NLP Word Prediction

Joey Husney

3/01/21

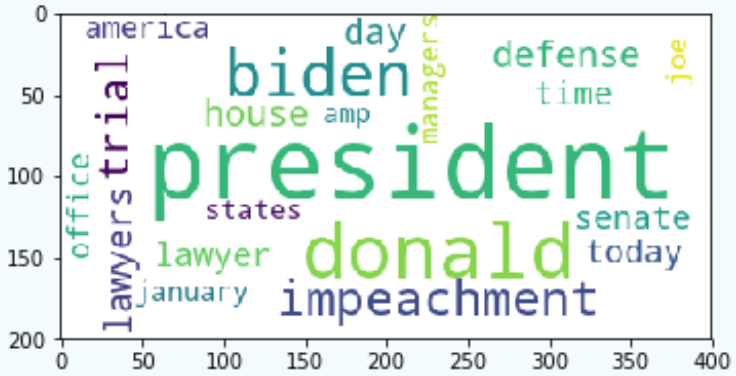
Overview

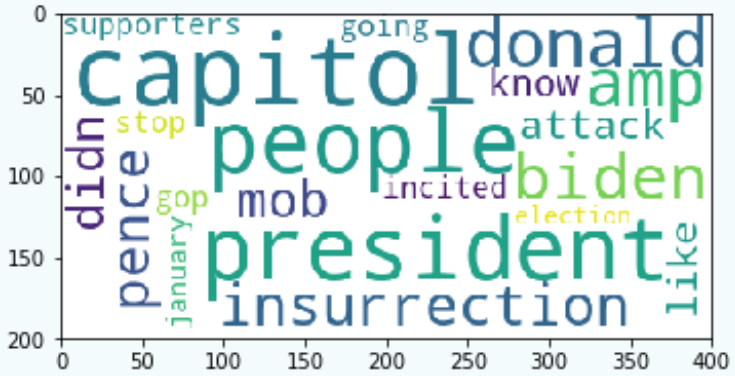
- Use tools given to us by NLTK, spacy, keras, and other libraries to create an application that can predict the next word in a tweet that the user wishes to type.
- The goal of this project isn't just to give one word to the user but a few different options ordered by the probability of each option being the desired next word.
- In order to achieve a high accuracy score, we will be deploying an unsupervised learning method prior to our supervised learning modeling.
- We will split the tweets into different categories using LDA. Only after that will we begin modeling but on each cluster individually.

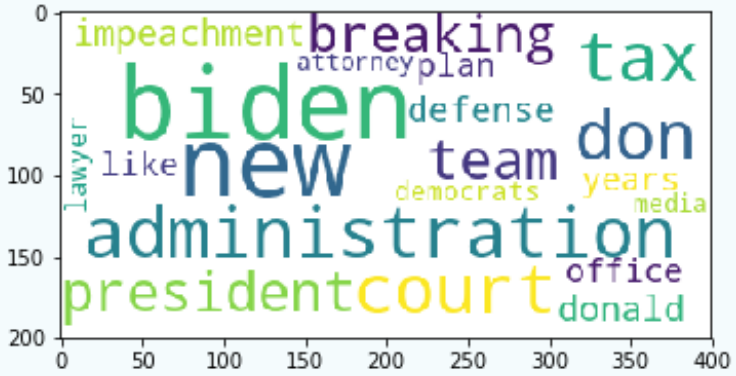
The Data

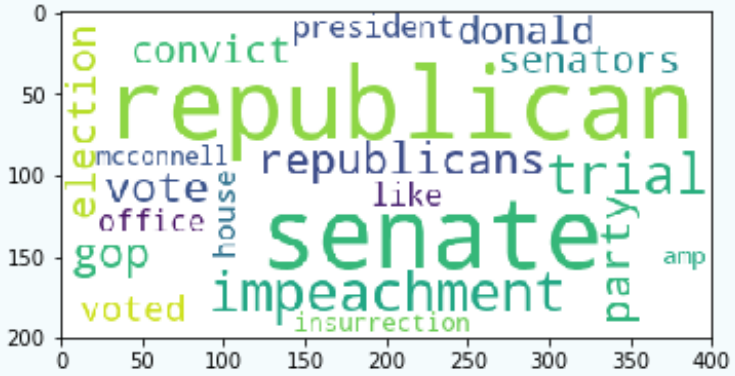
- The data we will be using consists of tweets from the twint API
- It contains not just tweets but loads of information about each tweet.
- Using tweet column to make prediction model.
- The Twint API uses web scraping to retrieve this data for us to access through its API.
- Scraping a list of the ten thousand most popular recent tweets (from scrape time) about former president Trump
- This data will be stored into a pandas data frame to be modified throughout this process.

