Thomas Butler, vra2cf
DS 5001
12/14/2021

Grouping Similar News Sources

Introduction

Online News Articles are an important source of consistent information about current events. However, current events can cover a wide variety of topics such as sports, local news, business, politics, and health issues. Various news companies also cover these current events with different frequency and depth. These articles are subject to the various biases of the writer and company either through political views or through ownership by a parent company.

The purpose of this project was to discover underlying knowledge about various news sources and apply this to knowledge to determine similarities and differences between papers and see if this knowledge is related to known information/biases about these papers.

A dataset with 986,461 online news articles from 15 different news sources was gathered to determine which news sources are like each other, what similar topics do they cover, and the various emotions each news source is associated with from the average news article.

The date of each article ranged from 11-2013 to 2-2020. This period includes the 2016 election and the beginning of the COVID pandemic.

The news sources varied greatly from US news on both sides of the political spectrum, to World news, to New York Times, a newspaper of record. A newspaper of record refers to a newspaper that is authorized by a government to publish public or legal notices; it also means that the newspaper is viewed to have high editorial independence, high journalism quality and large circulation. In general, a newspaper of record is a high-quality source.

Data & Methods

The Newzy dataset was taken from the UVA box holding potential datasets for this project. The columns that had NA rows were removed which returned a dataset that had 986,461 online news articles from 15 different papers. These news articles have either the first sentence or a piece of the first sentence in all the articles except PowerLine and Politico Magazine, which have the full article included. The article date ranges, exact article counts and a general description I researched for News sources are shown below in Table 1.

News Source	Documents	Date Range	Description
Breitbart	10011	2013-11-03 to 2020-02-27	American far- right
CNN	36184	2013-11-03 to 2020-02-27	Nonpartisan
Daily Kos	12307	2013-11-03 to 2019-10-31	Liberal American politcs
Drudge Report	77063	2013-11-03 to 2020-02-26	US Conservative
Fox	58077	2013-11-03 to 2020-02-27	US Conservative bias
Google News	64078	2013-11-03 to 2020-02-27	Aggregator of other news sites
Guardian	12392	2013-11-03 to 2020-02-26	UK News
NPR	24708	2013-11-03 to 2020-02-27	US Conservative non-profit
New York Times	21145	2013-11-03 to 2020-02-27	Newspaper of Record, Liberal
Politico Magazine	3324	2013-11-19 to 2020-02-27	US Conserative
PowerLine	9730	2013-11-03 to 2020-02-27	US Conserative
Real Clear Politics	34967	2013-11-03 to 2020-02-27	US Conserative bias Politics focused
Reuters	18506	2013-11-03 to 2020-02-27	Global Business Professional
UPI Latest	97804	2013-11-05 to 2020-02-27	US and Global News
US News	506165	2013-11-05 to 2020-02-27	US Ranking Services

Table 1 – General Information about News dataset

A violin plot was created to determine general distribution of number of articles published per day with a median of 879 articles per day, shown below in Figure 1.

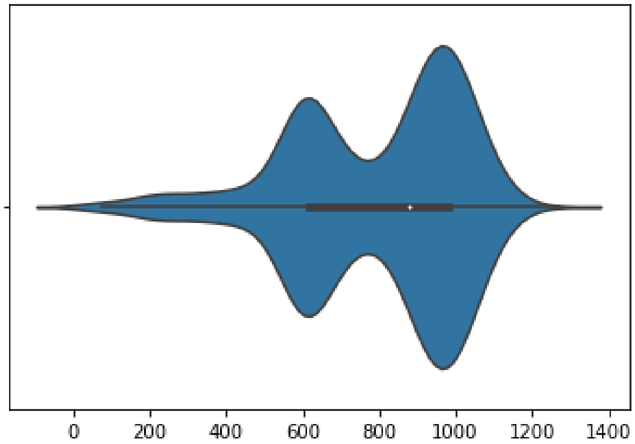
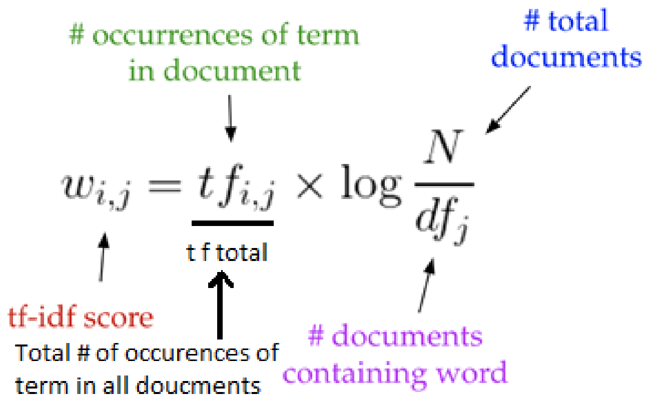


Figure 1 - Violin Plot of Number of Articles published per day.

I then cleaned and preprocessed the data into a OHCO format with a DOC, LIB, TOKEN and VOCAB table. The TOKEN table had Parts of Speech added with nltk.pos_tag. The VOCAB table had stop words identified, stems of words taken with nltk.stem.porter, and the most frequent Part of Speech tag attached to each word. The TOKEN table had term id added to the table based on the VOCAB table.

TFIDF (term frequency inverse document frequency) was calculated using the TOKEN table by creating a Bag of Words(BOW). Turning the BOW into a Document-Term matrix using the count of each token at the level of news sources. Then using the Document-term matrix one term per document / one term per all documents was done on each term * log(total documents/ documents containing word) = TFIDF. Figure 2 shows the equation written out.



# occurrences of term in document
# total documents
$w_{i,j} = \frac{tf_{i,j}}{\text{t f total}} \times \log \frac{N}{df_j}$
tf-idf score
Total # of occurences of term in all doucments
# documents containing word

Figure 2 - TFIDF equation

Basically, the more occurrences that term is in the document the greater the TFIDF is and the more times that word is in each different document the further TFIDF decreases all the way to 0 if the word is in every document. It is a calculation to determine how common this word is in this document discounted by every other document it is in. TFIDF will be used in a few models.

Model results

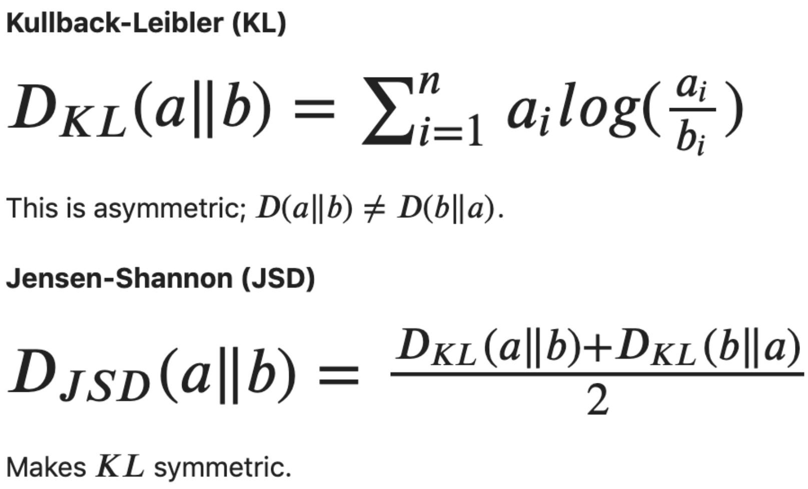
Hierarchical cluster diagrams

Distance between news sources was calculated using TFIDF by different metrics and plotted in a Hierarchical cluster. The height of lines denote distance from each node. Not all diagrams created will be examined, just cosine and Jensen-Shannon. The equations to calculated cosine and Jensen-Shannon are shown below in figures 3 and 4. The diagrams are in figures 5 and 6.

