# Final Term Report

**Tatev Gomtsyan (tmg6jda@virginia.edu (mailto:tmg6jda@virginia.edu)) DS 5001 Spring 2023**

# Introduction

Throughout the years, text analytics has become exceedingly capable of providing opportunities for gaining insight on some of the world's most interesting and pressing issues. The aim of this project is to analyze the data from news articles published by two specific channels, CNN and CNBC. CNN and CNBC are both U.S-based news organizations, but each of them focuses on different topics. CNN is known for primarily covering news related to politics, world events, and business. CNBC, on the other hand, focuses primarily on business and financial news, with a more special focus on financial markets and the economy. In our analysis, we expect to see these differences through the language that is used in the articles and which terms are more important and frequently used. We thoughT it would be interesting to explore two U.S-based sources that are similar enough but could have differences in the way that certain topics are represented.

For this analysis, we use clickstream data obtained by a company that tracks which websites users have visited. The text data is analyzed in this report using various methods suchs as TF-IDF, PCA, Topic Modeling, and Word2Vec. TF-IDF, Term Frequency-Inverse Document Frequency, is a statistical measure that can be used to evaluate the importance of a word in a document or collection of documents. PCA, also known as Principal Component Analysis, helps us reduce the dimensionality of the data for the purposes of clearer visualization. Topic Modeling is used to idenfity the underlying topics and themes in the textual data and help us understand what the most important content discussed is for people in a particular domain. Finally, Word2Vec is an algorithm that can generate word embeddings for purposes of semantic analysis and identifying relationships between words and concepts.

The analysis in this project is neither comprehensive nor conclusive. The nature of this research is long-term and ongoing. The ultimate goal is to assess the spread of misinformation on the internet through different news channels that are publicly trusted for the most part. This project gives us the chance to provide a general analysis of the data available and recognize opportunities for future and further exploration.

The timeframe selected for this data is September to November 2020, which was during the United States Presidential elections, as well as during a trying time for the whole world; the Covid-19 pandemic. It is thus interesting to explore different sentiments around both the pandemic and the election and how, if at all, the two compounded sentiments and were related in the way that people felt. Using the text analytics techniques described above, we hope to distinguish some patterns and trends within and between news articles from the CNN and CNBC and make general observations and insights. The results of this research could potentially be used for trend analysis and predictive modeling.

# Source Data

The general subject matter of the data used for this analysis is the 2020 elections in the United States.

The document types we are working with in this project are all CSV files. We are unable to provide URLs to the data source, since they were obtained from a professor conducting research on this topic. We received the datasets directly from her. The data was collected from a company that has collected search and click information from individuals that visit their website. The original data size is over 10TB, so for the purposes of this project, we will only focus on a specifc time frame of September to November 2020. While the large dataset contains data from different news organizations, we filtered the given dataset to only include CNN and CNBC. About 500 samples were taken from each source, giving us about 1000 news articles for this analysis. Articles were scraped and stored in a dataset. Relevant categories from the data included the URL, title, tag, source, etc. This is what the LIB table is comprised of. Since this dataset has been part of a research project that has spanned the majority of this semester, there was preprocessing and cleaning done initially before the beginning of this project. Thus, I began my application of text analytics tools starting from the LIB dataset that was produced by Tatev Kyosababyan, research assistant for the professor working on this project.

# Exploration

First, we worked on breaking down the structure of the dataset into paragraphs, sentences, and then words or tokens. The corpus was created using these components. We extracted vocabulary from the corpus to analyze the most frequently seen terms in the text. Then, we conducted our analysis based on the different techniques outlines: TF-IDF, PCA, and Word2Vec. The results can be seen below.

