Web Scraping Reddit Front Page

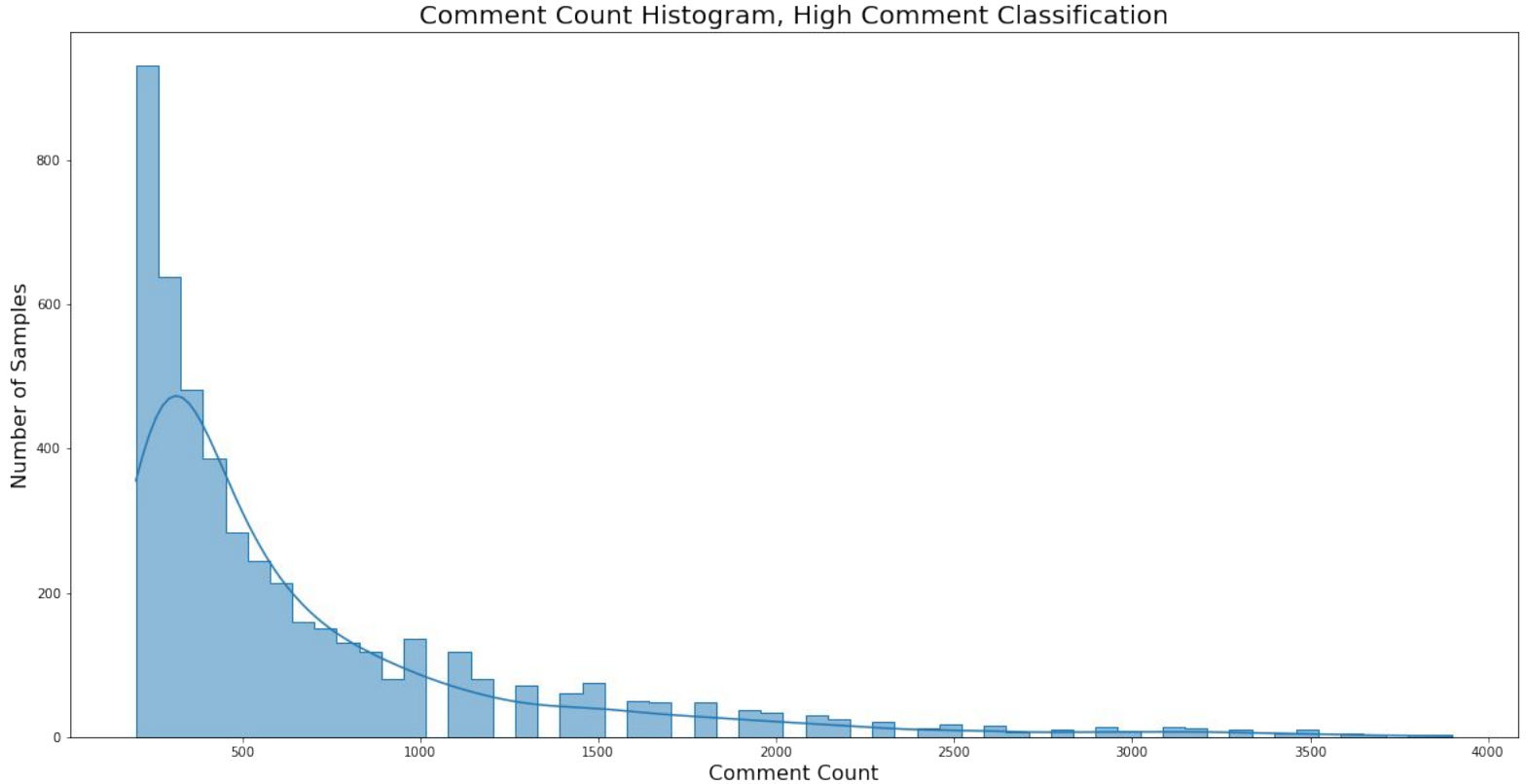
General Assembly Project Three
Peter Wentzel

