MATH513 Big Data and Social Network Visualization Coursework

kostas, victoria, sid

11 January 2021

1 Data Preparation

```
## # A tibble: 10 x 3
##    speech                                                location   date
##    <chr>                                                 <chr>      <date>
##  1 "So thank you Pennsylvania, very much. I'm thrilled to~ Latrobe    2020-09-03
##  2 "Well, thank you very much. Thank you. Thank you very ~ Winston    2020-09-08
##  3 "We brought you a lot of car plants, Michigan. We brou~ Freeland   2020-09-10
##  4 "Well, I thank you very much. So I want to start by sa~ Minden     2020-09-12
##  5 "Thank you, thank you. Wow. Wow, and I'm thrilled to b~ Henderson  2020-09-13
##  6 "Thank you, thank you very much. Thank you very much. ~ Mosinee    2020-09-17
##  7 "There's a lot of people. That's great. Thank you very~ Bemidji    2020-09-18
##  8 "What a crowd, what a crowd. Get those people over her~ Fayettevi~ 2020-09-19
##  9 "Wow, that's a big crowd. This is a big crowd. Thank y~ Ohio       2020-09-21
## 10 "Doesn't have the power. Doesn't have the staying powe~ Pittsburgh 2020-09-22
```

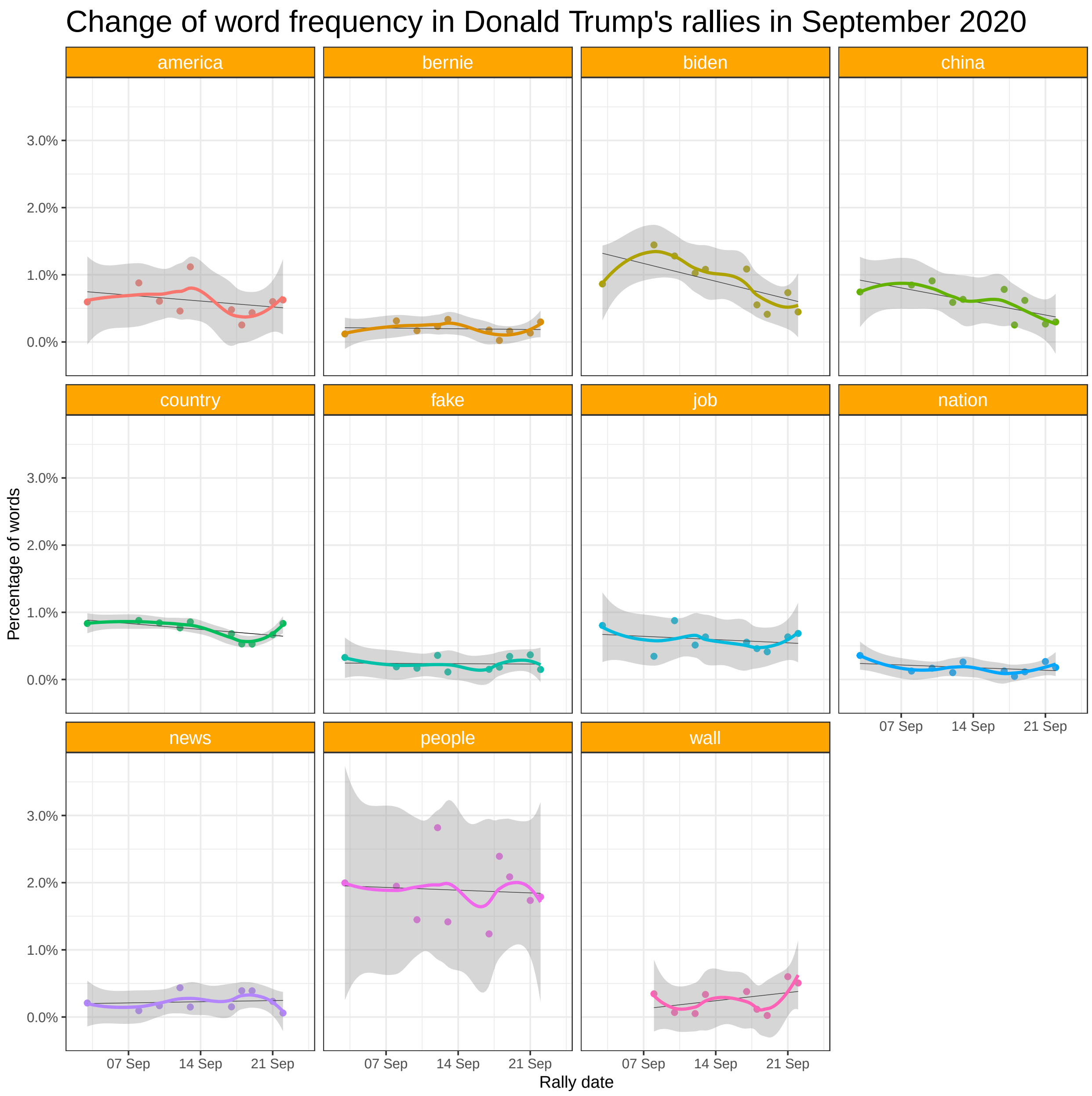
1.1 Discussion on Donald Trump's September 2020 rallies data

