Brandon Sung & Anna Zhang

MIS3640-01

Professor Zhi

21 October 2020

Assignment #2: Text-Mining

I. Project Overview

The project focuses on being able to use code to extract and analyze data from online resources to gauge the public's review on a film. We utilized IMDB and Twitter, as we wanted to extract both textual verbatim feedback and numerical ratings. We used basic dictionary and frequency models to capture the most frequently used words. Then, we leveraged natural language sentiment value analysis to draw insights on the positivity of each review. Lastly, we searched for a discrepancy between the two data source's audiences by comparing the sentiment analysis scores for each site on a scatter plot.

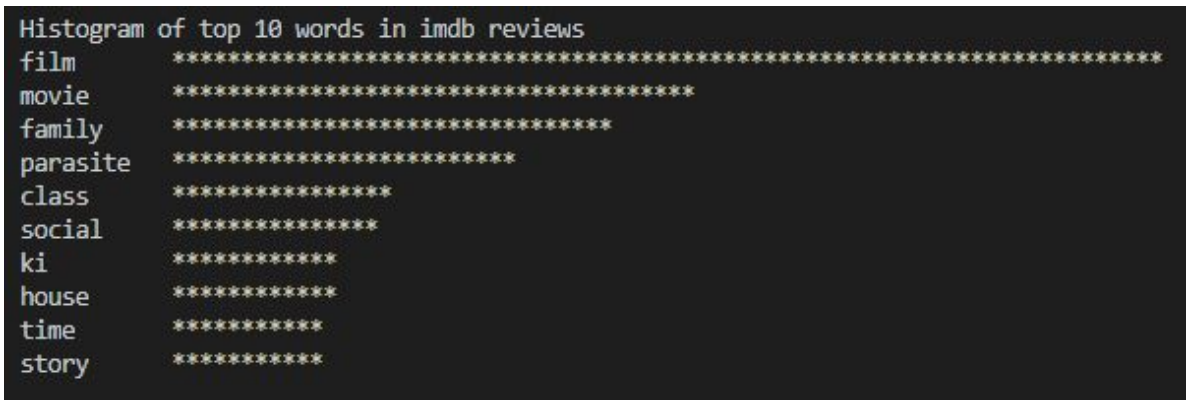
II. Implementation

To start off, we had to decide what data sources we wanted to implement and how we wanted to store them into a referrable document. We started by creating a pickle file for each data source so that we could have a referrable data source for our code that did not face API limitations. We then created two independent pickled data files for Twitter and IMDb. To analyze our IMDb source, we created a master function that analyzed the pickled file to provide us with an analysis of the top reviews. For our Twitter source, we applied the same function on the most recent tweets that include the film's title. The first portion of our master function creates a dictionary of the words in our data with a count for how many times it was iterated in the file. It then creates a histogram of the top ten words by the highest frequency and allows the user to visualize the discrepancy between the different words. The second portion of our master function is a function that runs the data file through a natural language sentiment analysis that provides us with the average sentiment compound score, as well as, a list of the individual scores for each IMDb review or tweet.

Next, we wanted to compare the two data sources to examine the differences between the two audiences. A common comparison point that captured the overall sentiment of each individual excerpt is the individual sentiment analysis compound scores. We wanted to dive down deeper into these values so we created a scatter plot that featured the positive sentiment score on the X-axis and the negative sentiment score on the Y-Axis. We decided to implement a scatter plot as it allows the user to visualize the sentiments for each individual entry as well as how it compares to the other entries from either source. We stored the sentiment values in empty lists and had it return the scores as (x, y) coordinates in our scatter plot function.