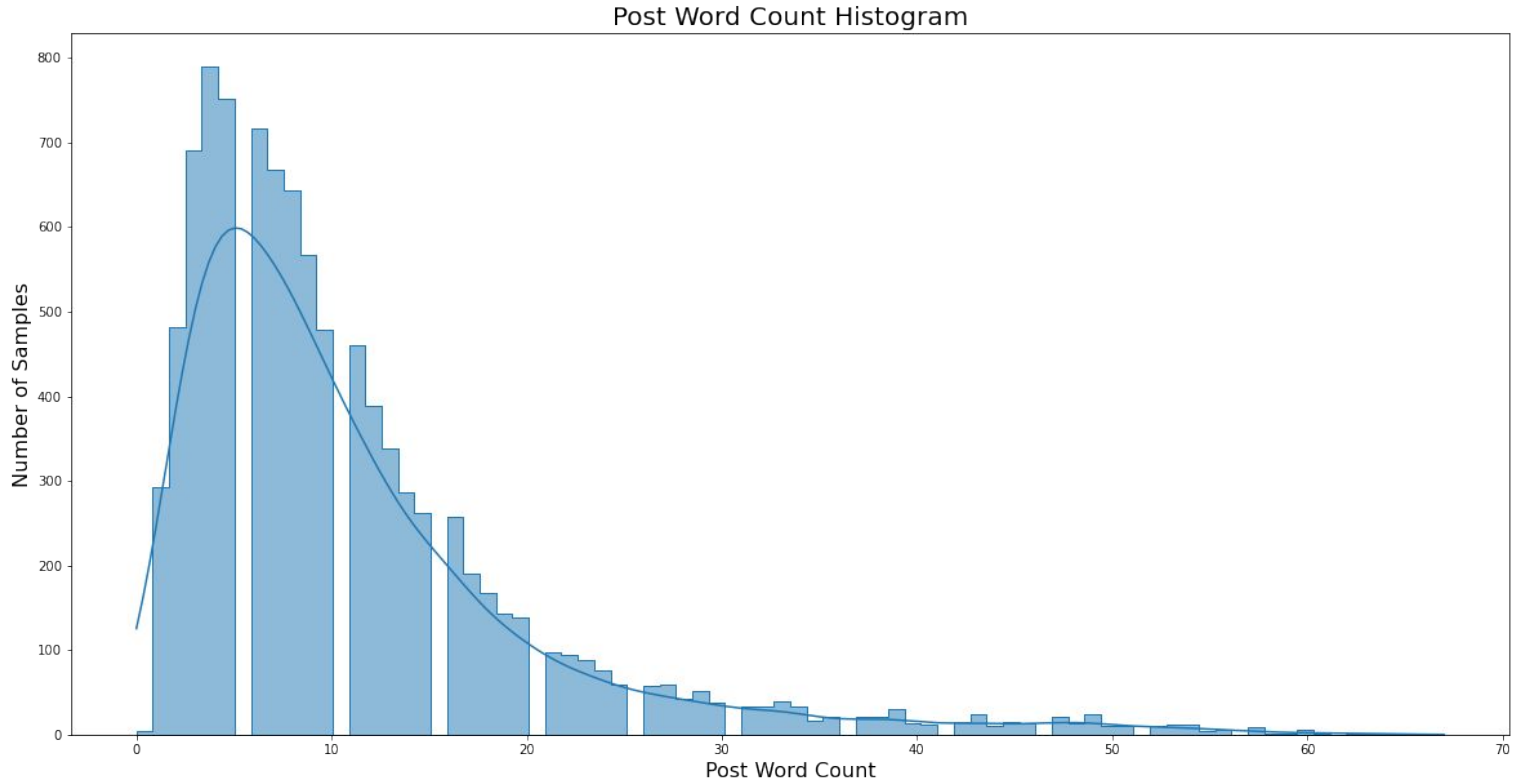
Problem Statement and Dataset/Acquisition

- By scraping information from the front page and performing data processing and modeling, the goal of this study is to provide more insight on the attributes of a high engagement post on the front page of Reddit.
- A web scraper was created in Python and in conjunction with Selenium to automate scrolling on the front page to gather data.
  - Five elements of information were gathered
    - Thread title
    - Upvote count
    - Comment count
    - Subreddit
    - Post time up
- Data was gathered from 23DEC2021 to 6JAN2022.
  - Data was gathered twice a day, ~1000 samples per
- After trimming dataset ~12000 samples used for observations and model creation.

