Part 3: Project Writeup and Reflection

1. Project Overview:

The data source that we used for this assignment is IMDb. By installing 'imdbpie', our team were able to import 'Imdb' module, from which we retrieved reviews of the movie 'Conjuring'. Our team conducted two kinds of analysis for this assignment. The techniques we used are as follows:

1. *Characterization by word frequencies*
2. *Sentiment analysis.*

The objective for this assignment was to figure out if there are certain words that appear more than the other, and to see if these words affect the overall sentiment score of the text. Overall, we wanted to see if the movie 'conjuring' is worth watching!

2. Implementation:

Our team used 'Imdb' module to retrieve reviews from the movie 'Conjuring'. The initial plan was to analyze a minimum of 100 reviews to get the clear sense of which word appears more than the other. We, however, were able to analyze only 10 reviews, which is the maximum number of reviews per page in Imdb website. Note that, 'Imdb' module is incapable of retrieving data that is on the other page.

The reviews that we retrieved were written and stored in a single file named 'review_combined.txt'. For word frequency analysis, we created a function 'process file(file name)' to remove punctuations, whitespace, and stop words that are considered irrelevant for the analysis. For the removal of stop words, we have used 'nltk.corpus', where we imported 'stopwords' module. Then, we created another function, 'most_common(hist)', to make a list of word-frequency pairs in descending order. Our team later used the list from this function to find 20 most common words that are used from the review of 'Conjuring'.

Prior to conducting sentiment analysis, our team installed ‘nltk’, a python package for natural language processing, through command prompt. For sentiment analysis, we have used ‘nltk.sentiment.vader’ to import ‘SentimentIntensityAnalyzer’ module, which provides an overall sentiment of the reviews. Note that, only 8 out of 10 reviews were used for this sentiment analysis due to the limitation on number of words that are allowed for this module.

3. Results:

We conducted two types of analysis, word frequency and sentiment analysis. First, we ran a sentiment analysis. Since our team expected the movie ‘Conjuring’ is a good movie that worth watching, we expected high compound from the sentiment analysis. Due to limitation, we were able to use only 8 reviews. The results come out as {‘Neg’: 0.164, ‘pos’: 0.14, ‘neu’: 0.696, ‘compound’: -0.9987}. Overall score of sentiment analysis was negative which means there were more of negative meaning words inside of reviews than positive words. First impression about the result is that reviews are full of negative words which possibly means that the movie ‘Conjuring’ is not a good movie.

For the word frequency, we eliminated stop words which filled the top 5 of the word frequency in the beginning. The result is as follows:

Top 20 Words		
Rank	Word	Count
1	Horror	47
2	Movie	30
3	It’s	27
4	Film	27
5	Conjuring	26
6	One	24
7	James	18
8	Wan	15
9	Story	14
10	Like	13
11	Would	11
12	Wan’s	11
13	Something	11
14	Best	11
15	Way	10
16	Seen	10
17	Films	10
18	Scary	9
19	Paranormal	9
20	much	9

•Note, that the words above are ranked in descending order

There still are useless words such as ‘it’s’ and ‘one’. However, as we were expected, the word describing the movie such as ‘horror’ and ‘movie’ were rated as top two. Also, the director of the movie, James Wan, was frequently mentioned as well.

The result of the word frequency explains the reason for negative score for sentiment analysis. For scary movies, reviews saying, “the movie is scary,” means the movie is good. Out of top 20 words, most of them are neutral words. One positive word, ‘Best’, and the rest are negative words which is complement for horror film. However, in the sentiment analysis, it only interprets every single words whether they have positive or negative meaning. We assume because of the word ‘Horror’, the sentiment analysis scored negative which did not give us the good insight for the movie as a result.

4. Reflection:

Our initial expectation from this assignment was to extract an insight from numerous reviews without having to read them all. Although our team was successful in both harvesting text from the data source and conducting appropriate analysis, we were not able to find any meaningful insight from this project. The major issue was that there was a limitation on the amount of data that we could retrieve and analyze with the modules that we used for this project. Our team later realized that web scraping could be a better fit for analyzing IMDb reviews in that it allows us to retrieve data that exist in multiple pages.

In addition, our team realized that sentiment analysis could be much more accurate and useful for shorter sentences, such as tweets from Twitter, but not suitable for IMDb reviews. Given that the IMDb reviews are much longer than the tweets which contain a sentence or two on average, we believe that the sentiment analysis is not suitable for capturing the context of multiple sentences or long documents. The final result from such analysis, which indicated that the overall sentiment for the IMDb reviews were more of negative, did not really give us a meaningful insight. Overall, the team process went smoothly throughout the project. We always paired to plan, program, and conduct analysis together.