INTRODUCTION

The complex task of understanding emotional content of unstructured and text-heavy data exploited by data-wrangling and visualisation is known as sentiment analysis. In this project, we take a tidy approach towards text mining in R and therefore try to understand the emotional intent of the text programmatically. The interpretations through the coding will give us a potential outcome of the emotions a book encompasses constructed on the intricacies of the verbiage.

We aim to perform sentiment analysis on the digital public domain texts from Project Gutenberg Collection. The two texts selected from the collection are The Wonderful Wizard of Oz by Lyman Frank and Emma by Jane Austen, each from the *Children's list* and *adult's list* respectively. The American fantasy-fictional children's novel, The Wonderful Wizard of Oz revolves around the story of a girl named Dorothy and her pet dog Toto who have been swept away from their home and their struggle against the Wicked Witch of the West. On the other hand, Emma is a fictional novel set in the 19^{th} century that embarks upon the youthful and romantic misunderstandings of the protagonist of the story. Our objective is to compare the result of sentiment analysis of the aforementioned titles and present with relevant interpretations.

METHODS

Step 1:- **Installing Packages**

The first step is to install the packages that are required to perform the text analysis.

Step 2:- **Downloading Text and Creating Data Frame**

The text of the selected titles from the "gutenbergr" library is downloaded and saved in a data frame.

Step 3:- **Text Tokenisation**

The data frame is converted into a tidy text format where the text is broken into individual tokens by the process of tokenisation. Tokenising text results in retaining line numbers, removing punctuations and converting all words to lowercase characters by default.

Step 4:- **Removal of Stop Words**

Using the built-in dataset of stop_words, extremely common words such as "the", "a", "of", "in", etc. which do not contribute much to the analysis are removed from the tidytext dataset by using anti_join().

Step 5:- **Word Count**

The count() function from the dplyr package is used to count the most frequently occurring words in the book as a whole. This is then visually represented by horizontal bar graph.

Step 6:- **Removal of Custom Stop Words**

Typically, custom stop words can be added which seem to be used multiple times in the text but can be misvalued during the sentiment analysis such as character names, titles or salutations.

Step 7:- **Word Frequencies**

The word frequencies for both the books are calculated by binding the data frames together. Thereafter, the word frequencies are plotted for comparison using ggplot(). Then, the correlation of word frequencies between the books is found out by cor.test().

Step 8:- **Sentiment Analysis with Tidy Data**

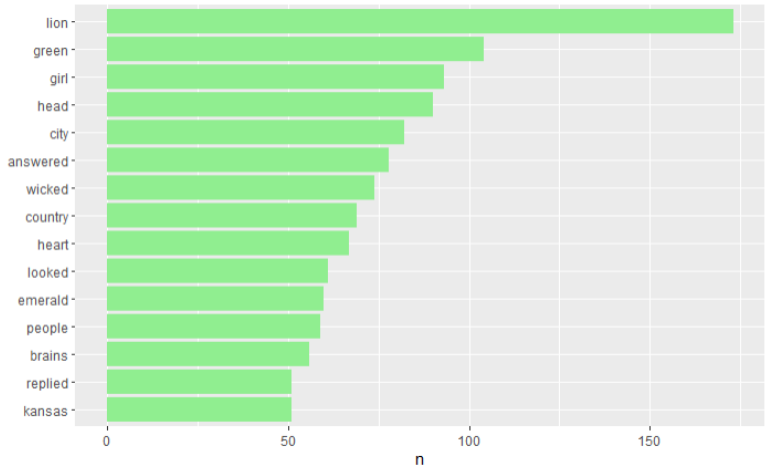
The emotions of the two texts are examined with respect to the three lexicons, namely, nrc, AFINN and bing which are available from the sentiment dataset. Relevant interpretation of the two texts are made based on their sentiment analysis.

RESULTS

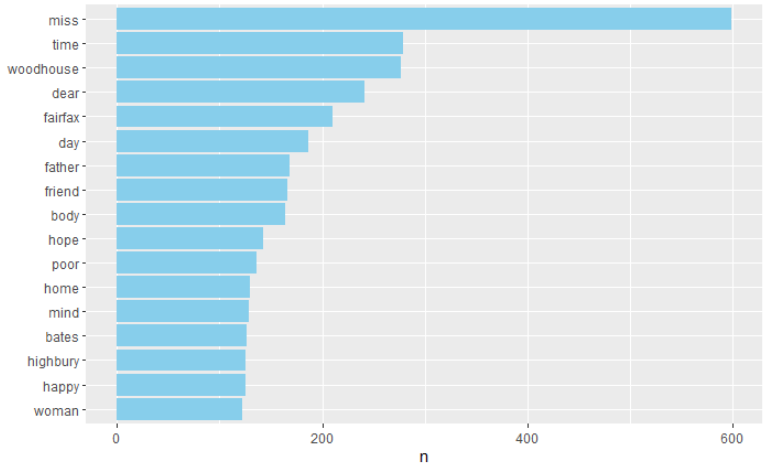
word <chr>	n <int>
dorothy	347
scarecrow	219
woodman	176
lion	173
oz	164
tin	140
witch	125
green	104
girl	93
head	90

word <chr>	n <int>
emma	786
miss	599
harriet	415
weston	388
knightley	356
elton	319
jane	282
time	279
woodhouse	277
dear	241

We count the most common words in each of the texts, and elect some custom stop words such as “dorothy”, “oz”, “emma”, “elton”, etc. and remove them from the tidy data.