The dataset contains 10 full speeches of Donald Trump's Rallies, from 2020-09-03 until 2020-09-22. Furthermore, the location and date columns state the where and when each speech took place. In general, we can notice from the above table, that in all speeches Trump uses similar wording. In all of them Trump praises the crowd. Also, in most of them we can notice repetition of words or phrases:

- ***WINSTON: "Well, thank you very much. Thank you. Thank you very much. . . "***
- ***HENDERSON: "Thank you, thank you. Wow. Wow. . . "***
- ***MOSINEE: "'Thank you, thank you very much. Thank you very much."***
- ***FAYETTEVILLE: "What a crowd, what a crowd. Get those people over here. See me. Let them come over. What is a big crowd, it's always big."***
- ***OHIO: "Wow, that's a big crowd. This is a big crowd. Thank you very much, everybody. Hello to Swanton."***
- ***PITTSBURGH: "Doesn't have the power. Doesn't have the staying power."***

Same trend is spotted at the closing part of the speeches as well, with words high repetition of the word "America". Furthermore, the phrases "we will make America safe again" and "we will make America great again" were being used in all closings the speeches.

2 Writing an R Function Showing the Change of Word Frequency Over Time



2.1 Discussion of the the Change of Word Frequency Over Time

In overall the word people has highest frequency, as well as with the lowest central tendency, since most of the points are more scattered and further from the regression lane, resulting high standard errors. In the other hand the word nation seems to be the word with the less frequency.

More analytically, we can assume that the reason the word people was used that much, is due to the nature of public speeches, since the speaker was addressing to the people. The slope of the regression line is negative and close to zero.

