Analyzing Social Trends on Recession conditions

Kanishk Barhanpurkar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
kbarhan1@binghamton.edu

Harshad Bhandwaldar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
hbhandw1@binghamton.edu

Nikita Mandlik
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
nmandli1@binghamton.edu

Brinda Eshwar
Computer Science
SUNY Binghamton University
Binghamton, New York, USA
beshwar1@binghamton.edu

ABSTRACT

Social Media platforms are widely used for transmitting data in different formats. We can generate a lot of information on recent trends. We have collected data from two social networking sites, Twitter and Reddit as public-based platforms. We have also collected data from new articles based on data sources such as New York Times API. Twitter is a public social networking domain and Reddit is a community-based social media platform. The primary aim of this project is to analyze the social media-associated data which predicts the conditions which lead to the Recession over time. Secondly, for the third dataset, we evaluated the news article which contains the topic of recession and economic crisis. techniques to understand the data in a better context. In project phase 3, we have developed a web-based dashboard that will allow for interactive querying.

1. INTRODUCTION

A recession is a business cycle contraction when there is a general decline in economic activity. In 2020, Covid-19 spread exponentially in the entire world due to which many nations imposed a lockdown on the entire nation. This decision affects international trade and decreases the currency flow among various countries [1]. The number of users using social networking sites makes it mandatory to analyze the trend of current affairs happening across the globe. In this project, we are using Twitter Data (Twitter API), Reddit Data (Reddit API) as public-opinion data, and NYTimes data as a news article dataset for recession conditions [2] [3] [4]. In Project phase 1, we collected and stored it in the MySql database. Also, in the second phase, we performed steps of data cleaning and data analysis which generated a lot of insights based on sentiment score analysis and keyword analysis. We have pre-defined the topic of the Recession and added a few keywords as filters to the data-scraping script. Additionally, we will be filtering the tweets and Reddit posts data based on the semantics of English languages. In Project phase 3, we have created a dashboard that shows the changes in several parameters over time. In this Section 2, we have provided the responses to the research questions we have mentioned in Proposal 3. Additionally, we have briefly described the tools and frameworks required for creating a web-based dashboard followed by the conclusion. We have created a separate section for the snapshots of the dashboard with log files for the Twitter, Reddit, and New York Times data.

2. RESPONSES TO RESEARCH QUESTIONS

The primary objective of the project phase is to describe the influence of the recession and economic conditions based on public-opinionated and newspaper articles-based datasets. We have answered the research questions of Proposal 3 as follows-

- **How do public opinion and sources of information (newspapers) associated with each other for recession topics?**

We have performed the keyword analysis based on the word cloud and character frequency analysis. We have found that public opinionated datasets (Twitter & Reddit) keywords changed more frequently as compared to the NYTimes articles data. More political and geographical significance has been observed in the information in the news articles dataset as compared to the Twitter and Reddit data. Additionally, the frequency of the characters per summarized news article is more as compared to public opinionated datasets.

- **How does the sentimental analysis score change for the recession conditions over time?**

We have analyzed this research question based on the sentimental analysis score calculated for each tweet, subreddit comment, and news article summary. We have used the sentimental analysis correlation scale for the analysis. The sentimental score ranges from -1 to +1 where -1,0 and +1 signify negative, neutral, and positive respectively. The NYTimes article data show extreme negativity and positivity compared to the public opinionated dataset. In the Twitter dataset, the sentimental score shows normal distribution which concludes that the user provides several types of opinions on the recession topic. In the Reddit data, we have observed that Reddit users follow the trend of an extremely negative score and positive comments are uniformly distributed on the Recession topic.

3. DASHBOARD

Project phase 3 involves the development of a web-based dashboard using Flask, a micro-web framework. It works in the Python language environment and the entire structure of the pipeline is developed in Python [5]. The dashboard has three tabs for the Twitter, Reddit, and New York Times datasets. Each HTML page shows three graphs as the part of analysis involving sentimental score analysis, and keyword analysis (using word cloud and word frequency analysis). We have also introduced the date slider which helps us to analyze the parameters based on the number of days. The graphs will change according to the dataset and the number of days selected. Thus, through our dashboard, we can make sense of the distribution of the information associated with three datasets. We have used Wordcloud [6], Seaborn [7], and Matplotlib libraries for the keyword analysis [8]. Additionally, for the second research question, we will use the Textblob library to perform sentimental analysis on the Twitter, Reddit, and New York Times data [9]. We have provided snapshots of the dashboard at the end of the report.