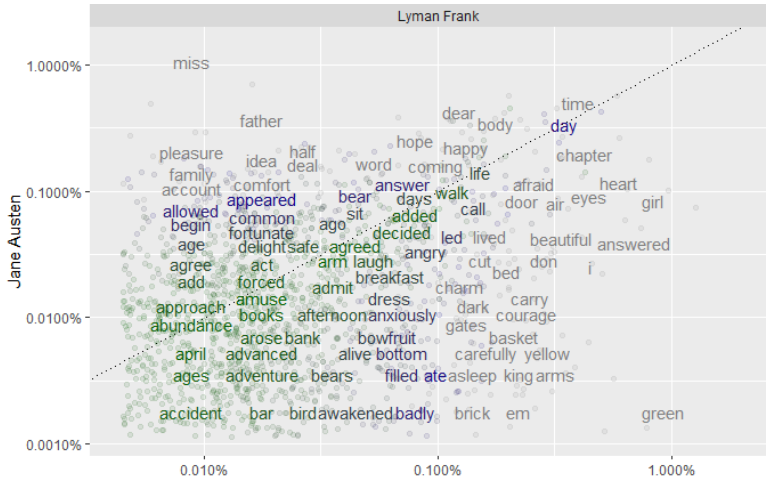
The Wonderful Wizard of Oz



Emma

After tidying the data, we see that the most common words in The Wonderful Wizard of Oz are “lion”, “green”, “girl”, etc. whereas in Emma it is “miss”, “time”, etc.

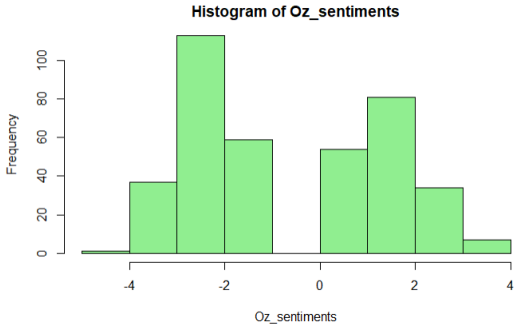
We compare the word frequencies of each text and plot it across a graph. In the graph below, the words that are close to the line have similar frequencies in both texts such as “day”, “time”, “life”, etc. Words that are farther away are found more in one set of text than the other, such as “green” in The Wonderful Wizard of Oz and “miss” in Emma as we have seen before during the word count. We also perform Pearson’s product-moment correlation and find that the word frequencies in both texts have a weak but positive correlation.

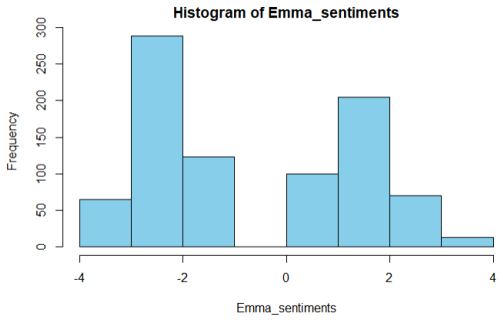



```
        Pearson's product-moment correlation

data:  proportion and Jane Austen
t = 10.214, df = 1410, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2132345 0.3103989
sample estimates:
      cor
0.2624819
```

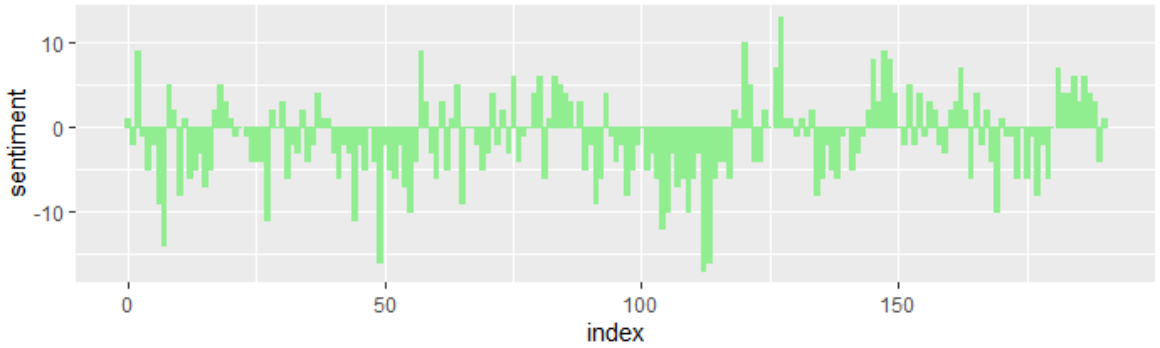
Next, we look into the sentiment content of each text based on unigram lexicons. At first, we consider the AFINN lexicon which assigns scores to each word between -5 to +5 indicating the sentiment transformation of negative to positive respectively.





