Ie Wha (Grace) Kim

Professor Li

Problem Solving & Software Design

27 March 2020

Assignment 2: Text Analysis and Text Mining
Sentiment Analysis of Movie Reviews

Project Overview:

The objective of this project was to find sentiment of movie reviews to compare with the movie rating. Using Imdb data source, I obtained 5 different movie reviews (2 best rated movie in 2019, 1 worst rated movie in 2019, 1 best rated movie of all time, and 1 worst rated movie of all time). By using polarity scores from Sensitivity Intensity Analyzer class that was part of NLTK library, I hope to learn whether sensitivity can give more insight in to movie review data as an individual's opinion might be hard to measure to really show the accuracy through simply rating.

Implementation:

In order to analyze the reviews, I had to first choose how I wanted to obtain the reviews. Through the Imdb data source, I was able to obtain the reviews. The source provides only few of the reviews not all as the volume of the reviews would be very large. So, I decided that given reviews were good amount of volume for me to use to test for sentiment analysis. The 5 movies I chose were *Parasite (2019), Joker (2019), Unplanned (2019), The God Father (1972), and Disaster Movie (2008).* I hoped that not just comparing the best and the worst rated movies but by looking at it from different timeline would also give an interesting insight into how rating and sentiment could be affected. Using the imdb.get_title_user_reviews, I retrieved reviews where I only extracted reviews using for loop function.

Initially, I simply found polarity scores that gave positive, neutral, negative, and compound valence. However, there were limitations in measuring valence as this technique allows to measure valences for long paragraphs but more accurate for sentences. For more accurate measurements, I decided to make another method of measuring sentiment while using same functions. The new method was, instead of combining all reviews together as a string to measure the sentiment as whole, to measure sentiment for each reviews within each movies. By doing so, the polarity scores can be more accurate with fewer sentences and words. Then after taking each scores, I took the average of each sentiments. Then to compare the results, combined all the scores to analyze each polarized sentiment scores.

Results:

Although limiting, the initial method where I measured combined reviews gave some insights. First exhibit is a simple example to use the scores as a reference. As for the movies chosen, the best rated movies are *Parasite (8.6), Joker (8.5),* and, *The Godfather (9.2)* and worst rated movies are *Unplanned (5.8)* and *Disaster Movie (2.0)* (Numbers in the parentheses are the ratings on Imdb).

```
Sentiment analysis has never been good.------------------------------ {'neg': 0.325, 'neu': 0.675, 'pos': 0.0, 'compound': -0.3412}
Sentiment analysis has never been this good!------------------------- {'neg': 0.0, 'neu': 0.621, 'pos': 0.379, 'compound': 0.5672}
Most automated sentiment analysis tools are shit.-------------------- {'neg': 0.375, 'neu': 0.625, 'pos': 0.0, 'compound': -0.5574}
With VADER, sentiment analysis is the shit!-------------------------- {'neg': 0.0, 'neu': 0.583, 'pos': 0.417, 'compound': 0.6476}
Other sentiment analysis tools can be quite bad.--------------------- {'neg': 0.351, 'neu': 0.649, 'pos': 0.0, 'compound': -0.5849}
On the other hand, VADER is quite bad ass--------------------------- {'neg': 0.0, 'neu': 0.423, 'pos': 0.577, 'compound': 0.802}
VADER is such a badass!--------------------------------------------- {'neg': 0.387, 'neu': 0.613, 'pos': 0.0, 'compound': -0.2244}
Without a doubt, excellent idea.------------------------------------ {'neg': 0.422, 'neu': 0.281, 'pos': 0.297, 'compound': -0.2235}
Roger Dodger is one of the most compelling variations on this theme.- {'neg': 0.0, 'neu': 0.834, 'pos': 0.166, 'compound': 0.2944}
Roger Dodger is at least compelling as a variation on the theme.----- {'neg': 0.0, 'neu': 0.84, 'pos': 0.16, 'compound': 0.2263}
Roger Dodger is one of the least compelling variations on this theme. {'neg': 0.132, 'neu': 0.868, 'pos': 0.0, 'compound': -0.1695}
Not such a badass after all.---------------------------------------- {'neg': 0.0, 'neu': 0.735, 'pos': 0.265, 'compound': 0.1139}
Without a doubt, an excellent idea.--------------------------------- {'neg': 0.37, 'neu': 0.37, 'pos': 0.26, 'compound': -0.2235}
--------------------------------------------------
```

```
------------------------------------------------------------------------------------------------------------------------

Movie Title: Parasite | Year: 2019

The Positive Valence for this movie is 0.157
The Negative Valence for this movie is 0.118
The Neutral Valence for this movie is 0.725
The Compoud for this movie is 0.9998

------------------------------------------------------------------------------------------------------------------------



------------------------------------------------------------------------------------------------------------------------

Movie Title: Joker | Year: 2019

The Positive Valence for this movie is 0.187
The Negative Valence for this movie is 0.132
The Neutral Valence for this movie is 0.681
The Compoud for this movie is 0.9998

------------------------------------------------------------------------------------------------------------------------



------------------------------------------------------------------------------------------------------------------------

Movie Title: Unplanned | Year: 2019

The Positive Valence for this movie is 0.132
The Negative Valence for this movie is 0.138
The Neutral Valence for this movie is 0.73
The Compoud for this movie is -0.981

------------------------------------------------------------------------------------------------------------------------
```

```
----------------------------------------------------------------------------------------------------------------------------

Movie Title: The Godfather | Year: 1972

The Positive Valence for this movie is 0.175
The Negative Valence for this movie is 0.085
The Neutral Valence for this movie is 0.741
The Compoud for this movie is 1.0

----------------------------------------------------------------------------------------------------------------------------


----------------------------------------------------------------------------------------------------------------------------

Movie Title: Disaster Movie | Year: 2008

The Positive Valence for this movie is 0.136
The Negative Valence for this movie is 0.178
The Neutral Valence for this movie is 0.687
The Compoud for this movie is -0.9999

----------------------------------------------------------------------------------------------------------------------------
```

