Subreddit Classification

Exploratory Machine Learning Models and Natural Language Processing

DISCLAIMER

This presentation includes potentially offensive or lewd language. Such is language as it is naturally spoken, and, this being the internet, such is language as typed by online savages.

The Goal

To build a series of classification models for predicting, given a reddit post's title, which subreddit the post belongs in:

r/askscience

r/shittyaskscience

r/askscience

“AskScience is a forum for answering science questions. It aims to promote scientific literacy by helping people understand the scientific process and what it can achieve.”

Why is nuclear fusion so much harder to achieve than nuclear fission?

r/shittyaskscience

Do you have a question that mainstream science refuses to answer? Are your theories and experiments so bizarre that sensible people tell you they will never be published? Do you need help building a doomsday machine, or shopping for a monstrous assistant? Then you've come to the right place! The esteemed panelists at Shitty Ask Science might be the only people who are willing to help you!

Which animal do guinea pigs use for their experiments?

Data Collection

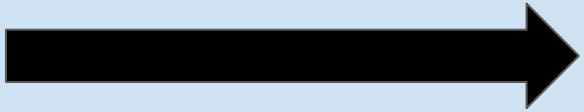
Using the Pushshift Reddit API

Time Anchor

(Last Saturday, Around Noon)

40,000 Posts

2,598 Posts



Data Cleaning

42,598 → 38,170

Flagged by Moderator

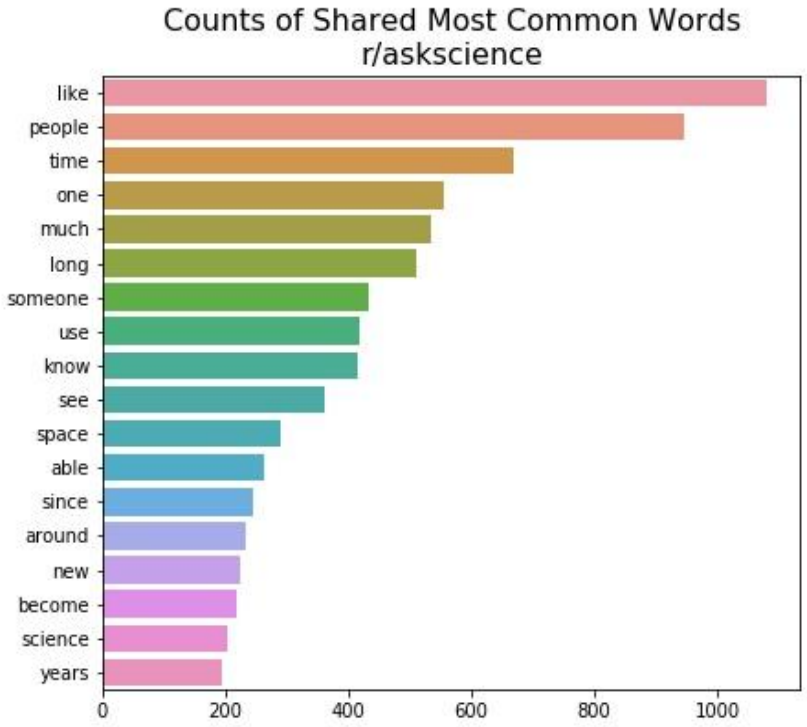
User Bans

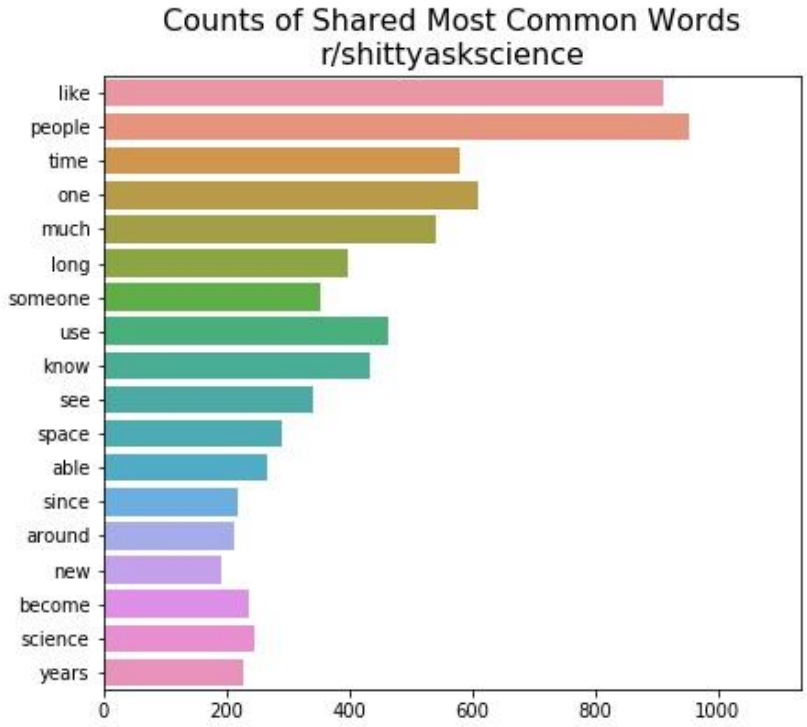
Deleted by User

'SPECIAL INFO FOR ANYONE WHO IS LOOKING TO BURN FATS AND LOSE WEIGHT'

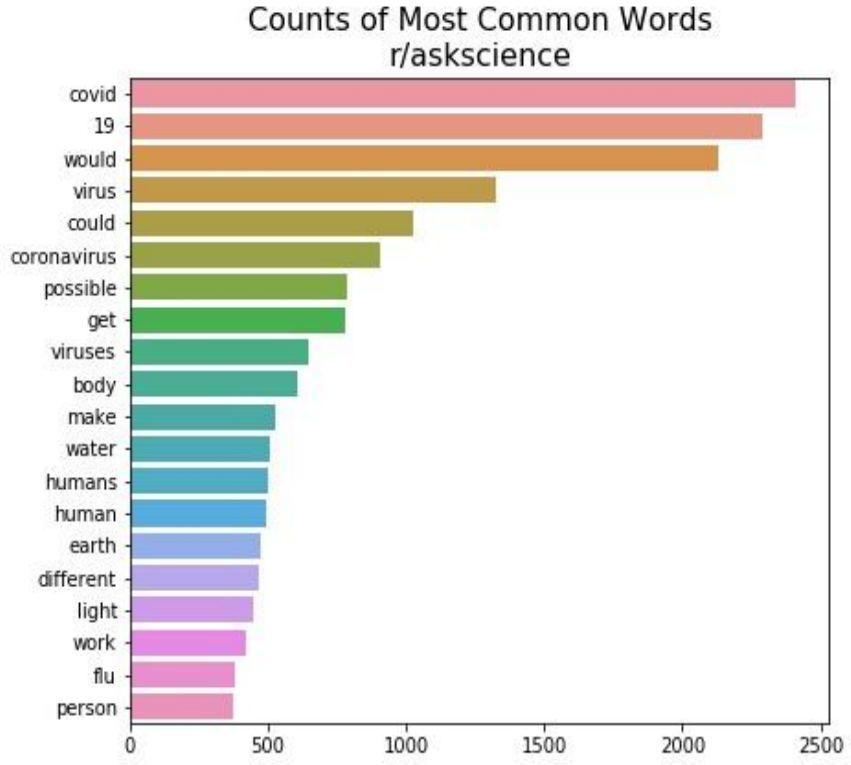
'Puffy Vest = Cold Arms!'