Cosine

$$sim(a, b) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} = \frac{a \cdot b}{\|a\|\|b\|}$$

Figure 3 – Cosine similarity equation



Kullback-Leibler (KL)

$$D_{KL}(a\|b) = \sum_{i=1}^{n} a_i log(\frac{a_i}{b_i})$$

This is asymmetric; $D(a\|b) \neq D(b\|a)$.

Jensen-Shannon (JSD)

$$D_{JSD}(a\|b) = \frac{D_{KL}(a\|b)+D_{KL}(b\|a)}{2}$$

Makes KL symmetric.

Figure 4 – Jensen-Shannon divergence equation

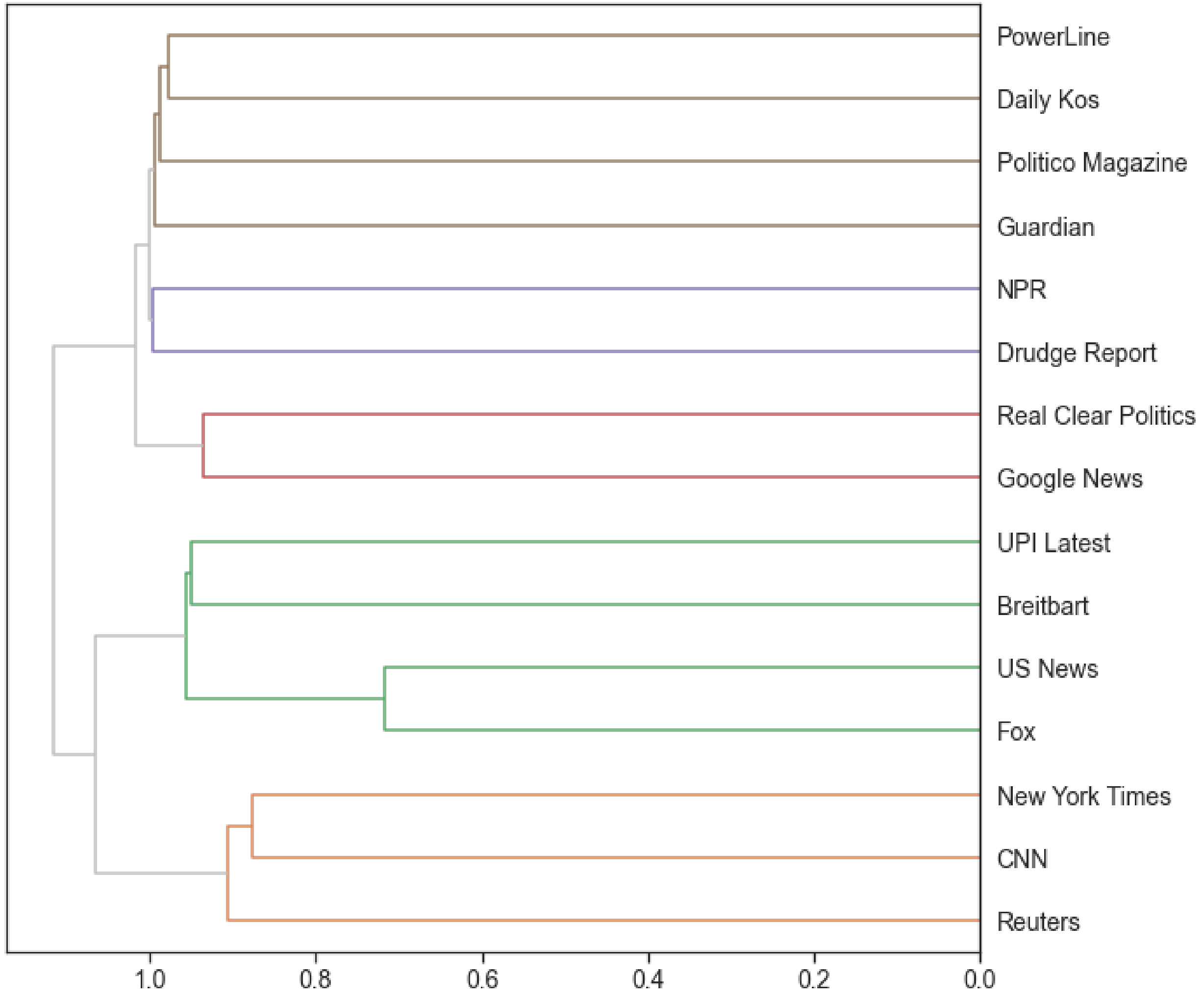


Figure 5 – Cosine Hierarchical Cluster Diagram

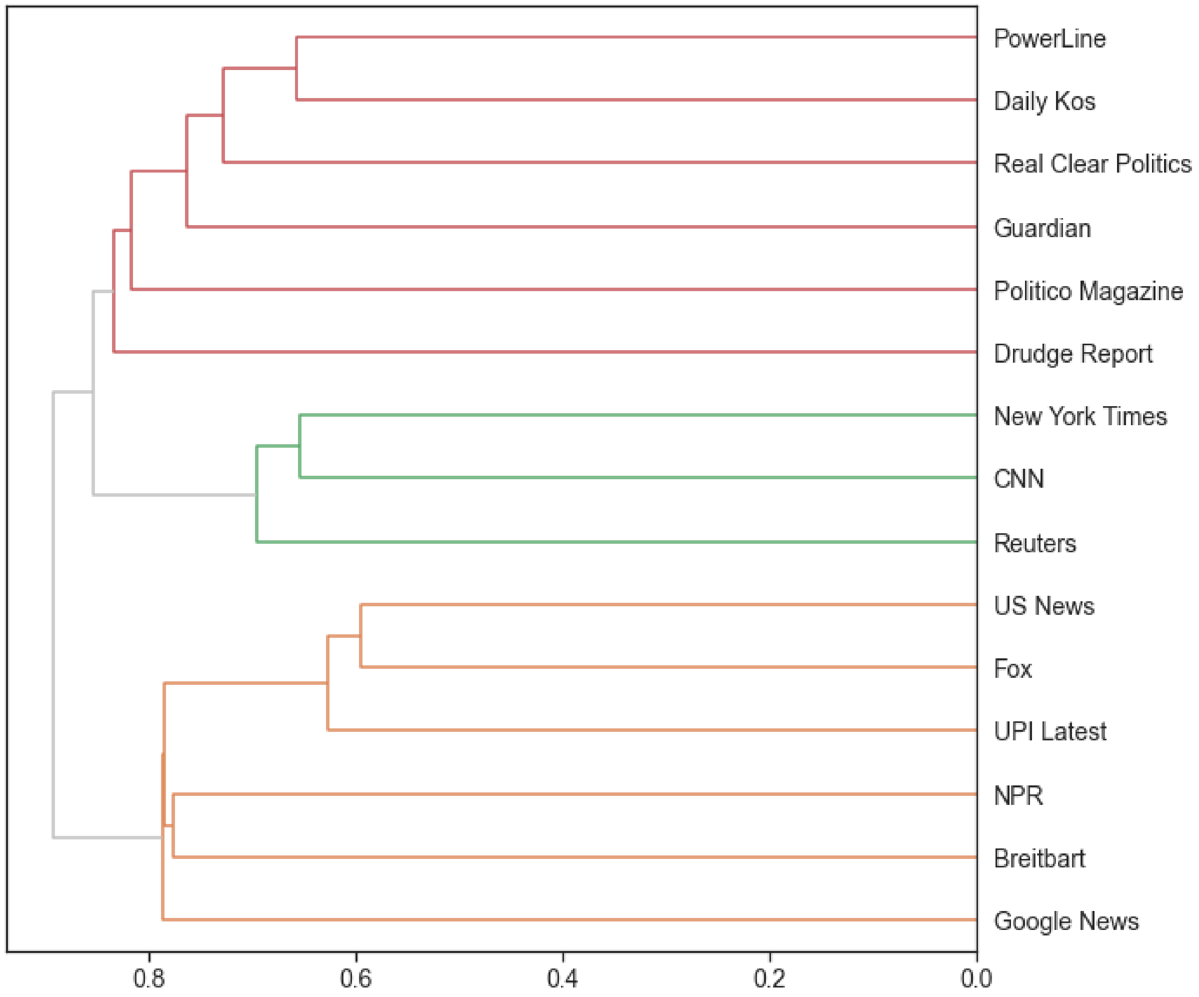


Figure 6 - Jensen-Shannon Hierarchical Cluster Diagram

Hierarchical cluster diagrams through similarity (cosine) and divergence (Jensen-Shannon). These two measurements were chosen because they were the most interesting clusters. NPR and Google News are the only real differences between the two but they still cluster in a like cluster. All the news sources clustered in similar political views with an insight into news sources with no found bias. The only outlier was Daily Kos which was in conservative clusters when it said it was liberal.

PCA

PCA was calculated using the largest 4000 TFIDF term sums from news sources. The data was normalized and centered, and the top 10 principal components were chosen with a total of 75% explained variance. A cumulative variance plot is show in Figure 7. Graphs to visualize PC 0 vs PC 1 and PC 1 vs PC 2 are shown in Figures 8 and 9.

