Assignment 5

**Haozhe Bai S5171326**

**Introduction**

I want to analysis the media tweet content between the two twitter corpus.I want to focus on the twitter content of Obama and Trump in the media to find out the patterns of use of these key terms, how these words have changed over time. As well as word preferences between the two former presidents.

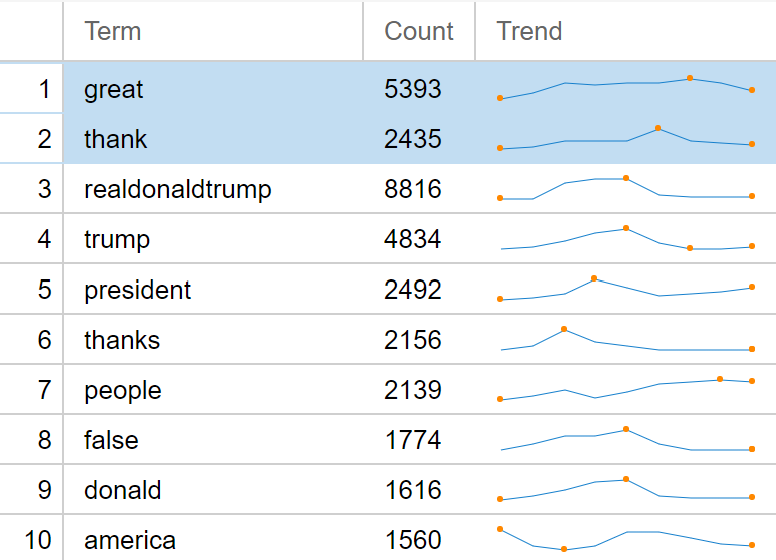
**Preprocess**

After importing the text file in txt format, I first needed to remove the deactivated words from the twitter text, the English stopwords that come with Voyant are not comprehensive. After application, the text also contained words or phrases such as http, t.co, etc. that had no practical or research value. I added the words http, https, t.co, â, rt, etc. to the deactivated words to enable better analysis of twitter text. After refining the deactivated words, I can get a clearer picture of the word frequency of twitter texts through the word cloud.

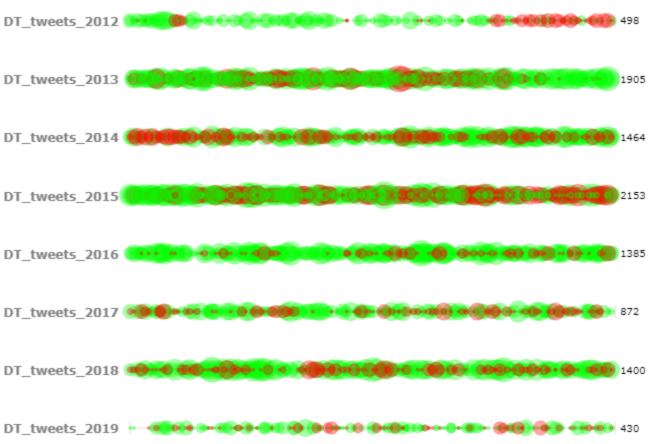
**Wordcloud**



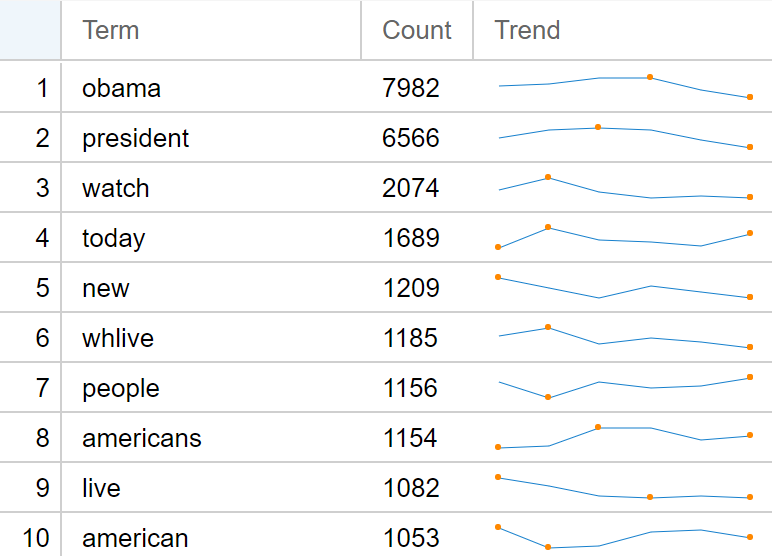
After removing the stopwords, we can compare the word clouds of the two corpus, Trump and Obama, to improve our understanding of the data. The most frequent words in both word clouds are the names of both presidents and 'president'. In addition, the most frequent words in Trump's word cloud are great, thank, people, America and false. The trend of the top 10 high frequency words shows an overall increase followed by a decrease. The most frequent occurrence of these words was in 2015-2016. It was during this time that he began his presidential campaign.



