The Speechanalysis Package

Authors: 10570155, 10696253, 10701983
Date: 14th January 2021

Usage:

```
library(speechanalysis)
```

Donald Trump Speech Data

The `speechanalysis` package comes with an example dataset. It consists of data detailing 10 speeches made by President Donald Trump throughout September 2020.

The dataset can be accessed like so:

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

Timing, location and election results

The speeches span from the 03/09/2020 to 22/09/2020 and take place in the following locations:

Date	City	State	Length of Speech	Election Result	Nearest Golf Course
03/09/20	Latrobe	Pennsylvania	01:31:29 - 1.52hrs	Biden	2.2mi - 4mins
08/09/20	Winston-Salem	North Carolina	01:13:55 - 1.23hrs	Trump	3.1mi - 7mins
10/09/20	Freeland	Michigan	01:21:06 - 1.35hrs	Biden	4.7mi - 8mins
12/09/20	Minden	Nevada	01:32:01 - 1.53hrs	Biden	7.2mi - 13mins
13/09/20	Henderson	Nevada	01:06:44 - 1.11hrs	Biden	4.7mi - 11mins
17/09/20	Mosinee	Wisconsin	01:32:36 - 1.54hrs	Biden	0.9mi - 4mins
18/09/20	Bemidji	Minnesota	01:48:20 - 1.81hrs	Biden	7.5mi - 14mins
19/09/20	Fayetteville	North Carolina	01:48:40 - 1.81hrs	Trump	9.1mi - 14mins
21/09/20	Vandalia	Ohio	01:05:53 - 1.10hrs	Trump	4.4mi - 11mins
22/09/20	Pittsburgh	Pennsylvania	01:29:51 - 1.50hrs	Biden	8.1mi - 10mins

Average legth of the speeches was 01:24:04 - 1.35hrs.

Average time to the nearest golf course by car from the airport was 9.6 mins at a distance of 5.19 miles.

Of the 10 places in the dataset only 3 of them ended up being won by Trump in the 2020 America Presidential Election.

Readability

The following table shows the readability calculation of each of Trump's speeches:

Date	City	Flesch-Kincaid Reading (Higher is easier)	New Dale Chall (Lower is easier)	Spache Readability (Lower is easier)	Flesch-Kincaid Grade (Lower is easier)	Gunning Fog Index (Lower is easier)	Coleman-Liau Index (Lower is easier)	SMOG Index Score (Lower is easier)	Automated Readability Index (Lower is easier)
14/09/20	Philadelphia	70	4.4	5	6.5	8.5	10.4	10	5.5
08/09/20	Winston-Salem	87.2	3.1	5	3.4	6	6.7	8	1.1
10/09/20	Freeland	82.5	3.5	5	4.3	6.9	7.3	8.8	2.2
12/09/20	Minden	82.2	3.6	5	4.2	6.7	7.7	8.6	2.3
13/09/20	Henderson	82.7	3.3	5	4.3	6.8	7.4	8.6	2.3
17/09/20	Mosinee	83.8	3.4	5	4	6.6	7.4	8.5	2
18/09/20	Bemidji	82.9	3.2	5	4.5	7.4	7.2	8.9	2.7
19/09/20	Fayetteville	84.7	3.3	5	3.7	6.3	7.2	8.3	1.6
21/09/20	Vandalia	85.5	3.1	5	3.7	6.4	7.1	8.2	1.7
22/09/20	Pittsburgh	79.9	3.5	5	5	7.7	7.6	9.2	3.1

We can then compare this to an equivalent speech by Joe Biden seen below:

Date	City	Flesch-Kincaid Reading (Higher is easier)	New Dale Chall (Lower is easier)	Spache Readability (Lower is easier)	Flesch-Kincaid Grade (Lower is easier)	Gunning Fog Index (Lower is easier)	Coleman-Liau Index (Lower is easier)	SMOG Index Score (Lower is easier)	Automated Readability Index (Lower is easier)
09/09/20	Warren	74.8	4.1	5	5.1	7.7	9	9.5	3
14/09/20	Philadelphia	70	4.4	5	6.5	8.5	10.4	10	5.5
21/09/20	Manitowoc	84.7	3.3	5	3.6	6.3	7.6	8.2	1.6
30/09/20	Greensburg	84.4	3.1	5	3.5	6.2	7.4	8.1	1.2