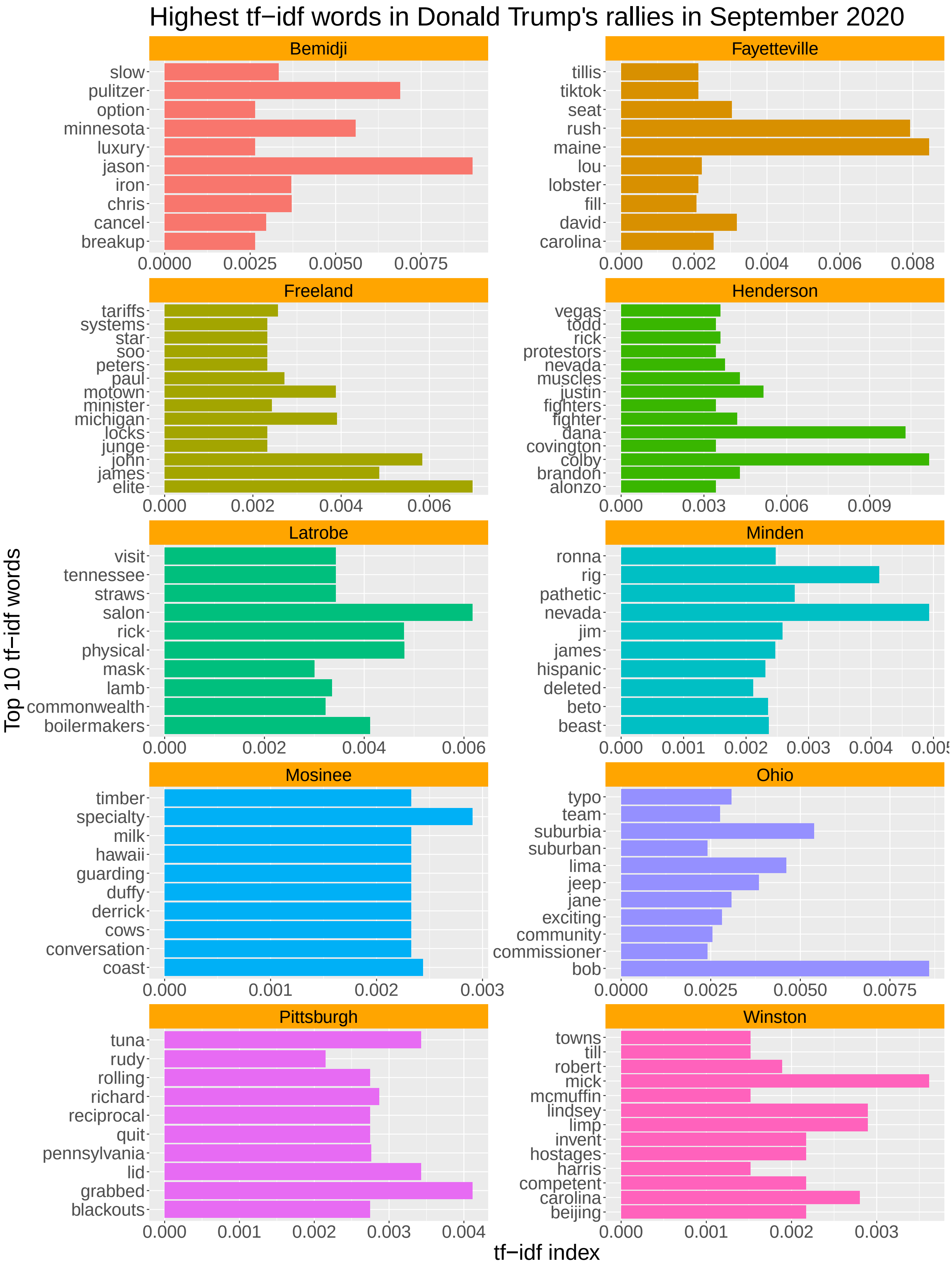
Furthermore Donald Trump, appears that he spent a lot of his speech time dealing with his political opponent Joe Biden and China. The word Biden is the next more frequent word. We can notice a slow fall in the usage of the word after its peak on the 8th of September, due to the slight negative slope of the curve. On this date ex-president Trump spent a lot of time attacking his opponent (1). He spoke negatively of him claiming Biden is ‘the dumbest of all candidates’ and pushed a false conspiracy theory that he uses performance enhancing drugs before attending debates. Also, the word China was used 206 times in total. According to the relevant graph, we can notice a moderate decrease in its overall usage, with a moderate variance since the distance of the points are not that away (low standard errors). China was accused by Trump, in most of his speeches, about Chinese tech companies funnel American citizens personal data (Huawei, Tik-Tok), as well as that China was incompetent to prevent the spread of Covid-19 to the rest of the world (2, 3). Trump spoke about ‘China’ when referencing COVID-19 calling it the ‘China virus.’

We can notice also that Trump preferred the usage of words country and America, compared to nation. We can assume that country had the most appearances (country:255, America:202) because it might refer to any country in general, while America and nation are referring to America. The word country appears also to have the lowest standard errors, since it had a stable percentage of usage in all speeches (slope close to zero), as well as nation. On the other hand, the word America seems to be overused Henderson speech on 13/09, and its points are more scattered. In general Trump talked about ‘America’ to all his speeches demonstrating his patriotism. This also referred to his famous slogan ‘Make America Great Again’ which was repeated many times.

The word ‘fake’ only has a 0.3%, however this seems mostly stable in all the speeches, since the slope of the line is close to zero. It is worth noticing that the word ‘news’ has almost the same pattern as the word ‘fake’. This due to his overuse of the phrase ‘fake news’. He has called journalists and news outlets “fake news” nearly 2000 times since the beginning of his presidency, during public speeches (4). Furthermore, seems that he spent equally amount of time mentioning the word ‘jobs’ to all his speeches since the slope is negative but close to zero, with a frequency from 0.6% to 0.5%. We can assume that he was referring to the creation of new ‘jobs’ if he were to be re-elected.

