CIS668- Natural Language Processing
Comparing Corpora with Corpus Statistics

Siri Chandana Sambatur
829955672

Overview

Gutenberg has over 54,000 free e-books. Among all these books available, the chosen two books belong to the genre ‘Psychology’. The books that have been chosen as corpora are-

1) Dream Psychology: Psychoanalysis For Beginners By Prof. Dr. Sigmund Freud

2) Ten Thousand Dreams Interpreted, Or, What’s in a dream. A scientific and practical exposition by Gustavus Hindman Miller

These two books that have been chosen are written about the same topic which is psychology behind dreams. Both these authors have a different perspective in this topic and hence, they have written their books using different approach where one discusses in depth about the reasoning and the other just interprets about what dreams mean.

During the comparison, the amount of difference that exists between these two books is to be found. Even though the topic is the same, the vocabulary of both these authors varies and I would like to find this difference in vocabulary. Also, I would like to know depending on the most common words they use if the genre of the book could be interpreted.

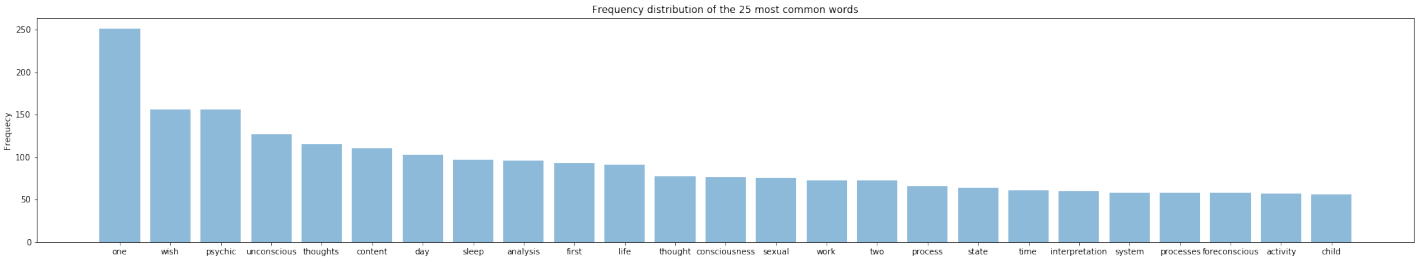
The procedure that has been used for analyzing the corpora is given below-

- Tokenized the text file that is extracting all the tokens using ntlk.word_tokenize().
- Converted all the tokens to a lowercase as the case of the words doesn't effect our comparison.
- Removed all the stop words from the tokens extracted by using the stop words present in the nltk corpus (nltk.corpus.stopwords.words(‘english’)).
- Removed all the non-alphabetical words from the modified list of tokens. For this I used regular expression to define if the word is non alphabetical.
- Created additional stop words removal based on the text files. These stop words are stored in the “stopwords.txt” file. Stop words like ‘dream’ and its variations have been included because this word is bound to occur more than other words as this is the basis of the topic. Other, four letter words and conjunctions which have been used frequently have been removed.
- Then listed the top 50 words by frequency and plotted a frequency distribution for the top 25 words.
- Calculated the bigrams and listed the top 50 bigrams based on the frequency.
- Also, listed the top 50 bigrams based on the pointwise mutual information.
- Then found the trigrams for the list of tokens for each corpora.
- Then listed the raw frequency and pointwise mutual information for the trigrams. For both bigrams and trigrams the frequency determines the occurrences of these words together whereas point wise mutual information tells about the association of these words. Both these measures are important and led to different conclusions.
- Finally, compared the information that was obtained for both the files.

Analysis of Dream Psychology: Psychoanalysis For Beginners By Prof. Dr. Sigmund Freud

The size of this text file is 319 KB and the number of tokens present in this file is 60324. Out of these tokens that where extracted the following is the list of preprocessing steps that is performed and the results that were obtained-

1) converted these tokens to lowercase.
2) removed stop words by comparing it to the nltk corpus stop words list. This reduced the number of tokens to 32076 (53.17% of the tokens remain).
3) removed the non-alphabetical tokens from the remaining ones. From this, the number of tokens obtained is 25617 reducing it further by removing 6,459 tokens (42.46% of the tokens remain from the total number of tokens).
4) additionally removed stop words from the stopwords.txt that has been created which is suitable for this corpora. This resulted in 22982 tokens and removed 2,635 tokens (38.09% of the tokens remain from the total number of tokens).
5) Then the most common words are found from this final extracted list of tokens along with its frequency and stored it in "DPMostCommonWords.txt" . Also, plotted the frequency distribution of the top 25 most common words as follows-



