Text Analysis of the Politics Subreddit

Cole Sanders

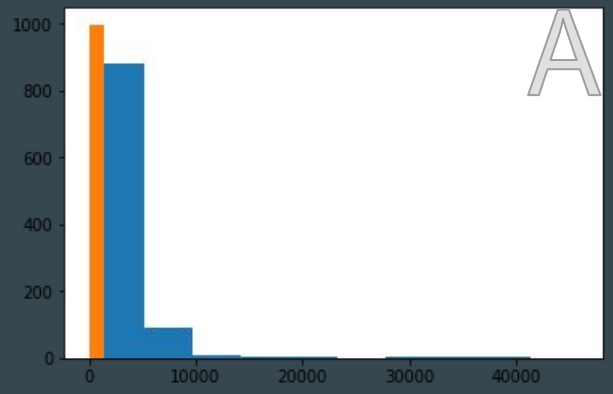
Proposal

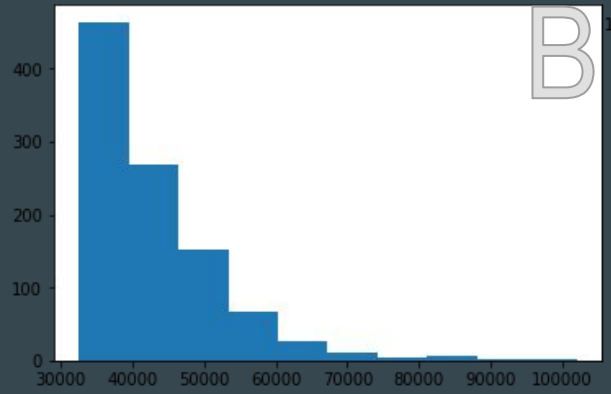
1. Pull data form the Politics Subreddit from both the Top and Controversial categories
2. Tokenize Text in Title and URL to look for common phrases
3. Build a model to show what combination of phrases, words, and websites are associated with higher scores and are more likely to appear in Top or Controversial
4. Build a Shiny Dashboard to help visualize the data and allow people to predict outcomes of using certain word and urls in a title

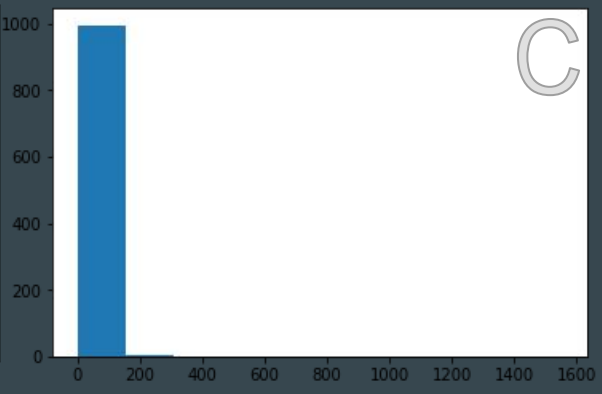
Controversial tends to have lower scores and comments than Top

- Graph A: Number of Comments, Blue is Top, Orange is Controversial
- Graph B: Score for Top
- Graph C: Score for Controversial

By mixing Top and Controversial we get a good mix of high and low scoring post, however this causes score and category to be strongly related. This is why these features will be part of our outcome variables only, and will not be used in Prediction.







Most Frequent Domains

Top:

1. ('thehill.com', 106),
2. ('newsweek.com', 69),
3. ('washingtonpost.com', 65),
4. ('self.politics', 41),
5. ('businessinsider.com', 37),
6. ('cnbc.com', 35),
7. ('cnn.com', 34),
8. ('independent.co.uk', 30),
9. ('thinkprogress.org', 28),
10. ('commondreams.org', 27),
11. ('nytimes.com', 26),
12. ('huffingtonpost.com', 24),
13. ('lawandcrime.com', 23),
14. ('vox.com', 22),
15. ('theweek.com', 18),

Controversial:

1. ('thehill.com', 98),
2. ('washingtonpost.com', 45),
3. ('politico.com', 32),
4. ('thedailybeast.com', 28),
5. ('commondreams.org', 26),
6. ('cnn.com', 26),
7. ('theintercept.com', 26),
8. ('washingtontimes.com', 25),
9. ('newsweek.com', 23),
10. ('nytimes.com', 22),
11. ('rightwingwatch.org', 20),
12. ('theguardian.com', 20),
13. ('nypost.com', 18),
14. ('washingtonexaminer.com', 16),
15. ('jacobinmag.com', 16),