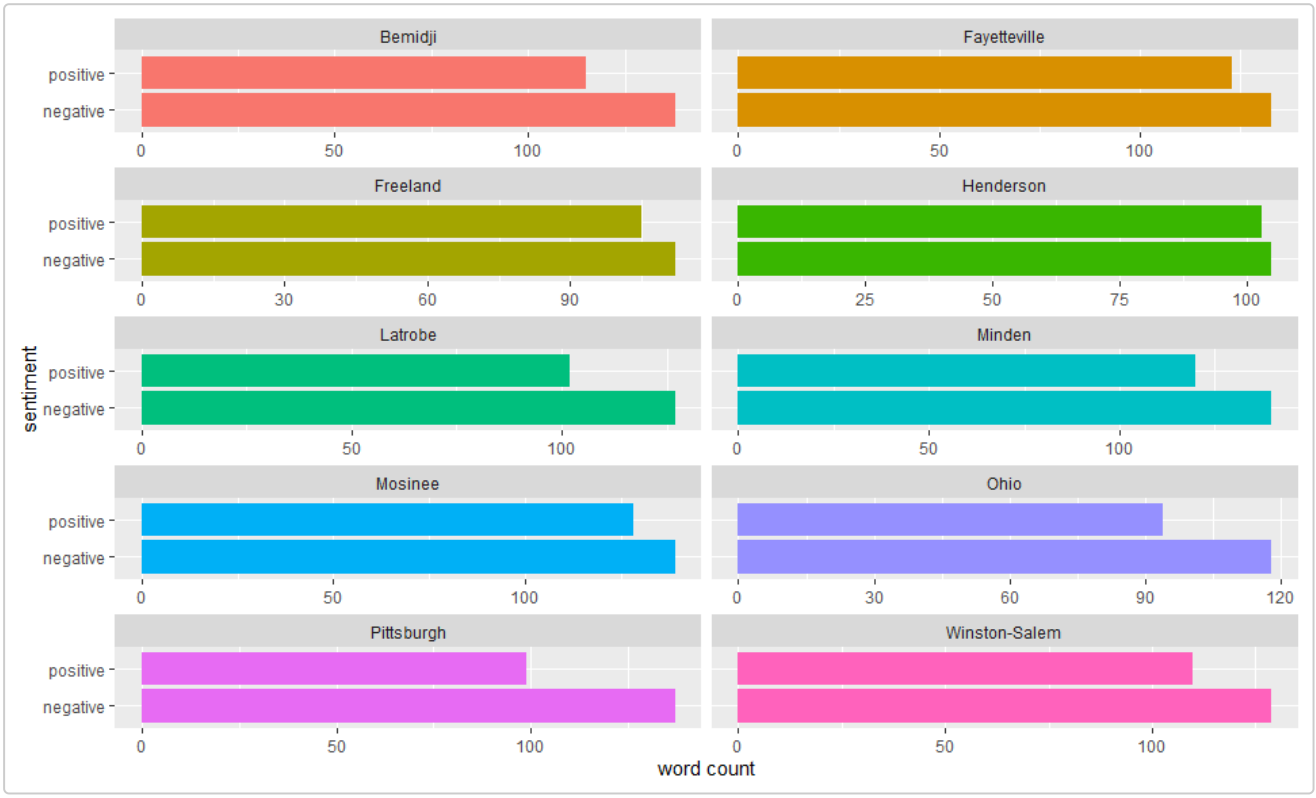
From this, we can determine the averages for each of the scores and compare them:

Candidate	Flesch-Kincaid Reading (Higher is easier)	New Dale Chall (Lower is easier)	Spache Readability (Lower is easier)	Flesch-Kincaid Grade (Lower is easier)	Gunning Fog Index (Lower is easier)	Coleman-Liau Index (Lower is easier)	SMOG Index Score (Lower is easier)	Automated Readability Index (Lower is easier)
Trump	84.1	3.3	5.0	4.0	6.7	7.2	8.5	2.0
Biden	78.5	3.7	5.0	4.7	7.2	8.6	9.0	2.8

We can see that the readability of Trump's speeches is on average easier than those of Biden's speeches. This could be due to them attempting to appeal to different voters. It is worth noting that Biden's speeches tend to be short, at an average of 26 mins.

Sentiment

We performed a sentiment analysis of Trump's speeches by individual words using the Bing lexicon. We removed stop words before processing and came up with the following chart.