In this list of words, the words unconscious, thoughts, foreconscious, state, consciousness show that the author is discussing about dreams. Analysis, interpretation, psychic and wish are some of the words that indicate that the dreams are being analyzed or understood which indirectly show the psychological aspect of this book. There are no direct psychological terms in this list which indicate this in the book. This is mostly because the author has not used many of these terms and has tired explain this concept using simple words that can be understood by common people.

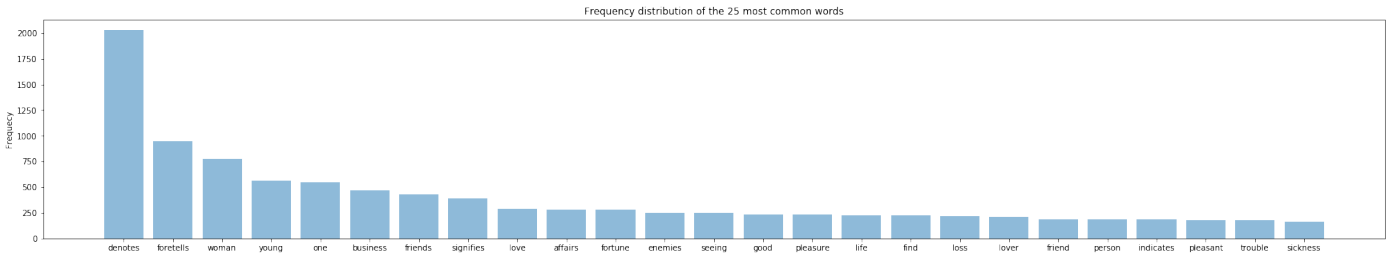
6) Then we computed the different bigrams from the extracted tokens list and calculated the raw frequency for each of these bigrams. The top 50 bigrams with high frequency are stored in "DPBigramsRawFrequency.txt". The mean value of raw frequency scores is 0.0003409.
7) Also, for the bigrams the mutual information score is calculated and the top 50 bigrams with high mutual information score are stored in "DPBigramsMutualInformation.txt". The mean value of mutual information scores is 6.60091.
8) Trigrams list from the extracted list of tokens is computed and stored in the "DPTrigramsList.txt".

9) Then, the trigrams were measured based on mutual information and raw frequency and the top 50 are stored in "DPTrigramsMutualInformation.txt" and "DPtrigramsRawFrequency.txt" respectively.
10) The mean value of trigram mutual information is 19.41665 and the mean value of the trigram raw frequency is 4.3754123107315034e-05.

Analysis of Ten Thousand Dreams Interpreted, Or, What's in a dream. A scientific and practical exposition by Gustavus Hindman Miller

The size of this text file is 901 KB and the number of tokens present in this file is 180688. Out of these tokens that where extracted the following is the list of preprocessing steps that is performed and the results that were obtained-

1) converted these tokens to lowercase.
2) removed stop words by comparing it to the nltk corpus stop words list. This reduced the number of tokens to 110224 (61% of the tokens remain).
3) removed the non-alphabetical tokens from the remaining ones. From this, the number of tokens obtained is 74761 reducing it further by removing 35,463 tokens (41.37% of the tokens remain from the total number of tokens).
4) additionally removed stop words from the stopwords.txt that has been created which is suitable for this corpora. This resulted in 67009 tokens and removed 7,752 tokens (37.08% of the tokens remain from the total number of tokens).
5) After that the most common words are found from this final extracted list of tokens along with its frequency and stored it in "DPMostCommonWords.txt" . Also, plotted the frequency distribution of the top 25 most common words as follows-



In this list of words, there are many words that indicate interpretations like denotes, signifies, foretells, indicates and in fact these words are synonyms of one other. So, this clearly shows that something is being understood or explained in this book by the author. Other words like loss, trouble, young, life, sickness show that these are the interpretations that are being made more frequently in this book. However, there are not many words in this top list that show that dreams are the ones that are being interpreted. This is mainly because we have removed the word 'dream' and its variations from the list of tokens by adding it to the stop words text file.