4. CONCLUSION

We build the third stage of the pipeline and developed a web-based dashboard using Flask. The entire data science pipeline was created in the Python environment which helps us to generate insights about the polarity of tweets, subreddit comments, and NYTimes articles over the particular timeline. Additionally, the word cloud and frequency of characters per row in the dataset helped us to analyze the difference between the public-opinionated and news articles dataset.

5. ACKNOWLEDGEMENTS

We would like to thank Prof. Jeremy Blackburn and TA Jay Patel for guiding and providing important insights on various steps that helped us to plan and design the project.

REFERENCES

[1] Feyisa, H. L. (2020). The World Economy at COVID-19 quarantine: contemporary review. International journal of economics, finance and management sciences, 8(2), 63-74

[2] TwitterAPI Documentation. https://developer.twitter.com/en/docs/twitter-api

[3] RedditAPI Documentation. https://www.reddit.com/dev/api/

[4] NYTimes API Documentation. https://developer.nytimes.com/docs/articlesearch-product/1/overview

[5] Flask Library Documentation. https://flask.palletsprojects.com/

[6] WordCloud Library Documentation. https://pypi.org/project/wordcloud

[7] Seaborm Library Documentation. https://seaborn.pydata.org/

[8] Matplotlib Library Documentation https://matplotlib.org/

[9] Textblob Library Documentation. https://textblob.readthedocs.io/en/dev/

Dashboard Snapshots

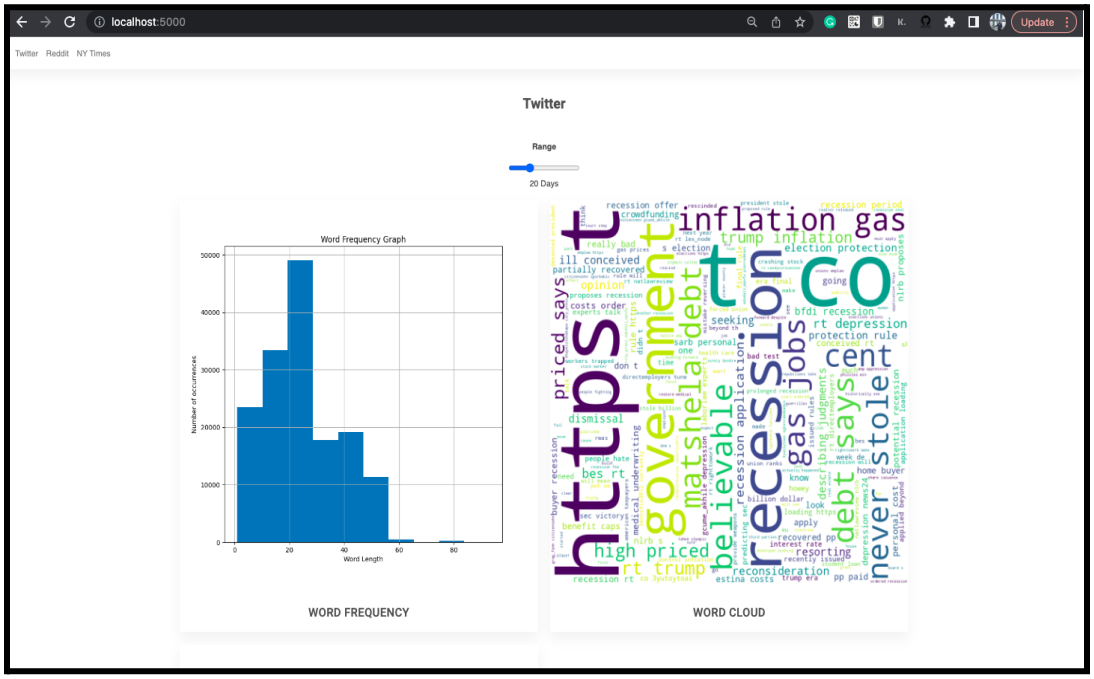

Figure 1: Word Frequency and Word Cloud analysis based on keyword analysis for the Twitter data with the days range of 20 days.

