Prepare for Class 7

Oliver Myers

What is Cross-Validation?

Cross Validation Is a procedure in which one will use to evaluate the effectiveness of their model by looking at “fit statistics” and effects of potential “overfitting”. Cross-Validation also helps in selecting the best model and also selecting the best features for a model.

Big Data Sampling

How to sample

Take a subset from the larger set and determine what the best variables are to use for predicting outcomes.

You want to designate one group left out when training a model so you can use the model on it later to evaluate the effectiveness of the model.

Why it’s important

This will allow us to make predictions of output data we don't have based on on input data. (ex cancer research use case)

Big Data Sampling

Issue in big data sampling

We have to account for different kinds of error in addition to the model making generalizable prediction to a larger population.

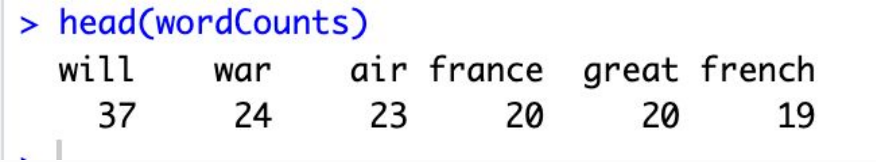
How to solve

Cross validation via using the "**holdout method**" to sample big data. (Training, tuning, testing) finds the best

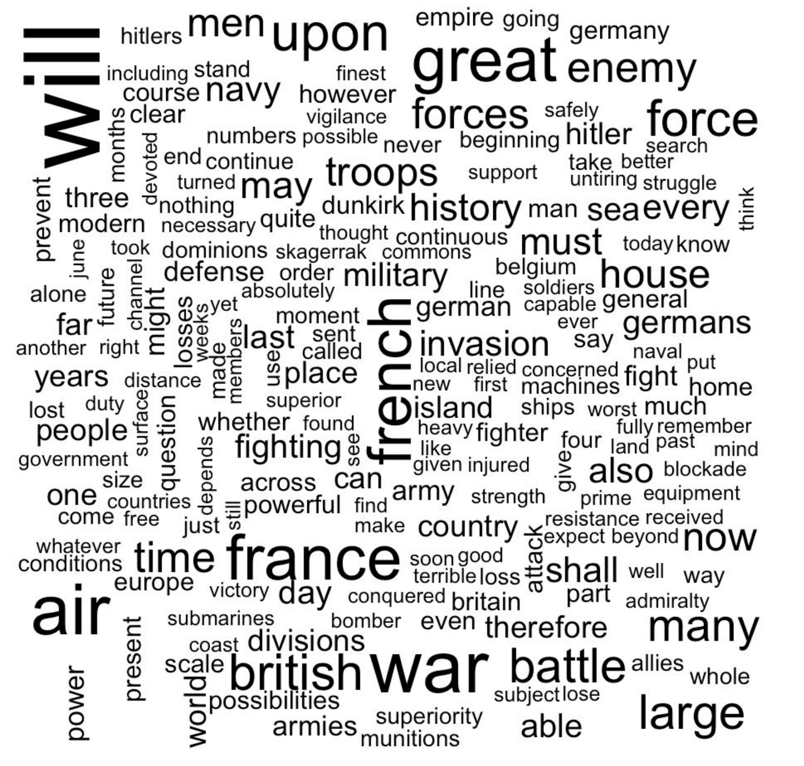
How to solve

Random placement of cases from the large dataset into various groups used for training, tuning and testing.

textdata_mining01



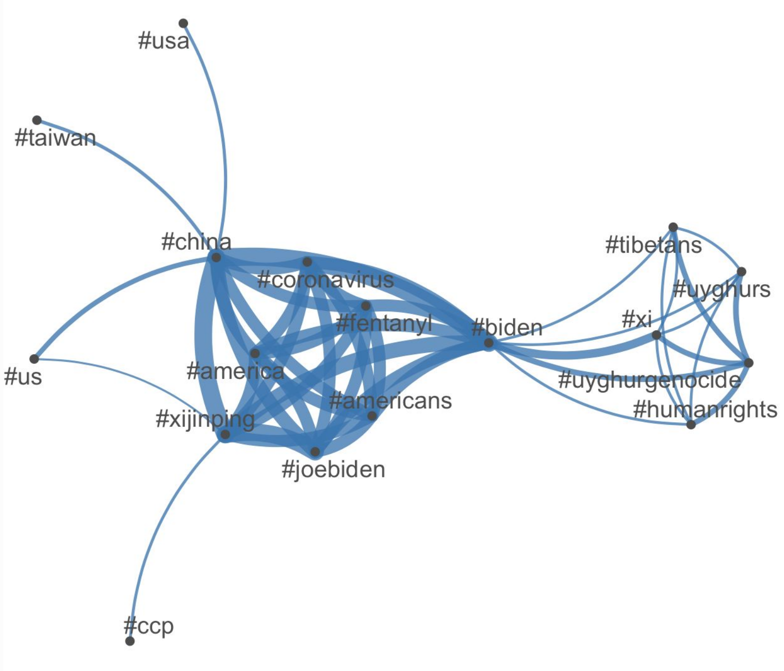
```
> head(wordCounts)
  will    war    air france  great french
    37     24     23     20     20     19
```

