sb2084@essex.ac.uk

SENTIMENT ANALYSIS OF THE WIZARD OF OZ AND JANE EYRE

INTRODUCTION

Sentiment analysis usually requires taking text, which is a sentence, a comment, or a document, and assigning it a sentiment score that indicates how positive or negative the text is. The rise of social media platforms such as forum discussions, blogs, Twitter, and social networks has increased the importance of sentiment analysis. This technique is useful in analyzing feedbacks, suggestions and summarizing text data. Understanding emotions is one of the important aspects of sentiment analysis. To conduct sentiment analysis, initially, subjective and objective text is distinguished. The emotions are only preserved in subjective text and objective text is about facts, therefore objective data is not required for sentiment analysis.

The given problem statement requires the sentiment analysis of two books, one book from child list and the other from adult book list. From the child book list, the wonderful wizard of Oz and Jane Eyre- an Autobiography from the adult list have been selected for the analysis. The main aim of this probe is to explore the sentiment scores of both the books and compare the results.

METHODOLOGY

Based on the emotions expressed, text can be divided into three groups. Positive, Negative and neutral. In text classification, each sentence is evaluated independently and categorized as negative, positive, or neutral depending on the sentiments of each word in the sentence.

The sentiment analysis on child book data is done in R as follows:

- The essential libraries like dplyr, stringr, ggplot2 are loaded into the R file.
- The data is read into R using read.csv () function. Empty strings in the data are replaced with NA and are removed to simplify the data.
- When the structure of the data is viewed, there are 2 columns – Gutenberg id and text. Id is a numeric column. Numeric data is not required, so only text column is used in this analysis.
- A corpus is built on the text data. Corpus refers to a collection of documents.
- The data cleaning is done for simplification like converting entire text into lower case since R is case sensitive.
- Punctuation marks are removed and replaced with empty strings. Stop words (common unnecessary words) are removed and white spaces are eliminated from the data.
- Some user defined stop word which don’t add any sentiments to the data are also deleted.
- The cleaned data is analyzed to check the number of times each word appeared. This can be achieved using the function TermDocumentMatrix() from text mining package.
- A barplot is used to analyze the most frequent words in the data.
- A word cloud is generated using the important words to visualize the data. The size of each word shows how many times it appears in the text.
- Sentiment scores for each sentence are obtained using the Syuzhet package which extracts sentiments using sentiment dictionaries like afinn, nrc , bing.

- To understand the emotions and sentiments of entire data, get_nrc_sentiments function is used.

The sentiment analysis of the book Jane Eyre-An Autobiography Is done as follows:

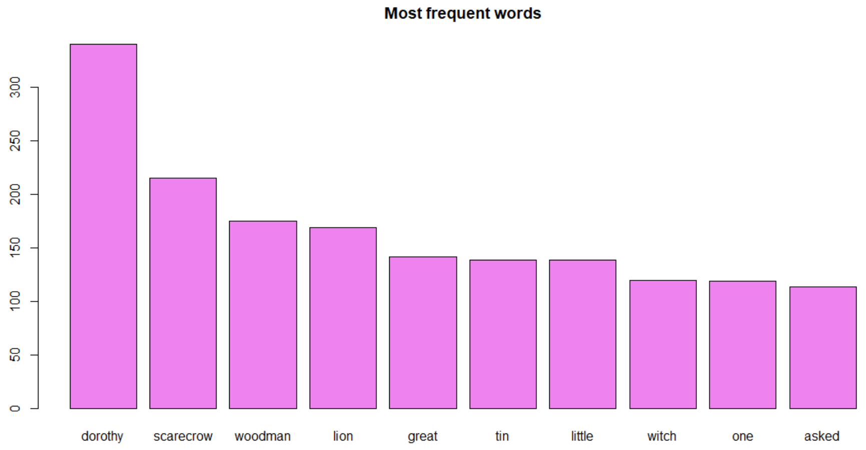
- The sentiments of word from dictionaries afinn, bing are obtained.
- The data from the book Jane Eyre is read into R and empty lines are omitted from the text and numeric column (Gutenberg id) is removed from the analysis
- Each word in the text is separated into individual tokens using unnest_tokens() function from the tidy text package. The function automatically converts the tokens into lowercase.
- A list of stopwords is loaded and are removed using anti_join(stop_words) function of tidy text package.
- The tokens are sorted and frequency of each word is calculated to know about the frequently occurring words.
- A word cloud is formed using significant words in the data using wordcloud() function.
- The positive and negative words are differentiated in the cloud using the cloud.comparison() function.
- The positive and negative words are distinguished and the bing dictionary values are assigned to the words.
- The total number of positive words and negative words is obtained.
- To understand the emotions apart from positive and negative words, get_nrc_sentiments function is used.