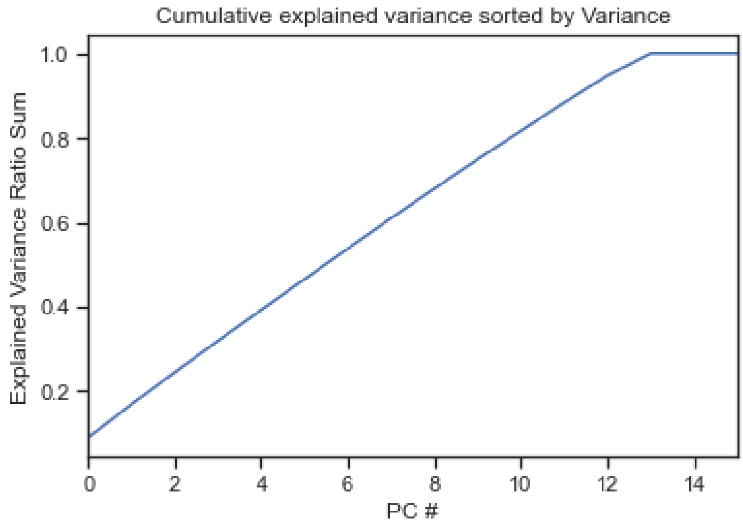


Figure 7 – Cumulative Variance Plot

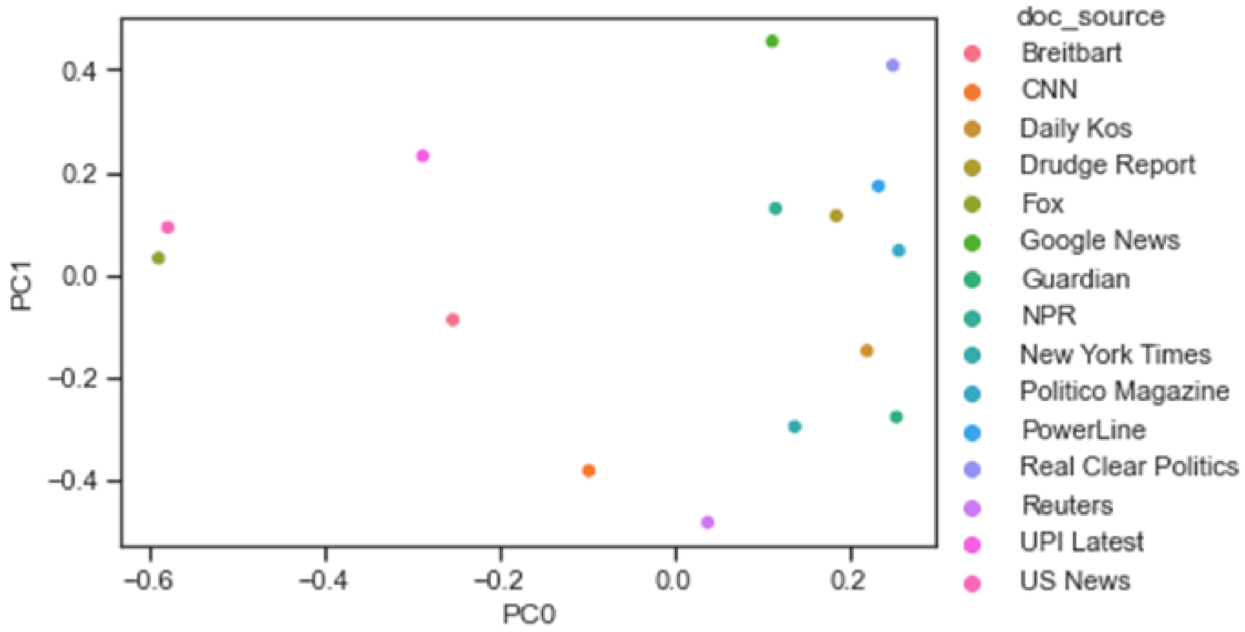


Figure 8 – PC0 vs PC1

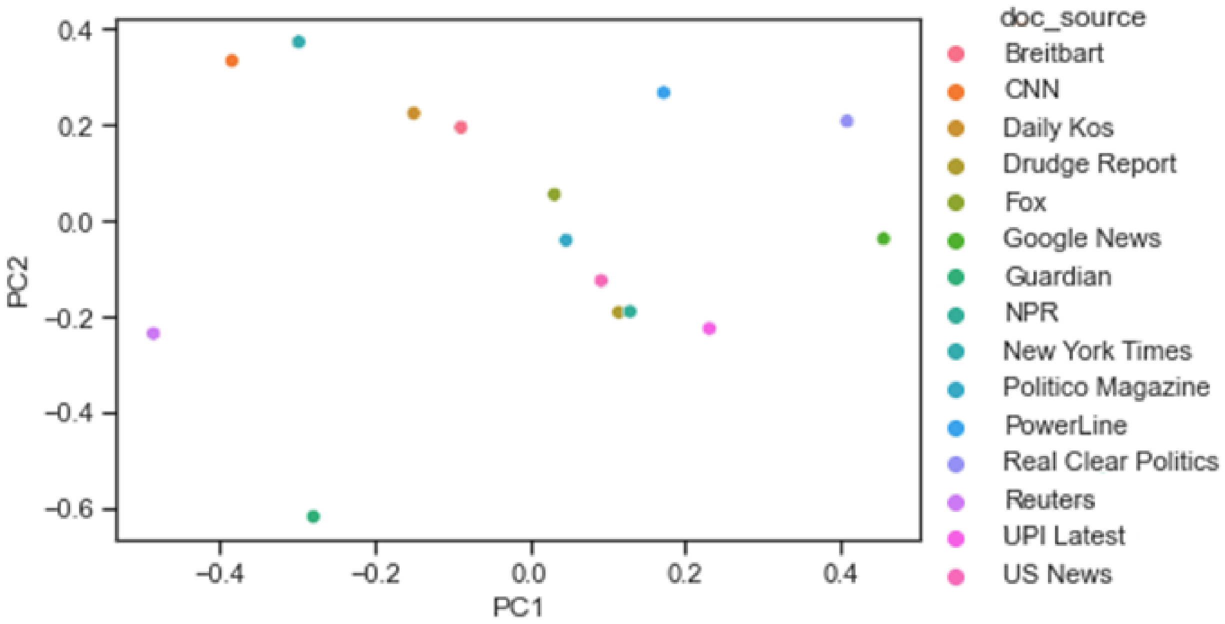


Figure 9 – PC1 vs PC2

PC1 seems to have an imperfect political bias where negative is more liberal and positive is more conservative.

LDA