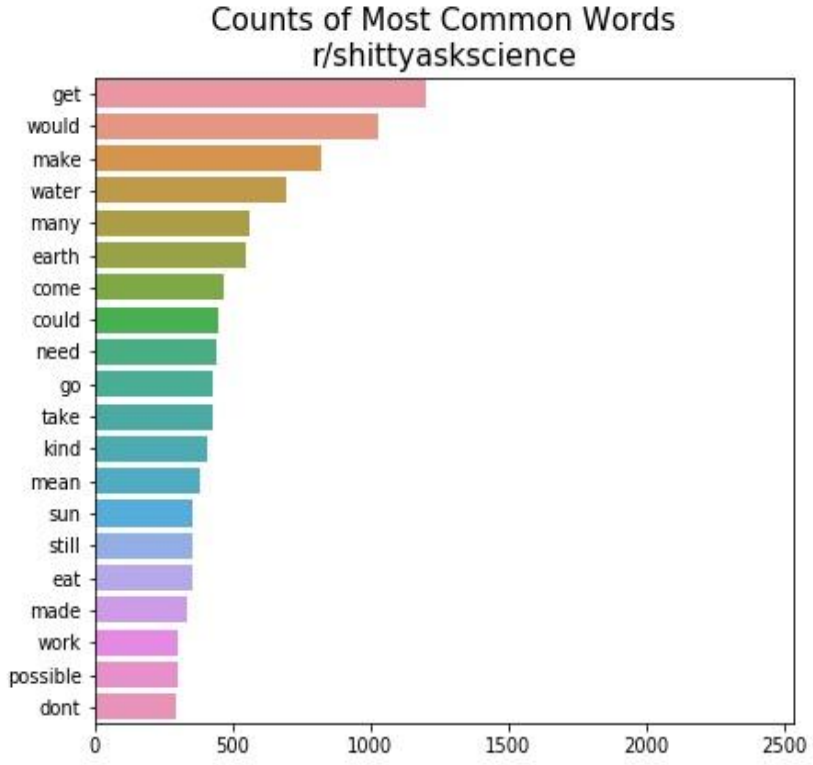
Stop Words


Counts of Shared Most Common Words
r/askscience
like
people
time
one
much
long
someone
use
know
see
space
able
since
around
new
become
science
years
0
200
400
600
800
1000


Counts of Shared Most Common Words
r/shittyaskscience
like
people
time
one
much
long
someone
use
know
see
space
able
since
around
new
become
science
years
0
200
400
600
800
1000

Most Frequent Words





Preprocessing

The Power of the Vectorizers

CountVectorizer and TfidfVectorizer can take objects as hyperparameters:

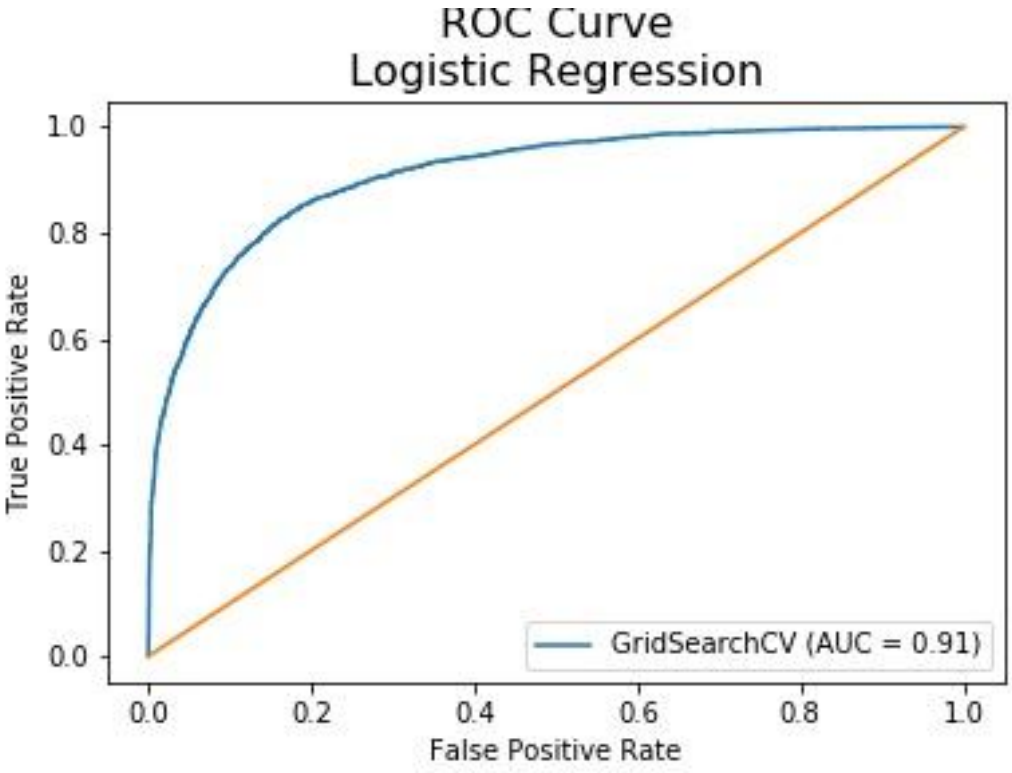
- all_the_stops (big list of stopword permutations)
- PorterStemmer (as a function)
- token_pattern = Regex Pattern

