Dushyanth Reddy B

DB20554@essex.ac.uk

SENTIMENT ANALYSIS ON BOOKS DATA

INTRODUCTION

Text mining involves extracting patterns from natural language text rather than from structured data. Sentiment analysis is viewed as an important application of text mining. Sentiment analysis emphasizes on detecting polarity and recognizing emotions. It is challenging because the system needs to understand the positive or negative emotions of each sentence as well as human beings. Sentiment analysis is applied in various instances like improving customer relationship by analyzing positive and negative feedbacks, gathering product reviews etc. The essential task of sentiment analysis is polarity classification which is classification of opposing sentiments such as positive and negative, like and dislike etc. Sentiment analysis is performed on text vector to explore the most prominent text features.

The aim of this analysis is to perform sentiment analysis and compare the results on 2 books choosing one from each of the given lists (Adult list and Child list). From children's book list Black Beauty has been chosen and from Adult list, Emma by Jane Austen has been chosen for the required analysis.

METHODOLOGY

The Black Beauty book consists of 5997 rows and 2 columns and Emma book consists of 16235 rows and 2 columns. one column represents the Gutenberg Id and the other represents the text data in both the books.

The sentiment analysis on child book data (Black Beauty book) is done in R as follows:

- The necessary libraries like tidytext, dplyr, stringr, ggplot2, wordcloud2, syuzhet are loaded into the R file.
- The csv file is read in to the r using read.csv() and the empty lines are replaced with NA.
- The null values are eliminated from the data as a part of initial data preprocessing.
- The text data is loaded into a corpus which is a collection of documents and is the main data structure used by tm (text mining) package.
- VectorSource() function is used with in the corpus to merge the complete text from different rows into a single vector.
- The inspect() function gives brief over view of the structure of the corpus.
- The following cleaning tasks are associated with sentiment analysis.
    a. Converting the entire content of the text into lowercase.
    b. Removing Punctuation marks.
    c. Removing the numerical data present in the text, because it is not required for text analysis.
    d. The previous transformations create extra white spaces which are removed using stripWhitespace().

e. Stop words are familiar words whose information value is low. Hence stop words are removed from text data.

- The term document matrix is created from the corpus which consists of the occurrences of a word in the text.
- from the matrix obtained we use subset function to filter the words whose frequency is more than 50. The obtained words are plotted using a bar graph and the presence of insignificant words is checked.
- Removing Commonly used insignificant words which do not add any sentiment or emotion to text, by checking the frequent words.
- After performing text cleaning techniques, the key words are visualized using a word cloud. The word cloud helps in discovering associations and patterns in the text.
- Word cloud2 package is used to show the frequency of words in the word cloud.
- The Syuzhet package which extracts sentiments using sentiment dictionaries like nrc is used to calculate sentiment scores for each sentence based on the values assigned to the words in the dictionary.
- To understand the emotions and sentiments of entire data, get_nrc_sentiments function is used.
- Sentiment score of a sentence is given by subtracting the sum of negative words from the sum of positive words.
- If the score is zero, it implies that there is no presence of opinion words in the sentence.

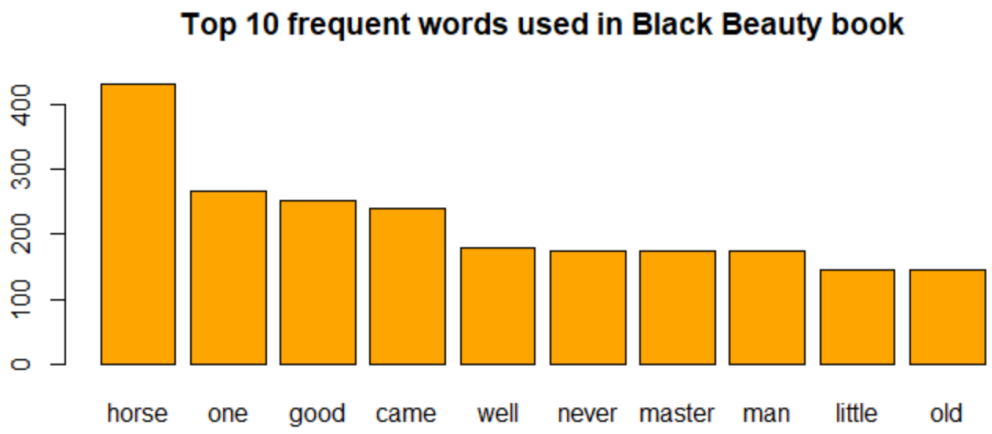
Similar analysis is carried out on Emma book from adult list and sentiment scores are calculated accordingly.