The top 4000 words by term frequency ignoring stop words were used for LDA. 30 topics were run on LDA with max iterations = 5. LDA was calculated using Gibbs Sampling. Figure 10 shows the 30 topics ordered by largest document weight. Figure 11 shows the topic score sums for each new source.

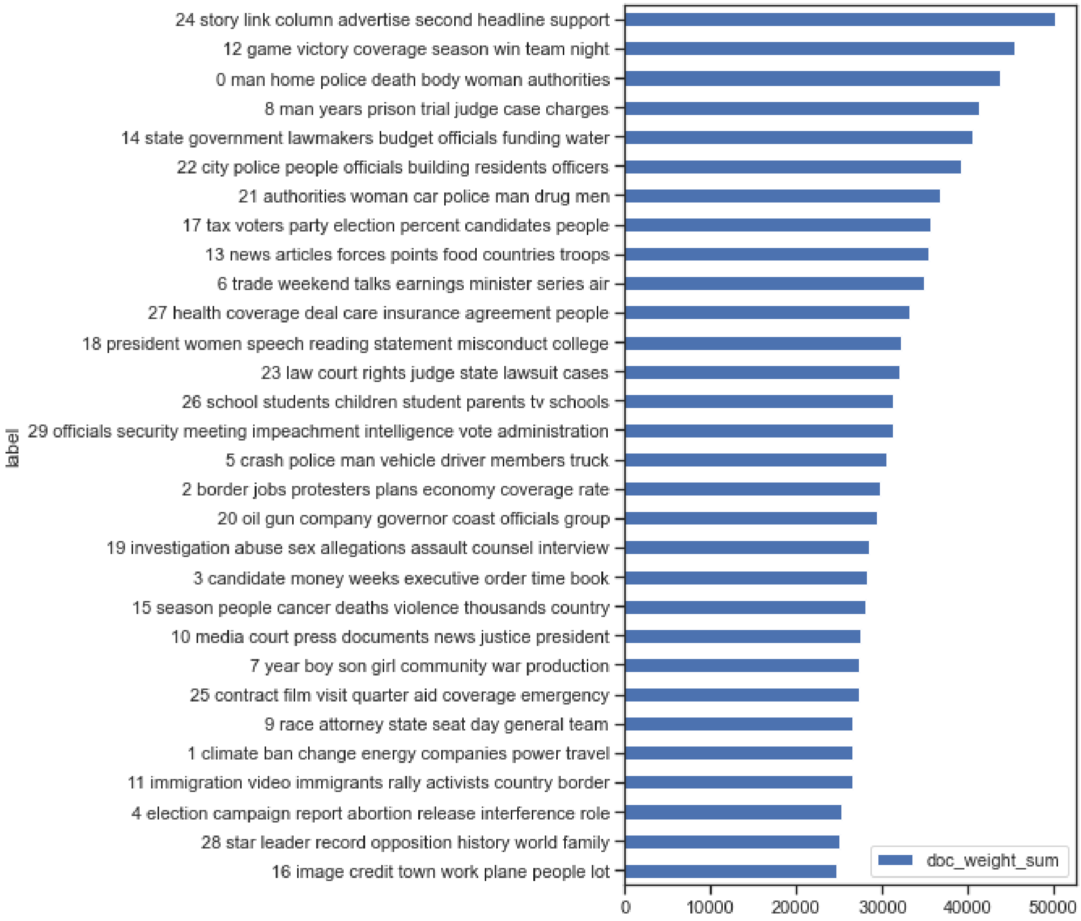
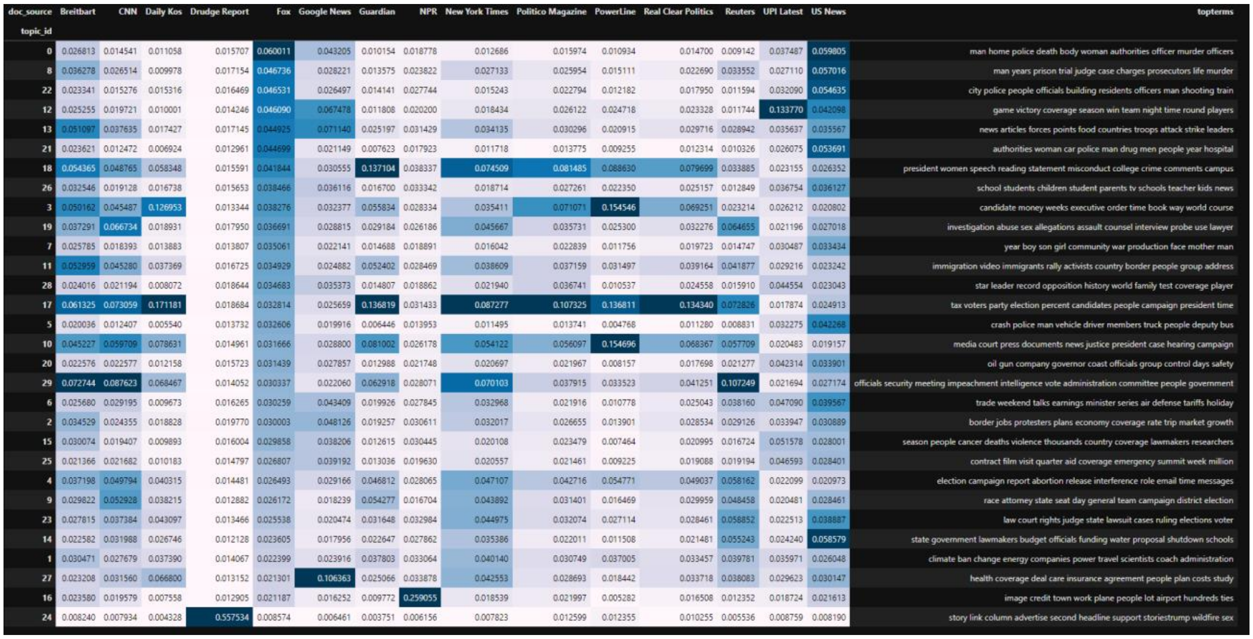