Model Metrics

Model	Training	Testing	Sens.	Spec.	ROC AUC
K_Nearest Neighbors	.666	.671	.382	.939	.74
Multinomial Naive Bayes	.823	.827	.785	.866	.90
Gaussian Naive Bayes	.754	.750	.646	.847	.79
Complement Naive Bayes	.825	.830	.792	.865	.90
Logistic Regression	.826	.830	.799	.858	.91
Ridge Classifier	.826	.832	.808	.855	.91
Perceptron Classifier	.784	.786	.795	.777	.87
Support Vector Classifier	.823	.826	.791	.859	.91
XGBoost Classifier	.815	.817	.780	.852	.90

The Chosen One

Logistic Regression with TfidfVectorizer



Heavy Weights

covid	702.4
viruses	463.7
19	405.6
virus	305.4
coronavirus	167.6
covid19	104.8
pandemic	67.8
masks	38.2
vaccine	36.7
affect	34.2

shit
poop
sex
fart
etc.

The Real Test

How does a waste disposal pyrolysis machine work?

T P

Why does spontaneous radioactive decay happen?

T P

Are there emotions humans can't feel?

T P

Why do some farts feel hotter than others?

T P

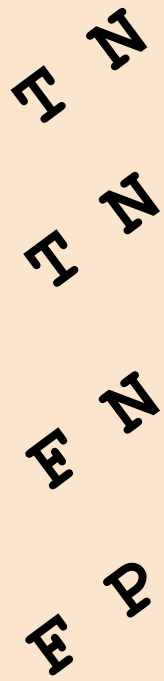
The Real Test

Is ice that’s made on the ocean salty?

Do crabs think fish can fly?

I’ve awaken coughing, farting, or snoring, but why I have I never awaken myself sneezing?

What frame rate am I seeing the world at?



The Real Test

If Oreo is milk's favorite cookie, what are some of milk's other favorite things?

Can somebody explain osmosis in middle school level terms?

If light travels fast and I travel light then why the hell is my plane late?

Why do birds sing so fucking much in the morning?

Conclusions

Language and Humor are Subjective

Plague-Skewed Results

“Where are my shoes?”

Special Thanks

The Official Documentation for SciKit Learn

"Applied Text Analysis with Python" by Benjamin Bengfort, Tony Ojeda, & Rebecca Bilbro

“Feature Engineering for Machine Learning” by Alice Zheng & Amanda Casari

"Doing XGBoost hyper-parameter tuning the smart way — Part 1 of 2" by Mateo Restrepo

https://towardsdatascience.com/doing-xgboost-hyper-parameter-tuning-the-smart-way-part-1-of-2-f6d255a45dde