Finally, is worth to mention that in the words we have chosen, the wall is the only with positive regression line slope. Also, the ex-President had not used this word until the 8th of September, however it gets more popular especially till the end of our search range which reached its peak at around 0.5%. Wall refers to the wall at the borders between the US and Mexico, which was the signature promise of President Donald Trump’s 2016 election campaign (5).

3 Plotting the Words with Highest tf-idf Value



3.1 Discussion of tf-idf

Tf-idf gives the weight of the word in an overall corpus. A word with high tf-idf, has a high term frequency but low document frequency (the don't appear a lot in an entire set of corpus). In our case, it can show if a word has meaning or relationship to a specific speech. As we can notice that names such as: Jason, Bob, Dana, Colby, Mick appear to have the highest tf-idf scores ranging from 0.007-0.010. This is not unusual since those names have a particular connection to their respective speech location.

In the Bemidji Minnesota, Jason refers to Jason Lewis, former U.S. Representative for Minnesota's 2nd congressional district from 2017 to 2019 and member of the Republican Party.

The name Bob in the Ohio speech is referring to Bob Paduchik who was the Senior Advisor of Trump's re-election campaign, who was quoted by Trump that they were going to win Ohio. Also, to Bob Latta U.S representative of Ohio and member of Republican party.

In Henderson, the word Colby, refers to the UFC fighter Colby Covington, who is an outspoken supporter of the Republican Party and President Donald Trump, who joined him at the speech, while Dana, to Data White, President of the Ultimate Fighting Championship (UFC), another outspoken supporter.

Furthermore, after checking the original speech text itself for all those names, it appears that there is a pattern in which Trump repeats himself when calling someone's name:

- ***We go to again win Ohio. I understand from Bob Paduchik, you know Bob? That we're going to win it by more than we did.***
- ***Representative Bob Latta. Thank you, Bob. Great job. They like you, Bob. That's very good. Good job, Bob. Thank you very much.***
- ***. . . they don't fight like Colby. You know who Colby is. You're going to say hello to Colby. They don't fight like Colby. . .***
- ***Let's say you'd had a fight and you happen to meet Colby Covington. You say, "What's your name?" And he said, "My name's Colby Covington." And the first time I saw Colby Covington. . .***
- ***But I'd like to introduce Colby Covington. Great fighter. Great, great fighter. Incredible. He is a great fighter.***

According to (6) When he needs to plan his next sentence he often buys time by repeating himself. This reinforces the impression that he is supremely confident and that what he is saying is self-evident.

In Freeland, Trump constantly talked about the crowd being 'elite' and above middle class. This implies that he was giving them a sense of superiority so they would vote for him in the upcoming election.