Initial EDA

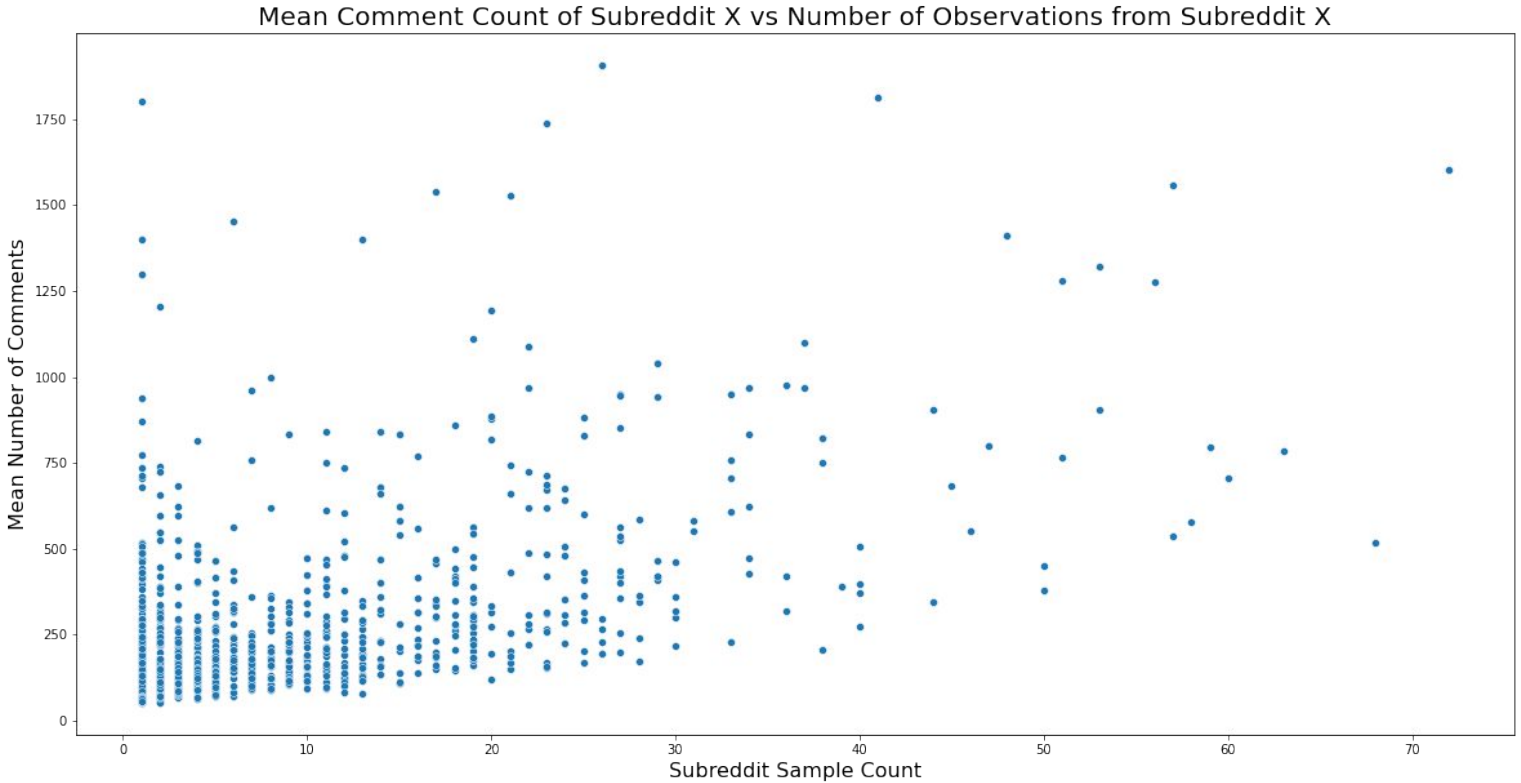
- Post initial cleaning, set was reduced to ~9900 samples
  - Duplicate entries lingered, Reddit front page for high traction posts does not completely change within 12 hours.
- Median comment count was found to be 200, which became the baseline metric.
- Easy answer for high comment post: run a giveaway!
- Ask a question that is easily relatable or on a controversial topic
- Keep the comment length somewhere between 8 and 20 words
- 1199 unique subreddits were observed in the cleaned dataset


Comment Count Histogram, High Comment Classification
Number of Samples
0
200
400
600
800
500
1000
1500
2000
2500
3000
3500
4000
Comment Count


Post Word Count Histogram
Number of Samples
Post Word Count
0
100
200
300
400
500
600
700
800
0
10
20
30
40
50
60
70

Powerful Metric, the Subreddit

Subreddit	Count in Data	Mean Comments
r/antiwork	72	1603
r/MadeMeSmile	68	517
r/next[REDACTED]level	63	785
r/Damnthatsinteresting	60	705
r/interestingas[REDACTED]	59	795
r/Unexpected	58	576
r/facepalm	57	1556
r/memes	57	538
r/WhitePeopleTwitter	56	1275
r/gaming	53	1321
r/HolUp	53	904
r/politics	51	1279
r/funny	51	764
r/aww	50	380
r/Superstonk	50	449
r/pics	48	1411
r/nba	47	800
r/AskReddit	47	10919
r/dankmemes	46	553
r/cats	45	681


Mean Comment Count of Subreddit X vs Number of Observations from Subreddit X
Mean Number of Comments
Subreddit Sample Count
0
250
500
750
1000
1250
1500
1750
0
10
20
30
40
50
60
70

Model Selection and Performance

- Two models were created to try and improve classification beyond the baseline.
  - Random Tree Classifier was used with engineered features and subreddit dummy variables
  - Multinomial Naive Bayes was used with engineered thread text
- Random Tree Classifier produced a model with 66% classification accuracy
  - This is an improvement over our baseline of 50% from the dataset median
- Multinomial Naive Bayes performed similar with cleaned thread text and tf-idf based input, around 57% accuracy.
- Further feature engineering and NLP will be required to increase performance of modeling.
- Image and video data processing of some degree.

69.7k

r/gaming · Posted by u/Agreeable-Giraffe623 16 hours ago 18 12 13

As it should be

When I start the game and see that the inventory space is unlimited

1.4k Comments Share Save Hide Report Tip

85% Upvoted

69.7k

r/gaming · Posted by u/Agreeable-Giraffe623 16 hours ago 18 12 13

As it should be

When I start the game and see that the inventory space is unlimited

1.4k Comments Share Save Hide Report Tip

85% Upvoted