Text Mining and Text Analysis Report

Project Overview

This project explores how computational techniques can be applied to analyze text and generate relevant information from it. The project is divided into two sections based on the data sources: Wikipedia and IMDB Reviews. The main objective of the Wikipedia analysis is to explore the idea of text similarity between two Wiki articles (based on the user's choice) and uncover the basic structure of the text, such as common words and word frequencies. On the other hand, the IMDB analysis will examine the relationship between the IMDB ratings and the score from sentiment analysis of user reviews, while also unpacking the basic structure of the text as well.

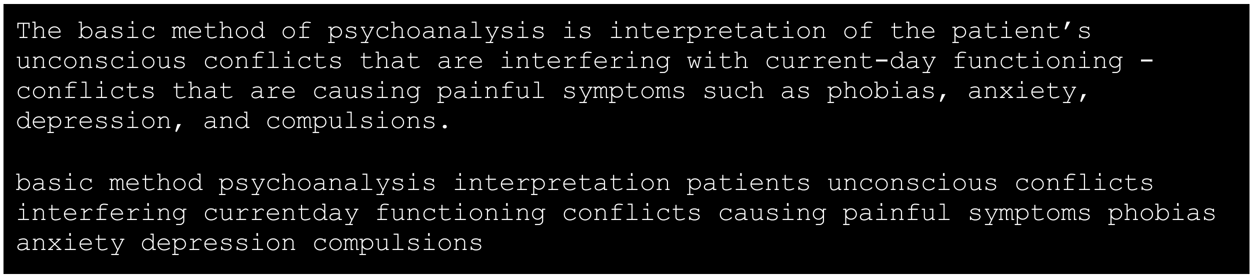
Implementation and Result

Wikipedia

The Wikipedia analysis consists of the fundamental text cleaning techniques, text analytics, and basic visualization. The program contains a function that allows access to the article from the Wiki webpage, which allows the user to grab any content from the website. For this assignment, two Wikipedia articles are being analyzed: "Psychoanalysis" and "Unconscious Mind."

- **Data Cleaning:**
    - Punctuation and stop words are not always useful in interpreting the meaning of texts, so they are removed prior to the text analysis process to generate only the relevant content of the text. This program contains functions that remove punctuation, convert text to lowercase, and remove stop words. By removing punctuation and converting words to lowercase, the program will recognize identical words, which will be useful later on. Additionally, stop words are considered as noise in the text. They may not carry any valuable information on their own. Some examples of stop words are articles ("the", "a"), conjunctions ("and"), and proposition ("with"). Punctuations and stop words are important to create comprehensive and completed sentences, but they may not be as relevant when it comes to interpreting the content of the text. Figure 1 displays the comparison of text prior to and after the cleaning process.

Figure 1: The comparison of text before and after the cleaning process



```
The basic method of psychoanalysis is interpretation of the patient's
unconscious conflicts that are interfering with current-day functioning -
conflicts that are causing painful symptoms such as phobias, anxiety,
depression, and compulsions.

basic method psychoanalysis interpretation patients unconscious conflicts
interfering currentday functioning conflicts causing painful symptoms phobias
anxiety depression compulsions
```

- **Text Analysis:**

  - **Word Frequency and Total Words:** One of the most common approaches to analyzing text is to count the number of times particular words appear in the text. This program contains a function that breaks down words in the sentence and counts their frequencies in the text. In addition, there is a function that counts the total number of unique words in the text. Figure 2 shows a dictionary of word frequency from the Wiki Article “Psychoanalysis.”

Figure 2: Dictionary of word frequency from the article “Psychoanalysis”

```
{'psychoanalysis': 103, 'greek': 1, 'ψυχή': 1, 'psykhḗ': 1, 'soul': 1,
'análysis': 1, 'investigate': 1, 'set': 6, 'theories': 19, 'therapeutic': 5,
'techniques': 17, 'related': 4, 'study': 12, 'unconscious': 41, 'mind': 17,
'together': 3, 'form': 5, 'method': 5, 'treatment': 29}
```

  - **Most Common Words:** Finding the most common words can also be relevant to analyzing the text as it allows the analyst to extract the main concept of the text. In this assignment, the article about psychoanalysis is being analyzed and it appears that the most commonly used words are unconscious, freud, and psychoanalytic, which all are related to the main concept of psychoanalysis. For instance, Sigmund Freud is the person who came up with the psychoanalysis theory, in which his name appeared several times in the article and the text analysis function was able to extract this information, without a reader having to read through the entire article. Figure 3 displays the top 20 most common words in the Wiki Article “Psychoanalysis.”