6) Then we computed the different bigrams from the extracted tokens list and calculated the raw frequency for each of these bigrams. The top 50 bigrams with high frequency are stored in "TTDBigramsRawFrequency.txt". The mean value of raw frequency scores is 0.000135.

7) Also, for the bigrams the mutual information score is calculated and the top 50 bigrams with high mutual information score are stored in "TTDBigramsMutualInformation.txt". The mean value of mutual information scores is 4.10657.
8) Trigrams list from the extracted list of tokens is computed and stored in the "TTDTrigramsList.txt".
9) After this the trigrams were measured based on the raw frequency and the mutual information and the top 50 were stored in "TTDTrigramsRawFrequency.txt" and "TTDTrigramsMutualInformation.txt" respectively.
10) The mean value of the trigram raw frequency is 4.3754123107315034e-05 and the mean value of the trigram mutual information is 19.4166.

Comparison between the two text files

After we performed the preprocessing steps for the 'dream psychology' text file, there is 38.09% of the tokens and for the 'ten thousand interpretations' text file, there is 37.08% tokens which will be used for our analysis. Even though both these documents were not of the same size and the ''ten thousand interpretations' text file is almost thrice as large as the 'dream psychology' text file, the percentage of useful tokens that are to be analyzed is proportional. The mean values of the raw frequency for trigrams has seem to drop for both the documents when compared to the bigrams and the point wise mutual information has increased for trigrams when compared to bigrams for both the documents. This clearly shows that there is more association between the words that are being used in the documents. After analyzing these two files the following is the answer that has been found for the questions that were to be addressed-

- How unique are the two text files written by different authors on the same topic "Dream Psychology"?

From this analysis the common words from the list of top 50 most common words used in both the text files are {new, one, life, work}. This shows that 92% of the most common words in both these files are unique. After analyzing these most common words, it can be said that the vocabulary or style of writing of both these authors is unique. The mean values of the raw frequency scores and the mutual information scores for the bigrams is also less for the document written by Gustavus Hindman Miller when compared to the document written by Dr. Sigmund Freud. This shows that variations in the bigrams are found to be more in the second document (ten thousand interpretations) when compared to the first document (dream psychology).

- Can the genre of the books be interpreted by the words that were being used by authors in their books?

In the first book (dream psychology) written by Dr. Sigmund Freud, the words used indicate that he was talking about dreams and also indirectly talking about psychology. Even though the proportion of psychology related terms was less when compared to use of general terms, the occurrence of these words as a whole lot definitely helped us in interpreting the genre of this book. The words which are selected as the most common were compared to the vocabulary of the

dream psychology flash cards(quizlet/glossary given in reference) and it was found that 30 words out of the 50 words belonged to this vocabulary and 60 words out of 100 words belonged to this vocabulary. This shows that the books genre is indicated with a measure of 60%.

In the second book (ten thousand interpretations) written by Gustavus Hindman Miller, the words that have been found to be most common indicated the concept behind the book which is that dreams are being interpreted. However, the genre of this book is 'psychology' and there are not many words in this list of words which indicate this. Out of the 50 most common words only 5 indicated this and out of 100 only 25 words could indicate this. So 25% of the book indicates that the genre is psychology. The main reason behind this could be that the author doesn't use many psychological terms when compared to the general terms. This happens because throughout his book he is trying to explain or interpret a dream rather than explain about dream psychology itself.

Conclusion

Even though both these books are written in the same domain, each author has a different style of writing it and this result in the 92% of uniqueness among these two books. Also, because of this variation in the style and conceptual approach of writing, the book written by Dr. Sigmund Freud could be identified as a psychological book with 60% accuracy and on the other hand the accuracy of identifying the book written by Gustavus Hindman Miller is just 25%.

References-

http://www.gutenberg.org/wiki/Romantic_Fiction_(Bookshelf)
http://www.gutenberg.org/wiki/Psychology_(Bookshelf)
https://quizlet.com/37439287/psychology-chapter-3-sleep-and-dreams-vocabulary-flash-cards/
http://www.apa.org/research/action/glossary.aspx