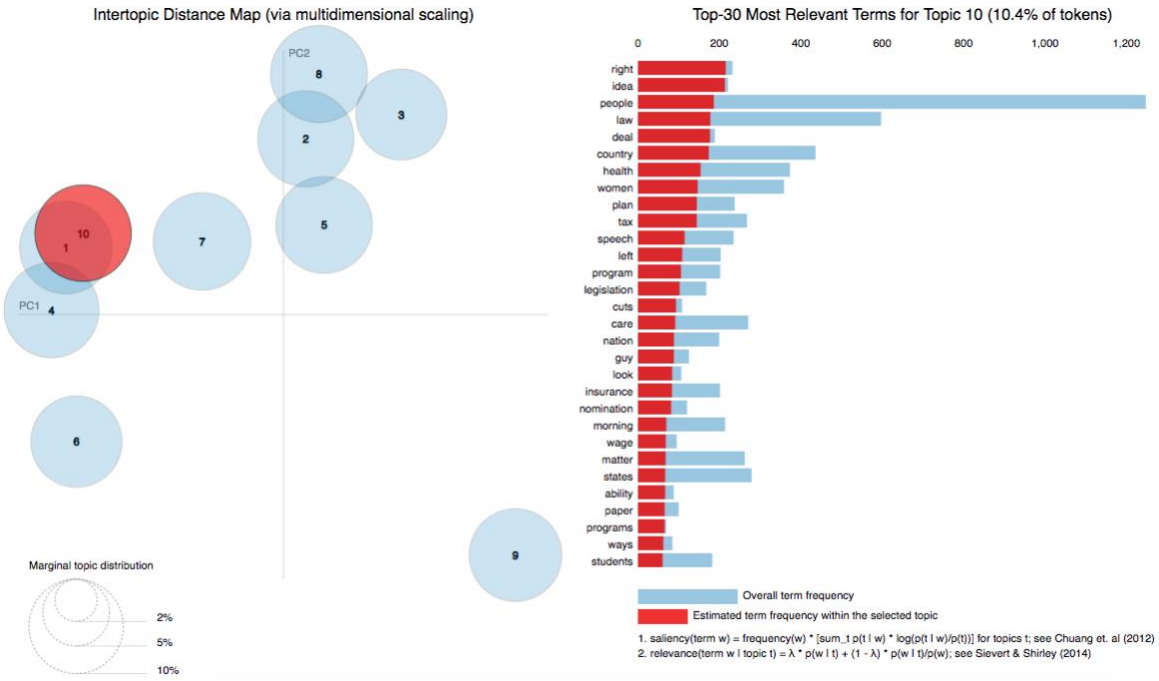
term_str	0	1	2	3	4	5	6	7	8	9
topic_id										
0	people	book	days	immigration	votes	number	reason	place	state	way
1	time	investigation	day	years	question	decision	today	staff	article	version
2	thing	team	law	game	friends	politics	information	sources	campaign	sense
3	members	things	person	health	care	press	media	job	administration	life
4	government	election	state	world	point	evidence	way	today	police	problem
5	percent	course	vote	policy	meeting	woman	states	poll	administration	death
6	administration	times	president	fact	week	history	statement	yesterday	men	people
7	campaign	year	president	man	work	money	school	debate	months	questions
8	case	story	voters	position	news	change	climate	matter	kind	party
9	right	idea	people	law	deal	country	health	women	plan	tax

If we split up the corpus by the sources of these news, we can see that the topics orders are also different. For Daily Kos, the most common topic distribution consists of words such as idea, law, country, health, women, plan, and tax. For PowerLine, it is investigation, question, decision, staff, article, version, etc. Having the different distributions show that these two news sources do tend to have different emphasizes when reporting news.

We then try to cluster the topics together by using the following dendrogram.



We also try out the pyLDA graph to have a more direct look at the distance among different topics. The index starts at 1, which is different than the dendrogram where index starts at 0. However, we can still see that both plots show the same results in terms of the similarity across topics. For example, topic 0 and 9 (1 and 10 in pyLDA) are very close towards each other in both plots. Therefore in the dendrogram, they are also one of the first ones that can be clustered together.



Conclusion

In this study, we present the four tables – LIB, DOC, TOKEN, and VOCAB, and also apply different models such as PCA, word embedding, LDA, and semantic analysis. The original motivation was to see whether there are some differences between news from sources that have opposite political stance and how these news coverage change over time. Throughout the analysis, I do get to find out some interesting differences between the two sources PowerLine and Daily Kos. For future study, I want further check whether these differences are indeed caused by their political differences, or other reasonings.