RESULTS

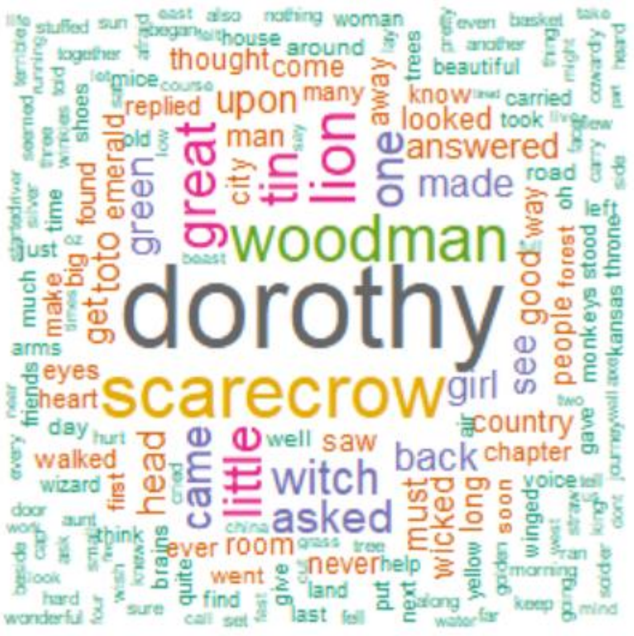
On performing Sentiment analysis on both the books the following results are obtained.

1)The Wonderful Wizard of Oz

The image below describes the most frequent words of The wonderful wizard of Oz. Dorothy is the most frequent word occurred. It represents the name of a girl in the book.

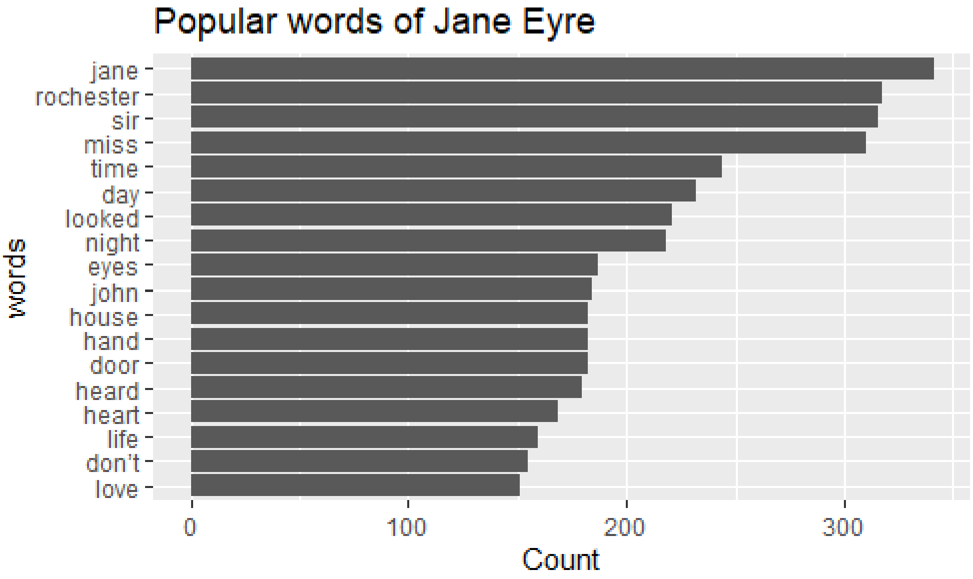


The word cloud below gives information about the key words in the book. The size of the words determines the frequency of the words. Bigger sized words appear more frequently than smaller sized ones.