Figure 3: The top 20 most common words in the Wiki Article “Psychoanalysis.”

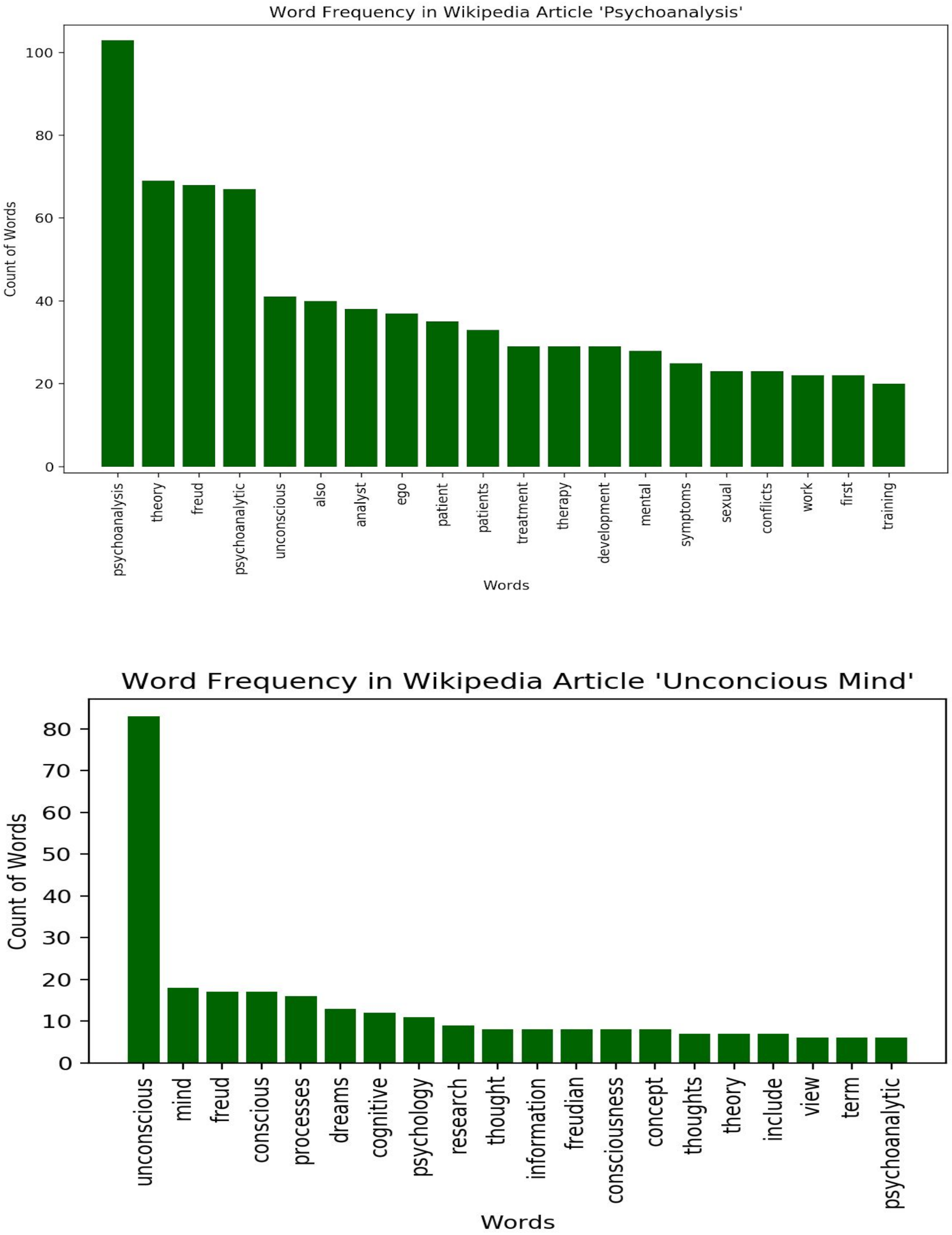
```
The most common words are:

psychoanalysis   103
theory             69
freud              68
psychoanalytic    67
unconscious        41
also               40
analyst            38
ego                37
```

  - **Similar Text Between Two Articles:** It can also be useful to compare the text similarities between two related articles. For instance, in this report, the article about psychoanalysis and unconscious mind are being analyzed. The result shows that the two articles are closely related in ideas since there are a total of 415 words in common and some of the common words include psychoanalysts, psychiatrist, freudian, phobias, and unconscious. This technique can be useful when trying to compare the main idea between two or more texts.