Figure 10 – 30 topics ordered by document weight sum



doc_source	Breitbart	CNN	Daily Kos	Drudge Report	Fox	Google News	Guardian	NPR	New York Times	Politico Magazine	PowerLine	Real Clear Politics	Reuters	UPI Latest	US News	topterms
topic_id																
0	0.026813	0.014541	0.011058	0.015707	0.060011	0.043205	0.010154	0.018778	0.012686	0.015974	0.010934	0.014700	0.009142	0.037487	0.059805	man home police death body woman authorities officer murder officers
8	0.036278	0.026514	0.009978	0.017154	0.046736	0.028221	0.013575	0.023822	0.027133	0.025954	0.015111	0.022690	0.033552	0.027110	0.057016	man years prison trial judge case charges prosecutors life murder
22	0.023341	0.015276	0.015316	0.016469	0.046531	0.026497	0.014141	0.027744	0.015243	0.022794	0.012182	0.017950	0.011594	0.032090	0.054635	city police people officials building residents officers man shooting train
12	0.025255	0.019721	0.010001	0.014246	0.046090	0.067478	0.011808	0.020200	0.018434	0.026122	0.024718	0.023328	0.011744	0.133770	0.042098	game victory coverage season win team night time round players
13	0.051097	0.037635	0.017427	0.017145	0.044925	0.071140	0.025197	0.031429	0.034135	0.030296	0.020915	0.029716	0.028942	0.035637	0.035567	news articles forces points food countries troops attack strike leaders
21	0.023621	0.012472	0.006924	0.012961	0.044699	0.021149	0.007623	0.017923	0.011718	0.013775	0.009255	0.012314	0.010326	0.026075	0.053691	authorities woman car police man drug men people year hospital
18	0.054365	0.048765	0.058348	0.015591	0.041844	0.030555	0.137104	0.038337	0.074509	0.081485	0.088630	0.079699	0.033885	0.023155	0.026352	president women speech reading statement misconduct college crime comments campus
26	0.032546	0.019128	0.016738	0.015653	0.038466	0.036116	0.016700	0.033342	0.018714	0.027261	0.022350	0.025157	0.012849	0.036754	0.036127	school students children student parents tv schools teacher kids news
3	0.050162	0.045487	0.126953	0.013344	0.038276	0.032377	0.055834	0.028334	0.035411	0.071071	0.154546	0.069251	0.023214	0.026212	0.020802	candidate money weeks executive order time book way world course
19	0.037291	0.066734	0.018931	0.017950	0.036691	0.028815	0.029184	0.026186	0.045667	0.035731	0.025300	0.032276	0.064655	0.021196	0.027018	investigation abuse sex allegations assault counsel interview probe use lawyer
7	0.025785	0.018393	0.013883	0.013807	0.035061	0.022141	0.014688	0.018891	0.016042	0.022839	0.011756	0.019723	0.014747	0.030487	0.033434	year boy son girl community war production face mother man
11	0.052959	0.045280	0.037369	0.016725	0.034929	0.024882	0.052402	0.028469	0.038609	0.037159	0.031497	0.039164	0.041877	0.029216	0.023242	immigration video immigrants rally activists country border people group address
28	0.024016	0.021194	0.008072	0.018644	0.034683	0.035373	0.014807	0.018862	0.021940	0.036741	0.010537	0.024558	0.015910	0.044554	0.023043	star leader record opposition history world family test coverage player
17	0.061325	0.073059	0.171181	0.018684	0.032814	0.025659	0.136819	0.031433	0.087277	0.107325	0.136811	0.134340	0.072826	0.017874	0.024913	tax voters party election percent candidates people campaign president time
5	0.020036	0.012407	0.005540	0.013732	0.032606	0.019916	0.006446	0.013953	0.011495	0.013741	0.004768	0.011280	0.008831	0.032275	0.042268	crash police man vehicle driver members truck people deputy bus
10	0.045227	0.059709	0.078631	0.014961	0.031666	0.028800	0.081002	0.026178	0.054122	0.056097	0.154696	0.068367	0.057709	0.020483	0.019157	media court press documents news justice president case hearing campaign
20	0.022576	0.022577	0.012158	0.015723	0.031439	0.027857	0.012988	0.021748	0.020697	0.021967	0.008157	0.017698	0.021277	0.042314	0.033901	oil gun company governor coast officials group control days safety
29	0.072744	0.087623	0.068467	0.014052	0.030337	0.022060	0.062918	0.028071	0.070103	0.037915	0.033523	0.041251	0.107249	0.021694	0.027174	officials security meeting impeachment intelligence vote administration committee people government
6	0.025680	0.029195	0.009673	0.016265	0.030259	0.043409	0.019926	0.027845	0.032968	0.021916	0.010778	0.025043	0.038160	0.047090	0.039567	trade weekend talks earnings minister series air defense tariffs holiday
2	0.034529	0.024355	0.018828	0.019770	0.030003	0.048126	0.019257	0.030611	0.032017	0.026655	0.013901	0.028534	0.029126	0.033947	0.030889	border jobs protesters plans economy coverage rate trip market growth
15	0.030074	0.019407	0.009893	0.016004	0.029858	0.038206	0.012615	0.030445	0.020108	0.023479	0.007464	0.020995	0.016724	0.051578	0.028001	season people cancer deaths violence thousands country coverage lawmakers researchers
25	0.021366	0.021682	0.010183	0.014797	0.026807	0.039192	0.013036	0.019630	0.020557	0.021461	0.009225	0.019088	0.019194	0.046593	0.028401	contract film visit quarter aid coverage emergency summit week million
4	0.037198	0.049794	0.040315	0.014481	0.026493	0.029166	0.046812	0.028065	0.047107	0.042716	0.054771	0.049037	0.058162	0.022099	0.020973	election campaign report abortion release interference role email time messages
9	0.029822	0.052928	0.038215	0.012882	0.026172	0.018239	0.054277	0.016704	0.043892	0.031401	0.016469	0.029959	0.048458	0.020481	0.028461	race attorney state seat day general team campaign district election
23	0.027815	0.037384	0.043097	0.013466	0.025538	0.020474	0.031648	0.032984	0.044975	0.032074	0.027114	0.028461	0.058852	0.022513	0.038887	law court rights judge state lawsuit cases ruling elections voter
14	0.022582	0.031988	0.026746	0.012128	0.023605	0.017956	0.022647	0.027862	0.035386	0.022011	0.011508	0.021481	0.055243	0.024240	0.058579	state government lawmakers budget officials funding water proposal shutdown schools
1	0.030471	0.027679	0.037390	0.014067	0.022399	0.023916	0.037803	0.033064	0.040140	0.030749	0.037005	0.033457	0.039781	0.035971	0.026048	climate ban change energy companies power travel scientists coach administration
27	0.023208	0.031560	0.066800	0.013152	0.021301	0.106363	0.025066	0.033878	0.042553	0.028693	0.018442	0.033718	0.038083	0.029623	0.030147	health coverage deal care insurance agreement people plan costs study
16	0.023580	0.019579	0.007558	0.012905	0.021187	0.016252	0.009772	0.259055	0.018539	0.021997	0.005282	0.016508	0.012352	0.018724	0.021613	image credit town work plane people lot airport hundreds ties
24	0.008240	0.007934	0.004328	0.557534	0.008574	0.006461	0.003751	0.006156	0.007823	0.012599	0.012355	0.010255	0.005536	0.008759	0.008190	story link column advertise second headline support storiestrump wildfire sex