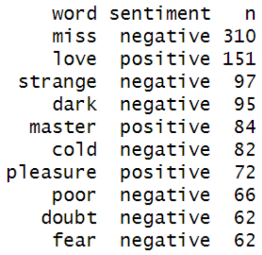


2)Jane Eyre- An Autobiography

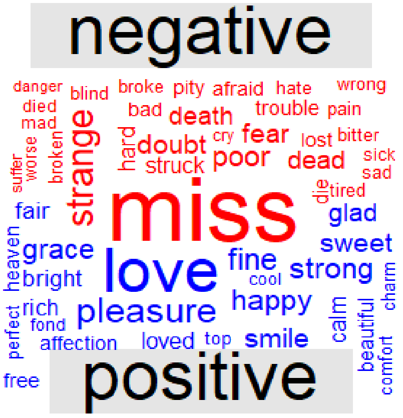
The image below describes the most frequent words of Jane Eyre. Jane is the highest occurring word which is the main character of the book.



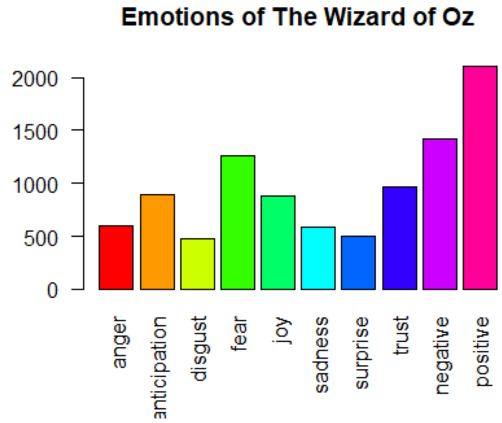
The below image gives the positive and negative words in the form of a word cloud. The sentiment of a word and the count of the word is provided.

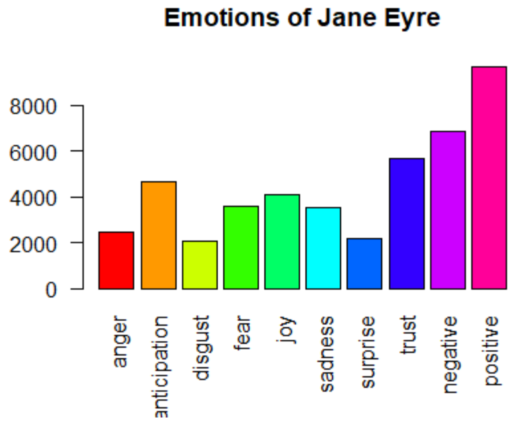


```
     word sentiment   n
     miss  negative 310
     love  positive 151
  strange  negative  97
     dark  negative  95
   master  positive  84
     cold  negative  82
 pleasure  positive  72
     poor  negative  66
    doubt  negative  62
     fear  negative  62
```



Comparing the emotions for both the books





It can be observed that in both the books, there is more of positive sentiment when compared to negative sentiment. There is less fear in the book Jane Eyre when compared to Wizard of oz which is a children's book. On the contrary, there is more trust in Jane Eyre than Wizard of Oz.

DISCUSSION

This analysis demonstrates text cleaning, transformation using different packages and creating a word cloud which is used to find popular trends or words in the data. The nrc dictionary was used to generate sentiment scores which is helpful in assigning values to the words and building an emotion classification. Sometimes, sentiment analysis can be challenging due to factors like negation words and words with ambiguous meanings affect the sentiment scores.