In addition, words such as great, thank, thanks and so on have a high frequency of occurrence. I used the category of words to classify the positive and negative words in the twitter texts. I used the four most frequent attitude words, great thank and false fake, to represent the distribution density of positive and negative words in these twitter texts respectively. Red is for negative words and green is for positive words. We can see that overall, Trump's twitter content uses more positive words than negative words. Negative words are mainly found from 2013 to 2016, and the proportion of negative words has decreased since he became the US President. Negative words were mainly used during the campaign period.

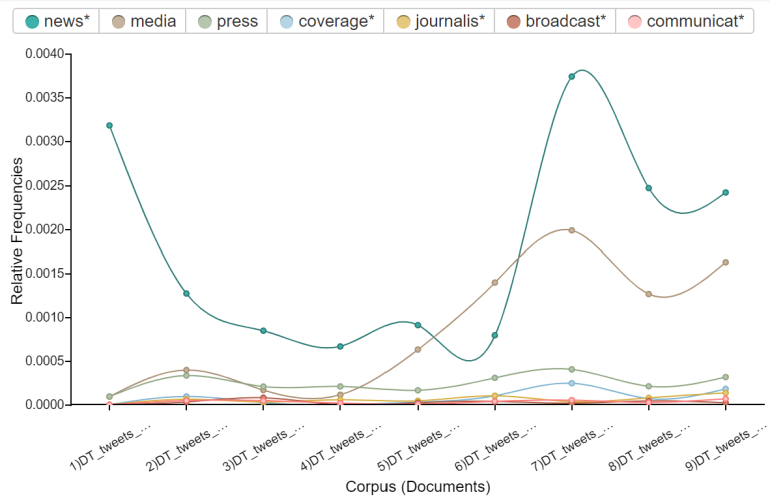


Let's shift our perspective to obama. The 10 most frequent words in the word cloud are obama, president, watch, today, new, whlive, people, americans, live, and america. these words are spread out more evenly over time, with peaks occurring mainly in the first to mid-term, around 2012---2013. Also, compared with Trump, he uses fewer positive or negative words in his twitter feed. His writing style is more down-to-earth. The most frequent word in his texts, good, is only 52nd.