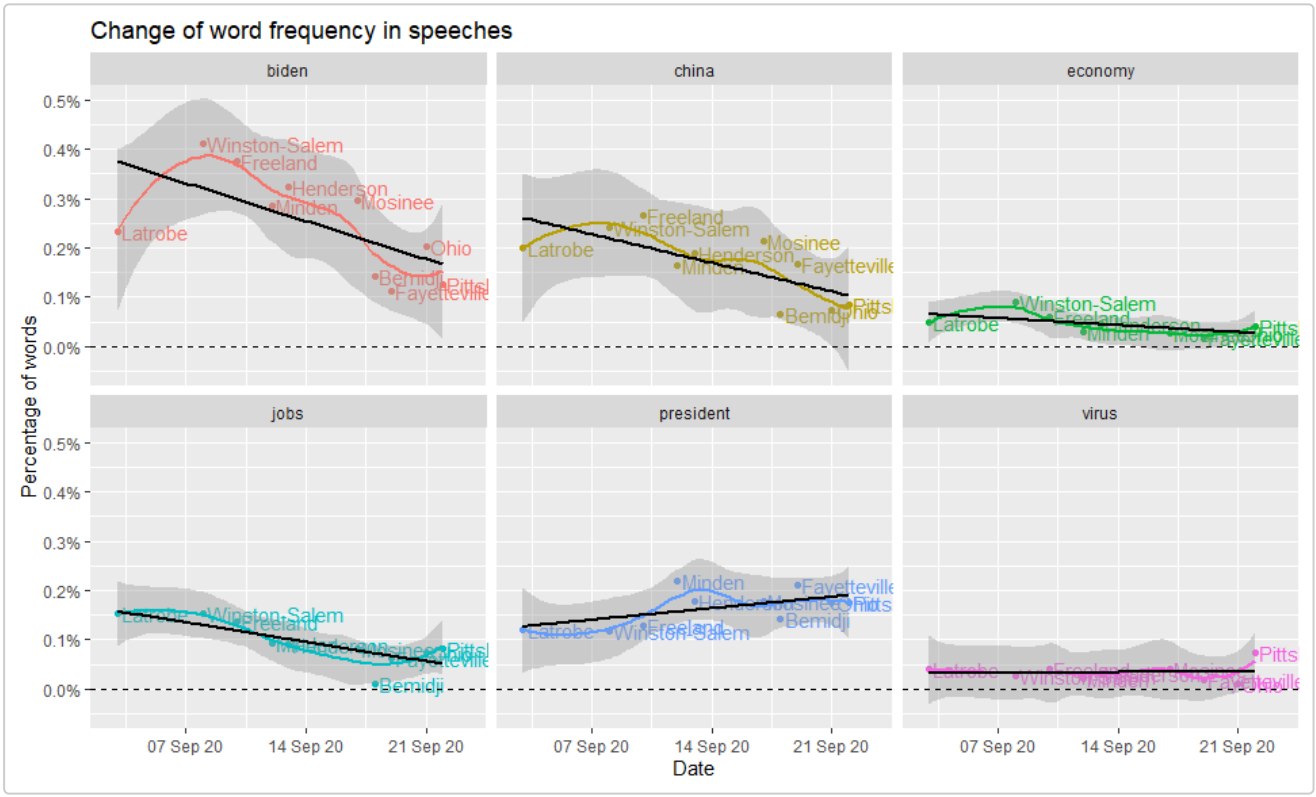
As you can see from the chart, Trump's speeches carry a more negitive sentiment. Out of the ten speeches none of them carry a mainly positive sentiment.

We believe that for better understanding, we could utilise the sentimentr package to analysis sentiment based on whole sentences.

The Change in Word Frequency Over Time

We can use the `word_frequency(df, c)` function, *where `df` is the data frame to apply to and `c` is a collection of words to check and plot ,* to plot a graphical summary of the frequency of words over time from the speeches.

```
word_frequency(trump_speeches, c("virus", "biden", "china", "president", "economy","jobs"))
```



The graphs show us the frequency of the words (y-axis) in each of the speeches (point on the graph) plotted over time (x-axis). A trend line is added to help identiy the usage of the words by Trump in his speeches. A horizontal line has been added to aid in seeing the percentage above zero for some words that are used very infrequently.

In the graph above, it is interesting to note the following:

- Virus
  - Trump rarely uses the word virus or in fact, when tested, any word we could consider, relating to the COVID-19 pandemic. However, it is still mentioned in each of the speeches in a very small amount
- Economy
  - Used only slightly more than the word virus but still not very often
- Jobs
  - Used less and less over the month in a downward trend

Based on these three words and their frequency, it could be concluded that he is trying to avert attention from these sensitive and current topics. The estimated cost to the Americans of the virus has been estimated at $16 trillion according to Gandal (2020) of CBS news.

Looking at the following words:

- President
  - Used more frequently as the election day approaches
- Biden
  - Used less frequently as the election day approaches

We can see that there is an increase in the usage of president and decrease in the word Biden. This could be in relation to Trump wishing to downplay his competitor in the race for the presidency and associate himself with the word president.

- China
  - Downward trend with Trump using it less frequently in his speeches towards the end of the month of September