## TF-IDF Analysis

After the text was tokenized, the Term Frequency-Inverse Document Frequency (TF-IDF) values for the entire corpus was calculated. This was a way to bring to light the most important words in the corpus overall. The top 20 words by TF-IDF score were found for the whole corpus. Some of the top terms I found interesting and most relevant include "trump", "biden", "shares", "election", "president", "covid", and "tax". We can agree that this seems to fit in with the general context of the text document in that it is very evidently about the election as well as covid and the president elects' stances on tax. The table below shows this list.

| term_str | mean_tfidf | max_pos |
|---|---|---|
| trump | 0.093882 | CD |
| you | 0.070173 | CD |
| your | 0.069067 | CD |
| biden | 0.064063 | CD |
| her | 0.060828 | CD |
| his | 0.057643 | CD |
| she | 0.055750 | CD |
| i | 0.055270 | CD |
| he | 0.054128 | CD |
| s | 0.050730 | CD |
| shares | 0.049535 | CD |
| election | 0.049435 | CD |
| president | 0.045963 | CD |
| ballots | 0.045220 | CD |
| company | 0.044109 | CD |
| covid | 0.042550 | CD |
| house | 0.041066 | CD |
| tax | 0.040644 | CD |
| said | 0.040518 | CD |
| its | 0.039436 | CD |

## Principle Component Analysis

PCA is a great tool for obtaining reduced dimensions without losing existing information. The main goal of this technique is to identify any possible relationships and patterns within the data. It functions by finding linear combinations of the original features explaining the most variance in the data. Then, the principal components are ordered in terms of their importance, so the one at the top explains the most variance in the data.

THis ethod was used to see if any interesting clusters would form from the different articles. We see from the following heatmap the strongest positive and negative weights that the corresponding words carry in each of the 10 principal components. The loadings matrix indicates the weights of each term in the principal components, which indicates how much a given term contributes to the variation captured by the components. From the figure below, we could say that the word "election" is strongly assiciated with the variation captured by PC0

since it has the highest value. This would also mean that the articles containing the term "election" would likely have high scores on PC0. The same word, however, does not have as strong of an association in PC2. This shows us the underlying structure of the data and helps identify the levels of association for each term.

| pc_id<br>term_str | PC0 | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|---|
| week | 0.001868 | 0.002108 | 0.036516 | -0.058619 | -0.000819 | -0.053904 | -0.022443 | -0.014334 | -0.038175 | -0.028635 |
| state | 0.045520 | 0.042799 | -0.072569 | -0.001265 | -0.004462 | -0.151065 | 0.010729 | 0.038729 | 0.069941 | -0.037990 |
| t | 0.045500 | -0.039933 | -0.067598 | 0.062410 | -0.018727 | 0.068526 | -0.063105 | -0.016228 | 0.030296 | -0.019008 |
| take | -0.005478 | -0.008777 | 0.003273 | -0.003357 | -0.014866 | 0.001260 | 0.012183 | -0.015889 | -0.024225 | 0.016974 |
| much | -0.007961 | -0.006942 | -0.015129 | -0.018898 | -0.021154 | 0.032664 | 0.013051 | -0.042359 | 0.013319 | 0.029199 |
| coronavirus | 0.012440 | -0.118362 | 0.078049 | -0.050216 | -0.034871 | -0.060054 | 0.028851 | -0.013087 | -0.003286 | -0.019029 |
| covid | -0.013102 | -0.196931 | 0.080955 | -0.060383 | -0.086501 | -0.105948 | 0.055684 | 0.035543 | -0.006755 | -0.021630 |
| house | 0.156552 | -0.088596 | 0.184633 | -0.160347 | 0.153851 | 0.062377 | -0.098784 | 0.031482 | 0.058511 | 0.005025 |
| election | 0.173554 | 0.155403 | -0.001324 | 0.071561 | -0.049573 | -0.231962 | -0.081477 | 0.036281 | 0.159299 | 0.054616 |
| make | -0.012259 | -0.009926 | -0.036853 | -0.014004 | -0.001221 | 0.036168 | 0.020021 | -0.063301 | 0.031325 | 0.028088 |

The table below gives us a different view from our PCA analysis, this time summarizing the positive and negative weights of the top 10 principle components we already looked at. PC0 is highly associated with the presidents and the election while it has very low association with business and market related topics. This helps us more easily visualize the different topics that our principal components are about in our data.

