

Online Science Forum Post Filtering



Rhoeun Park
dsi-flex-222



Brandeis University

Division of Science

We are planning to **build an online scientific forum** for students and faculties to freely exchange scientific questions of their interests.

This project seeks to find a strategy to **filter out 'troll' forum posts**, that are scientifically irrelevant.

By utilizing natural language processing and classification models on two different subreddits, '**AskScience**' and '**ShittyAskScience**', the project will explore the differences in the texts of troll and legitimate scientific questions and aim to classify between the two.

Datasets

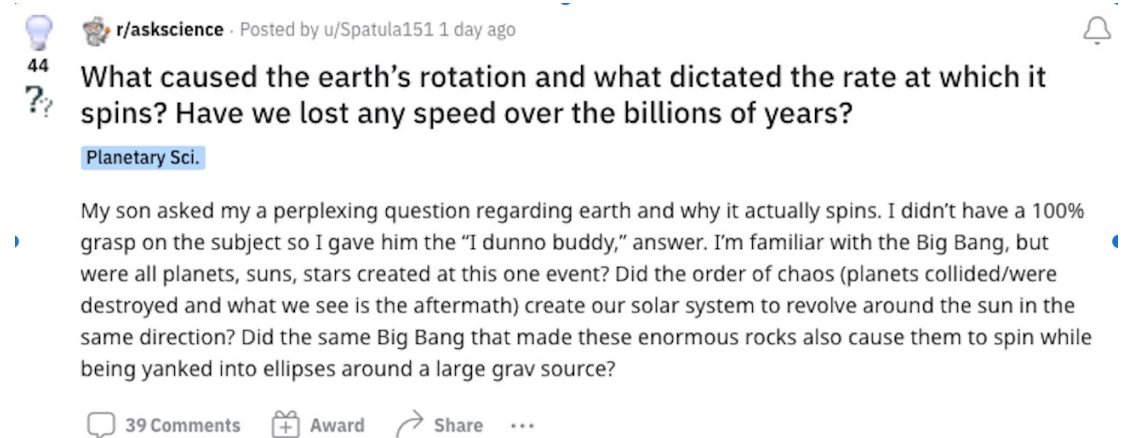


- ♦ **Pushshift API** to access and scrape texts from posts
- ♦ About 5000 posts scraped from each of the **AskScience** and **ShittyAskScience** subreddits
- ♦ Data cleaning
 - ♦ Remove URLs
 - ♦ Vectorizers to create bag of words

r/AskScience

Real science questions and real
scientific answers

Rules more strictly regulated



r/ShittyAskScience

Meant to be funny

Does not actually makes
scientific sense

14

r/

r/shittyaskscience

· Posted by u/AbouBenAdhem 1 day ago

If medical researchers keep developing antibiotics to target and destroy bacteria, will bacteria eventually adapt to target and destroy medical researchers?

4 Comments

Award

Share

...

8

r/

r/shittyaskscience

· Posted by u/alphabeticusername very human, yes 17 hours ago

I'm about to buy some uranium. What should I expect to happen when it converts to myranium?

3 Comments

Award

Share

...

2

r/

r/shittyaskscience

· Posted by u/FeebleKneevil 2 days ago

If a room is at 77°F and there is a heat source in said room that is at 98°F, what temperature will the room be when it reaches equilibrium?

The room is a 60x200 retail store with a drop ceiling, not insulated.

The heat source(s) are a metal roll up door on the west side of the building and large plate glass windows on the east side of the building. The humidity is at 88%. Temperature of the metal door is currently at 110°F.

The thermostat reads 77°F but it's located near a ceiling vent where the air is blowing on it. There are also 6 computers running but I don't know their running temperatures.

10 Comments

Award

Share

...

Drawbacks

Unable to access body of text using Pushshift API

Had to work with only the Title of the subreddits

Many posts in ShittyAskScience subreddit refers to other posts or GIF's, which are not able to be analyzed using Natural Language Processing

