NBA and NFL Prediction

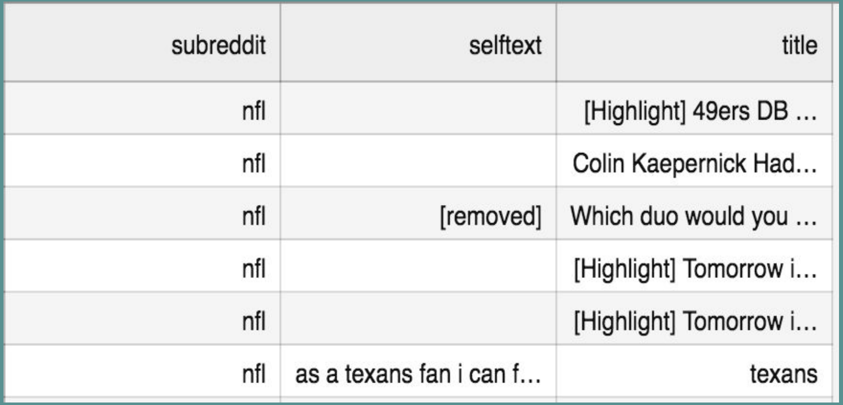
Prathik Chukkapalli

Purpose and Problem Statement

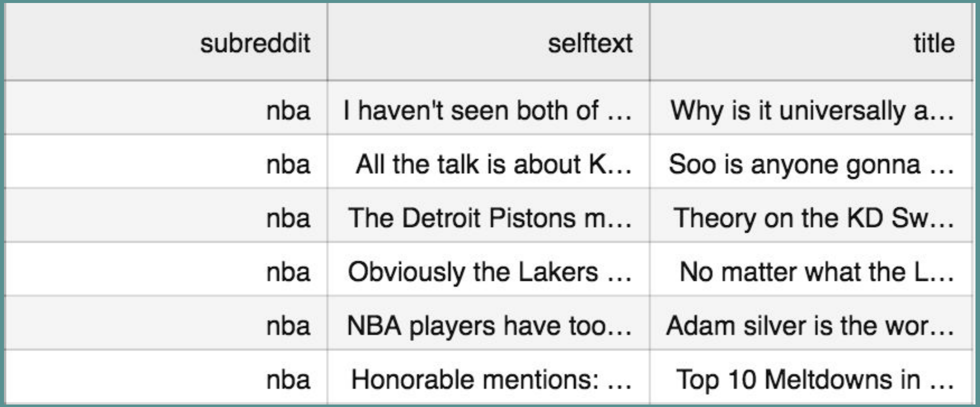
- I am a data scientist that wants to help workers that immigrated from other countries better communicate with their colleagues in America.
- Many workers talk about sports, the two most popular in America being the NBA and NFL.
- I want to create a model that can predict what sport is being talked about based on a text or message from a person.
- This will help immigrant workers from not mistaking what sport is being talked about and improve their sense of understanding about what the conversation they are in is about.

Gathering Data

- Web Scraping
  - pushshift
- Picking Columns
  - subreddit,selftext,title
- Clean-up,Concatenation
  - subreddit,title



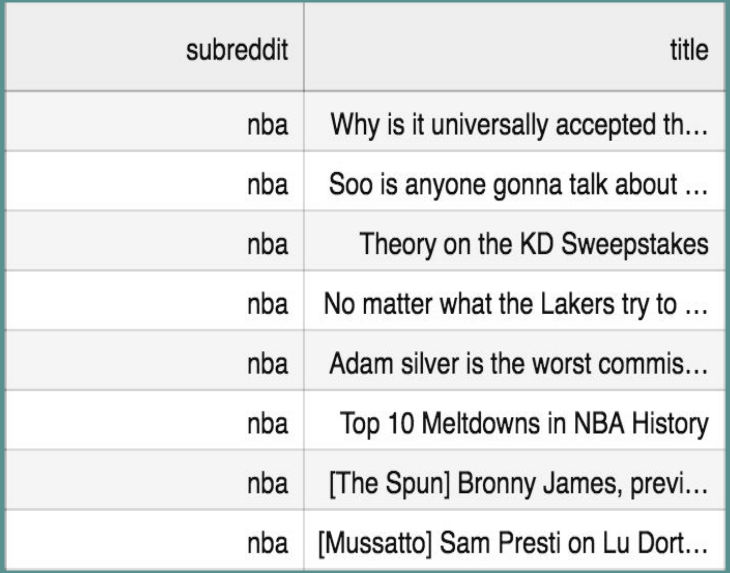
subreddit	selftext	title
nfl		[Highlight] 49ers DB ...
nfl		Colin Kaepernick Had...
nfl	[removed]	Which duo would you ...
nfl		[Highlight] Tomorrow i...
nfl		[Highlight] Tomorrow i...
nfl	as a texans fan i can f...	texans



subreddit	selftext	title
nba	I haven't seen both of ...	Why is it universally a...
nba	All the talk is about K...	Soo is anyone gonna ...
nba	The Detroit Pistons m...	Theory on the KD Sw...
nba	Obviously the Lakers ...	No matter what the L...
nba	NBA players have too...	Adam silver is the wor...
nba	Honorable mentions: ...	Top 10 Meltdowns in ...