EDA – Topic Modeling with LDA

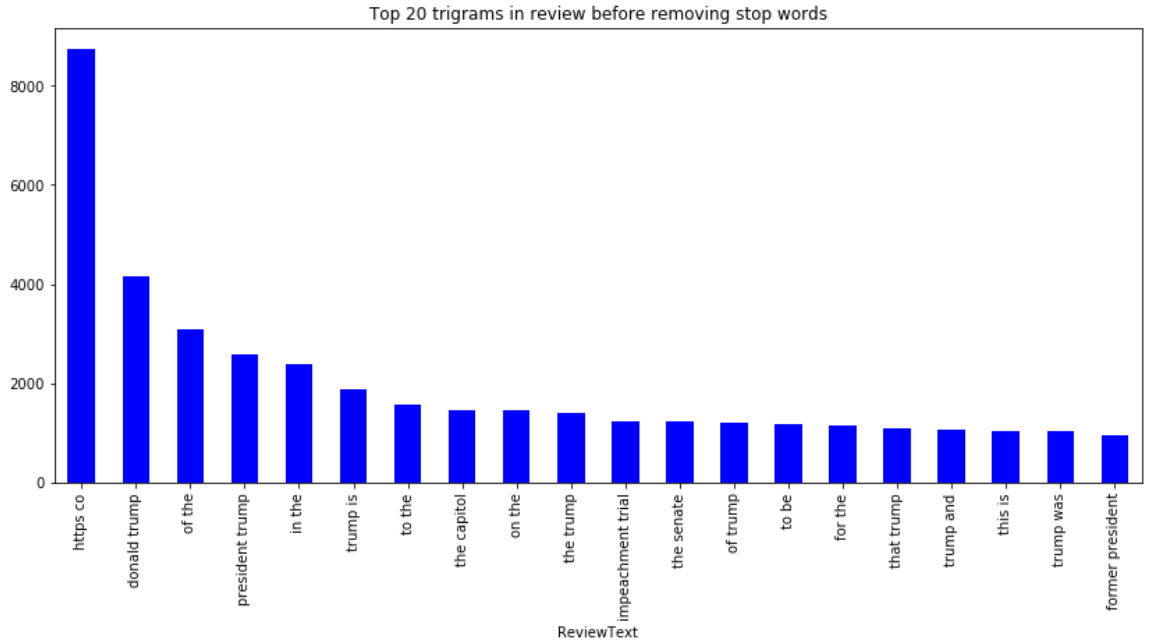








- Used LDA for 4 different topics:
- Capital Riot
- Trump impeachment
- Biden administration
- Trump in connection with republican party



EDA 2 – Common Bigrams Visualized

- No stopwords removed due to nature of next word prediction project
- No lemmatization for same reason
- Common phrases include ‘https co’, ‘Donald trump’, and ‘of the’

Final Model

```
In [107]: generate_text(model, tokenizer, seq_len, seed_text="donald", num_gen_words=1)
Out[107]: 'trump'
```

```
In [110]: generate_text(model, tokenizer, seq_len, seed_text="of", num_gen_words=1)
Out[110]: 'the'
```

- Hard to visualize neural network model
- Best way is to show through results
- Obtained accuracy score of roughly 14% after training on 500 epochs with 5 layers of kernels
- Used 100,000 tweets for model
- Sample predictions show on left hand side

