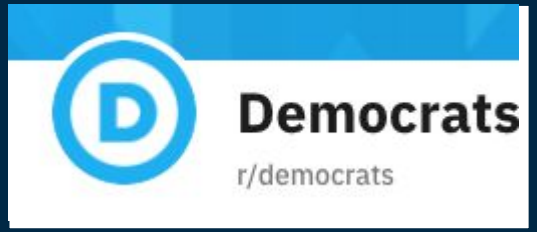
reddit



r/Politics vs r/Democrats

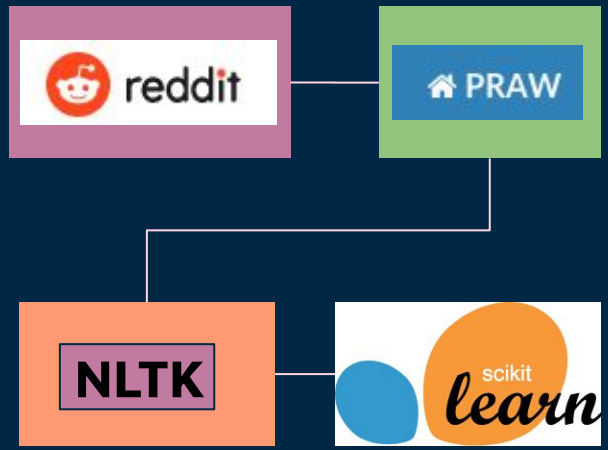
Natural Language Processing

Jared Delora-Ellefson
Data Scientist

How well can we build an NLP model to predict which subreddit a submission came from?

Subreddit Submission

Ex: "GOP operatives worry Trump will lose both the presidency and Senate Majority"



