Classifying r/GRE and r/GMAT posts

An Application of Natural Language Processing and Machine Learning

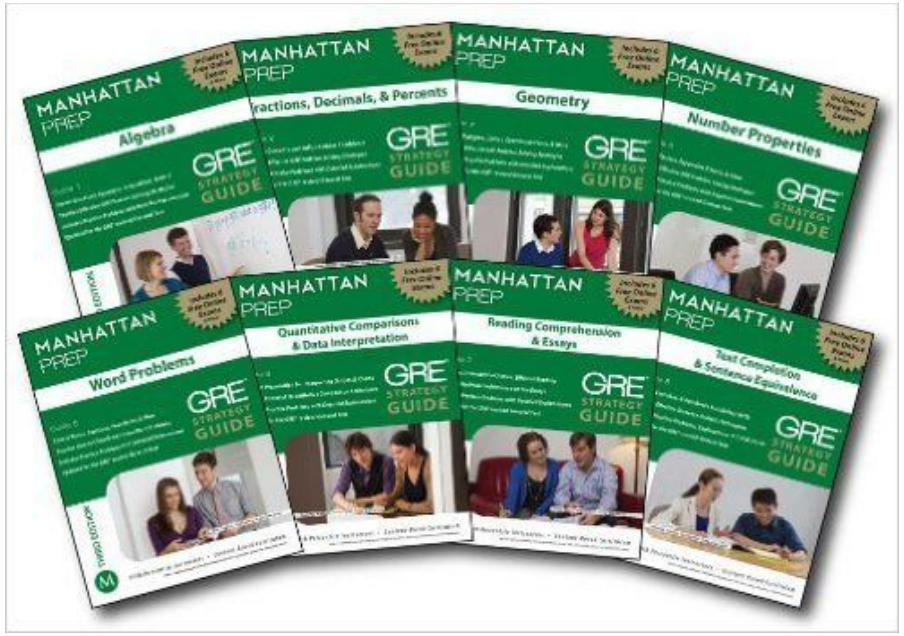
By

Nnenna Isigwe

DSIR 927, 11/5/2021

Problem Statement

How can Manhattan Prep track what people are saying about their products on Reddit?



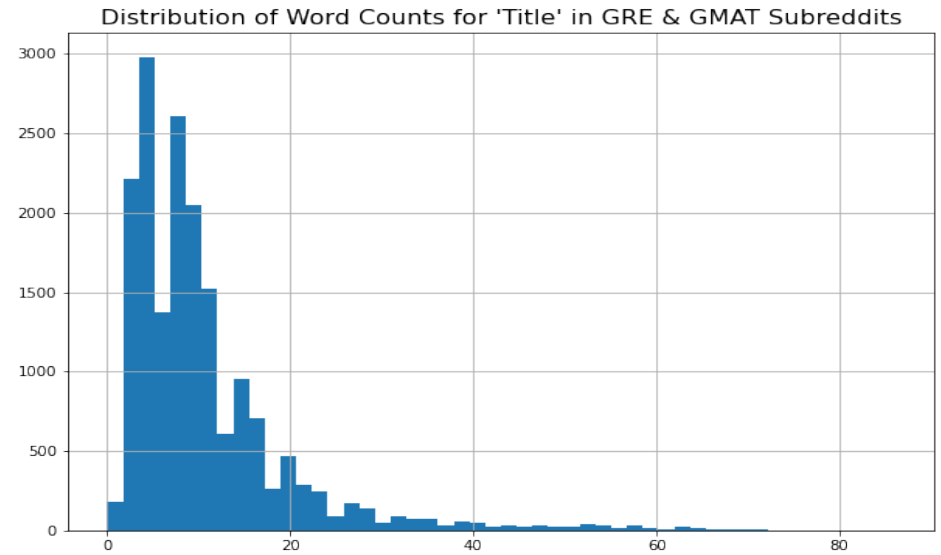
- The company would like to leverage NLP in improving UX.
- The aim is to remain a top choice for GRE and GMAT resources.