Features Compared



COUNTS

- Characters
- Words
- Numerical Texts



WORD FREQUENCIES

- Stop words?
- Lemmatized
- Bigrams



SENTIMENT SCORES

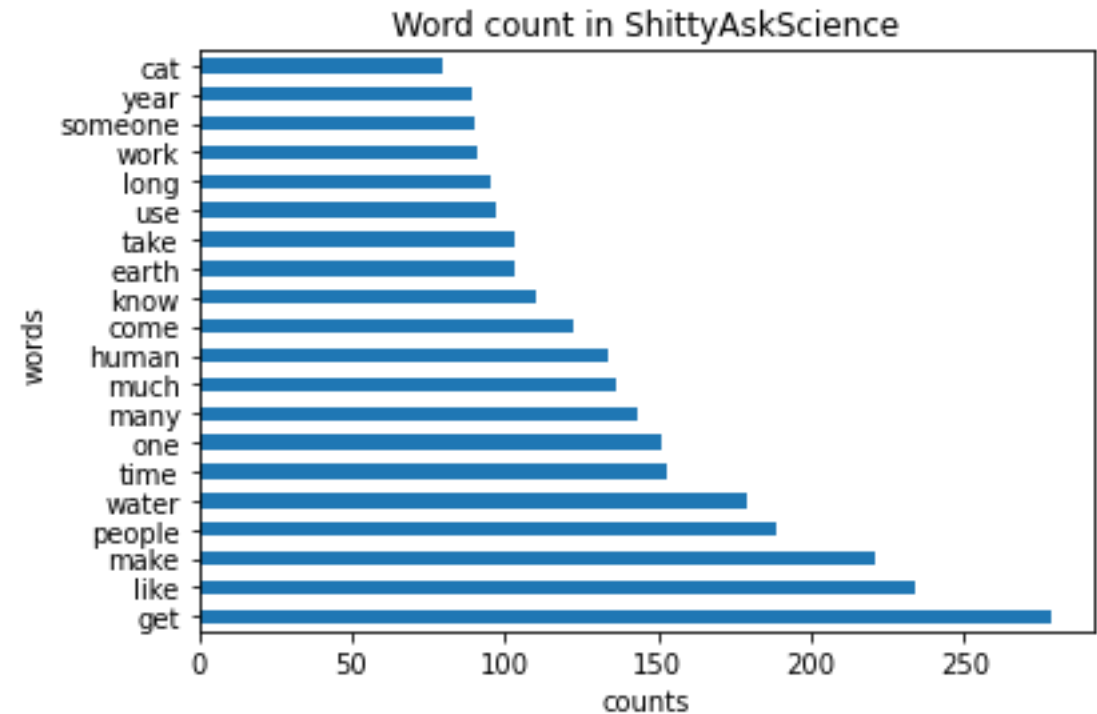
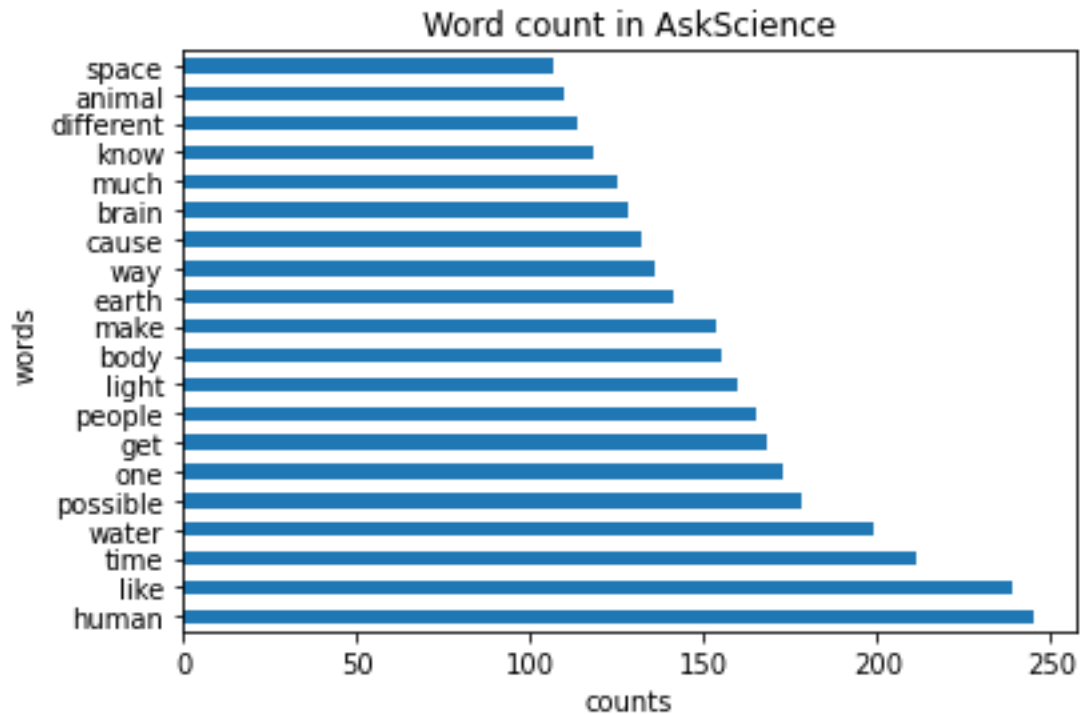
- Negative
- Neutral
- Positive