Figure 11 – Topic density scores

Topic 17, politics was a very popular topic for a large amount of news sources. Drudge Report's bias on topic 24 is most likely skewed since most of the words in that topic are stop words for news articles. NPR's bias on topic 16 is most likely skewed since image and credit are likely stop words. UPI Latest really likes topic 12, sports. Politico Magazine, PowerLine, Daily Kos, Guardian, and Real Clear Politics focused on politics. Google News has a lot of articles about healthcare, topic 27. New York Times, Fox, Breitbart, CNN, Reuters, and US News talk about a large range of different topics.

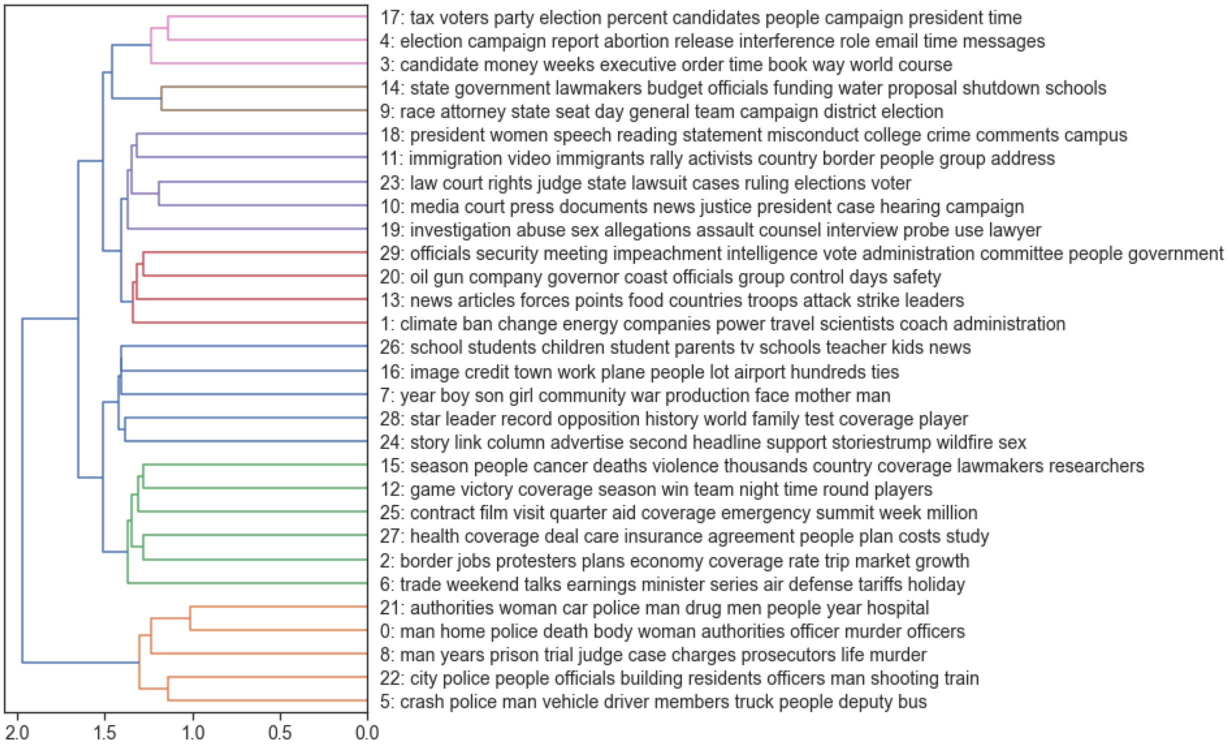


Figure 12 – Topic Cluster Diagram

Figure 12 shows how topics clusters. It does appear there is some overlap in topics and 30 topics might be too much.

Word2Vec

Implemented Skip-Gram Negative Sampling (SGNS) using Word2Vec with a window of 5 and used tSNE to visualize the data in two dimensions to determine similar clusters of words grouped by individual articles. Figure 13 shows a tSNE 2D visualization. There is a tSNE of all the words from each news source, but it is too crowded for this report. Please check the source code to view it.

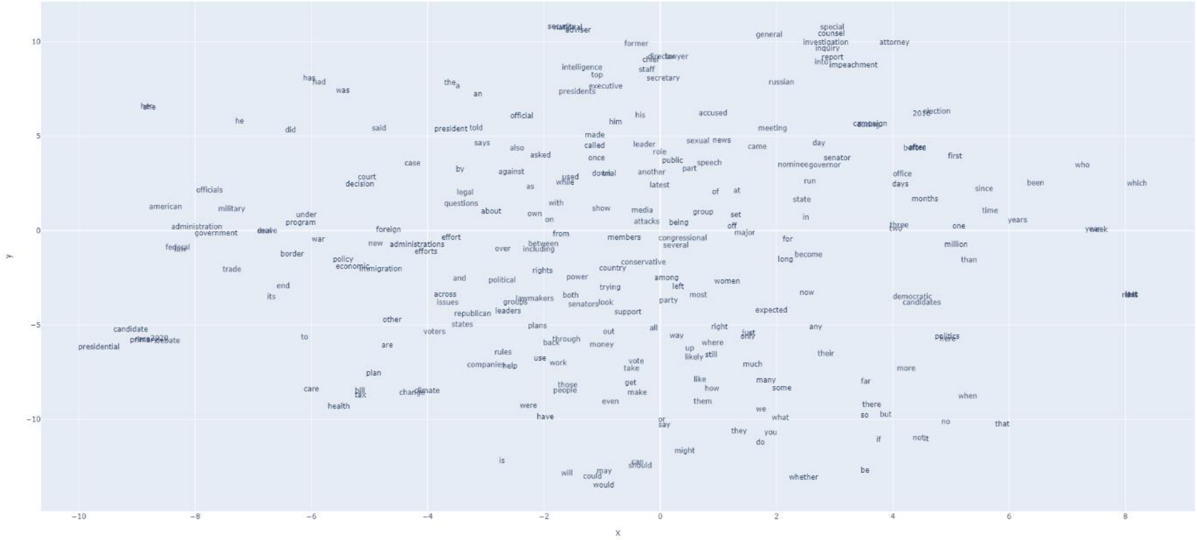


Figure 13 – tSNE of New York Times

Word2Vec and tSNE has worked well with grouping similar words and I see a few that match topics from the topic models. tSNE has confirmed that some of the topics shown in the topic models make sense. I have not found any further missing topics.