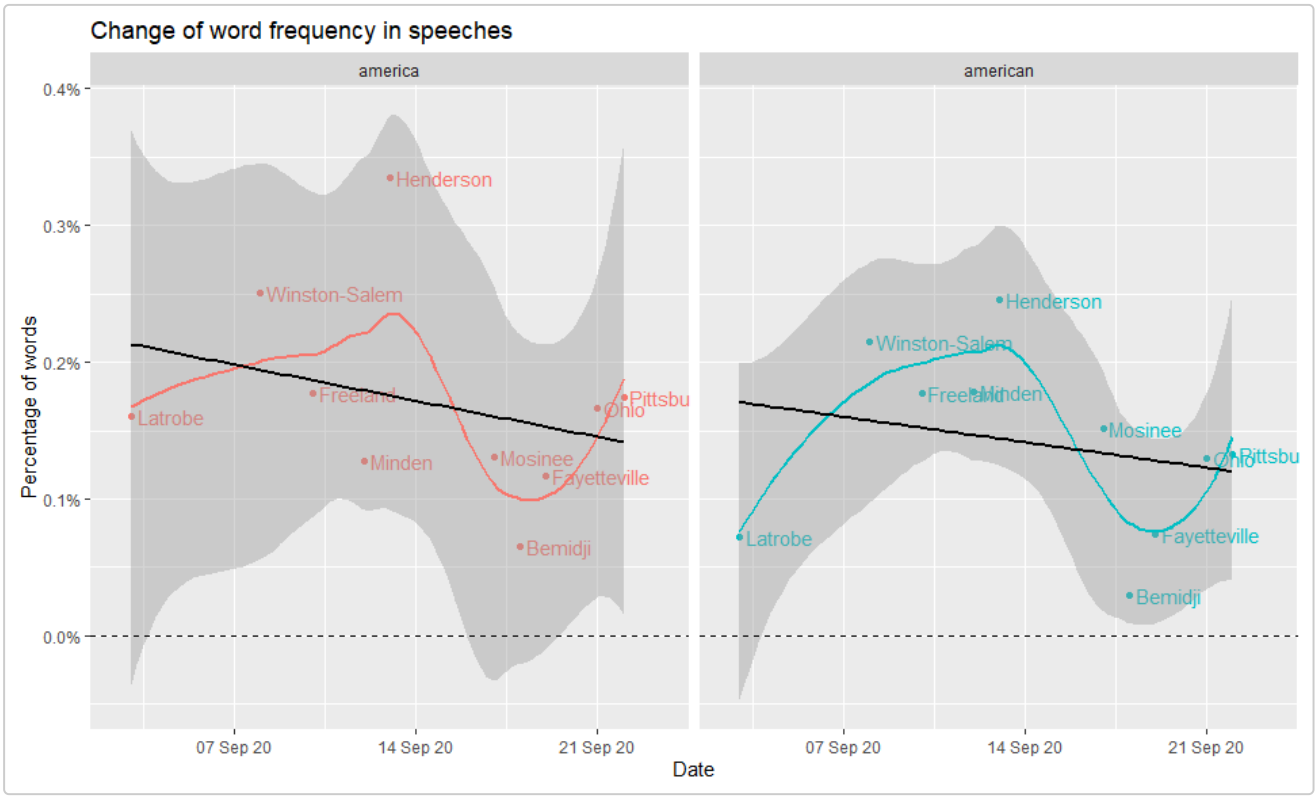
Tensions between the US and China were evident at the UN assembly on 15 September 2020 (BBC, 2020). Trump may be reducing the number of times China is mentioned in the same way that he is reducing the number of times he mentions Biden, the economy, jobs, and the virus. This is to distance himself from what could be seen to be negative.

One of the most commonly used words was America. This is due to each of the speeches finishing with the following phrases:

- make America wealthy again
- make America strong again
- make America proud again
- make America safe again
- make America great again

This can be seen reflected in the frequency graph below.

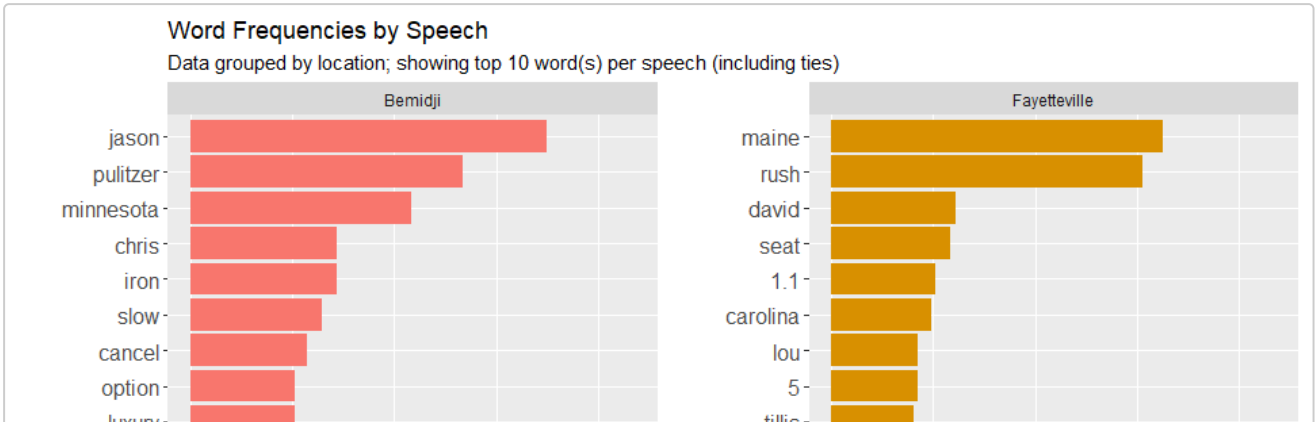
```
word_frequency(trump_speeches, c("america", "american"))
```

