Kimberly Park, Jasmine Gong

Project Reflection

Project Overview

This project aims to analyze texts from Carl Jung's and Freud's psychology papers to find similarities in two different arguments regarding the unconscious mind. The team used the Gutenberg Project to extract two books: Freud's "Dream Psychology" and Jung's "Collected Papers on Psychology." These two sources are used in the text analysis for frequency and finding a total of meaningful words. Additionally, the team extruded data from the Wikipedia article page of "Psychology" to conduct a similarity analysis with the cosine approach among data sources. The team hopes to learn the data scientist approach of analyzing texts from an online source and interpreting the result to create insights. Furthermore, the team hopes to create visual representatives of the data that are easy to comprehend.

Implementation

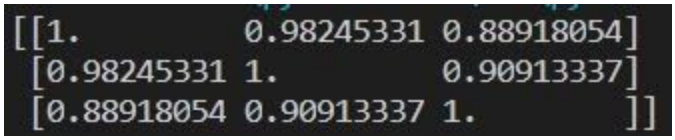
The project aimed to analyze the text of two books retrieved from the Gutenberg Project. We first downloaded the data from Gutenberg and saved it as a local file for a convenient location. Moving forward, we needed to process the data to gain the necessary information needed for our text analysis. We processed the raw data by filtering the words into two categories: meaningful words and meaningless words. Insignificant words were identified based on the word list “stopwords.txt” found under https://github.com/MIS3640/resources/blob/master/code/data/stopwords.txt. After completing categorizing the words, we computed the total number of words included in each book and the total number of non-repetitive words.

Next, we analyzed word frequency for each book and created a word cloud visual to gather insights in a convenient manner. Then we analyzed the similarities of the papers and the “Psychology” Wikipedia page through the cosine method. To gain a universal perspective on Psychology, we purposefully added a Wikipedia page to compare the influence of the two authors in the field of psychology. We deliberately chose to apply the cosine method of testing similarities between texts because we wanted to have consistent control variables to omit the varying length of each text.

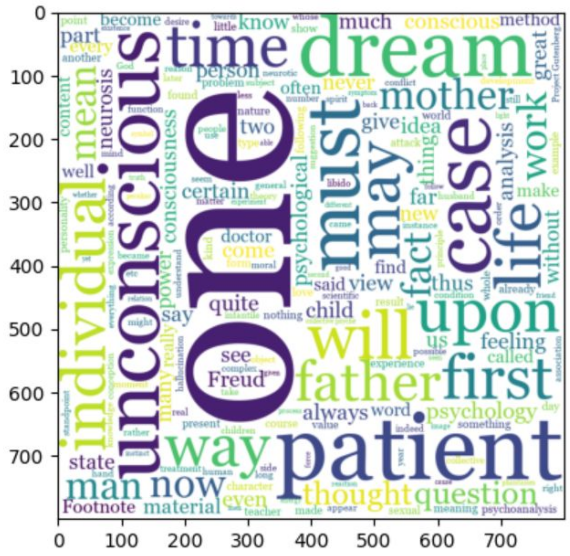
Results

For the first part of the project, we found Freud used 27,555 words in total, 6,413 words among which are unique. Carl Jung used 92,341 words in total and 15,311 unique words. After, we used word clouds to visualize the word frequency of each book. Based on the word cloud, the most frequently used word from Freud’s text was “dream” while it was “one” from Jung’s text.

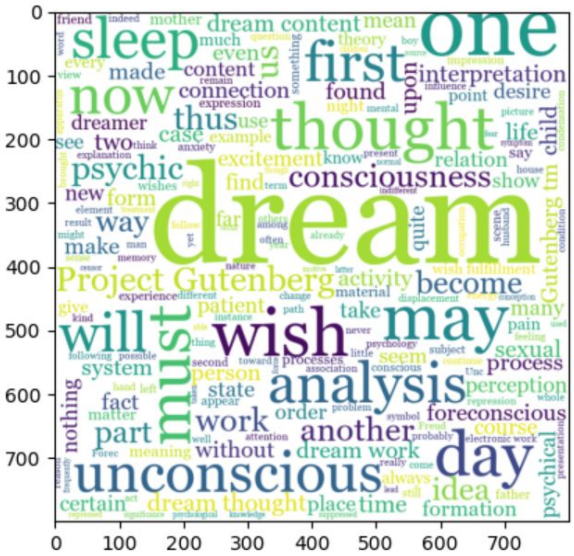
However, we could not get an objective perception of text similarity between Freud and Jung because the interpretation of the result is biased by the total volume of the book(the total number of words used in each work). Thus, we chose to use cosine similarity which compared text similarity on a percentage base. We



also chose to include extracted data from Wikipedia “Psychology” in the cosine similarity comparison, since both of them belonged to this scientific subject. From the matrix, we could tell there is a 98% similarity between Freud’s and Jung’s work and 89% similarity between Freud’s book and words used on the Wikipedia page. On the other hand, Carl Jung’s paper is 91% similar to the Wikipedia Psychology page.



Carl Jung’s “Collected Papers on Psychology”



Freud’s “Dream Psychology”

Reflection

From the process point of view, the team was effective in the preparation of the project by starting early and understanding the agenda for each meeting beforehand. The team had daily meetings via Webex and mobile communications to work on the project and employed a hybrid of delegating tasks and pair programming together. We took turn coding each step of processing data to analysis and paired program to fix errors. The team process was flexible since we pivoted from Twitter sentiment analysis to book text analysis in finding similarity between two different arguments by Jung and Freud regarding psychological consciousness. There were no issues that arose while working together and if there were a miscommunication, we made sure to explain to each other clearly and communicate more often. With more time, we could have improved on several things. For instance, when creating artistic visuals with word clouds, we could have generated a more visually appealing word cloud using jpg stencils. Also, we could have improved by writing codes in a concise manner for efficiency if we were on an advanced level. Going forward, we have learned a data analytics tool (text analysis) that we can use on the web scraping for other classes and in a corporate setting.