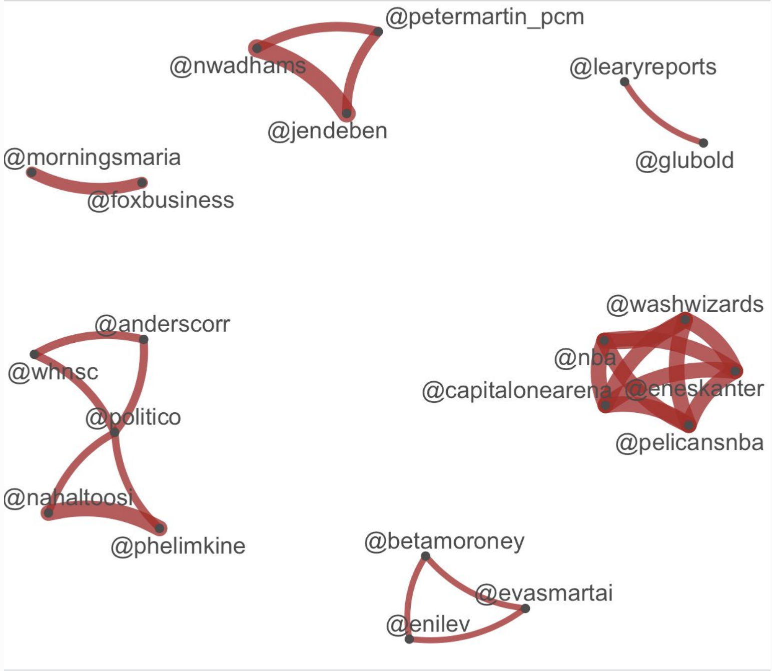


Text Data Mining on Winston Churchill's Speech

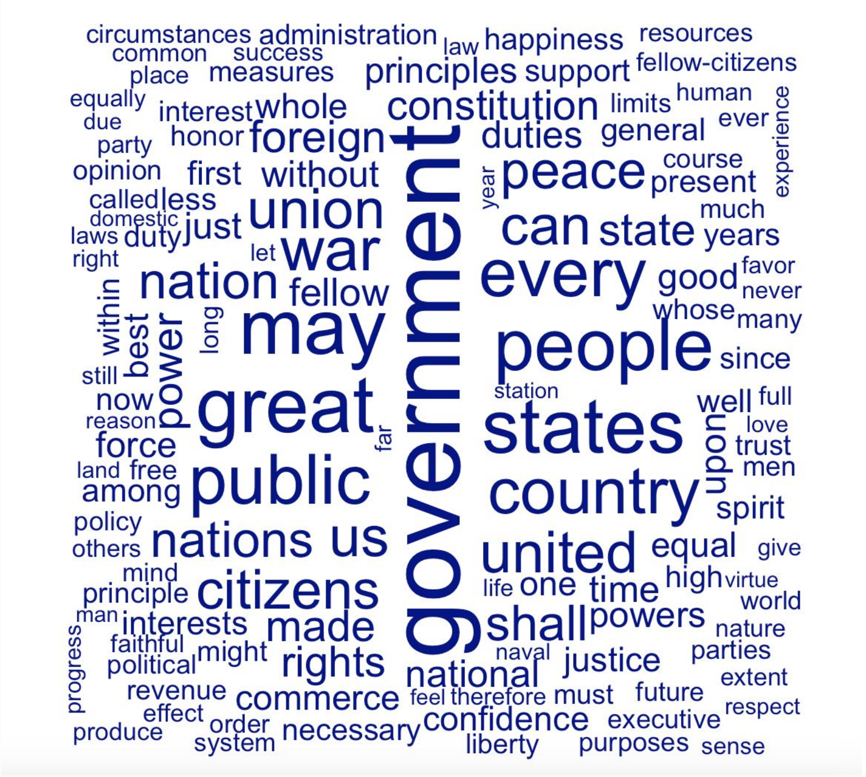
RQ: Look at speeches of a similar type by dictators and autocrats overtime as compared to democratic leaders.

Quanteda_text (1/12)

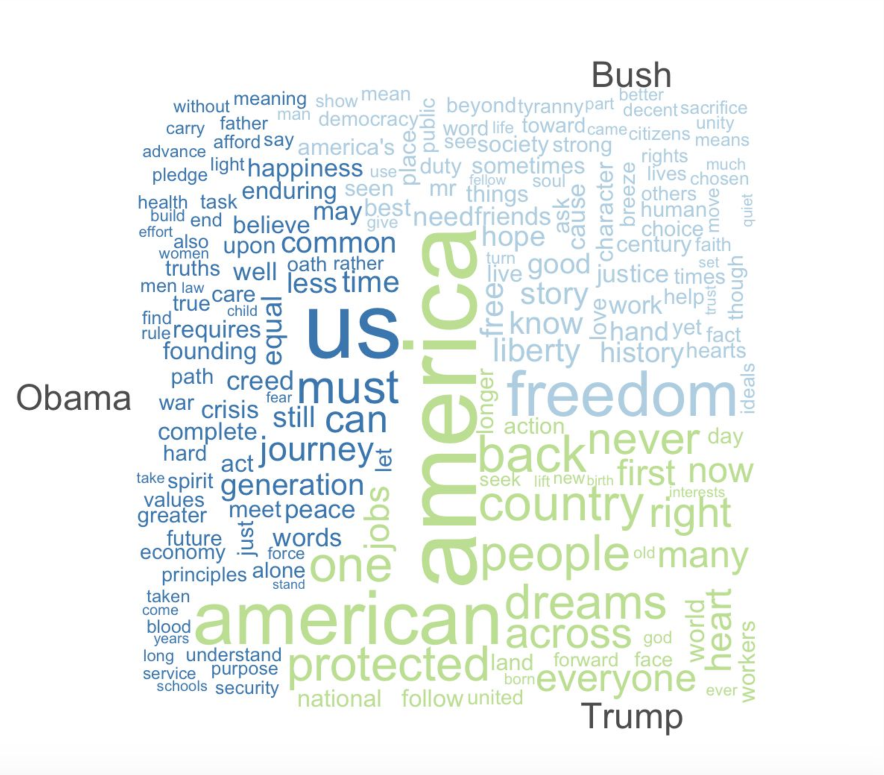




Quanteda_text (2/12)