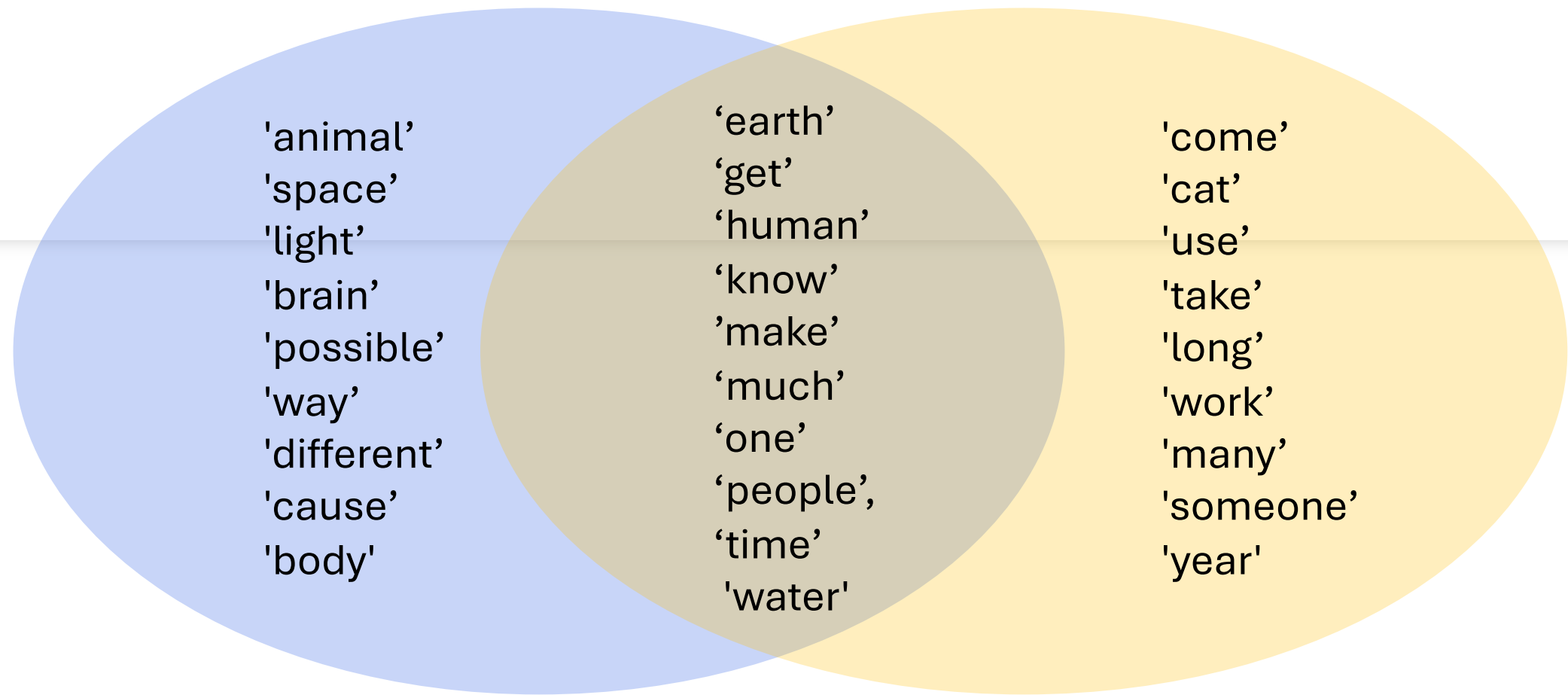


COMPLEXITY

- Ave word counts
- Flesch Reading Ease score (FRE)