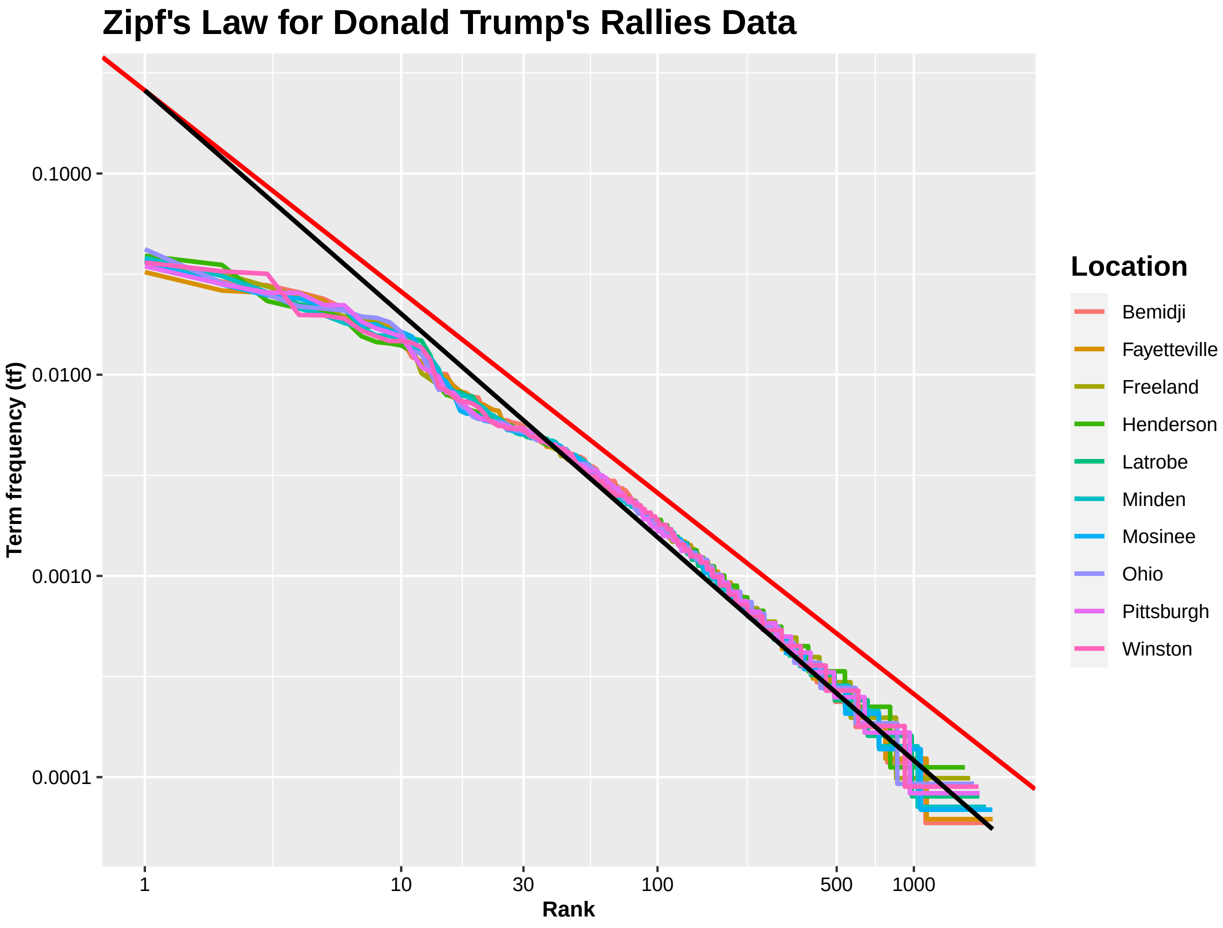
In the Minden rally, the name of the state 'Nevada' was mentioned a lot. This provides a sense of belonging in the (pro-Trump) crowd which reinforces group membership. He also mentioned the word 'rig' when talking about how the governor and Democrats were trying to rig the election as it would be the only way they could win.

Trump spoke about 'tuna' in the Pittsburgh rally and claimed that protesters were using cans and the fish to throw at law enforcement officers. There has been no evidence found of this claim (7).

Finally, we can notice the words with not that high score, such as tiktok, lobster, timber, cows, milk, hold a moderate score (0.02-0.03). The word tiktok appears to have a particular weight in Fayetteville, probably for one day before the speech (September 18), TikTok filed a lawsuit to Trump (8). Also, there Trump mentioned that he "opened the waters up for fishermen and for lobster men and women", referring to his policy regarding fisheries (9). The words timber, cows and milk show a relative weight in Mosinee, Wisconsin. This can be explained by the fact the Wisconsin has great agricultural industry and is a leading state of dairy products (10).

4 Zipf's Law on Trump's rally speeches

```
## (Intercept)    log_rank
##  -0.5871126  -1.1101344
```



4.1 Discussion of the graph

The plot above shows, the application of Zipf's law in the Trump's rally speeches. The scales are in logarithms with base 10. Furthermore, the black line is the regression line of our plot, while the red line is the theoretical regression line of Zipf's law. According to Zipf's law, the frequency of any word is inversely proportional to its rank in the frequency table (8).

From the graph we can notice that all the speeches follow almost the same trend, since all are concentrated together, especially for the ranks from 10 to 500. This can indicate that almost the same vocabulary, with the same word frequencies was used in all the speeches. However, for the words with lower rank (500-1000), can be distinguished. This means, that different words with greater meaning (words with lower frequencies tend to be words with higher meaning like nouns/verbs), were used in different locations. This is confirmed by the tf-idf graphs above (different words with high tf-idf appear on different locations).