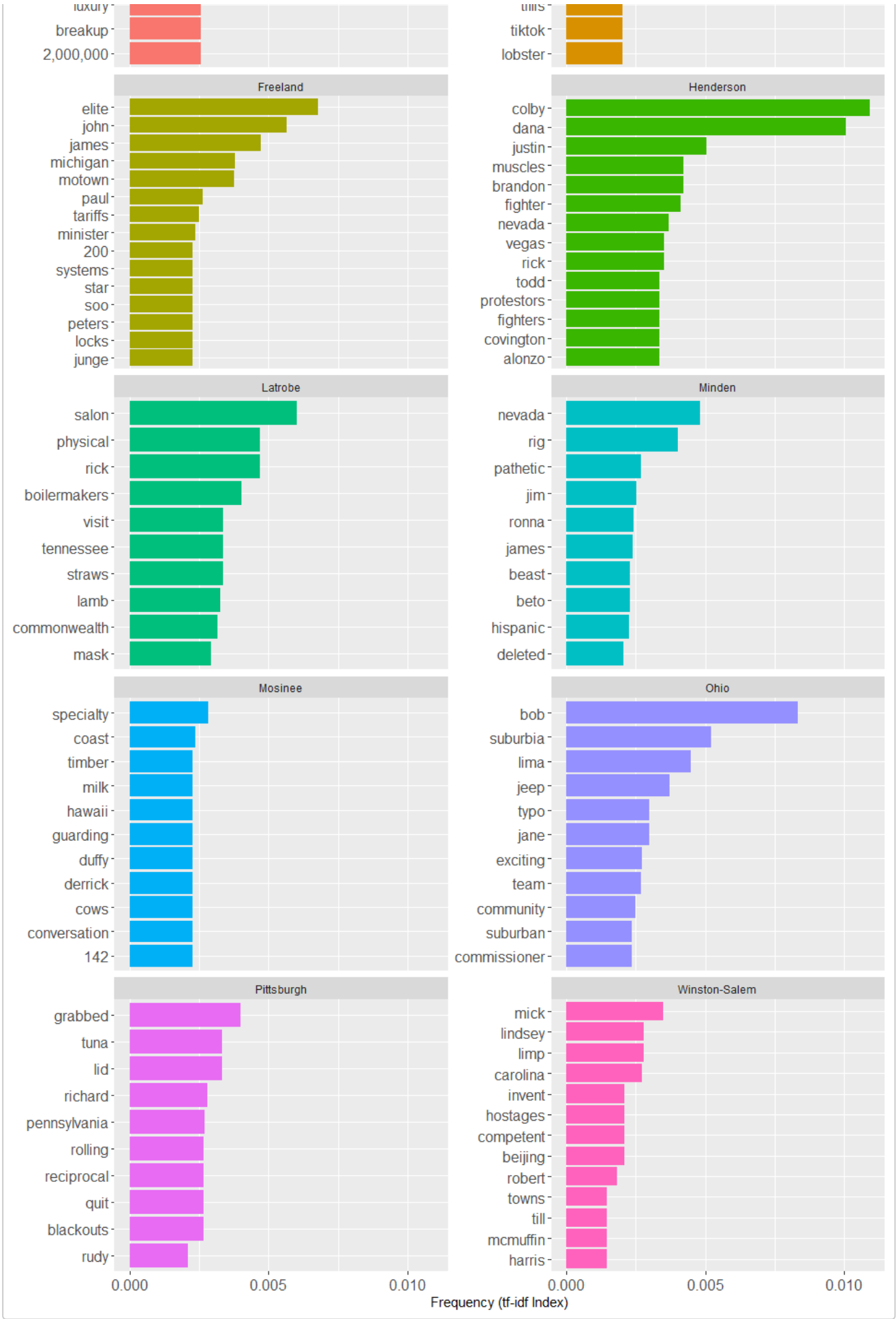


Frequency of Words per Speech

We can use the `plot_tf_idf(df, n)` function, *where `df` is the data frame to apply to and `n` is the number of 'top' words to show,* to plot a graphical summary of the most frequent words which occur in each speech. This function makes use of the **tf-idf index** which computes the frequency of a word adjusted for how rarely it is used.

```
plot_tf_idf(trump_speeches, 10)
```





Bemidji, Minnesota

‘**Jason**’ refers to Jason Lewis, a Republican who was running for Senate in September 2020 and has said that he *“won’t distance himself from President Donald Trump or his policies”* while running for federal office.

‘**Pulitzer**’ refers to the *Pulitzer Prize*, an award for achievements in newspaper, magazine, and online journalism. Trump explains how he doesn’t believe that organisations such as the New York Times and The Washington Post deserve the prizes they have been awarded.

‘**Chris**’ refers to Chris Vreeland, the mayor of Hoyt Lakes, Minnesota. All mentions of Chris are from a single statement where Trump is thanking him.

Fayetteville, North Carolina

‘**Rush**’ refers to Rush Limbaugh, an American radio personality, who Trump gave the Medal of Freedom to during Trump’s State of the Union address in February after the broadcaster was diagnosed with cancer. All mentions of Rush in this speech are instances where Trump is praising him for his support.

‘**David**’ and ‘**1.1**’ refer to a $1.1 billion construction project involving David Friedman, an American economist and lawyer.