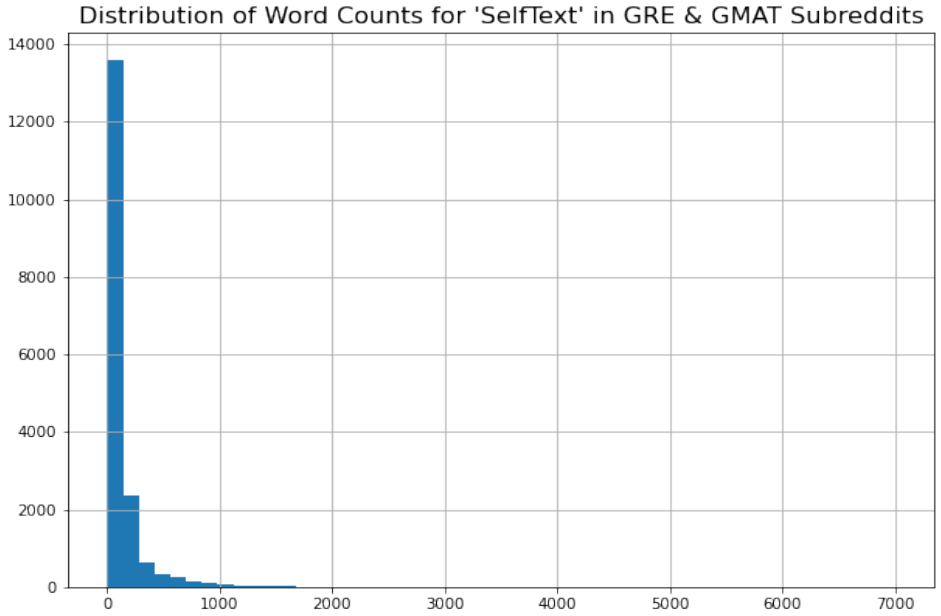
Data

	r/GRE	r/GMAT
Scraped	10,000 posts	8,500 posts
Unique Titles	9,523	8,149
Null values (self texts)	1,673	1,271
Final count (post-cleaning)	**9,447 posts**	**8,072 posts**

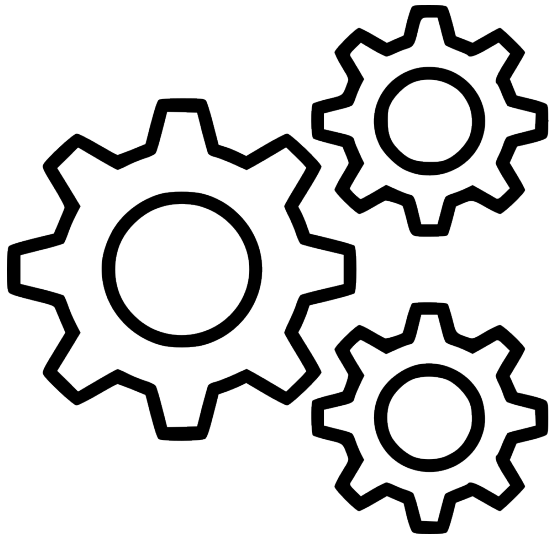
Data Cleaning

- Engineered word count variable for title and self text columns.
- Dropped entries where title word count > 60 words, self text word count > 1,700.
- Mean word counts did not differ significantly between both subreddits.
- For titles, 11 words and 10 words for r/GRE and r/GMAT, respectively. For self texts, 124 & 117 words for r/GRE and r/GMAT, respectively.
- Dropped duplicate titles, links, and replaced null self texts with corresponding entries in title column.