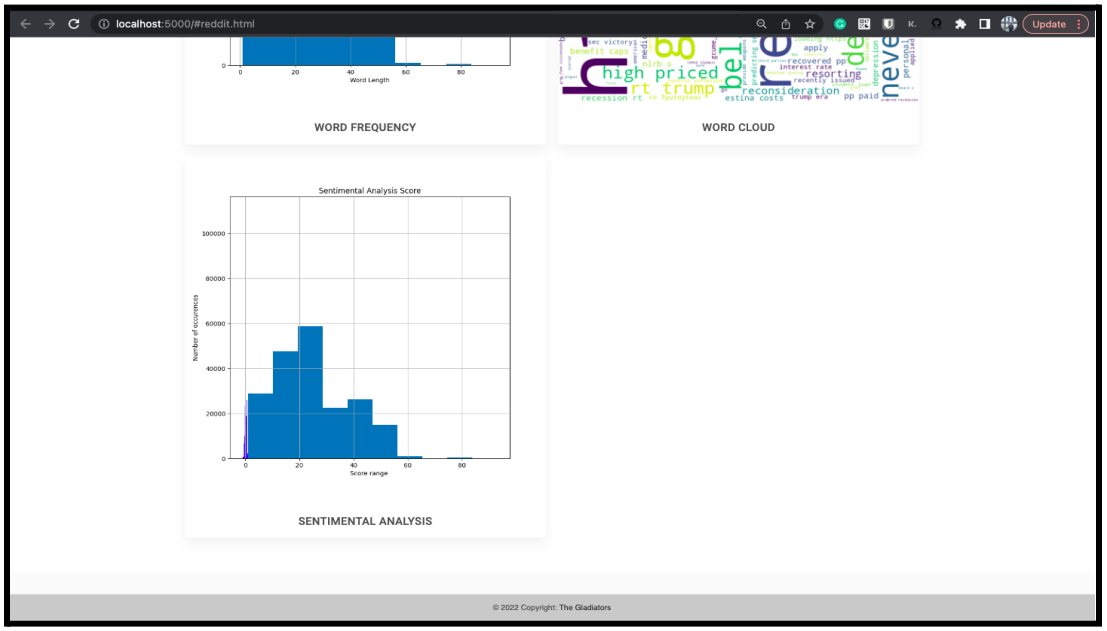

Figure 2: Sentiment Score Analysis for the Twitter data with the days range of 20 days.