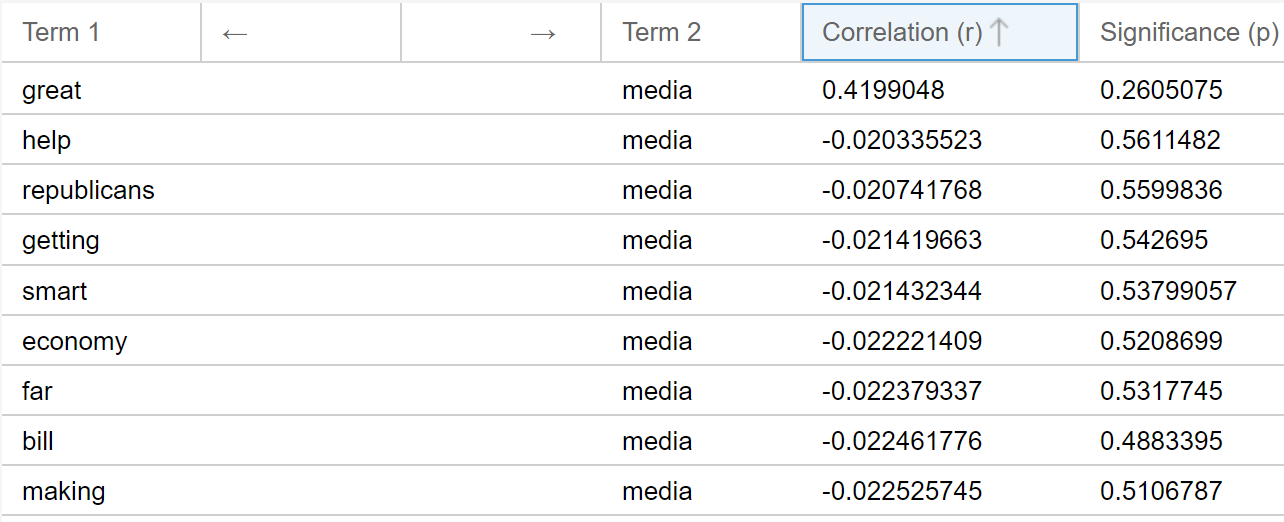
**Media——Trump**

Taking into account the frequency of words in the word cloud and the topic of media, I chose the seven words ‘media’ ‘news’ ‘broadcast’ ‘coverage’ ‘press’ ‘communication’ j’ournalism’ for this study of the topic of media.

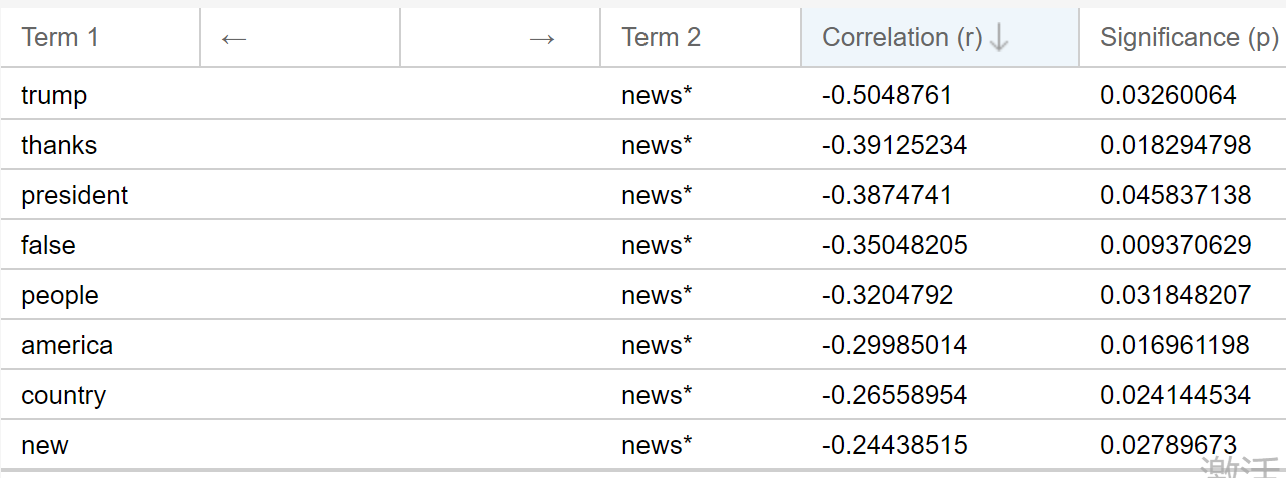


The words news and media appear most frequently in Trump's twitter corpus, far more frequently than any other words. Before we focus on these two words, we can focus on the words that occur less frequently. These words generally have a relative frequency below 0.0005. press is slightly more frequent than the other four words, but its relative frequency still does not exceed 0.0005. The temporal distribution of these words differs from that of the high frequency words. The frequency of the word Meida is generally increasing year by year, peaking in 2017. relative frequencies are close to 0.002, dropping slightly in 2018 and then rising again in 2019.

Having obtained the temporal distribution of the word media, I used Correlations to investigate which words media had a strong association with in the text. Interestingly, the only word with a positive pearson value for media was 'great'. This is different from our general understanding. We usually think of Trump as often taking to twitter to rebuke the media for broadcasting fake news. However, the word that is most relevant to the media is 'good'. This is despite the fact that the value of Significance(p) is a relatively large number, well above 0.05. The main reason this happens is that the word itself is a frequent word, ranking 18th in the word frequency list.