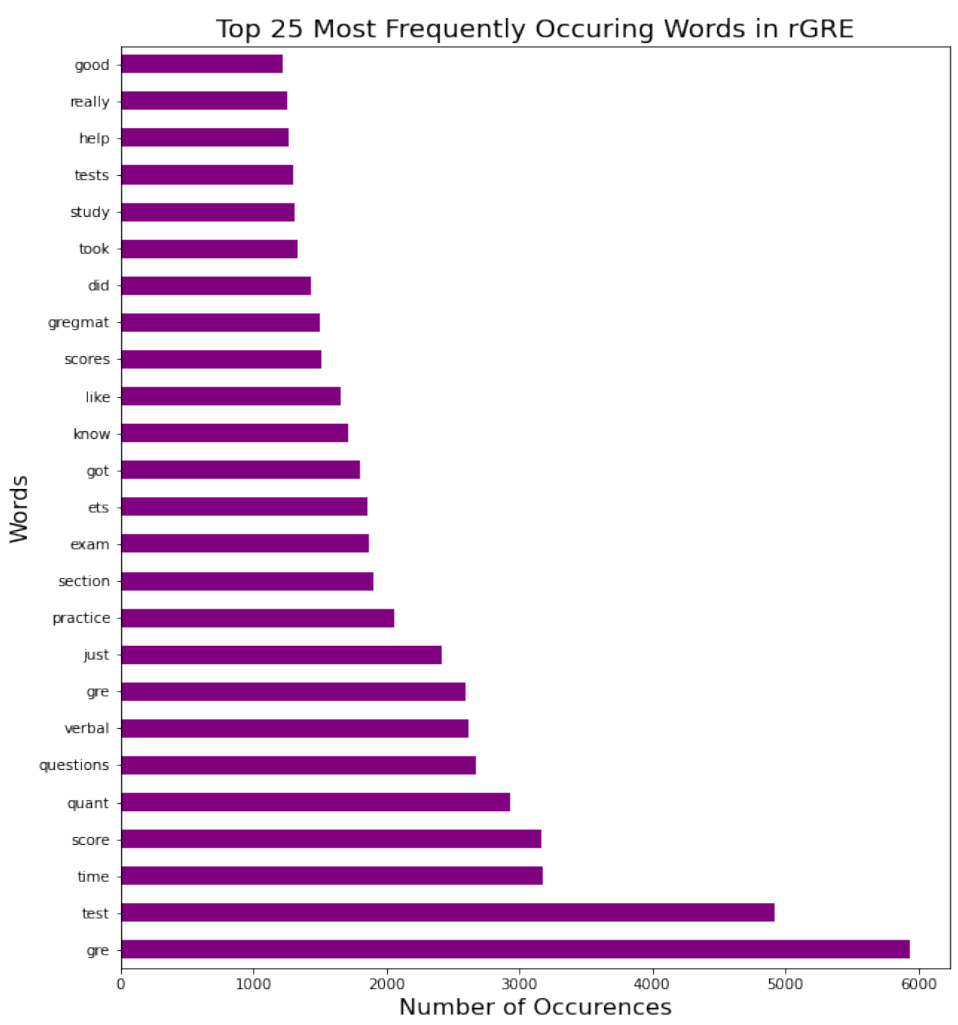


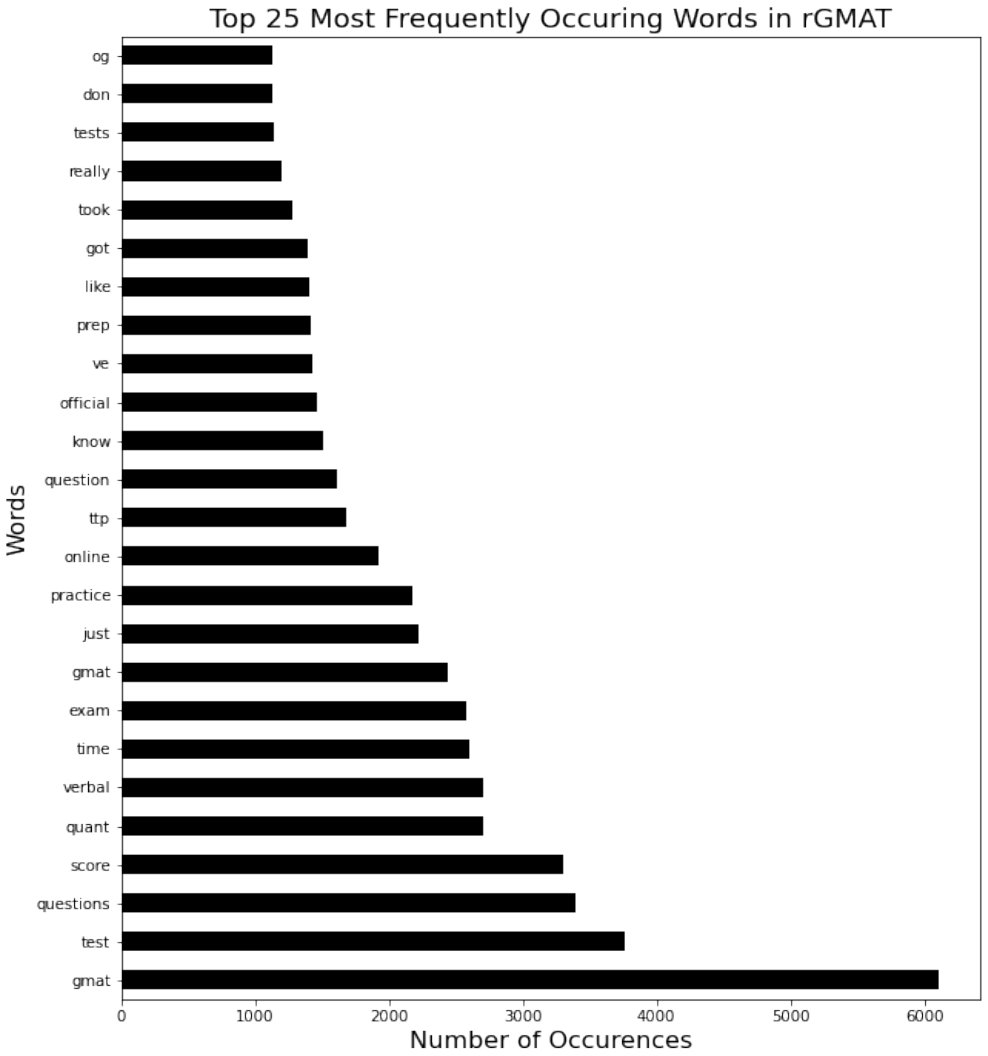
Preprocessing



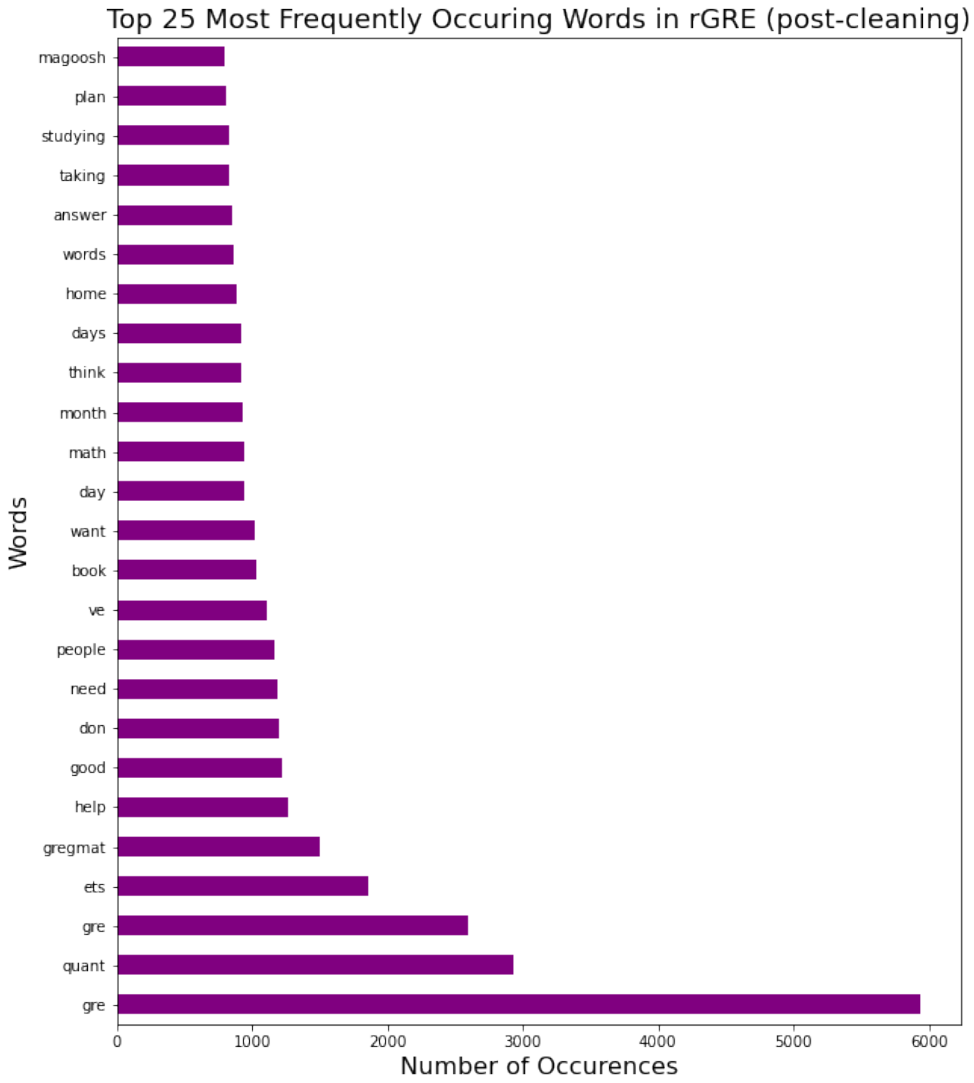
- $\boldsymbol{X} = title, self\ text$
- $\boldsymbol{y}$ = subreddit (1 = r/GRE; 0= r/GMAT)
- **Instantiated train/test split:**
  - Stratify = y
  - Test size = 0.25
  - Random State= 42
- **Instantiated Count Vectorizer** :
  - Stop words = “english”
  - ngram range = (1,3) for titles & (1,6) for self texts
  - Max df = 0.8 for both titles and self texts
  - Min df = 0.02 for titles, 0.05 for self texts
- Count vectorized features = 133