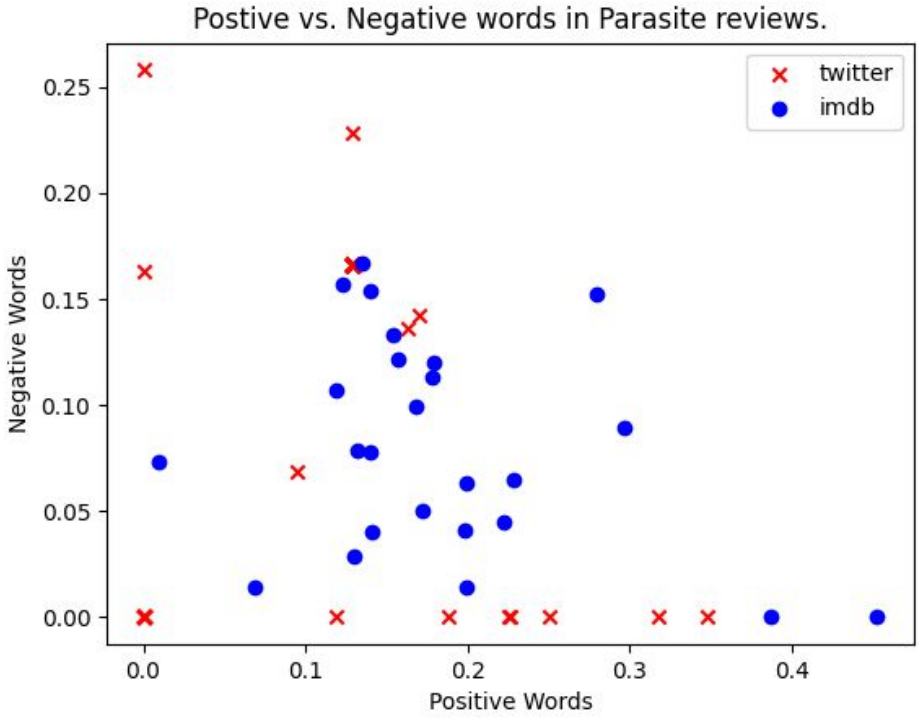
III. Results

For our text analysis, we began by taking a high-level approach of simply viewing the words with the highest frequency. Based on IMDB's most recent reviews of the film, *Parasite*, we found that within the most frequent words were "family," "social, and "class." This led us to believe that the aspects of the film that left a deep impression on the audience and their ratings, were related to family and the social class status of the characters. As we both had seen the film, we confirmed and found it interesting how the most frequently used words easily captured a major component of the film. For twitter, the most frequently used words only shared two words to IMDb's, which was "film" and "parasite." This led us to question whether the Twitter data output was collecting tweets that did not suggest the writer's rating or were not necessarily focused on the film.



For our natural language sentiment analyzer, we found that most of the reviews denoted positive connotations on IMDb. This leads us to infer that the film was generally liked by the

public as they digitally professed their appreciation for the film. The average sentiment compound score on IMDb was 0.557172. As this is positive and above an average of 0.50, we infer that on average, the IMDb audience enjoyed the film based on the connotation of their reviews. On a contrasting note, the average sentiment compound score on Twitter was -0.121689. This was interesting as we found that a large portion of the tweets denoted a neutral score of 0.0 which may have increased the sensitivity of the compounds.



Lastly, we compared the positive and negative sentiment scores for each review from both data sources. As seen above, we plotted the sentiment scores for each review and indicated the different sources by the different key icons. We see that the tweets typically denoted higher negative scores while the IMDb reviews is seen as having the two highest positive scores. Generally the IMDb results are clustered in the lower regiments on the scatter plot, while Twitter is a bit more spread out. This led us to infer that the IMDb audience enjoyed the film more than Twitter's.

IV. Reflection

Our team started the project by reviewing the different types of data sources to find areas of mutual interest, which brought us to our love of film. This sparked our decision to proceed by looking into IMDB ratings of the 2019 film, *Parasite*, which is known for exceeding the audiences' expectations and making revolutionary strides as a foreign film. A main pain point of this project was researching methods to output all of the film reviews into a referrable data source. We worked together by utilizing screen-sharing where we were able to collaborate live and brainstorm solutions to any errors. Live collaboration went extremely smoothly as we were able to consistently synthesize our knowledge and expedite the coding ideation process. We also consistently communicated during offline work sessions to ensure that our time and efforts were not redundant and we were continuously working as efficiently as possible. When we came across struggles, we decided to take a break by creating action items for the next agenda and reconvene after. We found this to be highly effective as each member took the necessary steps to find solutions to our stubborn obstacles. By using a mix of splitting tasks and working together, we found a balance that allowed us to learn from each other and strengthen our collaboration skills.