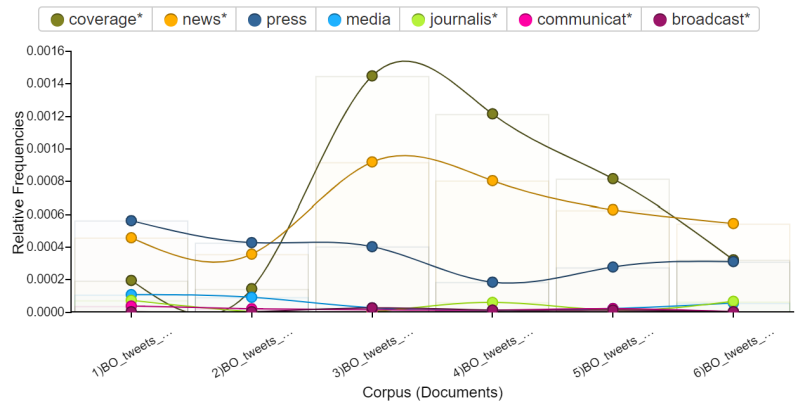


The relative frequency of News related words is higher in 2011 and 2017 - 2019. At its highest, the value reached 0.0037. There are some words that have a strong negative correlation with news, such as trump thanks. The value of Significance (p) is less than 0.05, which is significant.



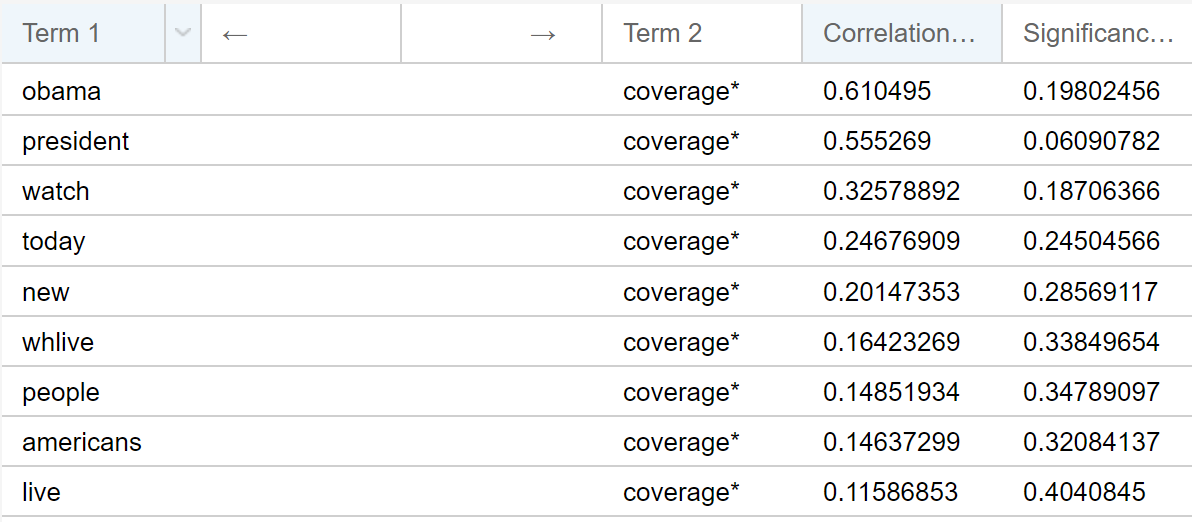
**Media——Obama**

Next, we use the same seven words in Obama's twitter corpus to analyse the media news broadcast coverage press communication journalism.



On the whole, the frequency of words in Obama's twitter corpus is lower than that of Trump's. Coverage peaked at 0.0015 in 2013. We can see that compared to Trump, Obama pays less attention to media issues. The media is much less concerned than Trump. Unlike Trump's media words, media is an extremely infrequent word in Obama's twitter corpus, peaking at around 0.0001. However, coverage and press, which have a low frequency in Trump's twitter corpus, have increased in frequency here.

The relative frequency of news increased from 2012 to 2013 and decreased from 2013 to 2016. And the word coverage has changed dramatically between 2012 and 2013. After 2013, the relative frequency of the word coverage has decreased every year. The correlations of coverage is shown in the diagram below.