| 1<br>comp_id | pos | neg |
|---|---|---|
| PC0 | trump biden president election white house cam... | company stock share trading market quarter rev... |
| PC1 | biden tax election stock trump share company r... | virus covid positive cdc health dr vaccine dis... |
| PC2 | stock house white positive trump share company... | police tax court black income attorney county ... |
| PC3 | police company black share stock biden county ... | tax stimulus relief income aid bill unemployme... |
| PC4 | quarantine police positive house white court m... | biden vaccine health debate u climate cdc covi... |
| PC5 | i security trump social tax white debate polic... | court election mail vote state vaccine voting ... |
| PC6 | tax quarantine trump york vaccine income manha... | black police relief stimulus pelosi senate aid... |
| PC7 | police vaccine tax security revenue social bla... | quarantine biden debate i negative guilty manh... |
| PC8 | tax mail election voting income county vote ea... | police vaccine biden department quarantine att... |
| PC9 | security social revenue age check share quaran... | tax police market i p black dow york u rate |

Further analysis warranted a comparison between keywords in both CNN and CNBC data sets. Below is a representation of the similarity between words of both channels. The tables reveal that many of the terms are repeated and show high similarity.

| | term | sim |
|---|---|---|
| 0 | biden | 0.860657 |
| 1 | pence | 0.783681 |
| 2 | president | 0.765098 |
| 3 | barr | 0.718398 |
| 4 | himself | 0.705618 |
| 5 | harris | 0.692974 |
| 6 | running | 0.674829 |
| 7 | obama | 0.669761 |
| 8 | fox | 0.667758 |
| 9 | kamala | 0.649452 |

| | term | sim |
|---|---|---|
| 0 | pence | 0.757725 |
| 1 | obama | 0.687855 |
| 2 | biden | 0.674829 |
| 3 | barr | 0.663307 |
| 4 | trumps | 0.648268 |
| 5 | bidens | 0.638223 |
| 6 | president | 0.634999 |
| 7 | harris | 0.621577 |
| 8 | mccarthy | 0.613487 |
| 9 | clinton | 0.586992 |

## Topic models (LDA)

For the topic modelling portion of this project, Latend Dirilecth Allocation (LDA) was used, which is a technique that helps identify underlying topics in a corpus of text. It assumes that each separate document in the corpus is a mix of many topics that are each probability distributions over words. When applying this method, we were able to produce the below matrix which shows the representation of articles belonging to different models. This heatmap graphic shows 20 topics and their associations with each of the topics from the articles divided by the two sources in question, CNN and CNBC.