- **Bar Graph - Word Frequency:** Graphs can be a useful technique to present information in an efficient way, as it allows the analyst to capture the main point of the data and communicate it through visualization. For this assignment, a bar graph is used to visualize the top 20 most common words in each article. Figure 4 illustrates the bar graph of word frequency from both articles.

Figure 4: The bar graph of word frequency from both articles



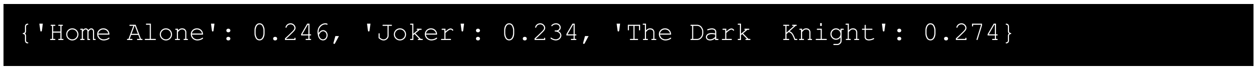
\IMDB Review

The main objective of the IMDB Reviews analysis is to examine whether there is a relationship between the users' reviews and the overall rating of the film. There is much information that can be extracted through natural language processing, however, the sentiment analysis is the most fitting for the purpose of this analysis. The sentiment analysis is implemented to compute the positive opinion from the users' reviews. The final result of this analysis will be a scatter plot of IMDB ratings and positive sentiment score from users' reviews. For the matter of efficiency, this report only contains the analysis of 20 movies.

- **Data Importation & Data Cleaning:**
  For this analysis, a function was built to take in user's input of any movie that a user would like to analyze. The function is connected to an API which allows it to import as many movie reviews and ratings from the IMDB website as the user wants. Similar to the data cleaning process in the Wikipedia analysis, punctuation and stop words are removed from the reviews to avoid having meaningless text in the analysis.

- **Text Analysis:**
  - **Sentiment Analysis:** Sentiment Analysis is a common Natural Language Processing (NLP) that tries to extract opinions within a given text across reviews and identify the emotion. The sentiment score consists of 3 main classifications, including "negative," "neutral," or "positive." All these three classifications add up to a total of 1, which means that the threshold for each classification is around 0.33. For instance, if the negative sentiment score of a movie is higher than 0.33, that means that the emotion of overall reviews for that particular movie is negative. However, the threshold is completely flexible and can be adjusted based on the dataset. Figure 5 displays the output of positive sentiment analysis from sample movies.