‘**TikTok**’ - The speech contains two references to how Trump is *“very close”* to a *“great deal”* with Tik-Tok.

Freeland, Michigan

’**Elite’** - Trump is referring to the audience; *“I think most of you are not middle class. You’re upper class”*.

‘**John**’ and ‘**James**’ refers to John James, the Michigan U.S. Senate candidate that Trump is endorsing.

‘**Motown**’ - *“That’s right, Motown. We love Motown. Michigan gave us Motown, gave us Motown”*.

‘**Paul**’ and ‘**Junge**’ refers to Paul Junge, a republican politician.

‘**Tariffs**’ and ‘**200**’ - Tariffs with Japan and China for imported cars was one of the key points of discussion during this rally. Trump mentions that his administration has added *“over 200 new auto and auto parts plants”*.

Henderson, Nevada

‘**Colby**’ and ‘**Justin**’ refers to Colby Covington and Justin Gaethje; two mixed-martial artists.

‘**Dana**’ is a reference to Dana White: the president of the UFC.

‘**Brandon**’ refers to Brandon Judd, president of the National Border Patrol Council. Trump invites Brandon to pick out a fighter, so he can see *“whether he can take Colby in a fight”*.

Latrobe, Pennsylvania

‘**Salon**’ refers to a controversy involving Nancy Pelosi, the Speaker of the United States House of Representatives. She was seen in a hair salon without a mask; the hair salon in question was pressured into shutting down as a result of this controversy. A GoFundMe raised $300,000 for the owner.

‘**Physical**’ - Trump discusses a physical he had from the White House doctor, Sean Conley.

‘**Rick**’ refers to Rick Perry, a republican who Trump was endorsing.

Minden, Nevada

‘**Rig**’ - Trump mentions how *“they”* are trying to rig the election. *“[the governor] is in charge of the election and the millions of ballots…he can rig the election”*.

‘**Pathetic**’ - Insults directed towards Joe Biden and the Mayor of Portland.

‘**Jim**’ refers to Jim Jordan, a U.S. representative who was awarded the medal of freedom by Trump.

‘**Ronna**’ refers to Ronna McDanial, the chair of the republican national committee. Trump is praising her.

‘**James**’ refers to James Settelmeyer, the Nevada state senator. Trump is praising him.

‘**Beast**’ refers to Air Force One.

Mosinee, Wisconsin

‘**Specialty**’ and ‘**Milk**’ - Specialty Milk.

‘**Coast**’ - Trump praises the U.S. Coast Guard.

‘**Timber**’ - The timber industry is prominent in Wisconsin; Trump mentions how he has saved *“many timber jobs in [Wisconsin]”*.

‘**Hawaii**’ - *“Remember the crazy senator from Hawaii?”*

‘**Guarding**’ - *“we right now have 27,000 soldiers from Mexico guarding our border.”*

‘**Duffy**’ refers to Sean Duffy, a former U.S. representative. Trump talks about how he was the world champion tree climber.

‘**Derrick**’ refers to Derrick Van Orden, the republican congressional candidate. Trump is praising him.

Vandalia, Ohio

‘**Bob**’ refers to a number of people: Bob Paduchik (senior advisor for Trump’s 2020 re-election campaign), Bob Latta (U.S. representative), and Bob Cupp (speaker of the Ohio House of Representatives)

‘**Suburbia**’ - *“I ended a regulation that will destroy suburbia”*.

‘**Lima**’ - Trump mentions how he *‘saved Lima’*: *“I came here and they were going to close that plant in Lima…and I said why are we closing a tank plant?”*

‘**Jeep**’ - Jeep originates from Ohio.

‘**Jane**’ refers to Jane Timken, the chairwoman of the Ohio republican party. Trump is praising her.

‘**Grabbed**’ - Trump talks about how the National Guard handled the riots: *“They grabbed one guy…they throw him aside like he was a little bag of popcorn…it’s a beautiful sight.”*

‘**Lid**’ - Joe Biden called a ‘lid’, which is a term used by reporters when a politician has finished travelling for the day.

‘**Richard**’ - Trump tells a story about himself and one of his friends.

Winston-Salem, North Carolina

‘**Mick**’ refers to Mick Mulvaney, an American politician. Trump is praising him.

‘**Lindsey**’ refers to Lindsey Graham, a U.S. senator. Trump is praising him.

Tf-idf Conclusion

Using the above insights and the tf-idf chart, it’s clear that Trump’s speeches were tailored to the location of the speech. Trump covers topics which are relevant to the audience, presumably to increase his appeal in order to maintain and improve upon his supporter-base.