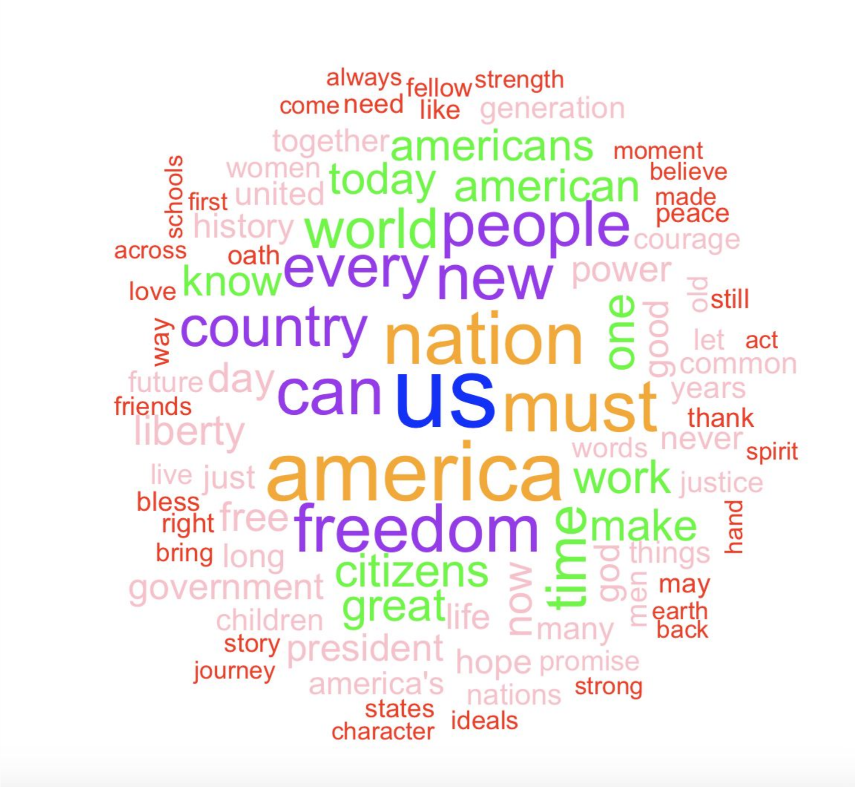


Base wordcloud

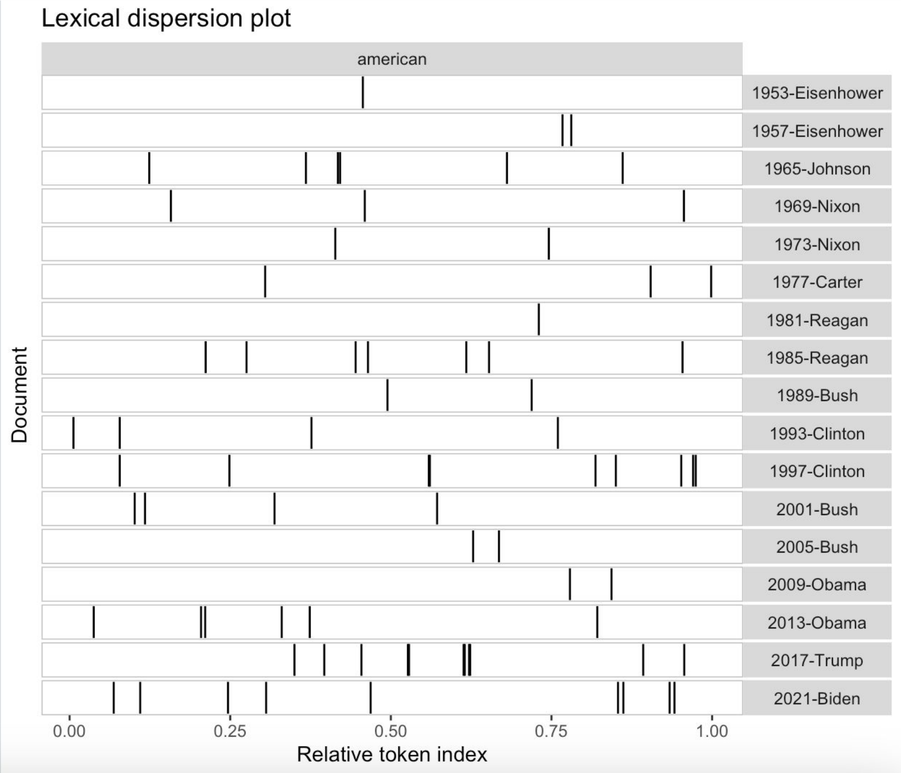


Wordcloud for Specific Presidents

Quanteda_text (3/12)