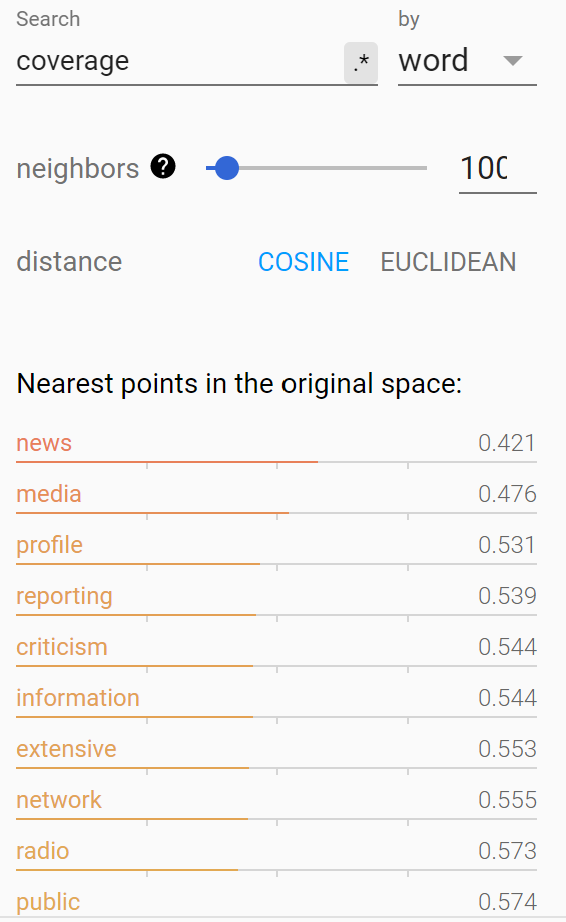
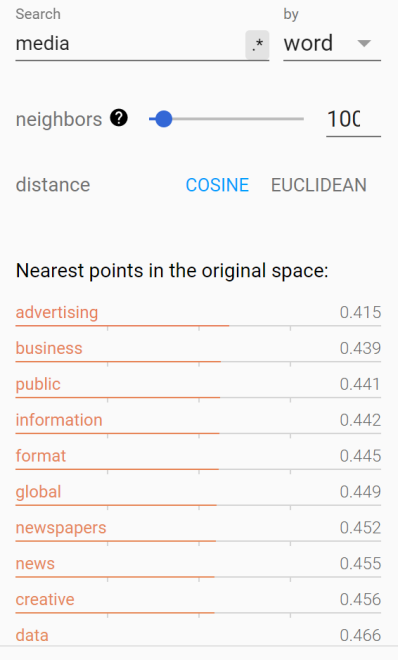
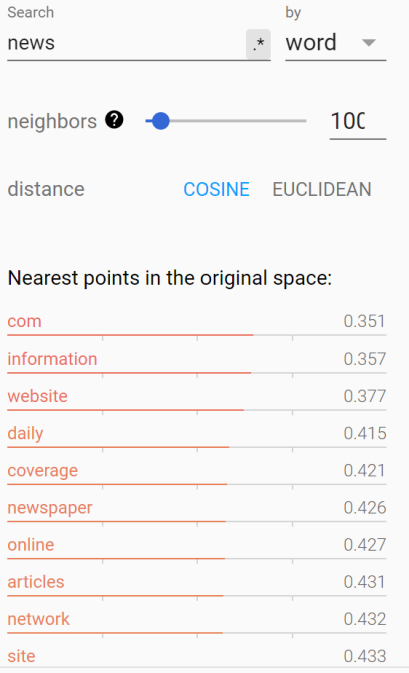
The relative frequency of news increased in 2012--2013 and decreased in 2013--2016.

**Conclusion**

In contrast, we can see that Trump is more concerned with media issues than Obama. He is more interested in this area. Most of the discussion was mainly focused on the period when he first started his election campaign, with the number dropping before 2016 and after he was elected president in 2018. trump was also more willing to use positive word or negative word, and his tweets were more emotionally adequate. I would also like to examine the extent to which they both have preferences for different media, but unfortunately there is no field for specific media names in the tweets.

**TensorFlow Embedding Projector**

Finally, I compared the selected key words with the word embeddings of these terms represented in TensorFlow's Embedding Projector with 'Word2Vec All' . I selected the words news, media and coverage, which appear more frequently in Obama's and Trump's twitter corpus, and looked for words in the model that were more closely related to these words.



In these pictures we can see that the words have a similar meaning. There are also many words that are repeated, such as information, newspaper, etc. These words are a good match for the keywords. They have similar meanings and are often used in the same sentence. When we compare the three words with each other, we can see that news has the highest degree of association with the nearest points words. The three words with coefficients below 0.4 are com, information and website, and we can assume that in this model the word news is usually used in the language of online news. This is the most common way of communicating news in this era. In contrast, the words that are more associated with these words in voyant are not strongly related in meaning, but are simply words that appear more often in corpus. When I removed more stopwords, the words that were more strongly associated were still words that were more common and not related to the media domain. This has a strong relationship with the model and the writing patterns in twitter.