| source | text_num | T00 | T01 | T02 | T03 | T04 | T05 | T06 | T07 | T08 | T09 | T10 | T11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CNN | 421 | 0.000260 | 0.000260 | 0.000260 | 0.000260 | 0.000260 | 0.000260 | 0.000260 | 0.204207 | 0.636600 | 0.000260 | 0.154766 | 0.000260 |
| | 74 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 | 0.000163 |
| CNBC | 735 | 0.000275 | 0.368333 | 0.000275 | 0.027587 | 0.000275 | 0.000275 | 0.000275 | 0.000275 | 0.514331 | 0.000275 | 0.000275 | 0.000275 |
| CNN | 265 | 0.057614 | 0.000806 | 0.432951 | 0.000806 | 0.000806 | 0.000806 | 0.000806 | 0.000806 | 0.000806 | 0.000806 | 0.000806 | 0.000806 |
| | 171 | 0.000182 | 0.000182 | 0.158095 | 0.000182 | 0.000182 | 0.000182 | 0.000182 | 0.000182 | 0.000182 | 0.000182 | 0.152535 | 0.000182 |
| | 677 | 0.000179 | 0.000179 | 0.028853 | 0.000179 | 0.000179 | 0.000179 | 0.085303 | 0.000179 | 0.000179 | 0.029255 | 0.000179 | 0.000179 |
| | 491 | 0.000355 | 0.000355 | 0.089565 | 0.000355 | 0.000355 | 0.000355 | 0.000355 | 0.000355 | 0.000355 | 0.000355 | 0.104892 | 0.000355 |
| CNBC | 877 | 0.912778 | 0.000316 | 0.000316 | 0.000316 | 0.000316 | 0.000316 | 0.081526 | 0.000316 | 0.000316 | 0.000316 | 0.000316 | 0.000316 |
| | 647 | 0.000199 | 0.000199 | 0.000199 | 0.000199 | 0.000199 | 0.996215 | 0.000199 | 0.000199 | 0.000199 | 0.000199 | 0.000199 | 0.000199 |
| | 713 | 0.000746 | 0.000746 | 0.000746 | 0.000746 | 0.000746 | 0.000746 | 0.000746 | 0.000746 | 0.000746 | 0.985821 | 0.000746 | 0.000746 |
| | 836 | 0.000455 | 0.000455 | 0.000455 | 0.000455 | 0.000455 | 0.000455 | 0.313764 | 0.000455 | 0.000455 | 0.000455 | 0.000455 | 0.000455 |
| CNN | 480 | 0.000148 | 0.000148 | 0.000148 | 0.000148 | 0.997189 | 0.000148 | 0.000148 | 0.000148 | 0.000148 | 0.000148 | 0.000148 | 0.000148 |
| CNBC | 772 | 0.000556 | 0.000556 | 0.235765 | 0.000556 | 0.000556 | 0.000556 | 0.585991 | 0.000556 | 0.000556 | 0.000556 | 0.000556 | 0.000556 |
| | 630 | 0.000370 | 0.000370 | 0.000370 | 0.364053 | 0.000370 | 0.262479 | 0.000370 | 0.000370 | 0.083869 | 0.000370 | 0.192260 | 0.000370 |
| CNN | 52 | 0.000172 | 0.000172 | 0.000172 | 0.000172 | 0.000172 | 0.000172 | 0.000172 | 0.000172 | 0.000172 | 0.113888 | 0.883008 | 0.000172 |
| | 707 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 | 0.002381 |
| CNBC | 516 | 0.000667 | 0.157123 | 0.000667 | 0.301155 | 0.000667 | 0.000667 | 0.000667 | 0.000667 | 0.111753 | 0.000667 | 0.000667 | 0.000667 |
| | 817 | 0.000318 | 0.000318 | 0.000318 | 0.000318 | 0.000318 | 0.000318 | 0.000318 | 0.054897 | 0.000318 | 0.817646 | 0.000318 | 0.000318 |
| CNN | 266 | 0.164988 | 0.000538 | 0.037945 | 0.000538 | 0.172085 | 0.000538 | 0.000538 | 0.000538 | 0.000538 | 0.000538 | 0.000538 | 0.000538 |
| CNBC | 700 | 0.900618 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 | 0.000758 |

The matrix below is similar, but hones in further on the words used in the articles within the channels and depicts the same kind of analysis as above. We can see that the word "forecasts" is strongly associated with topic 1, and "mcconnell" with topic 8.