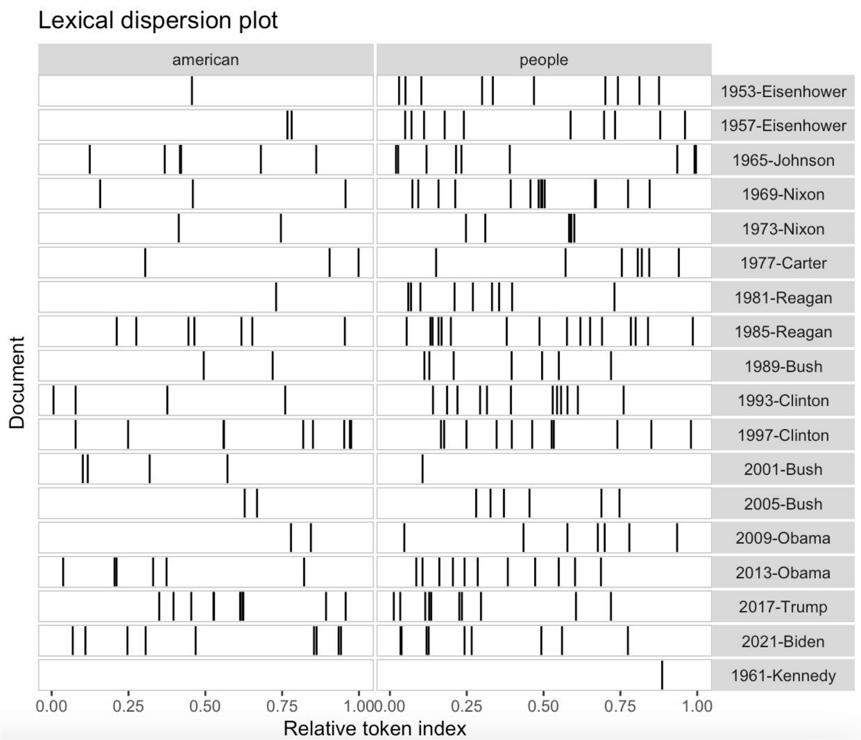


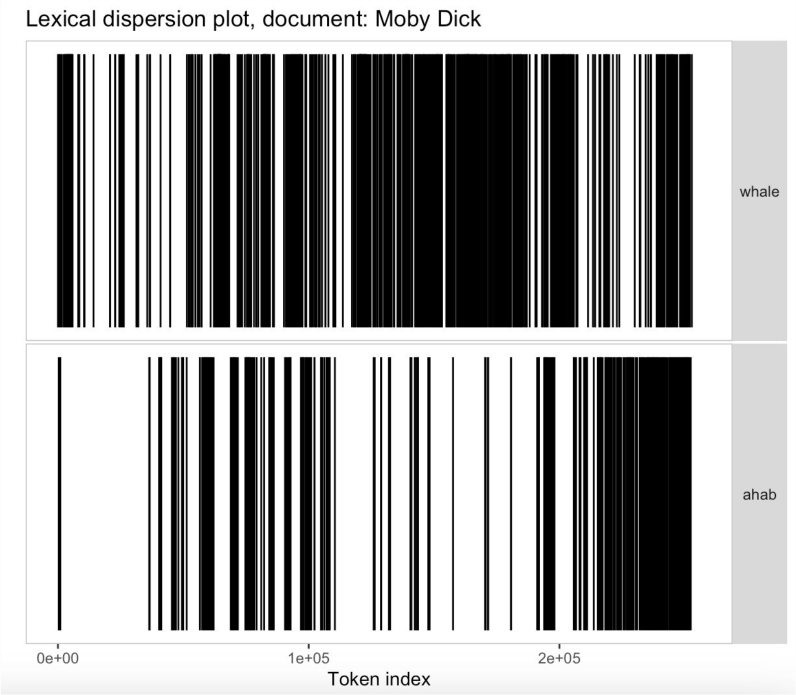
Colorized Word Cloud



Subset Corpus for Post-1949 Speeches

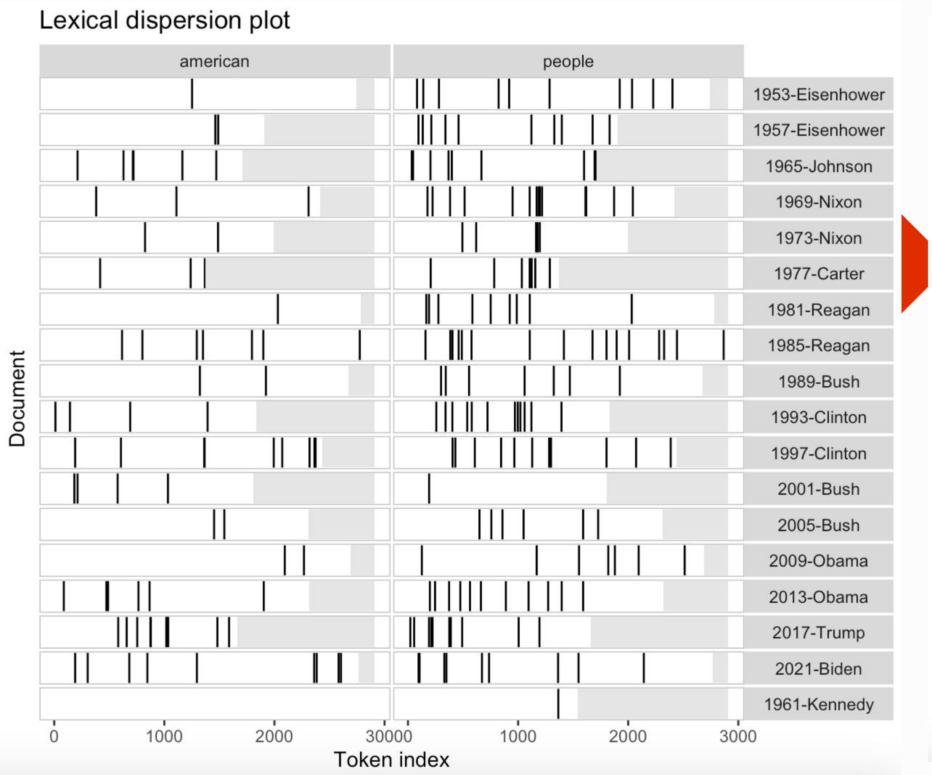
Quanteda_text (4/12)

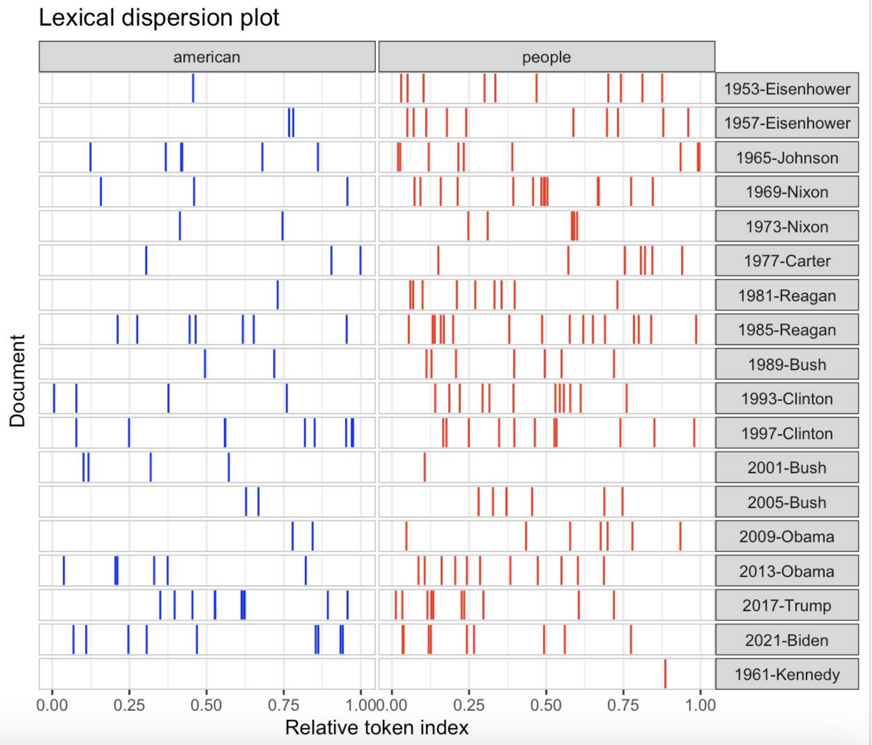




Multiple Keyword X-Ray Plot for “American, People”

Quanteda_text (5/12)