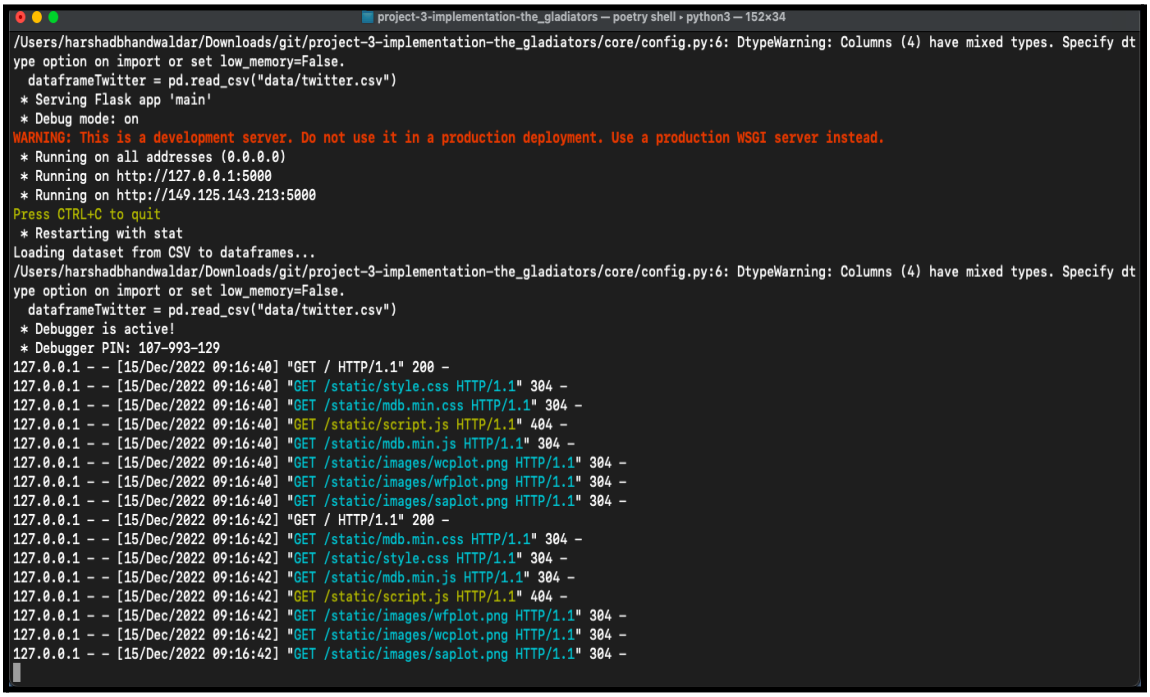

Figure 3: Log-file created while running the analysis and plotting the Twitter graphs accordingly.

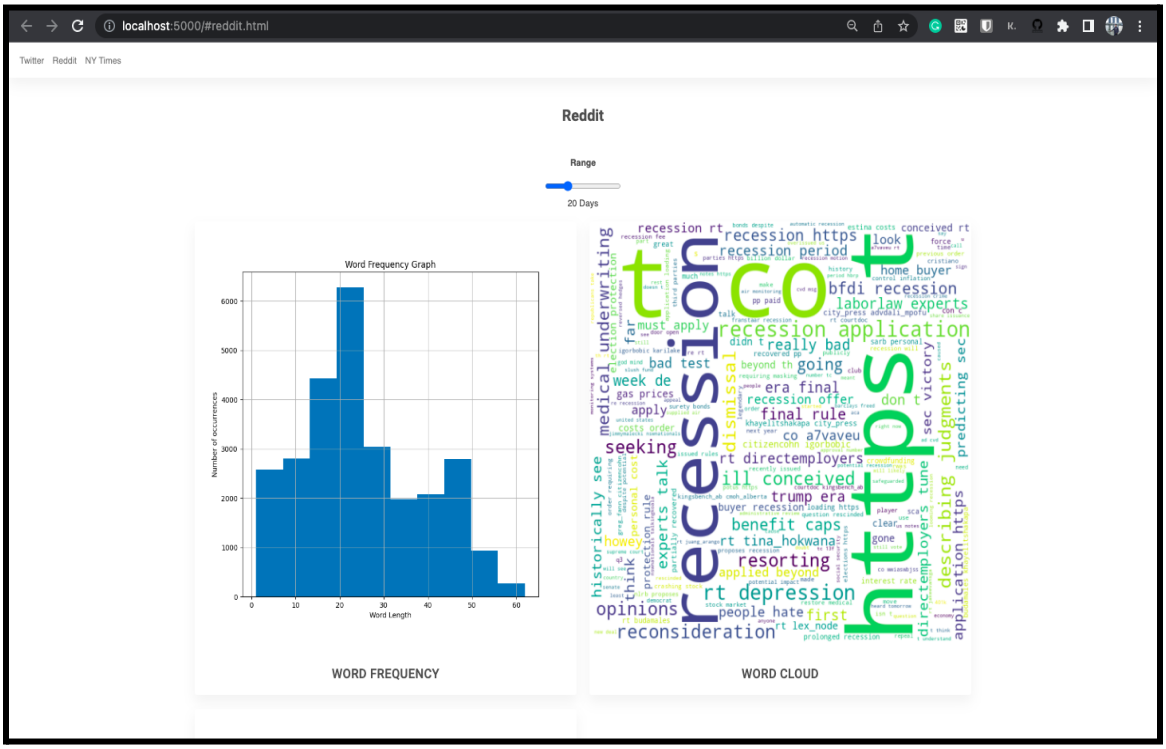

Figure 4: Word Frequency and Word Cloud analysis based on keyword analysis for the Reddit data with the days range of 20 days.