PRAW

API Wrapper for Reddit

NLTK

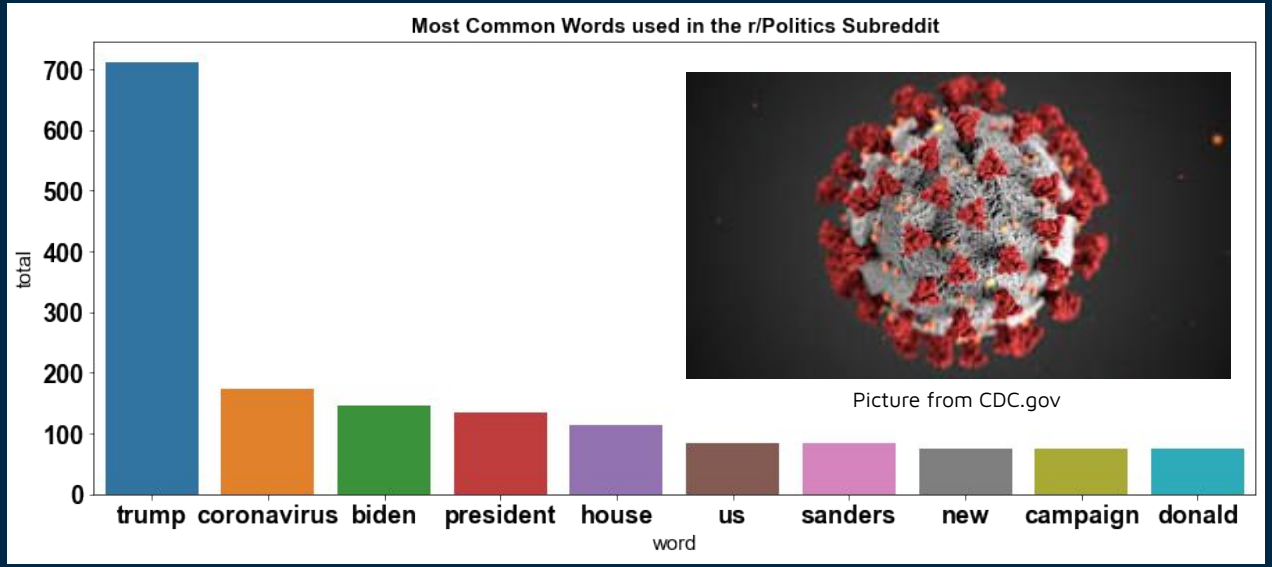
Tokenization, Lemmatization

scikit learn

Models, Predictions, Metrics

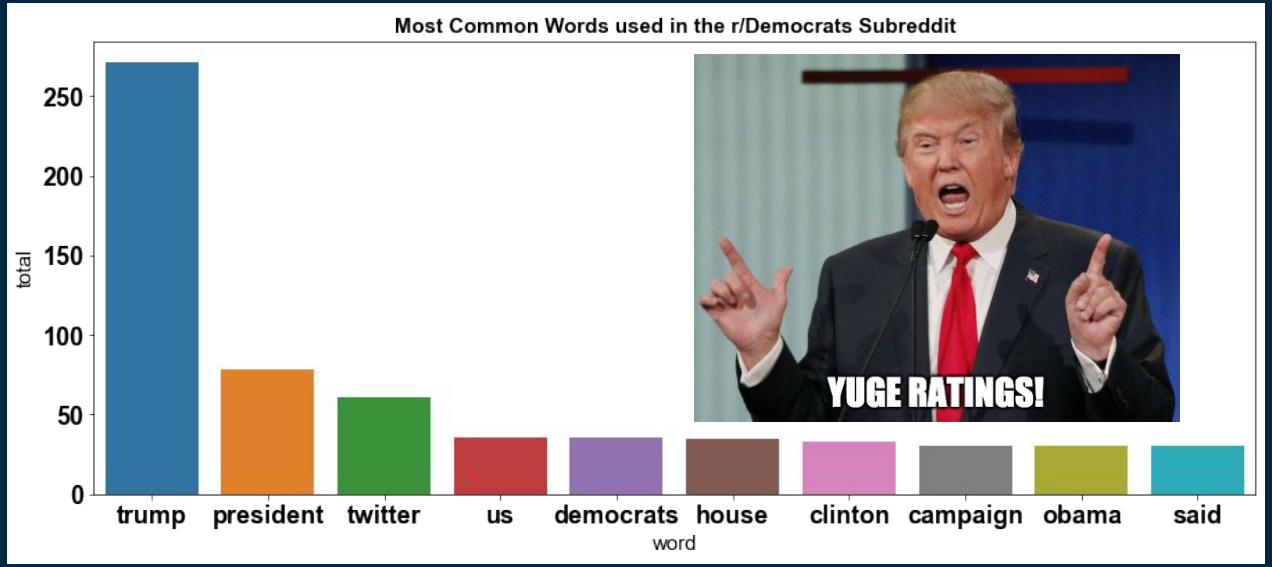
Logos taken from reddit.com, praw.readthedocs.io, scikit-learn.org

Distribution of Words Used on r/Politics



Submitters to r/Politics seem to be concerned with the Coronavirus and the coming US election

Distribution of Words Used on r/Democrats



Submitters to r/Democrats seem to be concerned with Trump's Twitter account, and the coming US election

Swear Word Frequency

Submitters to r/Democrats swear 6.5 times as much as submitters to r/Politics

Model Metrics

Count Vectorizer / Logistic Regression

	precision	recall	f1-score
r/Democrats	0.59	0.31	0.41
r/Politics	0.81	0.93	0.87
accuracy			0.78
macro avg	0.70	0.62	0.64
weighted avg	0.76	0.78	0.76

Count Vectorizer / Naive Bayes

	precision	recall	f1-score
r/Democrats	0.50	0.36	0.42
r/Politics	0.82	0.89	0.85
accuracy			0.76
macro avg	0.66	0.63	0.64
weighted avg	0.74	0.76	0.75

TfidfVectorizer / Logistic Regression

	precision	recall	f1-score
r/Democrats	0.62	0.15	0.25
r/Politics	0.79	0.97	0.87
accuracy			0.78
macro avg	0.70	0.56	0.56
weighted avg	0.75	0.78	0.72

TfidfVectorizer / Naive Bayes

	precision	recall	f1-score
r/Democrats	0.65	0.13	0.22
r/Politics	0.78	0.98	0.87
accuracy			0.78
macro avg	0.72	0.55	0.54
weighted avg	0.75	0.78	0.71

The F1 Scores for the TfidfVectorizer models are really poor for r/Democrats

My Experience Doing this Project

Q&A

jareddelora@gmail.com

https://www.linkedin.com/in/jareddeloraellefson/

THANKS