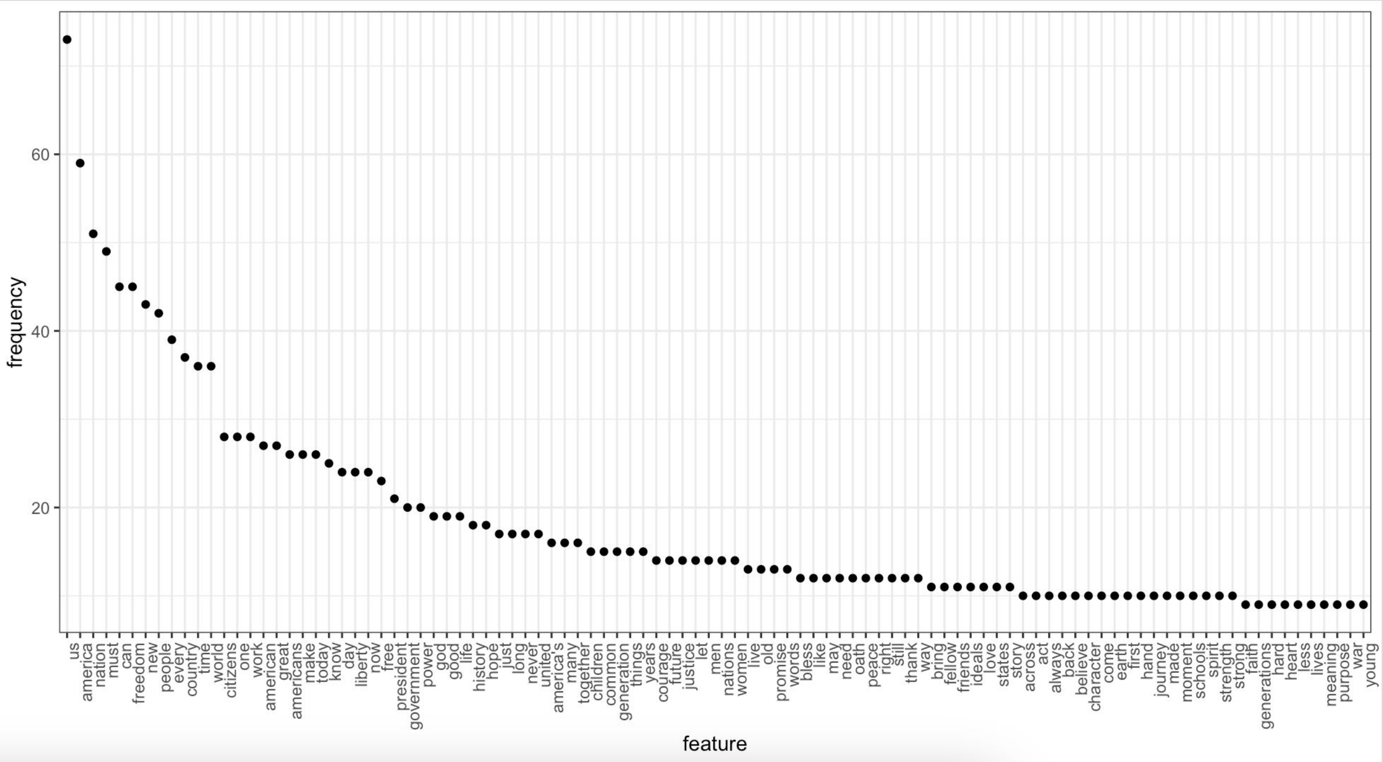




Lexical dispersion plot, w/t total words

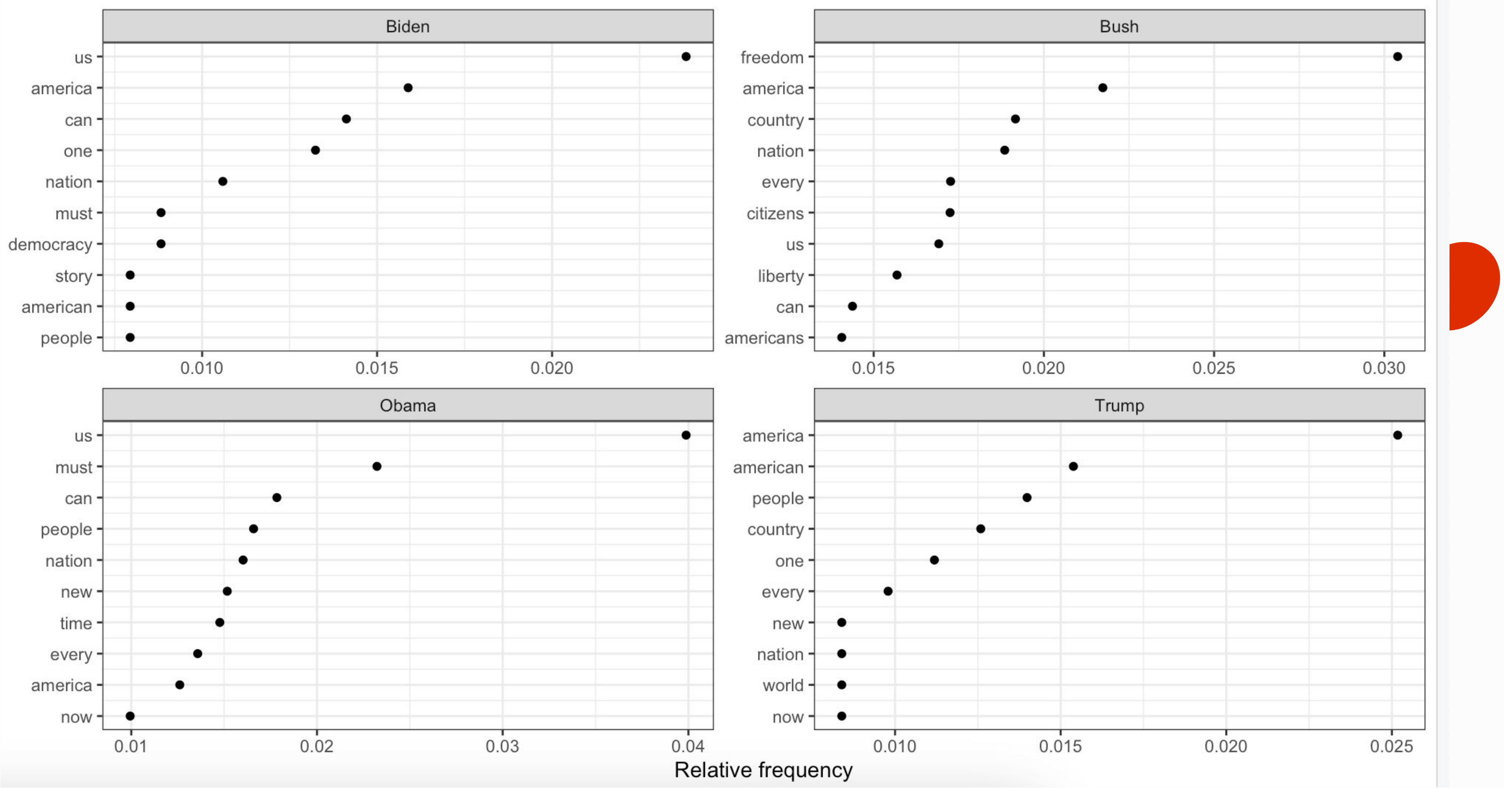
(now with party color)

