Reddit Analysis & NLP

Table of Contents

Main “Characters”

r/AskScience

r/AskScience aims to promote scientific literacy by helping people understand the scientific process and what it can achieve

r/AskPhilosophy

r/AskPhilosophy aims to provide serious, well-researched answers to philosophical questions

Problem Statement

Given 2 different specific reddit threads, what model should be chosen to differentiate between either, also, how should the model be evaluated?

Part 1

- **Web Scraping**
- **Data Wrangling**

Part l : Data Wrangling & Web Scraping

Web Scraping

Data was scraped using 3 libraries:
1) json
2) BeautifulSoup
3) requests

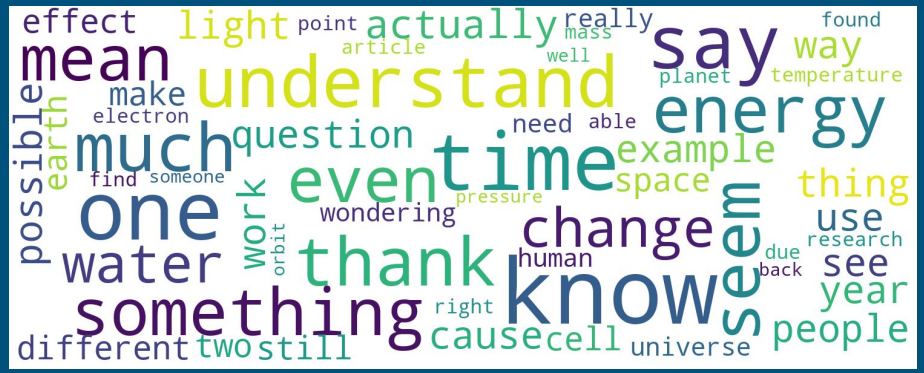
These were then imported in pandas dataframes after being exported as csv files

Data Wrangling

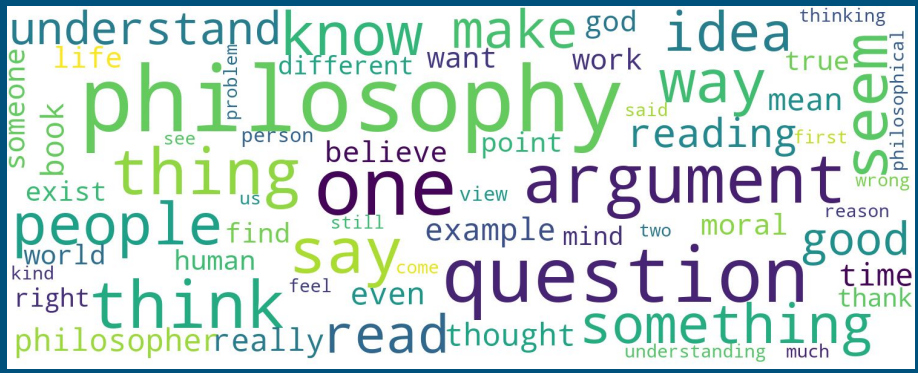
Preprocessing was done using regular expressions to remove unwanted words or signs.

Word Clouds to visualise top words of each thread

r/askphilosophy



r/askscience



Part ll

- Processing

Part ll : Preprocessing

Set up prediction column

Set up a binary variable column for predictions to be consolidated and trained on

Word Normalization

Word normalisation techniques such as lemmatization and tokenization applied to split the string to individual words, and the individual words were then lemmatized to avoid repetitiveness of certain words, and common words such as ‘the’, his’, ‘her’ etc. aka stopwords were removed.

Part lll

- Classification Modeling
- Interpretation & Model Selection

Part lll : Classification Modeling

Pipelines

Used to create a 2-step process:

1) The model selected
2) Word embedding method

Models Used

1) Multinomial Naive Bayes
2) k-Nearest Neighbors
3) Logistic Regression

Conclusions

- After comparing across the 3 models, MultinomialNB produces the highest accuracy scores consistently across both vectorisation methods.

- **Achieving an accuracy score of 0.92 would mean that we would be able to differentiate between a r/askscience between a r/askphilosophy thread 9 out of 10 times correctly.**

- A true positive meant that the model correctly predicted a *r/askscience* post, while a true negative meant that the model correctly predicted a *r/askhilosophy* reddit post.

- This rate compared to the baseline accuracy was within a 89% - 95% range

Thank you.