Text mining and sentiment analysis of the books 'Great Expectation & Rebecca of sunny brook farm' accessed through Gutenbergr library.

Muhammad Taimoor Khan Malik

Contents

SUMMARY

The purpose of this report is to describe the results obtained from data exploration, finding meaningful information with the help of text mining operations and sentiment analysis. Text mining operations were performed on two books from Gutenbergr library to find the most common and uncommon words in both the books and how many characters, words, lines, sentences and chapters does each book have. And the results have proved the hypothesis true, that child list book will be more positive and will be shorter in length as compared to the adult list book. Also, sentiment words were looked up by using the pre-stored sentiment dictionaries that are nrc, bing and afinn in order to find out the positive and negative words in both the books. It is therefore confirmed based on the result achieved that child book has more positive sentiment words as compared to adult sentiment words. And values from K-Test are obtained to find out that the difference of distribution of data is perfect and T-Test confirms that child book has more positive sentiment words as compared to the adult words. This study can be taken further to look at the book from different angle and to apply alternative hypothesis tests to find out something more and different.

Introduction

The purpose of the report is to present the findings of the results obtained through text mining and sentiment analysis of two books accessed from R integrated library ***"Gutenbergr".*** In the analysis, number of packages such as "DPLYR", "TIDYR", "STRINGR", "TIDYTEXT", "GGPLOT2", "GGTHEMES", "GGRAPH" and "RESHAPE2" are used for performing the mining operations as well as obtaining the sentiments. The analysis investigates the differences and the commonalities between two books which are from two different categories. One book belongs to the adult category while other belongs to the children category. Hypothesis are mentioned below which are under investigation for which the analysis is performed.

1. *Great Expectations book is for adults which makes up the assumption that it will be lengthy as compared to child book.*
2. *Child list book 'Rebecca of sunny brook farm' will have more positive sentiment expressions as compared to the adult list book 'Great Expectations'.*
3. *Finding the difference of distribution of data for two books to find if the data fits perfectly or not.*
4. *If 3 sentiment dictionaries produce the same result or not*

For performing sentiment analysis, "dplyr" library will be used to fetch the sentiments from three already stored dictionaries that are "AFINN", "NRC" and "Bing". These dictionaries will enable us to identify the positive and negative sentiments for both the books. Also, we will find out the similarities and differences between these 3 dictionaries. Statistical analysis that are performed are T Test and kolmogorov-smirnov test. "T-Test" is used to find out whether which book has more positive or negative sentiments. Kolmogorov-smirnov test is used to find the fit of the data. In addition, n-grams function is used to remove the common stop words and to find out the connection of words with each other.

Great expectation book revolves around the childhood and youth life of an orphan child pip working as an apprentice for a blacksmith who is his brother-in-law and how his life turned when he received the large fortune by anonymous benefactor and becoming the part of high society. Rebecca of a sunny brook farm is all about the story of searching for Rebecca living on a sunny brook after being rejected at an audition for radio advertising company.

Methods

Data is inserted into the variable from the book through reading csv function using the tidy verse library. Before processing the data, it is cleansed by using tidyr and stringr library and it enabled us to remove the non-meaningful information such as punctuations, null values and missing data.

After cleansing the data, the data is made tidy by using the *un-nest* function to distribute the data in a single word per row. *Stop words* are then removed from the tidy data for further cleaning of the data. *Regex* is used for the purpose of identification of chapters headings as well as chapter numbers in the book and to further look for the information within the book. Count function is used to identify the number of sentences and words in the book which will enable us to identify the length of the books.

Inner join is used to identify the common words between both books where *anti join* is used to identify the words that are exclusive to each book. *Afinn*, *nrc* and *bing* are three dictionaries that are used to perform the sentiment analysis and it enabled us to find the positive and negative sentiments for the books using inner join function and mutate function for adding the column in the tibble. Subsequently, difference between the 3 dictionaries is calculated by applying the functions of inner join, mutate, summarisation and then binding the row for the identification of value. Kolmogorov-test is performed to find out the difference in distribution of data using the function of *Ks.test*.