Quanteda_text (6/12)



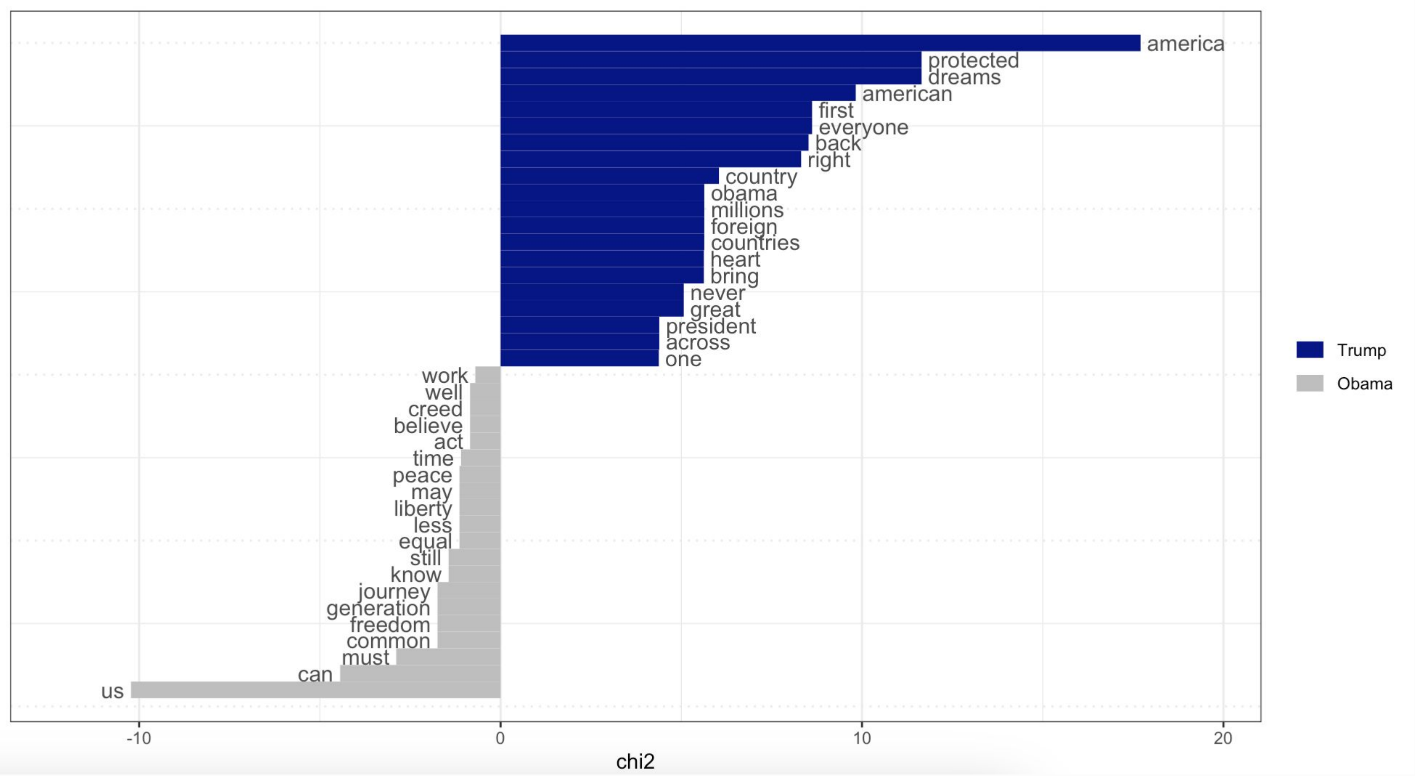
Sort by reverse frequency order

Quanteda_text (7/12)



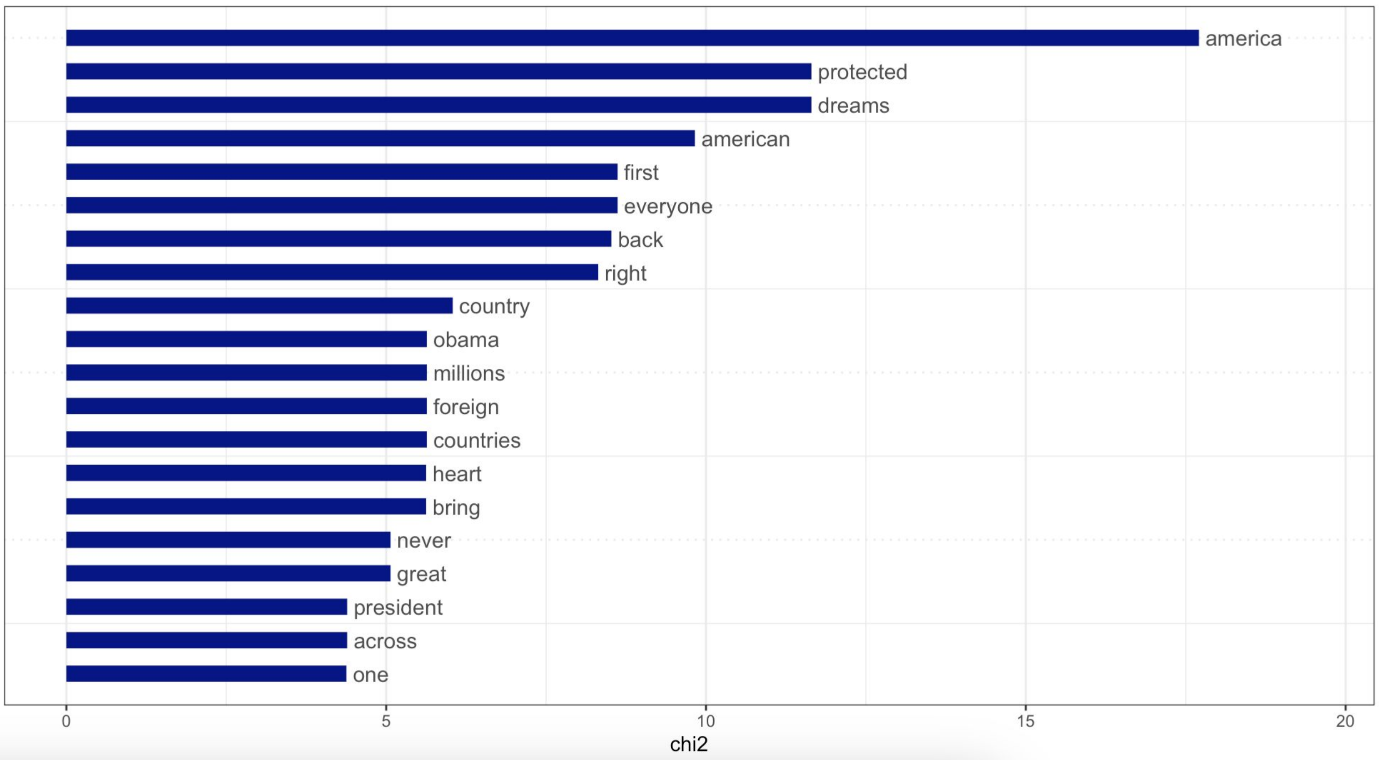
Calculate relative frequency by president