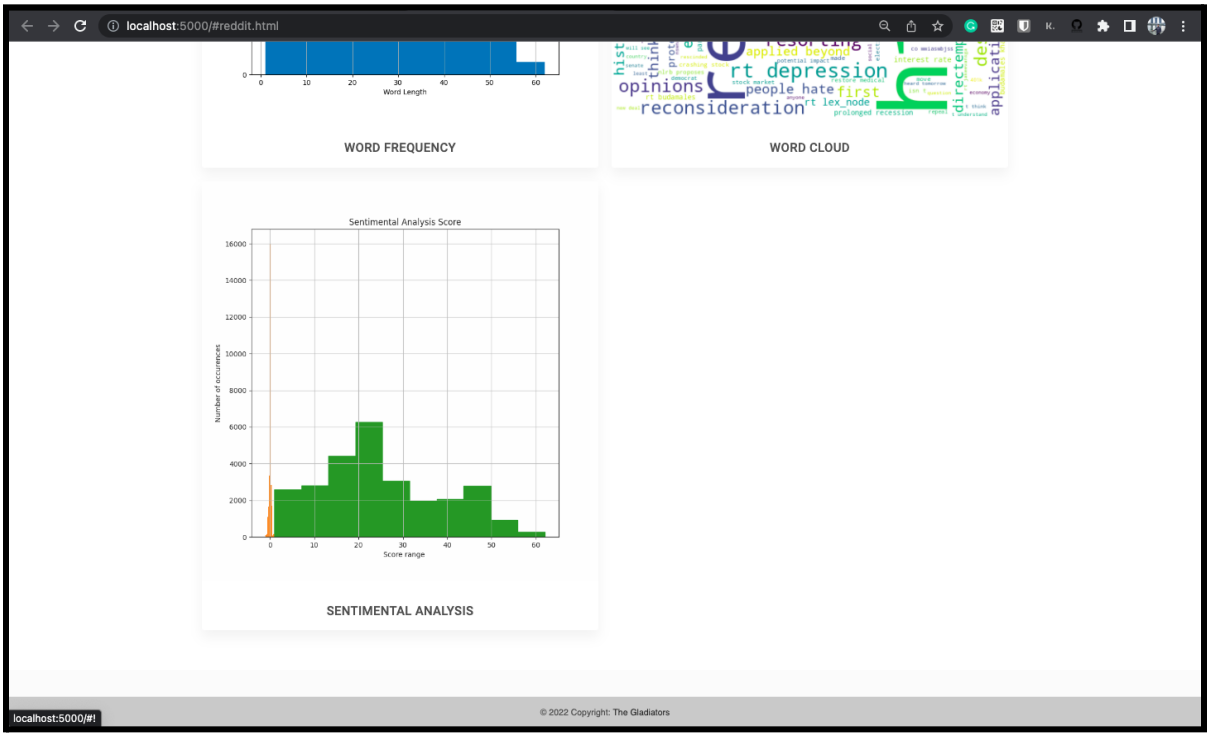

Figure 5: Sentiment Score Analysis for the Reddit data with the days range of 20 days.