If a large amount of data was gathered over a longer time, a potential use case is where articles are split up over years or decades to determine if/how word meanings have changed or new words with similar meanings.

Sentiment Analysis

The 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and polarity from the NRC lexicon were joined with the TOKENS table. The emotion averages for all articles, Fox articles, CNN articles and New York Times articles were determined. The top 4 emotions + polarity were plotted to visualize sentiment. The average emotions of all news sources were plotted, shown in Figure 14. The actual values are show in Table 2.

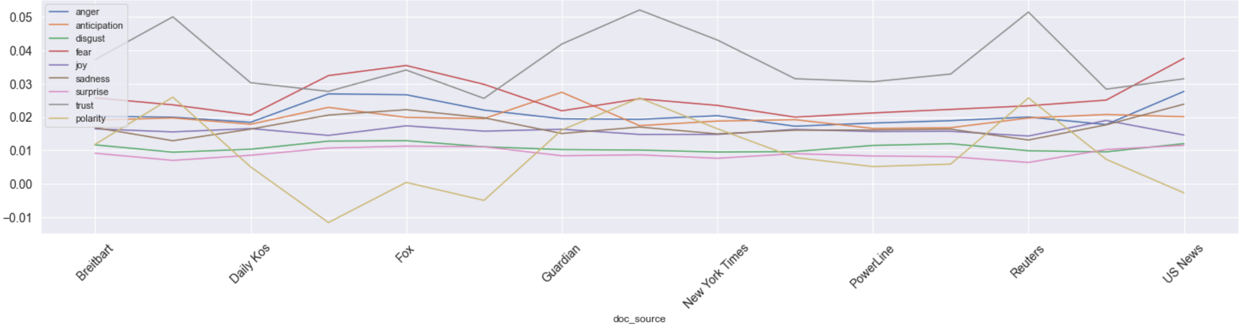
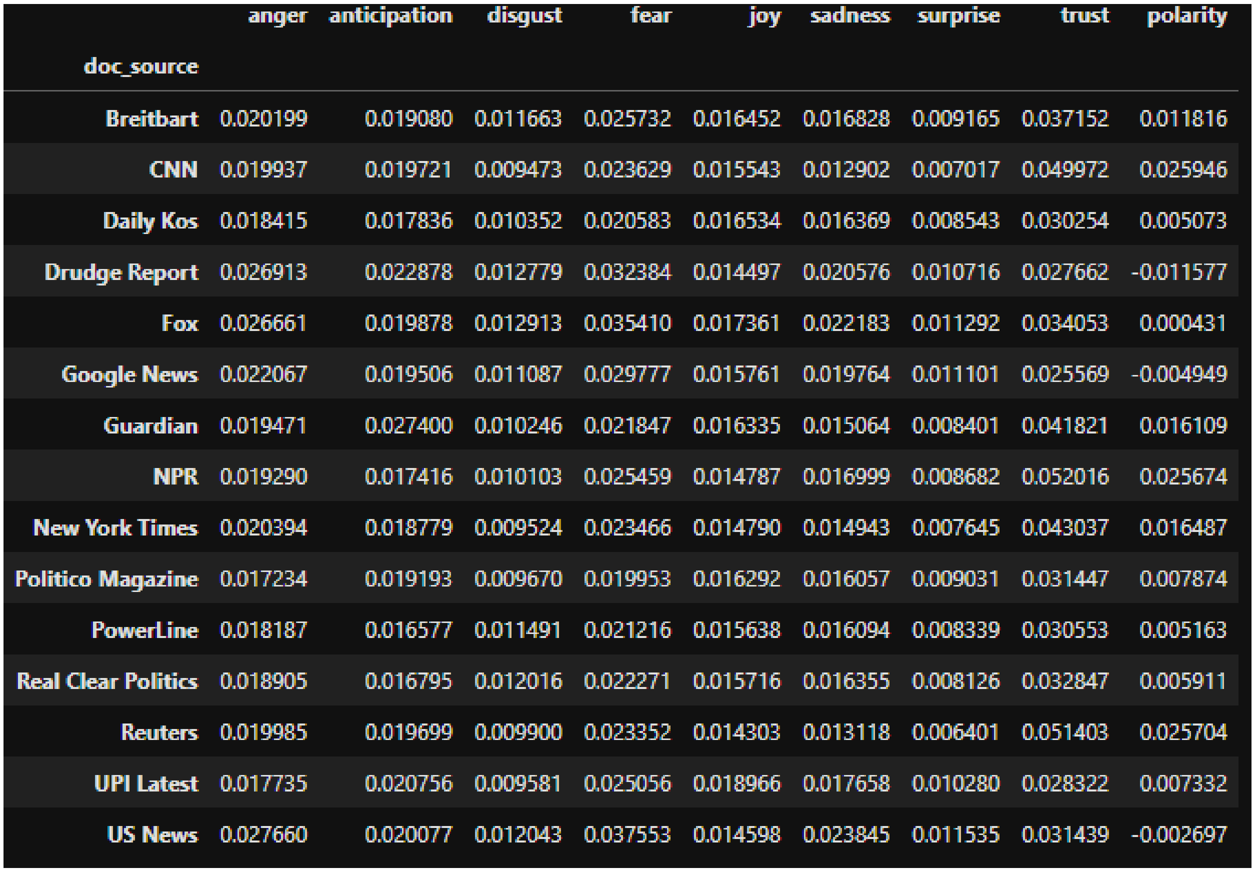
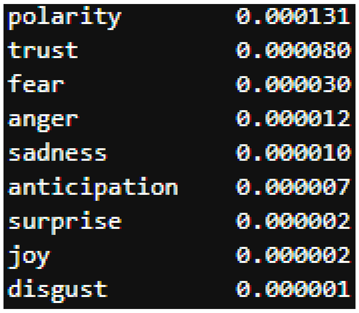