Zipf’s Law

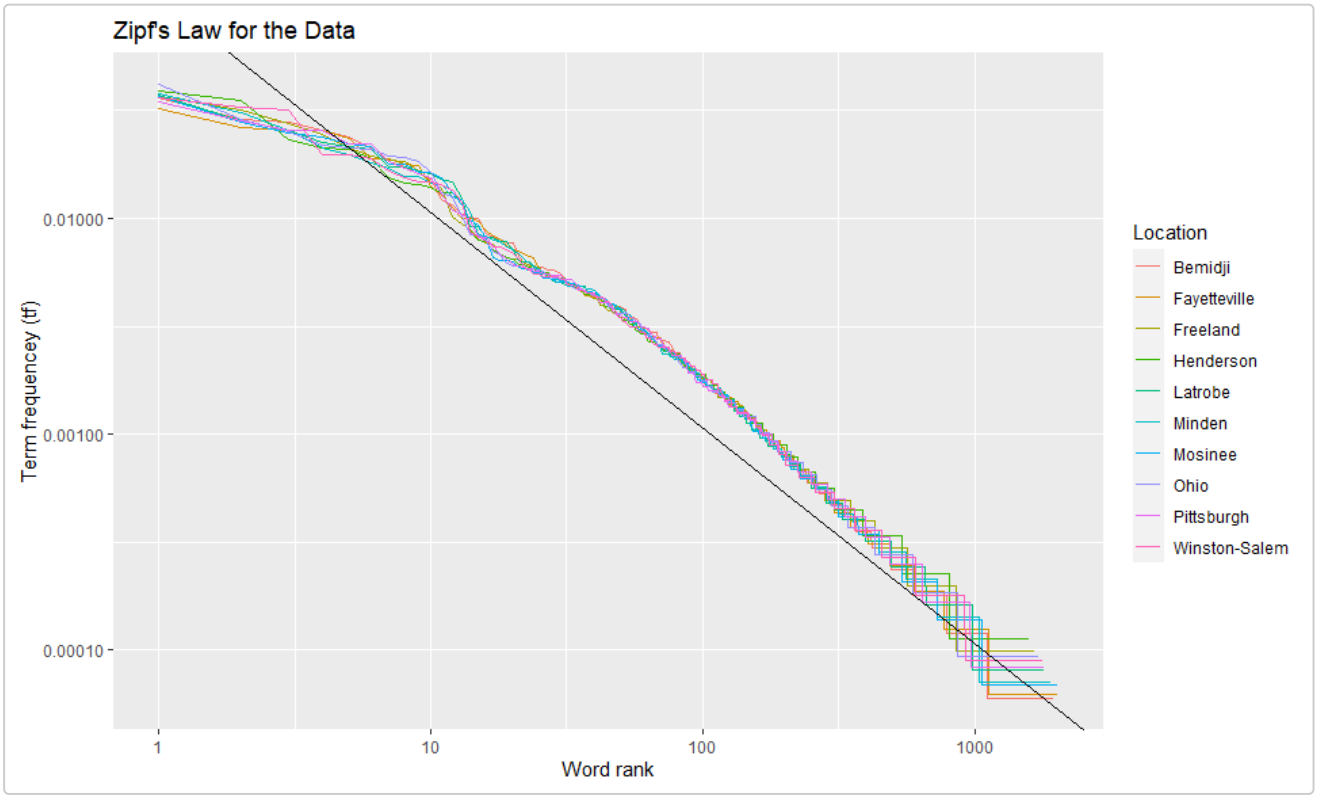
Zipf’s law states that the word frequencies in any text datasets will be very similar to Zipf’s distribution - the number of occurrences of the Nth most frequent word would be X/N, where X is the frequency of the most frequent word.

Most recent studies of this phenomena show that typically there exists a certain value of α which influences the distribution so that the frequency of the Nth word is described as: X/N^{α}. However, in order to explore

Zipf's law in its standard form, we defined $\boldsymbol{\alpha} = 1$.

Utilising Trump's speeches, grouped by the locations, we have plotted the relationship between the word frequencies and their respective ranks with a linear regression model.