RESULTS

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

1) Black Beauty (Child Book)

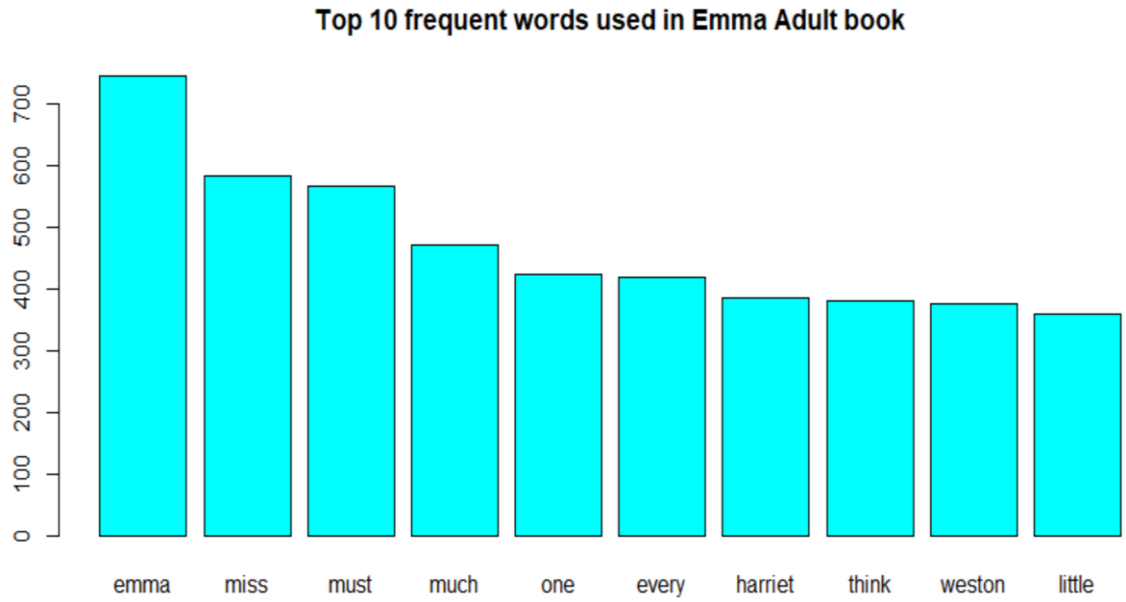
The below bar plot illustrated the most frequent words used in Black Beauty (child) book. We can observe that the word horse is used many times and it represents the character in the book. On the other hand, old and little are the two words which are used less frequently.



Using word cloud and wordcloud2, we can retrieve the information about the key words used in the books. The size of the words determines the frequency of the words. Bigger sized words will appear frequently, and small sized words appear less frequently. We can use worldcloud2 for any kind of different shape and image representation of the words.

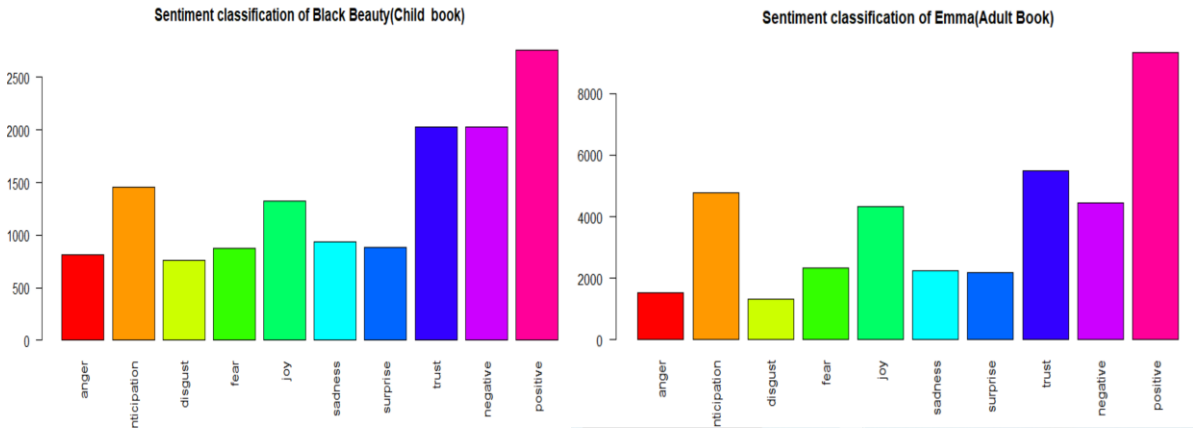
2) Emma (Adult Book)

The below bar graph demonstrates the use of top ten words in Emma adult book. We can observe that emma word is used most of the times, as it represents the character in the book and miss and must words also used more. Comparatively to other's little word is used very less.



Using word cloud and wordcloud2, we can retrieve the information about the key words used in the books. The size of the words determines the frequency of the words. Bigger sized words will appear frequently, and small sized words appear less frequently. We can use worldcloud2 for any kind of different shape and image representation of the words.

Comparing the emotions of both books (Black Beauty child and Emma adult book)



From above emotion classification graphs, it is evident that in both the books positive sentiment is more than negative sentiment. Emotions like anger and disgust are very less in Emma (Adult book) when compared to Black Beauty (Child Book). Anticipation and joy are carried out similarly in both the books. The difference between positive and negative sentiments is less in black beaty when compared to that of Emma book.

DISCUSSION:

Sentiment analysis is a technique for interpreting the emotions in documents. This analysis demonstrated how to approach sentiment analysis using tidy data concepts. It is observed that word clouds can be used in variety of ways to provide analysis by extracting text of most frequently occurring words. Sentiment analysis presents several difficulties such as, a word that is positive in one context can be considered negative in another. Moreover, people do not express their views in the same way.