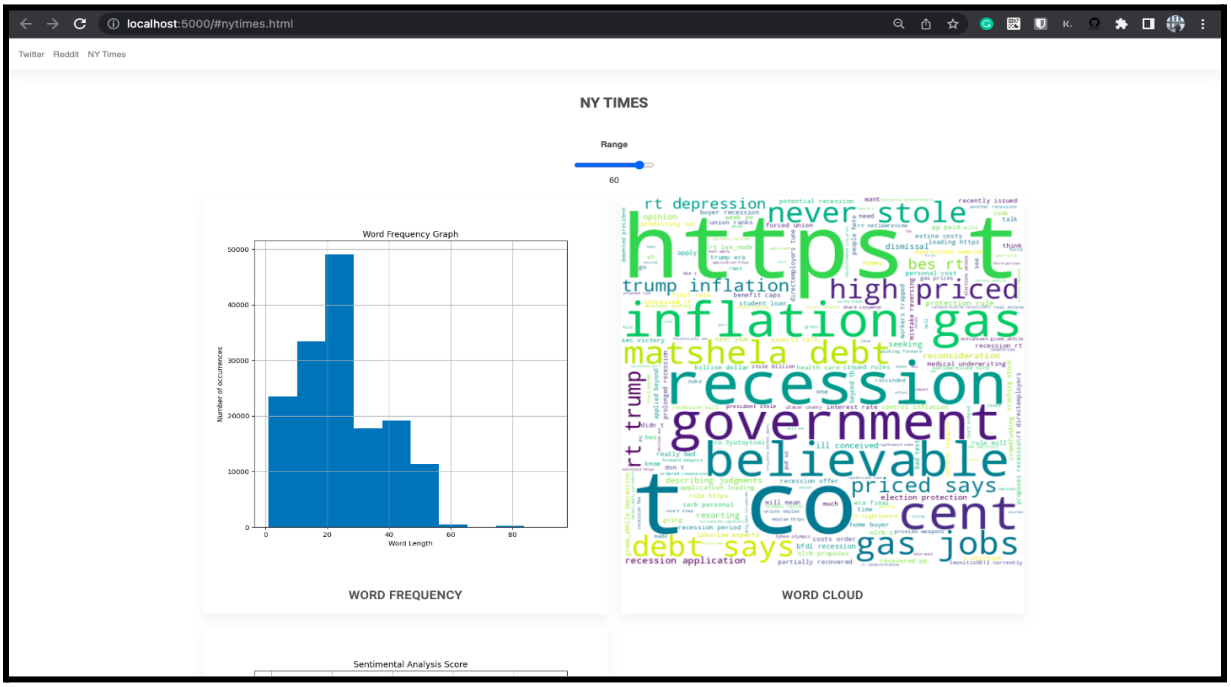

Figure 6: Word Frequency and Word Cloud analysis based on keyword analysis for the NYTimes article data with the days range of 60 days.