Due to the technique having to analyze large volume of words, as shown in the exhibits, the compound scores are skewed either to very positive are negative. So, using this method is not intuitive, but it does clearly distinguish between well reviewed and badly reviewed as it stayed constant with the rating given. Therefore, this method can be used to categorize good or bad movie considering the review sentiments.

Below are the new methodology where I took the average of the polarity score of each reviews. The alternative method clearly gave more of an intuitive results as shown.

```
----------------------------------------------------------------------------------------------------------------------------

Movie Title: Parasite | Year: 2019

The Positive Valence for this movie is 0.17208
The Negative Valence for this movie is 0.08835999999999998
The Neutral Valence for this movie is 0.7396400000000001
The Compoud for this movie is 0.48668400000000006

----------------------------------------------------------------------------------------------------------------------------


----------------------------------------------------------------------------------------------------------------------------

Movie Title: Joker | Year: 2019

The Positive Valence for this movie is 0.16372000000000003
The Negative Valence for this movie is 0.10547999999999998
The Neutral Valence for this movie is 0.7308399999999999
The Compoud for this movie is 0.5323999999999999

----------------------------------------------------------------------------------------------------------------------------


----------------------------------------------------------------------------------------------------------------------------

Movie Title: Unplanned | Year: 2019

The Positive Valence for this movie is 0.12179999999999998
The Negative Valence for this movie is 0.12528
The Neutral Valence for this movie is 0.7528799999999998
The Compoud for this movie is -0.09405200000000001

----------------------------------------------------------------------------------------------------------------------------
```

```
----------------------------------------------------------------------------------------------------------------------------

Movie Title: The Godfather | Year: 1972

The Positive Valence for this movie is 0.18156000000000005
The Negative Valence for this movie is 0.06436
The Neutral Valence for this movie is 0.75408
The Compoud for this movie is 0.7383159999999998

----------------------------------------------------------------------------------------------------------------------------


----------------------------------------------------------------------------------------------------------------------------

Movie Title: Disaster Movie | Year: 2008

The Positive Valence for this movie is 0.11844000000000002
The Negative Valence for this movie is 0.15592000000000003
The Neutral Valence for this movie is 0.7256
The Compoud for this movie is -0.26186000000000004

----------------------------------------------------------------------------------------------------------------------------
```

Especially comparing the data with the examples mentioned above, the new scores seem to be more accurate where the compound scores are versatile by each movie. Overall, the scores seem to match the rating, but we can see that between *Parasite* and *Joker* where *Parasite* had higher rating by 0.1, while *Joker* actually had a better compound score. However, when observing positive, negative, and neutral valence, *Parasite* actually had higher positive sentiment on the reviews. Ultimately, rating and reviews seem to be parallel in portraying the audience's satisfaction of the movie, but through the sentiment analysis, the scores give more of an intuition through the proportion between the positive and negative views on each movie.

Reflection:

In process of generating this analysis, I was able to evaluate my original method and create better alternative for the objective that I set out. Through multiple trial and error as well as numerous researches, I was also able to learn through this experience by mining texts from internet and using python to analyze the text. Especially, as texts tends to be unstructured data, generating it into a quantitative data to analyze the sentiment was a new learning experience. If this project can be extended to further the analysis, I would develop on other ways to analyze the movie reviews to evaluate feature vs. aspect based identification to distinguish what specific aspect of movies audiences particularly like and dislike. By distinguishing those categorize, then possibly movie directors can easily pin point on what to emphasize and what to look out for.