Quanteda_text (8/12)



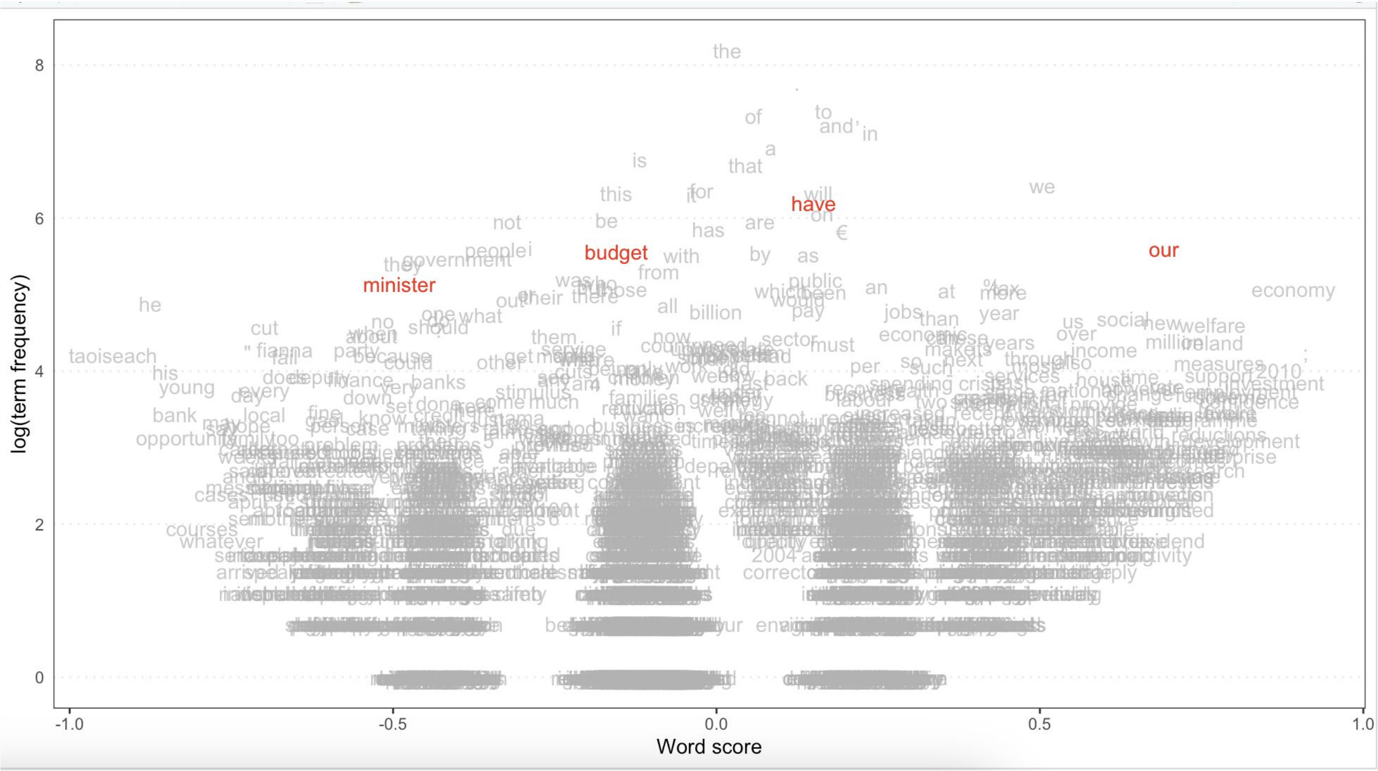
Plot estimated word keyness

Quanteda_text (9/12)