Results

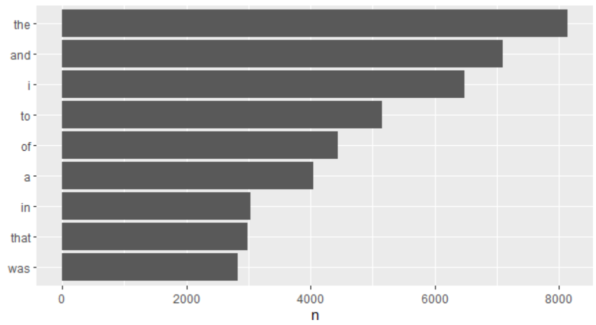

Figure 1: Adult book

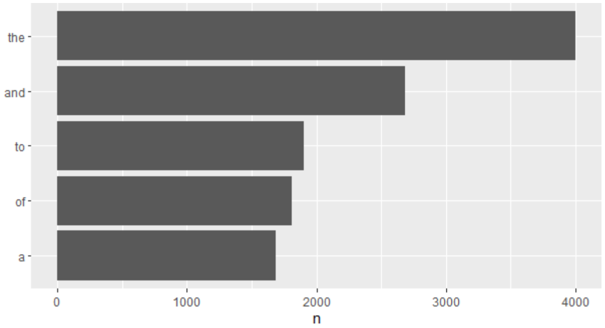

Figure 2: Child book

Figure 1 and Figure 2 shows the most common words which enabled us to build a custom stop words table in order to remove them from the books for making it tidier.

Text mining operations enabled us to identify the lengthy book by providing us with the stats and the stats show that adult book is lengthy as it has more chapters, words, sentences, lines and characters. Also, Figure 3 supports the fact claimed.

Description	Adult Book	Child Book
Chapter Count	59	31
Words Count	55,575	26,592
Sentences	15,949	8,969
Characters	761,768	318,093
Lines	15,939	6,612

Table 1: Comparison

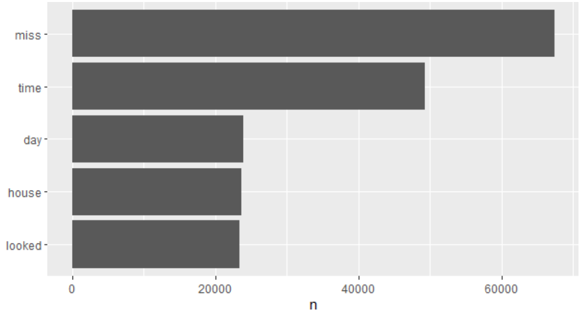

Figure 4: Common words

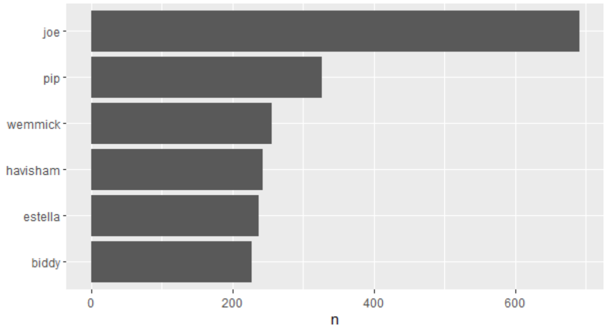

Figure 5: Uncommon words

Figure 4 shows that “’miss’, ’time’, ’day’, ’house’, ’look’” are the most common words and “’joe’, ’pip’, ’wemmick’, ’havisham’, ’estella’, ’biddy’” are the most uncommon words in the book.

