Web APIs & NLP

Sara Zhou

Problem Statement :

Given a post from either r/OutOfTheLoop and r/explainlikeimfive, how well can we predict which subreddit it came from?

Background on the Subreddits

"A subreddit to help you keep up to date with what's going on with reddit and other stuff."

"A subreddit to help you keep up to date with what's going on with reddit and other stuff."

How well could the model perform?

The Process

Obtaining Posts → Analysis & Preprocessing → Modeling

Grabbed 100,000 Post

After cleaning out all the null and deleted posts I ended up with around 25,000 - 30,000 per subreddit.

Tokenizing

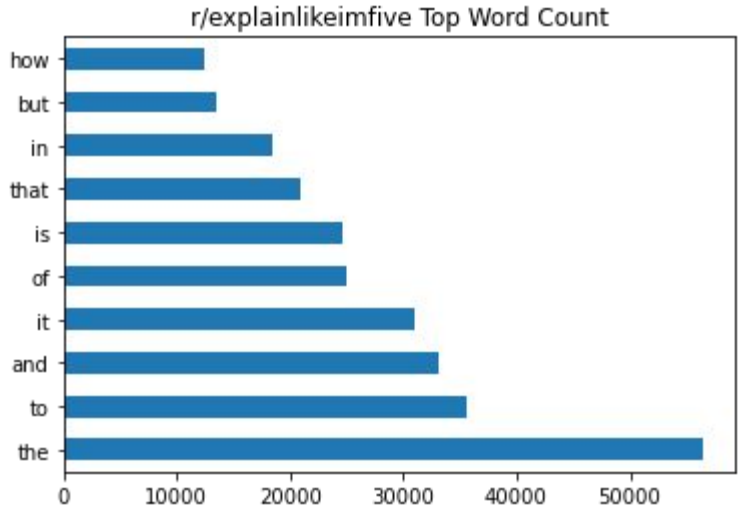
I tokenized all the self text and removed common words and decided to not look at titles because in ELIF users must have ELIF in the title name.

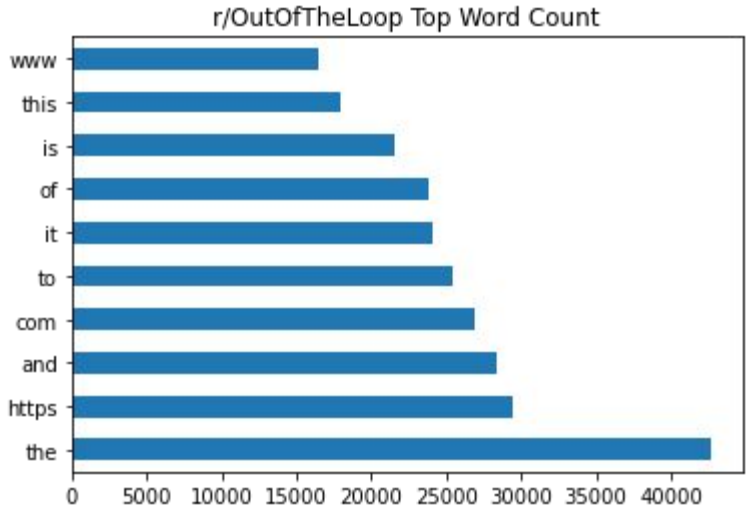
I used 3 different models

1. GS Over CountVectorizer BernoulliNB
2. GS Over TF-IDF Vectorizer and Multinomial NB
3. Random Forests

Model Performance

GS Over CountVectorizer BernoulliNB	GS Over TF-IDF Vectorizer and Multinomial Naive Bayes	Random Forest
Train: 94.16%, Test: 93.81%	Train: 94.53%, Test: 94.26%	Train: 99.93%, Test: 93.81%


r/explainlikeimfive Top Word Count
how
but
in
that
is
of
it
and
to
the
0
10000
20000
30000
40000
50000


r/OutOfTheLoop Top Word Count
www
this
is
of
it
to
com
and
https
the
0
5000
10000
15000
20000
25000
30000
35000
40000

Conclusion:

My training and testing score stayed in the 90s. When choosing these subreddits I expected a lot of overlap because they both seemed very similar but after running these models it's clear that they had no problem differentiating the two. The list of top words for either were different enough for the most part, it seems that the users of these subreddits were both looking for answers for their questions but the types of questions asked differed enough to produce a fairly accurate model.