Data Dictionary

Feature	Type	Dataset	Description
subreddit	string	nba_nfl_modelling_data	The values are either nba or nfl, the two subreddits the data was scraped from.
title	string	nba_nfl_modelling_data	The title of posts that came from the nba and nfl subreddits.

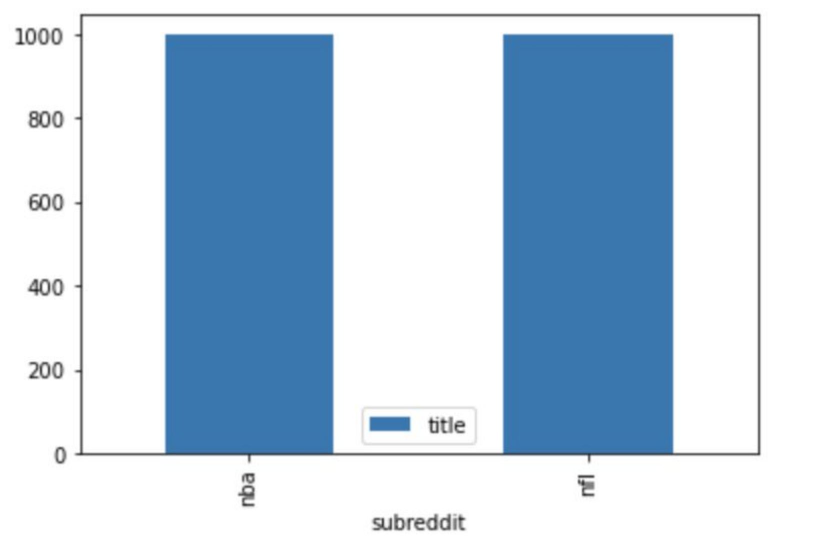


subreddit	title
nba	Why is it universally accepted th...
nba	Soo is anyone gonna talk about ...
nba	Theory on the KD Sweepstakes
nba	No matter what the Lakers try to ...
nba	Adam silver is the worst commis...
nba	Top 10 Meltdowns in NBA History
nba	[The Spun] Bronny James, previ...
nba	[Mussatto] Sam Presti on Lu Dort...

Analysis

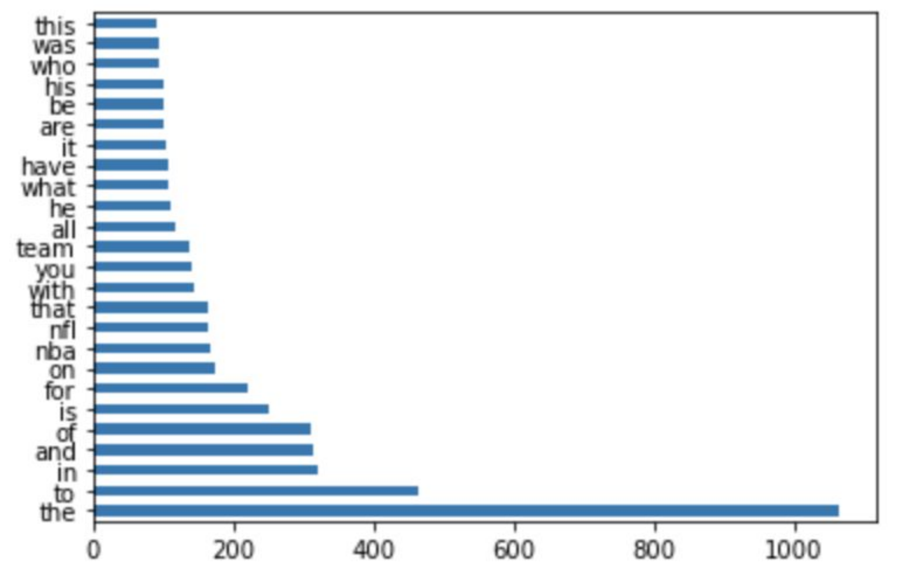
- Bar Graph
    - Total amount of each subreddit
- Average Word Count

Subreddit	Average Word Count
NBA	14.6 words
NFL	17.1 words



CVEC Analysis

- Top 25 words in all titles



Models using CVEC

Scores	MultiNomialNB	KNN	LogisticRegression
Best Params	Cvec__max_df: 0.5, Cvec__max_features : 250, Cvec__min_df : 3 Cvec__ngram_range: (1, 1)	Cvec__max_df: 0.5, Cvec__max_features : 370 Cvec__min_df : 8 Cvec__ngram_range: (1, 1) Cvec__stop_words : ‘english’ Knn_n_neigbors : 8 Knn_weights : ‘distance’	Cvec__max_df: 0.2, Cvec__max_features : 1200 Cvec__min_df : 2 Cvec__ngram_range: (1, 1) Cvec__stop_words : ‘english’ Log__max_iter : 100 Log__fit_intercept : False Log__multi_class : ‘auto’
Best Score	0.83	0.75	0.89
Training Score	0.86	0.96	0.98
Testing Score	0.80	0.76	0.93
Cross Val Score	0.82	0.77	0.87
Specificity	0.87	0.89	0.94

Conclusion

- The classifier that did the best with Countvectorizer was Logistic Regression.
- This model can be used to predict if messages are talking about the NBA or NFL.
- In the future it would be good to delve deeper into the parameters of classifiers to improve the predicting ability of the model as well as create a model that not only predicts if a message is about the NBA or NFL, but can help craft a message that could be a response to the one received.