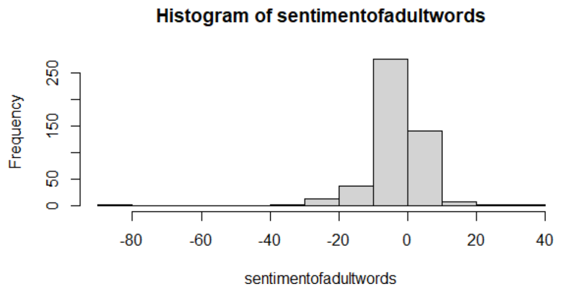

Figure 6: AFINN Child book sentiments

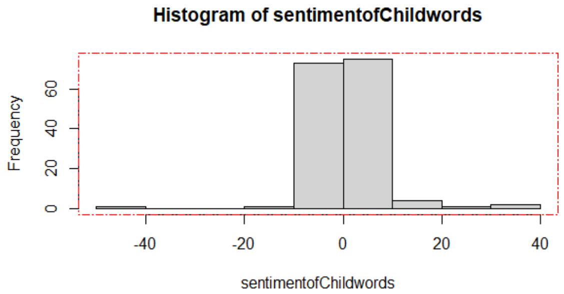

Figure 7: AFINN Child book sentiments

Figure 6 shows that child book has relatively more positive sentiments whereas it is opposite for the adult book as shown in Figure 7 that the value is more skewed towards the negative side and have more negative sentiments. Also, the *"T-Test"* values supports the argument by giving us the value of 0.86 for child book which describes that the book has more positive sentiments and -2.43 value for adult book shows that adult book has more negative words. "**Ks-Test**" gave us the value of **1** for both the books which defines that the data fit is perfect.

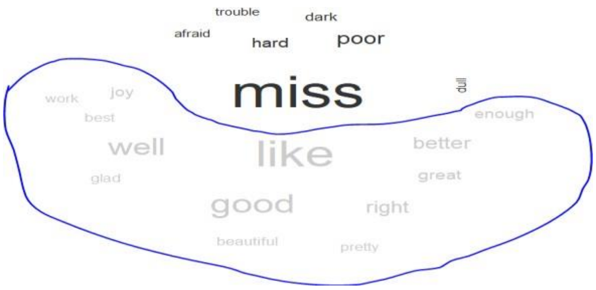

Figure 8: Adult book positive words

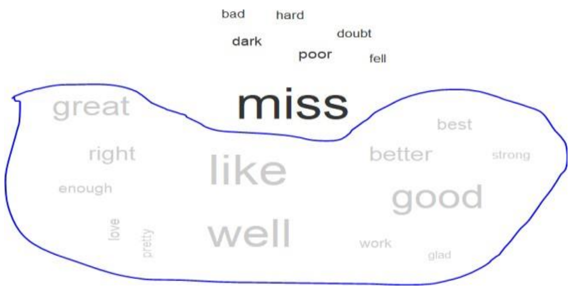

Figure 9: child book positive words

Figure 8 shows the top 20 positive words and negative words for Adult book where encircled ones are positive, and Figure 9 shows the same for child book. And difference between 3 dictionaries is that "*nrc*" has 2308 positive words and 3318 negative words whereas "*bing*" has 2005 positive words and 4781 negative words. Therefore, using *bing* will result in identifying more negative sentiments as compared to *nrc*.

Conclusion

Assumed hypothesis and questions under investigation appeared to be true that child book is not lengthy as compared to the adult book and has more positive sentiment words. Correspondingly, data that is used for analysis was also the perfect fit and have good balance of distribution and this is proved by Kolmogorov test. Likewise, 3 dictionaries of sentiments provide different results due to their own tibble reference data. This exploratory research can be carried out further to look for more useful information and how the results can be different if other tests were used.

References

Robinsen DD. *Download and Process Public Domain Works from Project Gutenberg.* 498 & 1400. R Studio. Library(Gutenbergr); 2021 Available from; https://github.com/ropensci/gutenbergr[Accessed 18th December 2021].