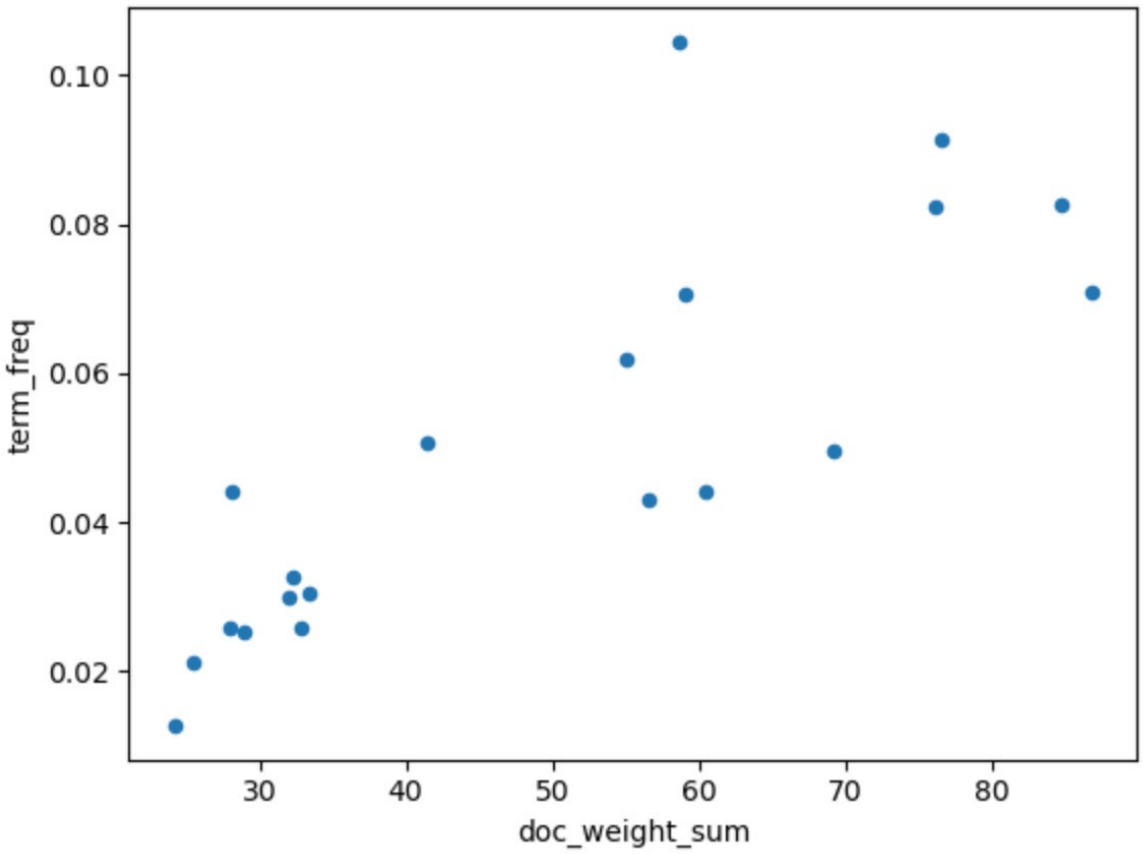
| topic_id | T00 | T01 | T02 | T03 | T04 | T05 | T06 | T07 | T08 | T09 | T10 | T1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| term_str | | | | | | | | | | | | |
| worlds | 1.717856 | 0.050000 | 0.050000 | 6.468524 | 0.050000 | 4.808049 | 13.127132 | 0.383793 | 0.050001 | 4.578936 | 0.050000 | 0.05000( |
| victims | 0.050000 | 4.796803 | 4.124852 | 2.303197 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 2.739717 | 0.050000 | 0.05000( |
| committees | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 7.050000 | 0.436627 | 0.050000 | 0.050000 | 0.05000( |
| earnings season | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.05000( |
| aircraft | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 3.811485 | 0.050000 | 0.05000( |
| capacity | 0.290295 | 0.239459 | 0.050000 | 0.050021 | 0.050000 | 5.715156 | 0.050000 | 8.304709 | 9.010101 | 16.756059 | 3.764000 | 0.05000( |
| mcconnell | 0.050000 | 0.050000 | 0.200551 | 0.051420 | 0.050000 | 0.050000 | 0.050017 | 27.325789 | 59.901536 | 0.050000 | 13.722792 | 0.05000( |
| covid relief | 0.050000 | 0.050000 | 0.050000 | 0.050278 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 30.848478 | 0.050000 | 0.050000 | 0.05000( |
| year date | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 3.278813 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.05000( |
| polling | 9.558934 | 0.050000 | 0.096615 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050029 | 0.050000 | 0.050000 | 0.050000 | 0.05000( |
| friends family | 5.402554 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 4.142879 | 0.050000 | 0.050000 | 4.604500 | 0.050000 | 0.05000( |
| circumstances | 5.816276 | 6.575299 | 0.350153 | 0.050000 | 0.050000 | 0.050000 | 0.052517 | 0.050000 | 0.050076 | 0.050000 | 11.242058 | 0.05000( |
| endorsement | 0.050000 | 2.049999 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 6.194633 | 0.050000 | 0.057296 | 1.05000( |
| roberts | 0.050000 | 0.050000 | 23.540454 | 0.050000 | 2.095946 | 2.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.05000( |
| imbert | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.05000( |
| hearing | 1.147152 | 0.050000 | 16.851161 | 0.050000 | 1.379737 | 2.170706 | 0.050000 | 11.530951 | 1.726604 | 0.050000 | 0.302375 | 0.05000( |
| person matter | 0.050000 | 0.050000 | 0.050000 | 0.626888 | 0.050000 | 0.050000 | 0.050000 | 12.528065 | 0.050000 | 0.606193 | 26.551831 | 0.05000( |
| execution | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 3.332727 | 0.05000( |
| john mccain | 0.050000 | 0.050000 | 4.451927 | 0.050000 | 2.037396 | 0.050000 | 0.050000 | 0.050000 | 0.050000 | 0.165158 | 1.934738 | 0.05000( |
| forecasts | 0.050000 | 48.004389 | 0.050000 | 0.050000 | 0.050000 | 0.548758 | 5.078699 | 0.050000 | 0.050001 | 2.263246 | 0.050000 | 0.05000( |