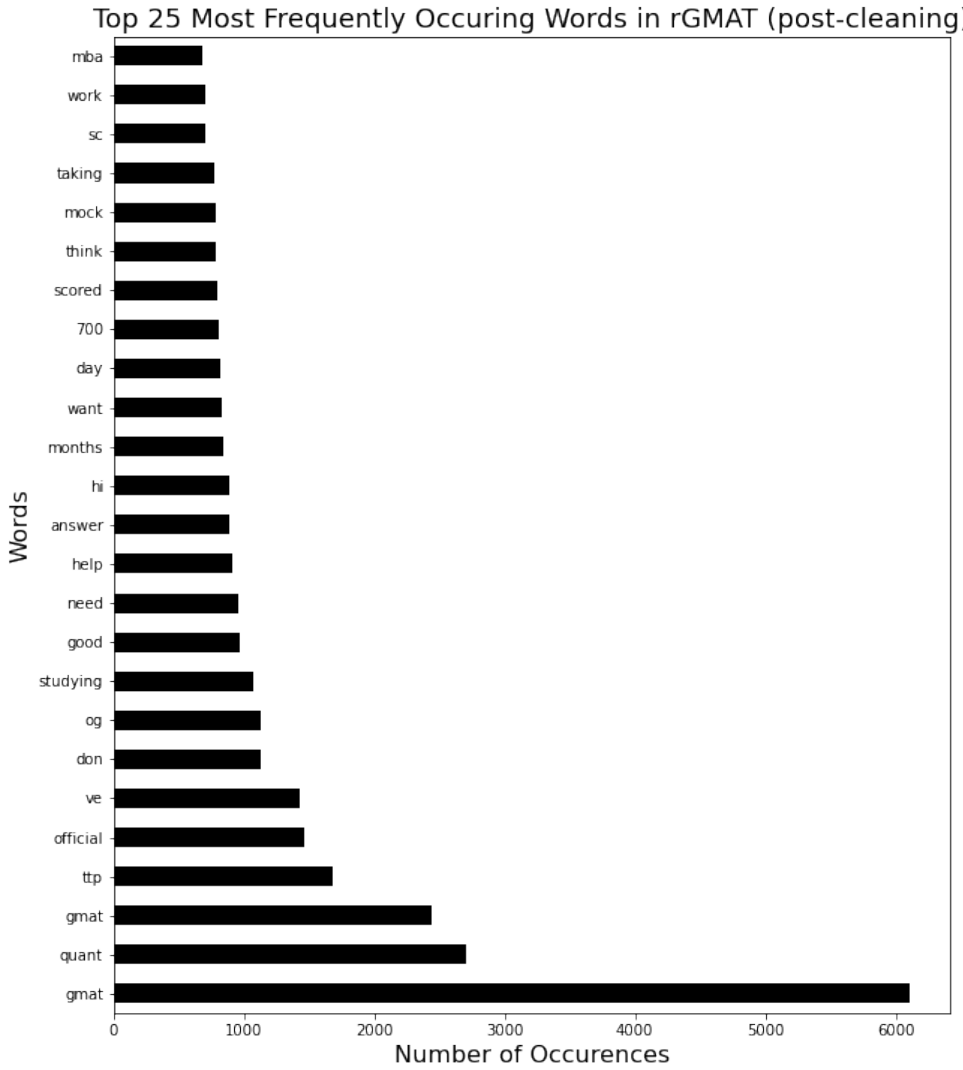
EDA


Top 25 Most Frequently Occuring Words in rGRE
Words
good
really
help
tests
study
took
did
gregmat
scores
like
know
got
ets
exam
section
practice
just
gre
verbal
questions
quant
score
time
test
gre
0
1000
2000
3000
4000
5000
6000
Number of Occurences


