Web scraping and NLP

By

Douglas Hundley

Problem Statement

- Need a story pitch for Fivethirtyeight
- Need to web scrape a certain amount of data in a limited time
- Need to make classification models with NLP

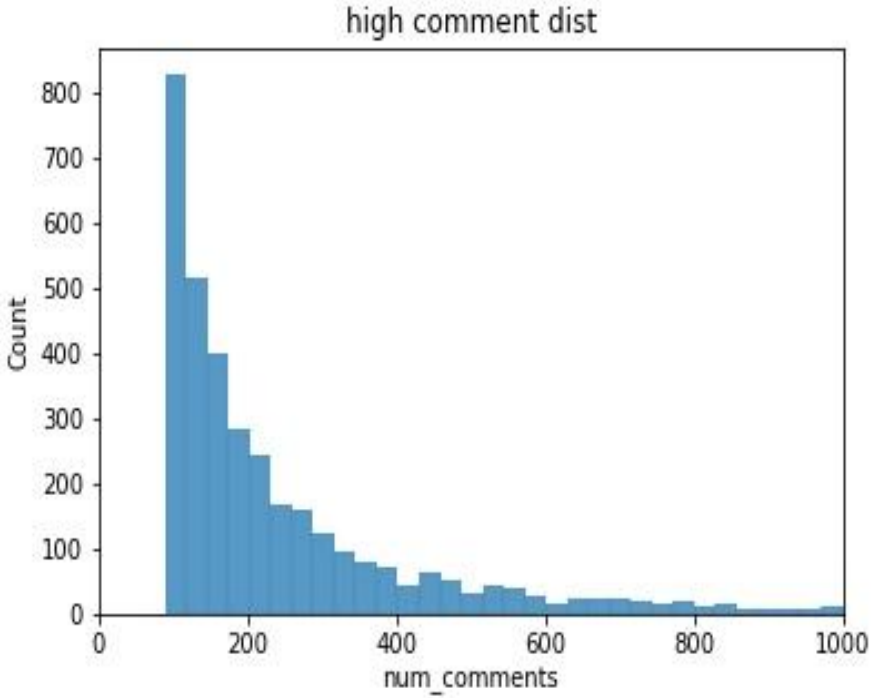
Discoveries while Collecting the Data

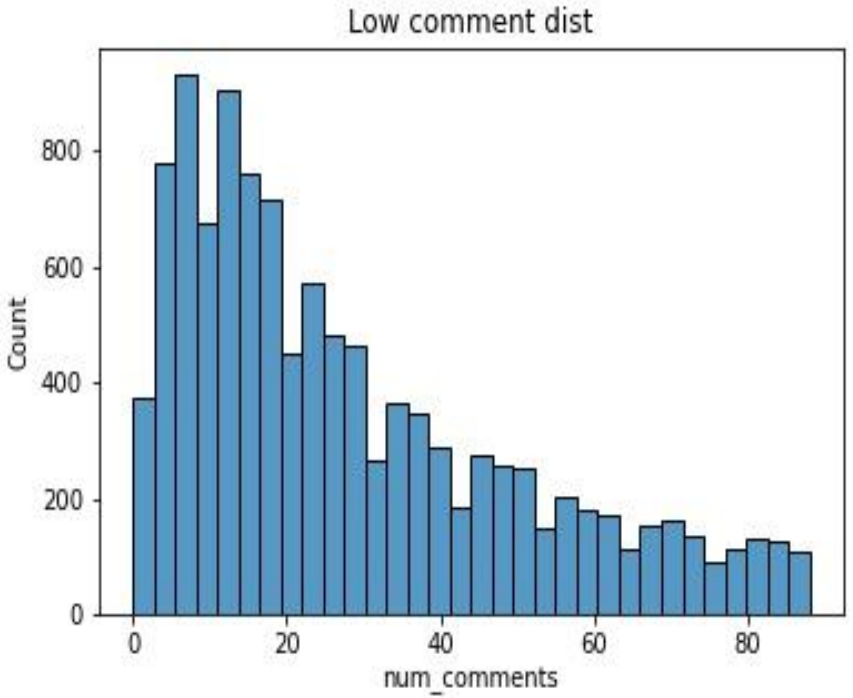
- New “hot” posts are cycled out much slower than expected
- Things could have moved along faster using an API
- Reddit is very lenient when you are web scraping