The following is an interesting comparison between different topics within articles and the weight and term frequency. The top 8 topics seem to have to do with the election, but also include words such as "police", "justice", "taxes", and "health". From how I interpret it, it seems as though these articles discuss different political viewpoints and topics that are perhaps

included in the president elects' speeches or ideologies. The topics aren't placed in order of decreasing doc weight and term frequency, so the next chunk after this table shows the actual relationship between the two.

| term_str<br>topic_id | 0 | 1 | 2 | 3 | 4 | 5 | 6 | label | doc_weight_sum | term_freq |
|---|---|---|---|---|---|---|---|---|---|---|
| **T00** | police | people | family | taylor | officers | cnn | life | T00 police, people, family, taylor, officers, ... | 69.222423 | 0.049714 |
| **T01** | share | checks | people | cents | revenue | premarket | cents share | T01 share, checks, people, cents, revenue, pre... | 33.457869 | 0.030492 |
| **T02** | trump | court | president | justice | department | law | case | T02 trump, court, president, justice, departme... | 59.098367 | 0.070633 |
| **T03** | tax | taxes | world | states | year | scientists | people | T03 tax, taxes, world, states, year, scientist... | 32.798044 | 0.025902 |
| **T04** | people | group | children | health | activity | women | adults | T04 people, group, children, health, activity,... | 27.990425 | 0.025968 |
| **T05** | biden | president | policy | women | trump | years | country | T05 biden, president, policy, women, trump, ye... | 32.265093 | 0.032680 |
| **T06** | companies | company | cramer | year | time | business | market | T06 companies, company, cramer, year, time, bu... | 60.442268 | 0.044070 |
| **T07** | president | house | trump | presidents | health | president donald | donald | T07 president, house, trump, presidents, healt... | 28.054252 | 0.044080 |

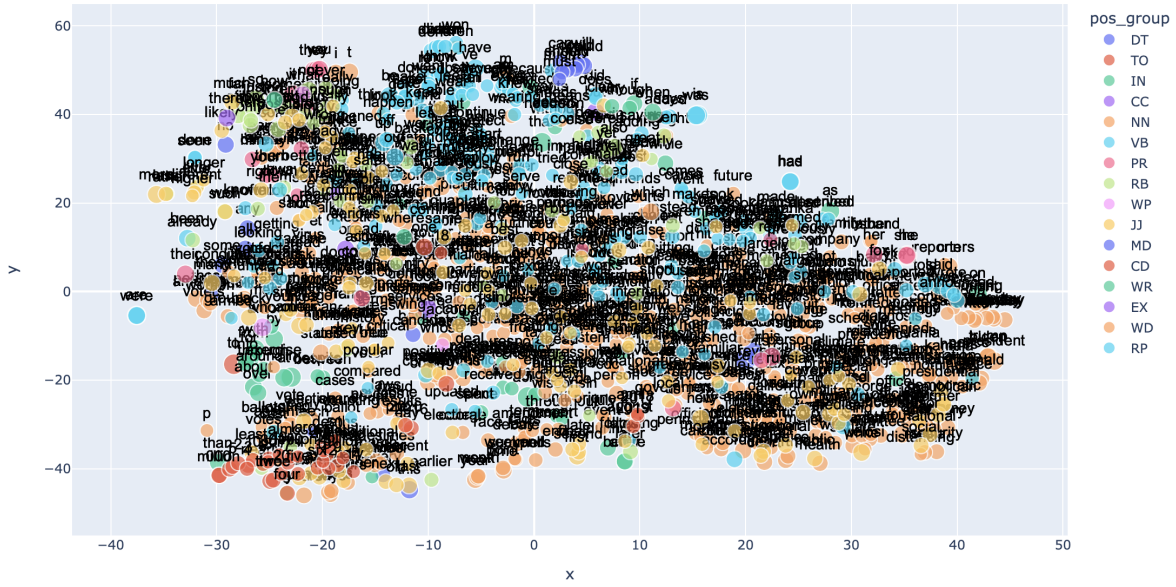
This scatterplot is just an interesting concept to point out as it shows how the weight of the document is correlated with the term frequency in a document. There is a strong positive correlation between the two.