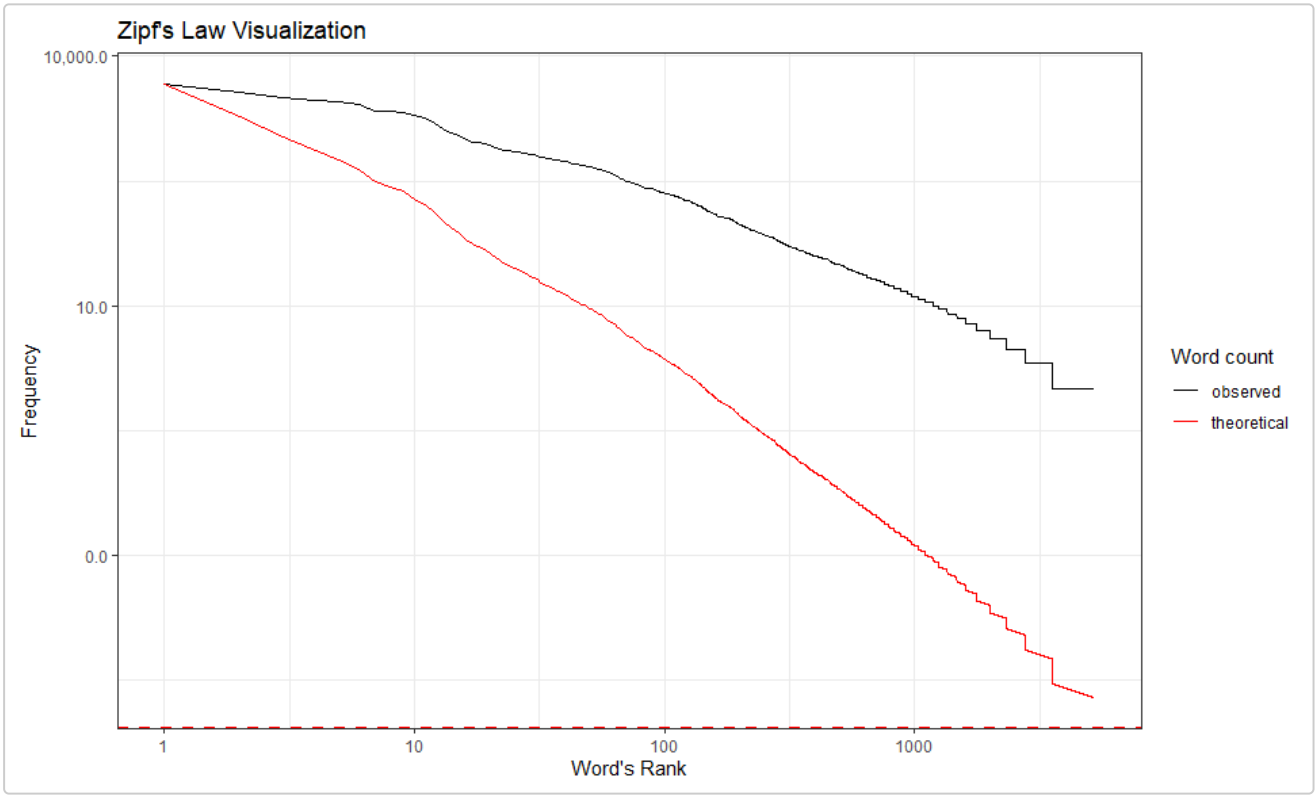
```
zipfs_law(trump_speeches, 2)
```



Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data (Yale University, 1998). In our case, we're trying to determine a relationship between the word rank and the term frequency.

In general, the 'term frequency - word rank' curve follows the regression line trend. However, we can see that Trump uses words with ranks of 10-100 more frequently than predicted by the linear regression model. On top of this, the predicted term frequency for the most frequent word (which has rank No1) is supposed to be bigger according to the linear regression model. This shows us that Trump's speeches do not follow the model exactly.

```
zipfs_law(trump_speeches, 3)
```



In order to explore how much Trump's speeches deviate from the theoretical Zipf's law, we counted frequencies for each word across the dataset of all the provided speeches and plotted them alongside the word frequencies calculated using Zipf's law. The red dash line indicates the absolute 0-point on Y-axis.

As it can be seen, Trump's word frequencies are overall higher than the predicted ones as the line with observed values declines more slowly. This means that he uses more popular and simple words (such as 'the', 'and', 'a', 'you', 'up', 'to', etc.) more often than they should normally be used according to Zipf's law.

References

- BBC (2020) UN General Assembly: US-China tensions flare over coronavirus. Available at: https://www.bbc.co.uk/news/world-54253408 (Accessed: 13 January 2021)
- CNN Politics (2020) Presidential Results. Available at: https://edition.cnn.com/election/2020/results/president#mapmode=lead (Accessed: 20 December 2020)
- Gandel, S (2020) Coronavirus pandemic to cost Americans $16 trillion, study finds. Available at: https://www.cbsnews.com/news/coronavirus-pandemic-cost-americans-16-trillion/ (Accessed: 20 December 2020)
- Maj, M (2019) Investigating words distribution with R – Zipf's law. Available at: https://www.r-bloggers.com/2019/02/investigating-words-distribution-with-r-zipfs-law/ (Accessed: 10 January 2021)
- Rev (2020) Joe Biden Climate Change Speech Transcript in Wilmington, Delaware September 14. Available at: https://www.rev.com/blog/transcripts/joe-biden-climate-change-speech-transcript-september-14 (Accessed: 21 December 2020)
- Rev (2020) Joe Biden Campaign Speech Transcript Manitowoc, Wisconsin September 21. Available at: https://www.rev.com/blog/transcripts/joe-biden-campaign-speech-transcript-manitowoc-wisconsin-september-21 (Accessed: 21 December 2020)
- Rev (2020) Joe Biden Train Tour Campaign Speech Transcript Greensburg, PA September 30. Available at: https://www.rev.com/blog/transcripts/joe-biden-train-tour-campaign-speech-transcript-greensburg-pa-september-30 (Accessed: 21 December 2020)
- Rev (2020) Joe Biden Speech Transcript Warren, Michigan September 9. Available at: https://www.rev.com/blog/transcripts/joe-biden-speech-transcript-warren-michigan-september-9 (Accessed: 21 December 2020)

- Yale University (1998) Linear Regression. Available at: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm (Accessed: 13 January 2021)