Data Dictionary

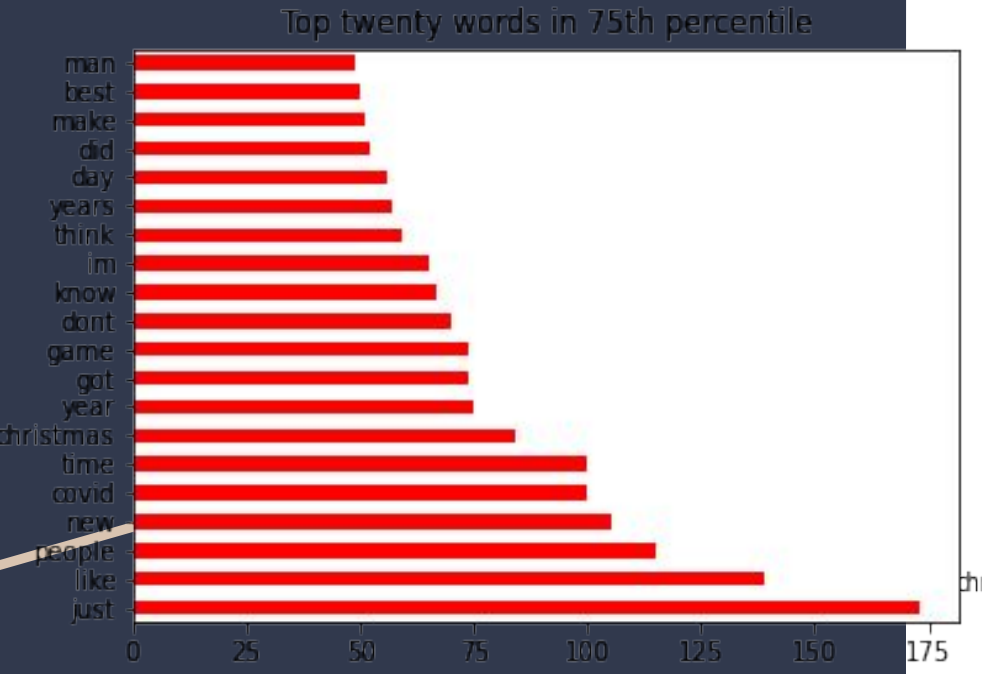
Feature	Type	Dataset	Description
subreddit	*obj*	backup_data.csv	The title of the subreddit
title	*obj*	backup_data.csv	Title of post
num_comments	*int*	backup_data.csv	Total number of comments on post
upvotes	*int*	backup_data.csv	Total number of upvotes on post
created_at	*int*	backup_data.csv	Epoch time of post creation
high_low_count	*int*	backup_data.csv	Binary column of upper and lower 75th percentile
title_length	*int*	backup_data.csv	Total characters of post
title_word_count	*int*	backup_data.csv	total count of words in title column

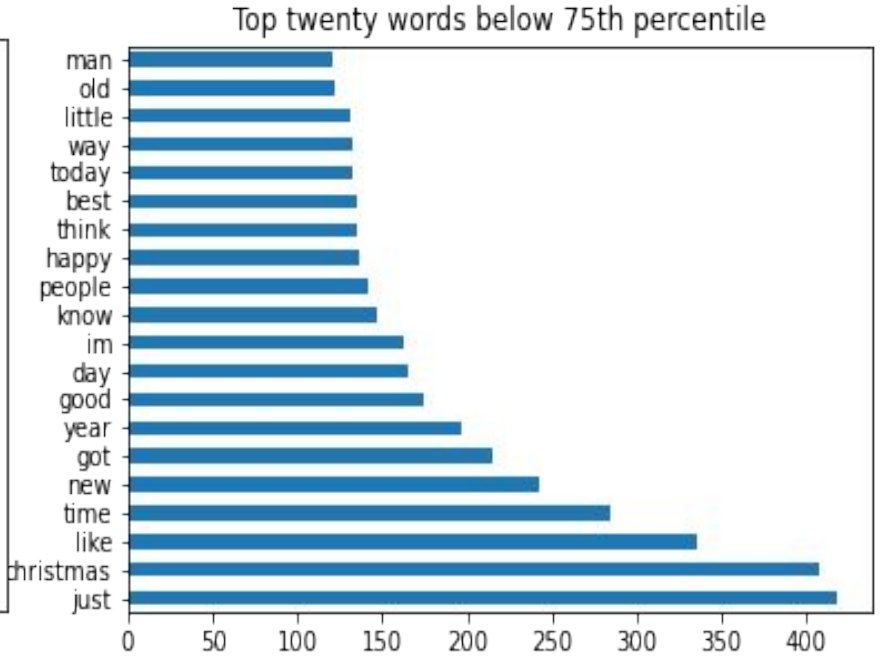
Distribution of comments





Top twenty words





The Models

- Logistic Regression score on both attempts stayed at 75 percent accuracy
- Random Forest score on both attempts stayed at 100 percent accuracy

Conclusion

- r/Askreddit seems to be the best place for exposure
- Logistic Regression was the better model
- Need to try less features when modeling
- Could explore other NLP methods