Project Overview

For this project, I used data source from IMDB Movie Reviews. I analyzed the source using four techniques: characterizing by word frequencies, computing summary statistics, and doing natural language processing using sentiment analysis and text summarization. Through this project, I hope to learn to directly obtain information from online data sources, store and clean up the data, and conduct text analysis using Python.

Implementation

I focused on getting Joker (2019)'s movie reviews form the IMDB website, so that I implemented the project first by writing the reviews into a file called "review_output.txt". Since reviews are written by viewers instead of scholars, there might be a lot punctuations, unexpected white spaces, and also capitalizations. In order to accurately check each's word's frequency, I removed all the obstacles mentioned above and stored each work into a dictionary as key and stored its number of occurrence as the value. Then, I checked for the top ten most frequency words in the movie reviews.

After conducting basic text analyzing, I did natural language processing first using sentiment analysis to figure out whether the reviewer has a negative, positive, or neutral feeling toward the movie through their words. I particularly checked the first three reviews, rated by helpfulness, to find out general public's attitude toward the movie because I think the top rated reviews are representative. Then, I did text summarization to summarize each review since many times the review is lengthy.

Throughout the implementation, when I obtained data from the IMDB website, I debated whether I should store the reviews data in a list or I should create a new file to store the data. I chose to store in a file because the size of the data is large. In addition, it is easy for me to check the data if it is in the file form.

Results

After I characterized by word frequencies and computed summary statistics, the results were not helpful as it shows that most frequent words are "the", "a", "and", "I", etc., which are connection words instead of sentimental words.

So, in order to obtain some useful information, I conducted the sentiment analysis on the first three reviews and defined a function to translate the result into useful information. The results are shown in the figure below. It shows a mixed reactions to the

Joker (2019), positive and negative. Since the movie was rated at 8.8/10, relatively high score, I wondered what the third review is about so I printed out the review in the console. As I read through it, I think the reviewer's attitude is positive instead of negative. I think the sentimental analysis showed negative is because the review consists of a lot of negative words, such as "hype", "complaints", "criminal", etc. This shows that the sentimental analysis is not always correct.

```
Overall sentiment dictionary is :  {'neg': 0.062, 'neu': 0.701, 'pos': 0.237, 'compound': 0.9913}
sentence was rated as  6.2 % Negative
sentence was rated as  70.1 % Neutral
sentence was rated as  23.7 % Positive
Sentence Overall Rated As Positive
Overall sentiment dictionary is :  {'neg': 0.095, 'neu': 0.733, 'pos': 0.172, 'compound': 0.8462}
sentence was rated as  9.5 % Negative
sentence was rated as  73.3 % Neutral
sentence was rated as  17.2 % Positive
Sentence Overall Rated As Positive
Overall sentiment dictionary is :  {'neg': 0.2, 'neu': 0.617, 'pos': 0.182, 'compound': -0.4817}
sentence was rated as  20.0 % Negative
sentence was rated as  61.7 % Neutral
sentence was rated as  18.2 % Positive
Sentence Overall Rated As Negative
```

In addition, I did text summarization on the first review because the review is long and time consuming to read, so I summarized the main argument in the review. The result is shown below. I think it is very accurate and concise.

```
Joaquin bleeds, sweats, and cries his every drop into this magnificently dedicated performance. Heath Ledger would be proud. This is undoubtedly the greatest acting performance since Heath's joker. Believe the hype. This is going to be revered as a transcending masterpiece of cinema.
```

Reflection

Overall, processing the text from online source went well in terms of obtaining the information, check word frequency, conducting natural language processing. One thing that I can improve upon is that I was only able to obtain 25 reviews instead of all the reviews for Joker (2019) because I was not able to figure out how to press the "load more" button through Python.