Predicting Subreddit from Post Title

Chetan Birthare, Zion Steiner

Can Naive Bayes Predict Subreddit from Post Title?

Reddit - common interest communities divided into “subreddits”

Five subreddits:

- AskReddit
- WritingPrompts
- TodayILearned
- Worldnews
- UnethicalLifeProTips

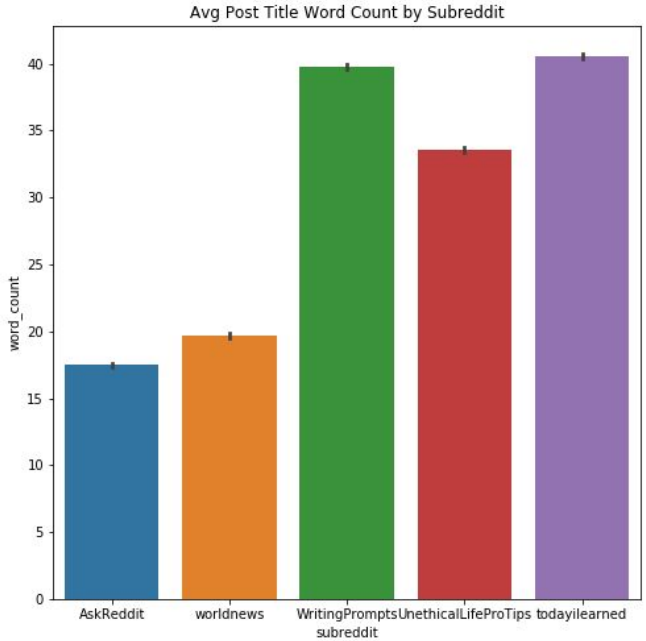
Collecting Subreddit Post Titles

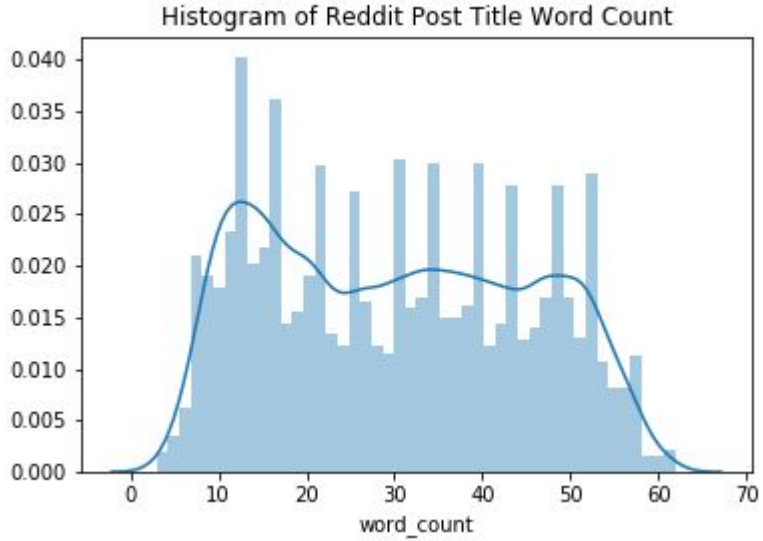
Python API Wrapper PRAW

The 500 most popular of all time post titles were collected from each subreddit

	subreddit	title	word_count
346	AskReddit	bar staff of reddit, have you ever had a man u...	21
125	AskReddit	what was the best moment you've seen where the...	16
2110	UnethicalLifeProTips	ulpt: you can copy off wikipedia if you transl...	23
2186	UnethicalLifeProTips	ulpt: going to a baseball game? keep an old ba...	56
1658	worldnews	rushed amazon warehouse staff reportedly pee i...	29
446	AskReddit	what quote has always stuck with you?	7
312	AskReddit	askreddit has hit 25,000,000 subscribers! (ins...	9
31	AskReddit	what looks easy peasy lemon squeezy but is act...	13
2439	UnethicalLifeProTips	ulpt: if you contact adobe customer care and a...	54
1149	todayilearned	til that sir christopher lee(who played saruma...	47

Data Exploration

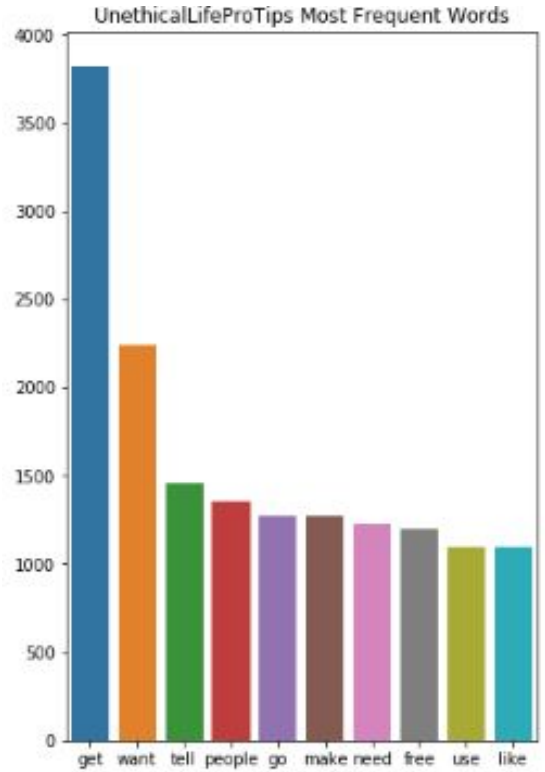




Data Exploration

Top 5 Words for ULPT:

- Get
- Want
- Tell
- People
- Go



Text Feature Engineering

- Word count vectors

“This sentence is a sentence” = [this: 1, sentence: 2, is: 1, a: 1]

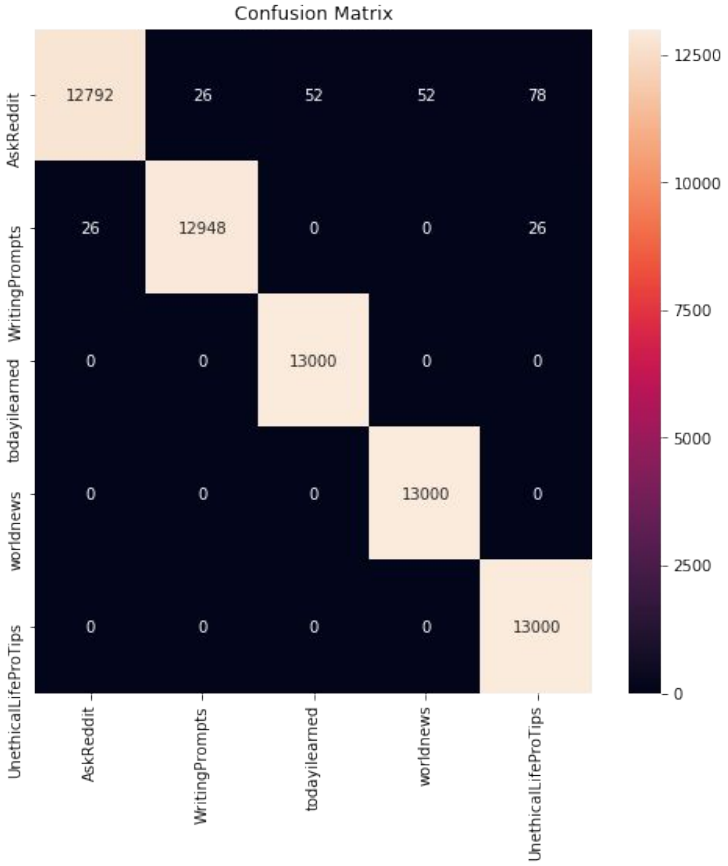
- Term Frequency-Inverse Document Frequency
  - High importance for words that occur frequently in a document
  - Penalization term for how frequent words occur throughout entire corpus

- Remove stopwords
  - “A”, “the”, “you”, etc.

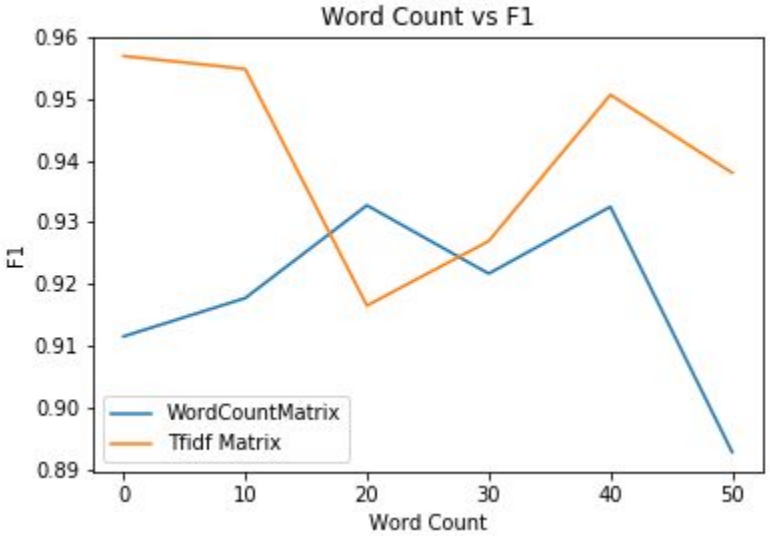
Naive Bayes Performance

Method	**F1**	**Precision**	**Recall**
Count, stopwords included	0.929	0.934	0.930
Count, stopwords removed	0.823	0.831	0.828
Tfidf, stopwords included	0.887	0.894	0.888
Tfidf, stopwords removed	0.819	0.827	0.821

Confusion Matrix for Count Matrix on Cleaned Titles



How Do These Methods Work on Different Title Lengths?



13k titles from each subreddit

Conclusions

- Word Count Matrix works better than Tfidf for this dataset
- Best model f1: 82.5%