We calculated the slope of our regression line in order to figure out the degree of deviation from Zipf's theoretical line (red line, slope = -1). In theory a Zipf's law curve will have a -1 slope, starting from the top left corner to the bottom right (9). In our case we notice that all the curves have some degree of deviation from this theory, especially for words with high rank. For words with rank 1 to 10, the curve is flatter, with a slope between -1 and zero. A flatter curve can indicate broader vocabulary for those ranks (9). However, the words in high ranks tend not to be that meaningful, mostly consisting of conjunctions.

In the other hand we can notice that from rank 10 the speeches are almost sticked to their regression line. The regression line of the speeches (black line) has a slope of -1.11 indicating a sharper degree of change. That means the lower the rank of a word (greater x-value), the less term frequency it has, compared to the Zipf's law theory. So, a sharper slope (less than -1) can indicate a poorer vocabulary, for the words from ranks 10 and after (more meaningful words).

5 References

1. Lybrand H, Subramaniam T. Fact check: Trump makes 10 false and misleading claims about Biden during North Carolina rally [Internet]. CNN. 2020 Sept 9 [cited 11 January 2021]. Available from: https://edition.cnn.com/2020/09/08/politics/donald-trump-north-carolina-rally-fact-check/index.html

2. McGraq M. Trump's first TikTok move: A China quagmire of his own making [Internet]. POLITICO. 2020 Sept 11 [cited 11 January 2021]. Available from: https://www.politico.com/news/2020/09/11/trumps-tiktok-china-412053

3. Saletan W. Trump's Excuse for His Coronavirus Lies Is Even More Incriminating [Internet]. Slate Magazine. 2020 Sept 10 [cited 11 January 2021]. Available from: https://slate.com/news-and-politics/2020/09/trump-woodward-book-panic-coronavirus-china.html

4. Woodward A. 'Fake news': A guide to Trump's favourite phrase – and the dangers it obscures [Internet]. The Independent. 2020 Oct 2 [cited 11 January 2021]. Available from: https://www.independent.co.uk/news/world/americas/us-election/trump-fake-news-counter-history-b732873.html

5. Poole T. Trump's border wall: A broken promise, a second term, or both? [Internet]. BBC News. 2020 Oct 7 [cited 11 January 2021]. Available from: https://www.bbc.co.uk/news/world-us-canada-49805982

6. Donald Trump's language offers insight into how he won the presidency [Internet]. The Economist. 2020 Aug 8 [cited 11 January 2021]. Available from: https://www.economist.com/books-and-arts/2020/08/08/donald-trumps-language-offers-insight-into-how-he-won-the-presidency

7. Dartunorro C. Bumble Bee responds to Trump's claim about protesters throwing cans of tuna [Internet]. NBC News. 2020 Sept 23 [cited 11 January 2021]. Available from: https://www.nbcnews.com/politics/politics-news/bumble-bee-responds-responds-trump-s-claim-about-protesters-throwing-n1240780

8. Yaffe B.T, Patterson E. TikTok Sues Trump Administration to Block U.S. Ban [Internet]. Bloomberg. 2020 Sept 19 [cited 11 January 2021]. Available from: https://www.bloomberg.com/news/articles/2020-09-19/tiktok-sues-trump-administration-to-block-u-s-ban

9. CARES Act provides more than $5M for North Carolina commercial fishing industry - Saving Seafood [Internet]. Saving Seafood. 2020 May 14[cited 11 January 2021]. Available from: https://www.savingseafood.org/news/state-and-local/cares-act-provides-more-than-5m-for-north-carolina-commercial-fishing-industry/

10. Wisconsin.Gov Home [Internet]. Wisconsin.gov. 2021 [cited 11 January 2021]. Available from: https://www.wisconsin.gov/Pages/Home.aspx

11. Piantadosi S. Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review. 2014;21(5):1112-1130. doi: 10.3758/s13423-014-0585-6

12. Allen DE, Mcaleer M. Fake News and Propaganda: Trump's Democratic America and Hitler's National Socialist (Nazi) Germany. Sustainability 2019;11(19):5181. doi: 10.3390/su11195181