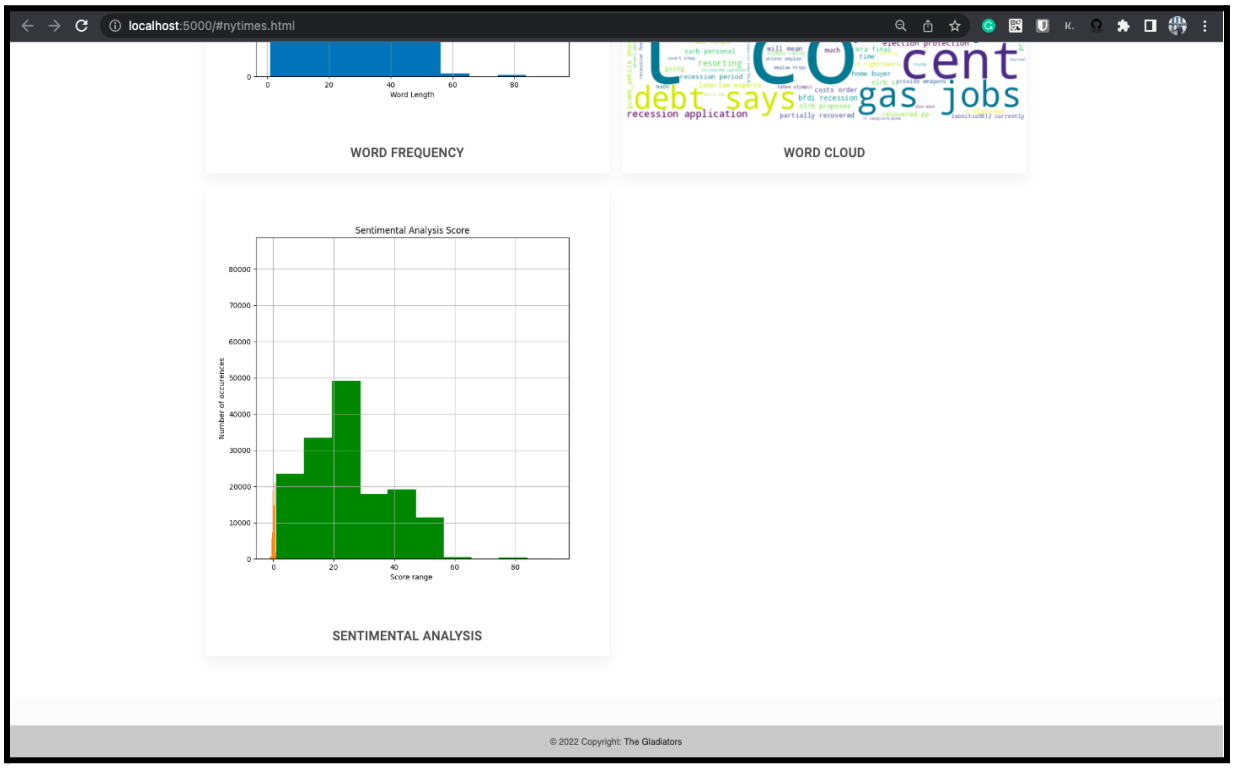

Figure 7: Sentiment Score Analysis for the NYTimes article data with the days range of 60 days

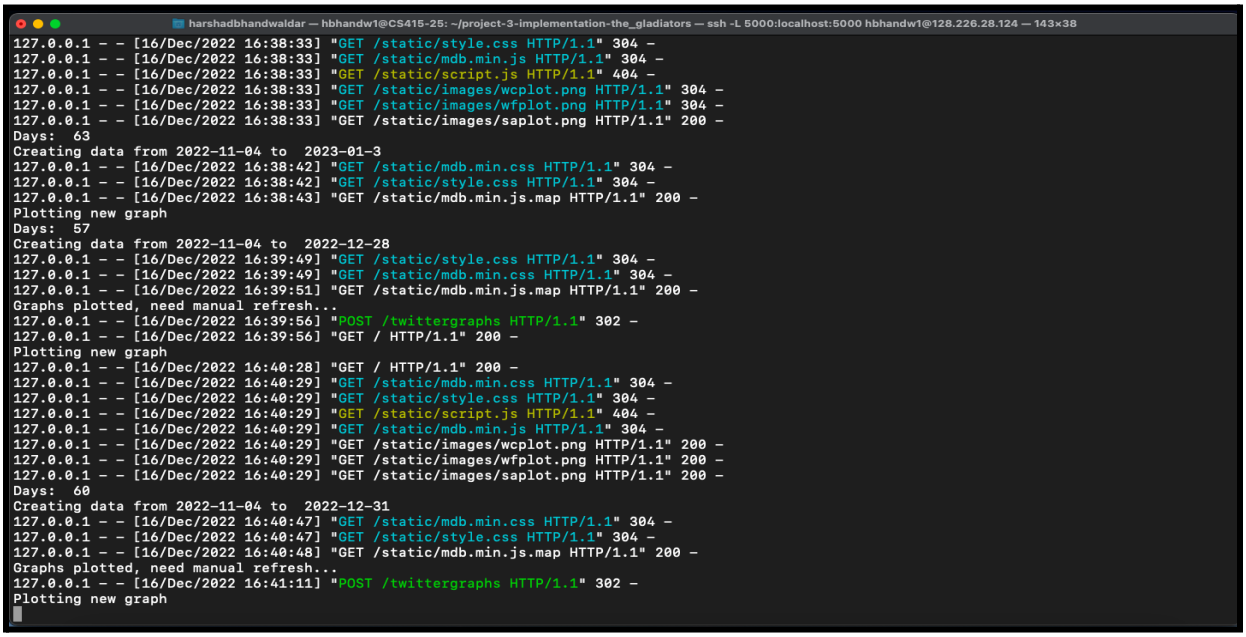

Figure 8: Log-file created while running the analysis and plotting the Reddit and NYTimes graphs accordingly.