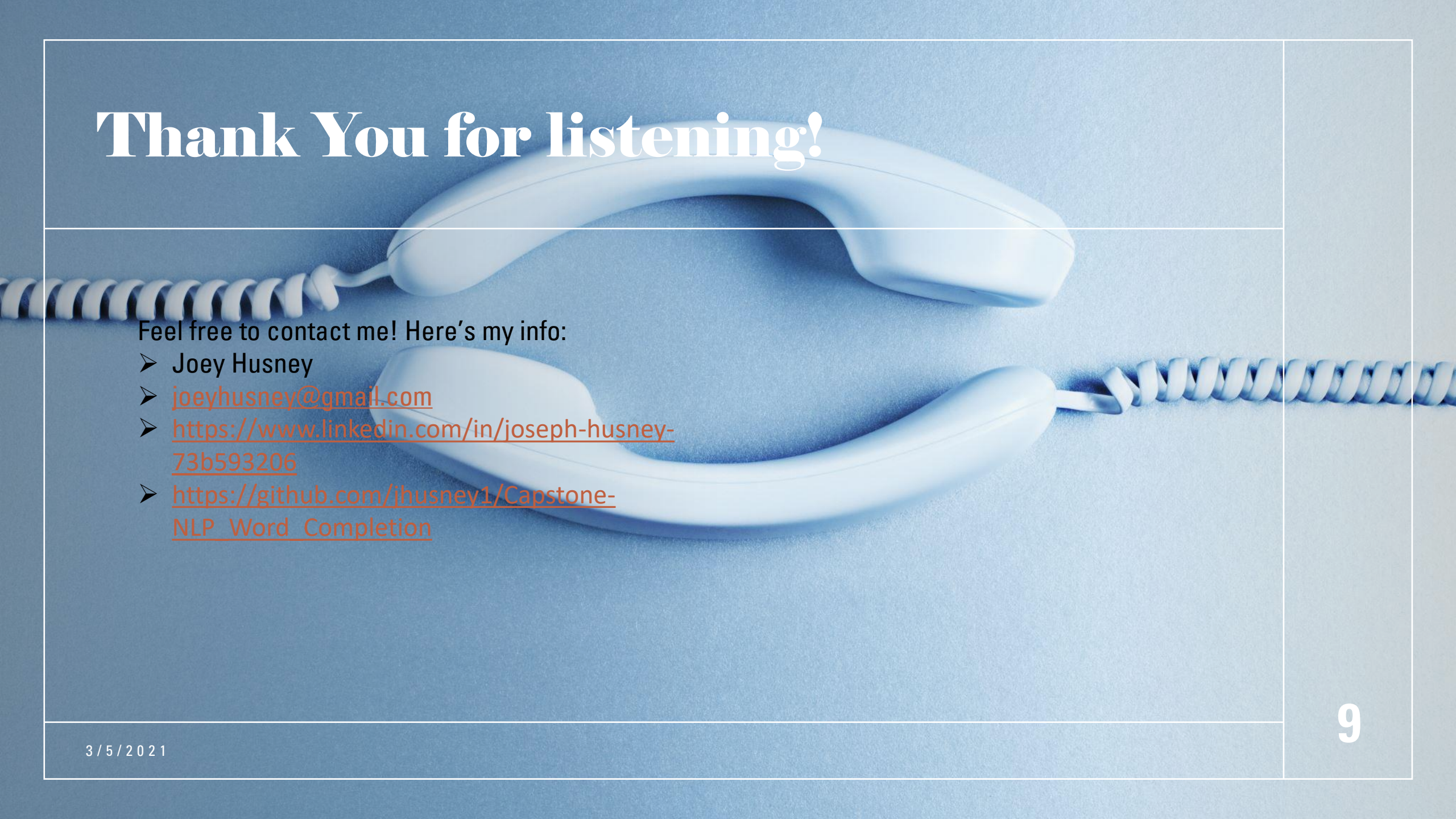
Conclusions and Recommendations

- Neural Networks are cool to work with because of high accuracy abilities
- Caveat is the difficulty to interpret results
- Very useful to visualize with LDA in conjunction with word cloud

Future Work

- Trying other models besides for neural network models
- Figuring out a way to incorporate LDA into project (besides for visualization purposes)
- In process of creating a front end for visualizing word predictions

Thank You for listening!



Feel free to contact me! Here's my info:

- Joey Husney
- joeyhusney@gmail.com
- https://www.linkedin.com/in/joseph-husney-73b593206
- https://github.com/jhusney1/Capstone-NLP_Word_Completion