Word Frequencies

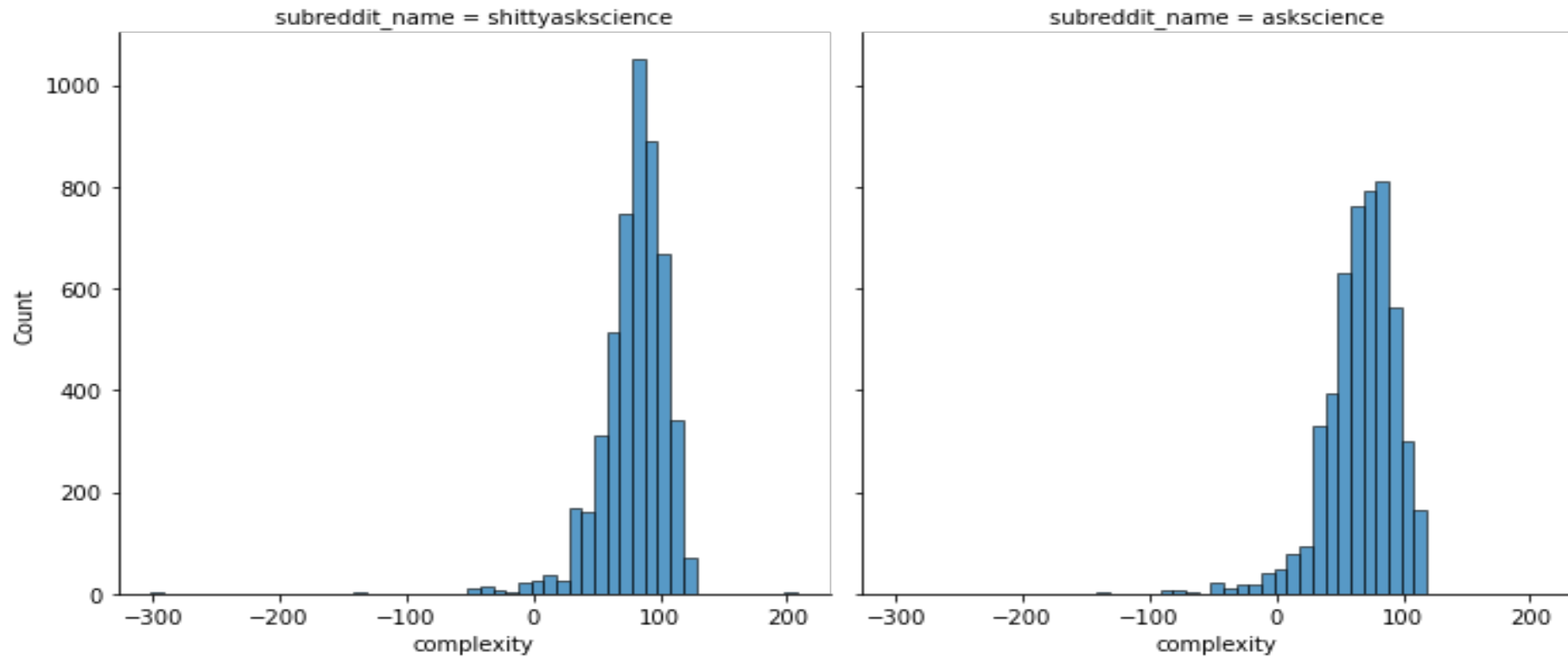




AskScience

ShittyAskScience

Text Complexity



$$FRE = 206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Subreddit	Average Word Length	FRE Score
AskScience	6.04	66.08
ShittyAskScience	5.49	79.00

Modeling: Metrics

	MultinomialNB	Random Forest	Logistic Regression	KNN Classifier
Accuracy	0.7201	0.7262	0.7069	0.6109
Recall	0.7082	0.6488	0.6984	0.8236
Specificity	0.7320	0.8035	0.7154	0.3986
Precision	0.7251	0.7672	0.7102	0.5776
F1	0.7166	0.7031	0.7042	0.6790

accuracy: percentage of correct subreddit classification

Sensitivity(Recall): Among the actual ShittyAskScience submissions, how many(or proportion) did I predict correctly

Specificity(True negative rate): Among the actual AskScience submissions, how many(or proportion) did I predict correctly

Precision(Positive predictive value): Among the posts I predicted as ShittyAskScience, how many did I get correct?

Conclusion

- Overall, model improvement is needed as the model with highest accuracy predicts less than a third of the texts correctly classified
- Possibly able to enhance the model if I can scrape the text body, not just the title
- Will return with a better model!