Figure 14 – Average emotions of news sources



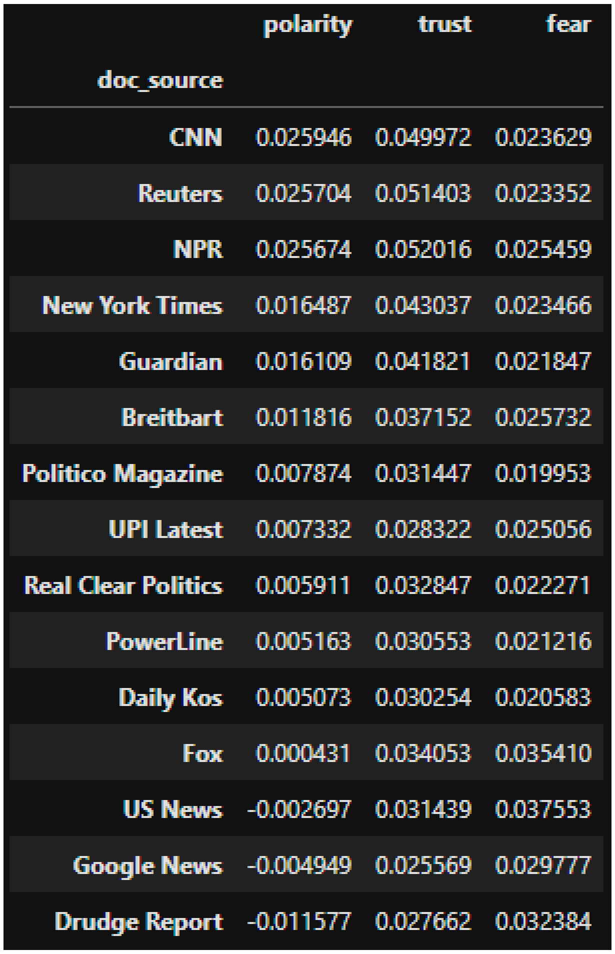
	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	polarity
doc_source									
Breitbart	0.020199	0.019080	0.011663	0.025732	0.016452	0.016828	0.009165	0.037152	0.011816
CNN	0.019937	0.019721	0.009473	0.023629	0.015543	0.012902	0.007017	0.049972	0.025946
Daily Kos	0.018415	0.017836	0.010352	0.020583	0.016534	0.016369	0.008543	0.030254	0.005073
Drudge Report	0.026913	0.022878	0.012779	0.032384	0.014497	0.020576	0.010716	0.027662	-0.011577
Fox	0.026661	0.019878	0.012913	0.035410	0.017361	0.022183	0.011292	0.034053	0.000431
Google News	0.022067	0.019506	0.011087	0.029777	0.015761	0.019764	0.011101	0.025569	-0.004949
Guardian	0.019471	0.027400	0.010246	0.021847	0.016335	0.015064	0.008401	0.041821	0.016109
NPR	0.019290	0.017416	0.010103	0.025459	0.014787	0.016999	0.008682	0.052016	0.025674
New York Times	0.020394	0.018779	0.009524	0.023466	0.014790	0.014943	0.007645	0.043037	0.016487
Politico Magazine	0.017234	0.019193	0.009670	0.019953	0.016292	0.016057	0.009031	0.031447	0.007874
PowerLine	0.018187	0.016577	0.011491	0.021216	0.015638	0.016094	0.008339	0.030553	0.005163
Real Clear Politics	0.018905	0.016795	0.012016	0.022271	0.015716	0.016355	0.008126	0.032847	0.005911
Reuters	0.019985	0.019699	0.009900	0.023352	0.014303	0.013118	0.006401	0.051403	0.025704
UPI Latest	0.017735	0.020756	0.009581	0.025056	0.018966	0.017658	0.010280	0.028322	0.007332
US News	0.027660	0.020077	0.012043	0.037553	0.014598	0.023845	0.011535	0.031439	-0.002697

Table 2 – Average emotion values of news sources



```
polarity        0.000131
trust           0.000080
fear            0.000030
anger           0.000012
sadness         0.000010
anticipation    0.000007
surprise        0.000002
joy             0.000002
disgust         0.000001
```

Table 3 – Emotion variance of news sources



	polarity	trust	fear
doc_source			
CNN	0.025946	0.049972	0.023629
Reuters	0.025704	0.051403	0.023352
NPR	0.025674	0.052016	0.025459
New York Times	0.016487	0.043037	0.023466
Guardian	0.016109	0.041821	0.021847
Breitbart	0.011816	0.037152	0.025732
Politico Magazine	0.007874	0.031447	0.019953
UPI Latest	0.007332	0.028322	0.025056
Real Clear Politics	0.005911	0.032847	0.022271
PowerLine	0.005163	0.030553	0.021216
Daily Kos	0.005073	0.030254	0.020583
Fox	0.000431	0.034053	0.035410
US News	-0.002697	0.031439	0.037553
Google News	-0.004949	0.025569	0.029777
Drudge Report	-0.011577	0.027662	0.032384

Table 4 – Average news source emotion values sorted by polarity

Fear, Trust and Polarity are emotions with the most variance. Looking at those values we can group news sources with similar average emotional values.

Emotion Groups

- CNN, NPR, and Reuters
- New York Times and Guardian
- Politico Magazine, Real Clear Politics, PowerLine, and Daily Kos
- Fox and US News

Discussion

Hierarchical Clustering, Topic modeling, and Sentiment Analysis have shown clusters of similar news sources. The various clusters are listed out in Table 4.

Clusters of Similar News Sources
Fox and US News
CNN, New York Times and Reuters
Daily Kos, Politico Magazine, PowerLine and Real Clear Politics

Table 4 – Clusters of Similar News Sources

These clusters are based on how the news sources cover topics. There is some bias with the political views of each news source in how they cover topics, but it is not always correct. Daily Kos, Politico Magazine, PowerLine, and Real Clear Politics all clearly have political focused content, and therefore they are grouped together. CNN, New York Times, and Reuters all have a liberal bias and are grouped together. US News most likely has a conservative bias which is why US News and Fox are grouped together.

Future Work

These are all ideas worth exploring but for lack of time and/or domain knowledge were not implemented.

Running PCA or LSA on doc id. My computer didn't have the ram required to process TFIDF on document id so I could only process it on each news source instead and run PCA. It would be more interesting to see which articles don't cluster in their news source and see the range of documents a news source has rather than a general overview of all the documents for each news source and seeing an average score between all articles inside that news source.

Removing stop words in a more efficient manner as this negatively affected the results for LDA.

Try running LDA on less than 30 topics and determine an optimal number of topics.

Reducing unique words or reducing dimensionality would likely improve our model in a better manner than PCA or at least keep interpretability while providing more information. This can be done several ways. Distributional Hypothesis or terms that occur in similar term context have similar meaning, to combine like terms can reduce dimensionality. TFIDF and domain knowledge to determine non-significant words will also further reduce dimensionality. Run LSI on individual words, this is a good way to determine word similarity and could be used in place of the above distributed Hypothesis idea, especially as a more quantitative approach. Can use Word2Vec and t-SNE to determine similar words through clustering. Word2Vec and t-SNE with PCA or topic model to input new word embeddings to determine similar words though clustering.

Word similarity and removing non-significant words can both be used in combination to further reduce dimensionality. This can be done by combining some of the methods described in the above paragraph.

Conclusion

986,461 online news articles from 15 different papers were examined from the Newzy dataset. Several models were built with the aim of finding similar news sources. The methods were not perfect, and more work could be done to classify more news sources, but 3 groups were found in the end. 9 of the news sources were grouped into 3 separate groups based on both observations discovered with models and explained with some domain knowledge.