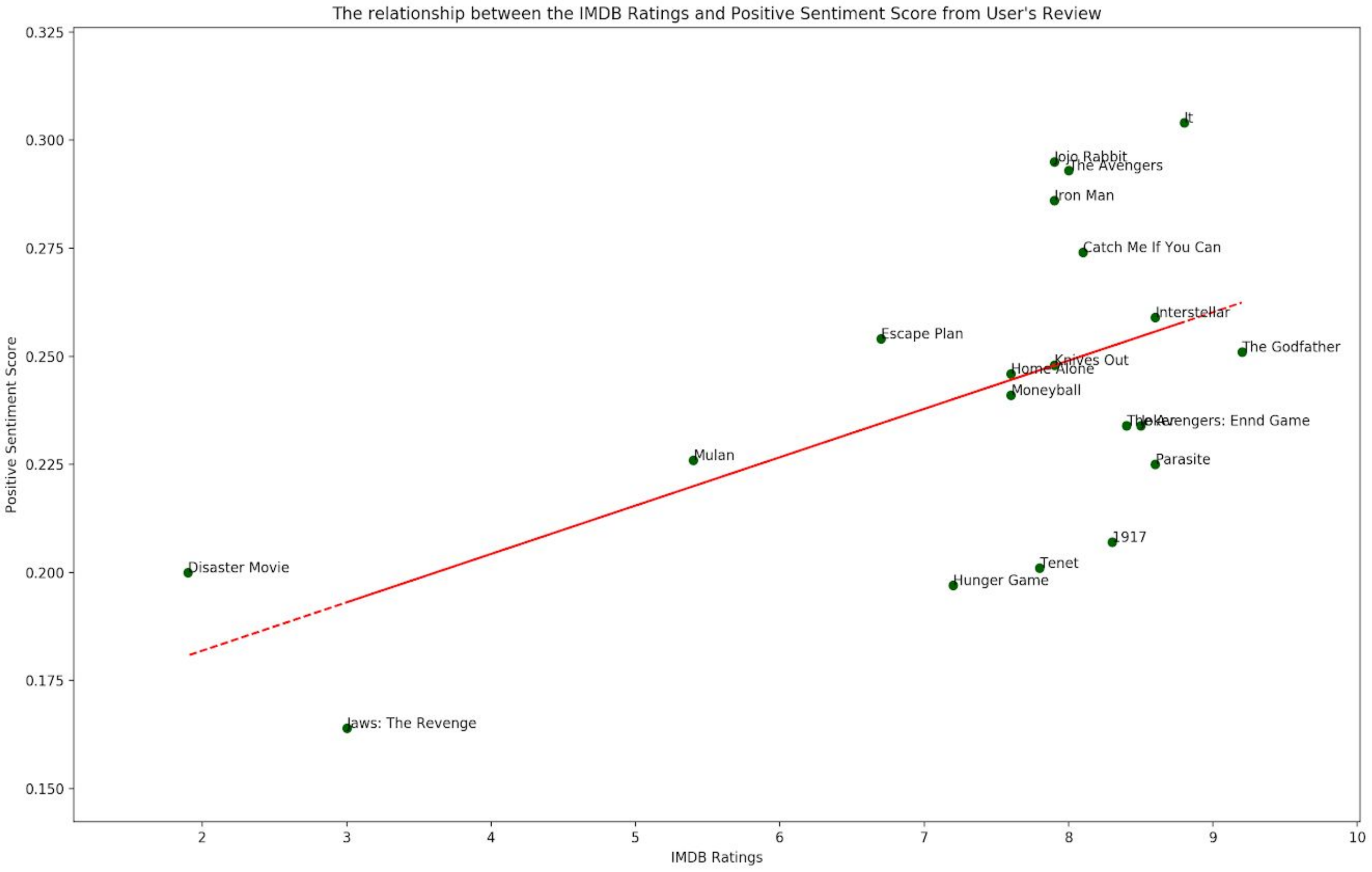
Figure 5: The output of positive sentiment analysis from sample movies.



```
{'Home Alone': 0.246, 'Joker': 0.234, 'The Dark  Knight': 0.274}
```

  - **Scatter Plot Between Overall IMDB Rating and Positive Sentiment Score:** Scatter plot can be used to examine the relationship between the ratings and the sentiment score. Based on the result, there are some movies like Tenet, Hunger Game, and 1917 which have a relatively high rating, but low positive sentiment score. Additionally, the trendline in Figure 6 suggests that there may be a weak relationship between the ratings and the sentiment score. However, it is important to note that this report is limited by the amount of data, therefore, the result may not be accurate.

Figure 6: The scatter plot between the ratings and the positive sentiment score



Third Party Library and Resources

Python has a large standard library with many useful tools, however, it also has a large collection of third-party modules that are relevant to build a functional program. To successfully employ all of the functions in this analysis, the following package are required:

- Mediawiki : Allows access and curates content from Wikipedia article.
- String : Contains a number of functions to process standard Python strings. In this analysis, string.punctuation is used to get all sets of punctuation.
- Nltk : Consists of tools that allows us to work with human language data. In this analysis, nltk.sentiment.vader is used to conduct a sentiment analysis on users' reviews.
- Pickle : This module converts a Python object into byte stream to store it in a file/database. In this analysis, pickle was used to export a dictionary of word frequency and stored it as a file.
- Matplotlib.pyplot : This module creates interactive plots.

- Numpy : A Python library used to work with arrays. In this analysis, numpy was used to create a best fit trendline for the graph.
- Imdbpie : Allows access and curates reviews from IMDB website.

Additional open resources that were used to complete this project include online documentation such as StackOverflow and online tutorials on text analysis.

Reflection

Text analysis is such a powerful tool as it creates structured data out of free text content. I would say that I find all of the skills and techniques in this assignment, such as data extraction and content scraping to be relevant to my future career. However, since there are many techniques that can be used when analyzing the text, choosing the right tool for the data can be difficult. In the beginning, I found it quite challenging to choose the right technique for my chosen data sources. However, the more I studied pieces of information contained in my text, the easier it was for me to select the scope of my project. Overall, I was very satisfied with my results, but I wish I had more background knowledge on text analysis prior to the start of the project.