Top 25 Most Frequently Occuring Words in rGMAT
Words
og
don
tests
really
took
got
like
prep
ve
official
know
question
ttp
online
practice
just
gmat
exam
time
verbal
quant
score
questions
test
gmat
0
1000
2000
3000
4000
5000
6000
Number of Occurences

EDA (2)


Top 25 Most Frequently Occuring Words in rGRE (post-cleaning)
magoosh
plan
studying
taking
answer
words
home
days
think
month
math
day
want
book
ve
people
need
don
good
help
gregmat
ets
gre
quant
gre
Words
0
1000
2000
3000
4000
5000
6000
Number of Occurences


Top 25 Most Frequently Occuring Words in rGMAT (post-cleaning)
mba
work
sc
taking
mock
think
scored
700
day
want
months
hi
answer
help
need
good
studying
og
don
ve
official
ttp
gmat
quant
gmat
Words
0
1000
2000
3000
4000
5000
6000
Number of Occurences

Model Summary Results

Baseline Accuracy
1 = 0.5392
0 = 0.4608

	Naïve Bayes	Random Forest	Logistic Regression
Parameters	α = 1.0	Max_depth= 19 Max features ='auto' Min samples leaf = 2 Min samples split = 4 No of estimators = 750	Max iter = 20,000 Penalty = l2 Solver = 'lbfgs'
Accuracy (Train | Test)	0.89 |0.90	0.90 | 0.90	0.91|0.90
Sensitivity	0.85	0.98	0.85
Specificity	0.94	0.79	0.94

Sentiment Analysis

- $\boldsymbol{X} = self\ text$; $\boldsymbol{y}$ = subreddit (1 = r/GRE; 0= r/GMAT)
- 1,288 posts where “manhattan” is mentioned at least once.
- Instantiated Count Vectorizer with same parameters.
- Instantiated Sentiment Intensity Analyzer with default parameters.
- “Sentiments” associated with Manhattan Prep in r/GRE & r/GMAT were neutral on average.

Compound Score	r/GRE	r/GMAT
Positive	504 posts	410 posts
Negative	164 posts	139 posts

Model Summary Results (2)

Baseline Accuracy for ***Manhattan Subsample***
1 = 0.5512
0 = 0.4488

	Naïve Bayes	Random Forest	Logistic Regression	Logistic Regression (Manhattan Subsample)
Parameters	α = 1.0	Max_depth= 19 Max features ='auto' Min samples leaf = 2 Min samples split = 4 No of estimators = 750	Max iter = 20,000 Penalty = l2 Solver = 'lbfgs'	**Max iter = 10,000** **Penalty = l2** **Solver = 'lbfgs'**
Accuracy (Train | Test)	0.89 | 0.90	0.90 | 0.90	0.91 | 0.90	**0.99 | 0.94**
Sensitivity	0.94	0.98	0.94	**0.93**
Specificity	0.85	0.79	0.85	**0.95**

Conclusion & Recommendations

- Estimated 4 classification models, which have accuracy scores of approx. 90% in distinguishing between r/GRE and r/GMAT posts.
- Recommended model is Logistic Regression, due to better model performance in this case.
- All 4 models are low bias and low variance models.
- Sentiments associated with Manhattan Prep in both subreddits were **neutral** on average, and did not differ significantly between both subreddits.
- Deeper analysis should be carried out on posts where compound sentiment scores were negative.

References

- https://towardsdatascience.com/classifying-reddit-posts-with-natural-language-processing-and-machine-learning
- 695f9a576ecbhttps://towardsdatascience.com/classifying-reddit-posts-with-natural-language-processing-and-machine-learning-695f9a576ecb
- https://stackoverflow.com/questions/66783488/code-efficiency-performance-improvement-in-pushshift-reddit-web-scraping-loop