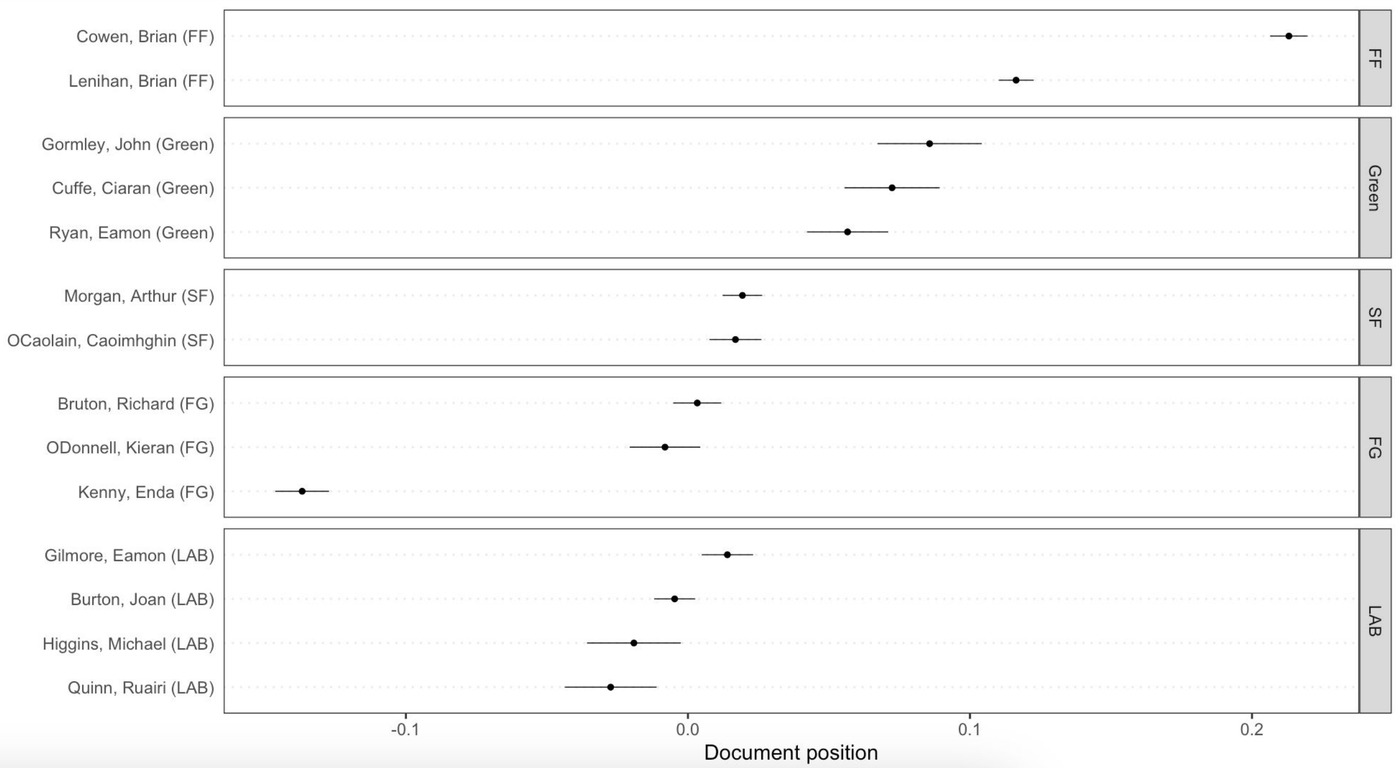
# Plot without the reference text (in this case Obama)

Quanteda_text (10/12)



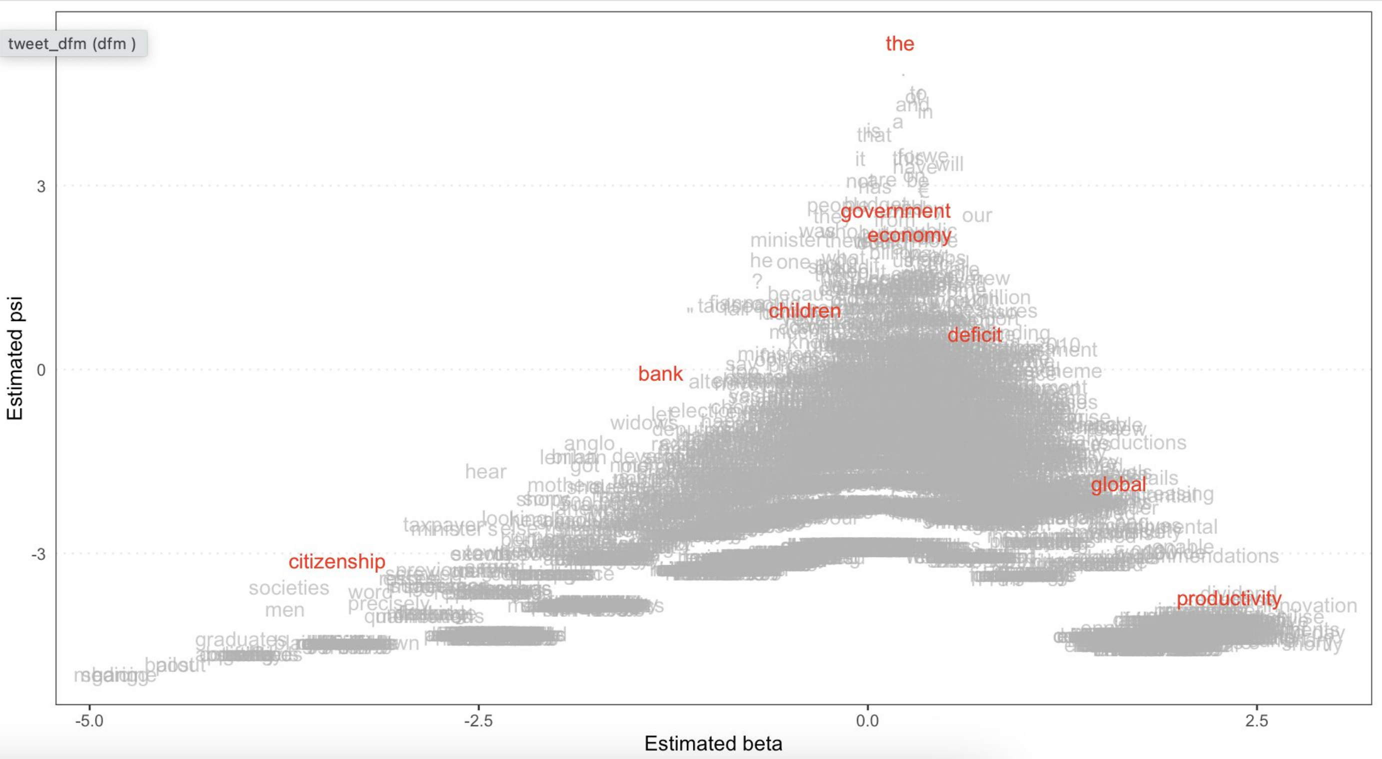
estimated word positions (highlight words and print them in red)

Quanteda_text (11/12)



Plot estimated document positions and group by "party" variable

Quanteda_text (12/12)



estimated word positions (Wirdfish model)