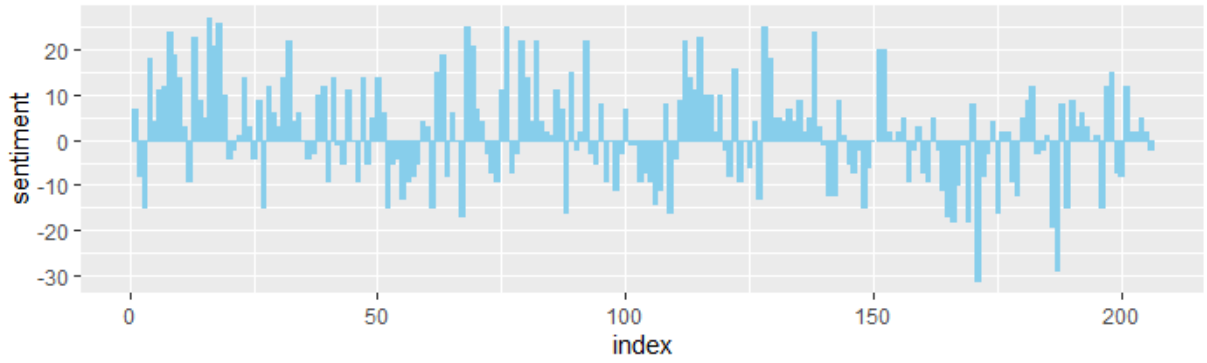
By plotting histogram, we witness that there are more negative valued words in the children's book ranging from -5 to +4 as compared to the adult book whose value ranges from -4 to +4. Further to assess the situation, we find the weighted mean of the sentiment values for both the texts and realize that Emma has a positive score of 0.3130081 whereas the children's book, The Wonderful Wizard of Oz tends to have an overall negative sentiment as it has a mean of -0.1231842.

We can further evaluate the sentiments using the bing lexicon which categorises words into binary positive or negative. We plot the net sentiment (i.e., positive – negative sentiment) across the plot trajectory of both the books as too less or too many words can result in wrong sentiment estimates or may affect the narrative structure.



The Wonderful Wizard of Oz

We infer that there is more negative sentiment in the text of The Wonderful Wizard of Oz as there is more dips than peaks; whereas in Emma, the sentiments seem to be equally distributed along the trajectory of the story.

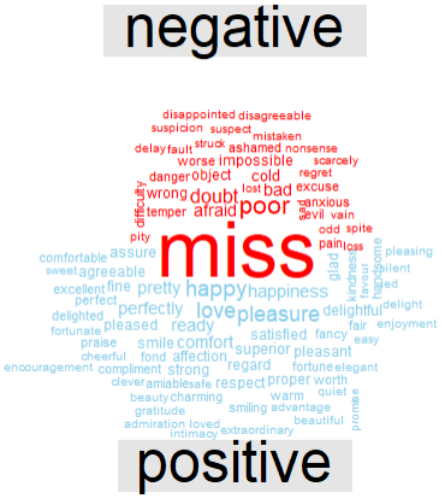


Emma

Thus, a word cloud of negative and positive words can be presented referring to the bing lexicon with the help of joins, piping and dplyr on the tidy data. It can be used to visualise the most positive or negative words.

The Wonderful Wizard of Oz

Emma



In The Wonderful Wizard of Oz, “wicked” and “beautiful” are the most negative and positive words respectively. While in Emma, the words “happy”, “love”, “pleasure” have been classified as positive but the word “miss” has been misclassified as negative because “miss” is used more as a salutation in the book rather than a negative word.

DISCUSSION

Sentiment analysis is an important aspect of Natural language Processing that helps computers comprehend and interpret human language. It makes path for understanding the attitudes and opinions expressed in texts and gives a measure of the emotional content. On performing sentiment analysis on the two titles under consideration, The Wonderful Wizard of Oz by Lyman Frank from the *Children’s list* and Emma by Jane Austen from the *Adult’s list* we conclude that the children’s book has more negative sentiments involved when compared to the adult’s book. This is an interesting finding as one expects the children’s book to have more positive words such as “compassion”, “happy”, “kindness”, “joy”. Sentiments are inherently subjective. Hence, we come across the limitations of sentiment analysis as it is programmatically difficult to recognise jokes, irony, negations and exaggerations. We witness an anomaly- how the word “miss” is grouped under negative words wherein in reality it is simply used as a title for young, unmarried women. Although, this limitation can be overcome by adding “miss” to the list of custom stop words and removing it from the text.