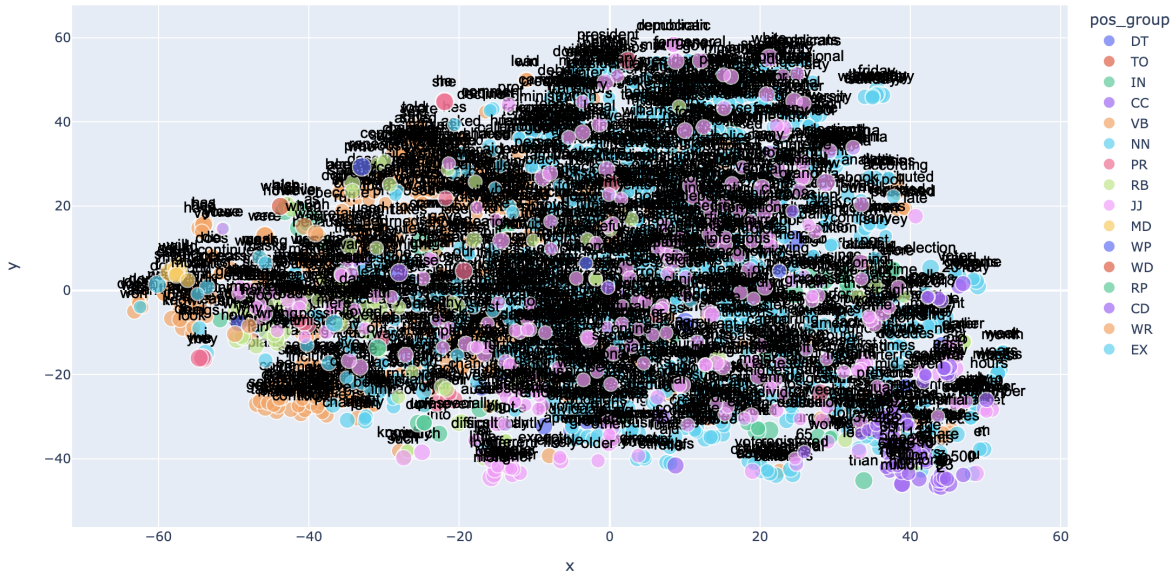


## Word2Vec

Lastly, we move into word2vec, which is a library we used to generate word embeddings and create models for the news articles from CNN and CNBC using high-dimensional vectors. We plotted the words on an x and y space to conduct a comparison and capture the semantic meanings and context behind them. Two scatterplots were produced and the coordinates were obtained through the application of t-SNE dimensionality reduction technique to the word vectors. The part of speech that the words represent are indicated by the different colors.

The first one represents CNN and the 2nd, CNBC.




# Interpretation

What was most interesting to me throughout this project was the comparison aspect. I was curious to see the different results produced by our analysis for CNN and CNBC and be able to make some assumptions or conclusions from that.

One of the main things I noticed was that the terms most associated with certain topics within CNN were more about the election and the candidates, while for CNBC, I was seeing more words about the pandemic and vaccines, police, and the word "black". It seems to me like the CNBC was more mistrusting of the situation in the world. Given that the channel in general is business and finance related, this would be something I would want to explore more to see how users interact with the content and feel.

# Conclusion

This project serves as an initial exploration of articles published by news channels and the potential to spread misinformation, influencing the sentiments that users feel about these topics. Through various methods, we have been able to extract the most important terms and concepts from these documents and conduct a primitive comparison of the different articles and channels in our dataset. It was evident through our analysis that there was a strong association with topics and words related to the elections and the individuals involved. This was expected. Overall, applying text analytical techniques to various articles and news channels offered interesting insights into the content of the texts and the sentiments of users around the elections and pandemic. In the future, it may be beneficial to incorporate more news channels into the exploration to get some variety. It would be cool to compare news channels that aren't based in the same country in order to test the differences in sentiment then and explore this idea further.