Registration # 2101142

Report on
Sentiment Analysis on two books
"Great Expectation and Treasure Island".

Sentiment analysis is a contextual mining of text which use text analysis, natural language processing, linguistics and biometrics to identity and extracts subjective information in sources of two books. It helps organizations in understanding the social sentiment of their products, brands and services. In this particular, sentiment analysis, two books treasure Island and great expectations have been thoroughly analyzed, the responses and feelings, most positive and negative words, emotions of joy, fear, anger have been examined. Sentiment analysis is very important and can tell the human behavior. 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.

Treasure Island is the book which revolves/reflect the desires and greed of the characters whose greed is to get the treasure. Also, the writer has also emphasized that some adventures are important for our growth and development. Jim, in her journey learns many important lessons such as personal integrity, self-confidence, and maturity. It could be argued that the purpose of this novel is to show how people's expectations in life are often unrealistic, uninformed, or unreasonable.

Great Expectations has been chosen in adult category. The moral theme of Great Expectations is quite simple: affection, loyalty, and conscience are more important than social advancement, wealth, and class.

Theme of the report: AIMS TO INVESTIGATE:

1) Exploratory data analysis
2) Classification of lexicons with respect to two books
3) Statistical analysis, hypothesis testing for the fit of the distribution

A. PACKAGES USED:

library	description	Functions of library
dplyr	used for data manipulation such as to manipulate, clean and summarize unstructured/structured data	Mutate, select, filter, %>% , summarise, arrange
tidytext	used in conversion of text to and from tidy formats	
gutenbergr	a library of many texts	
stringr	contains a cohesive set of functions to manipulate strings	str_detect, str_locate_all, str_subset, str_trim, str_c
tidyr	it contains tools for reshaping (pivoting) and hierarchy (nesting and 'unnesting') of a dataset.	
ggthemes	used for look and feel of graphs, visualization	
tm	text mining package used for data wrangling	Gsub, removeWords, lemmatize_words
ggplot2	it is used for plotting graphs, data visualization.	
SnowballC	used for stemming of words, i.e., changing words to its root elements	
textstem	tools for Stemming and Lemmatizing Text	wordStem

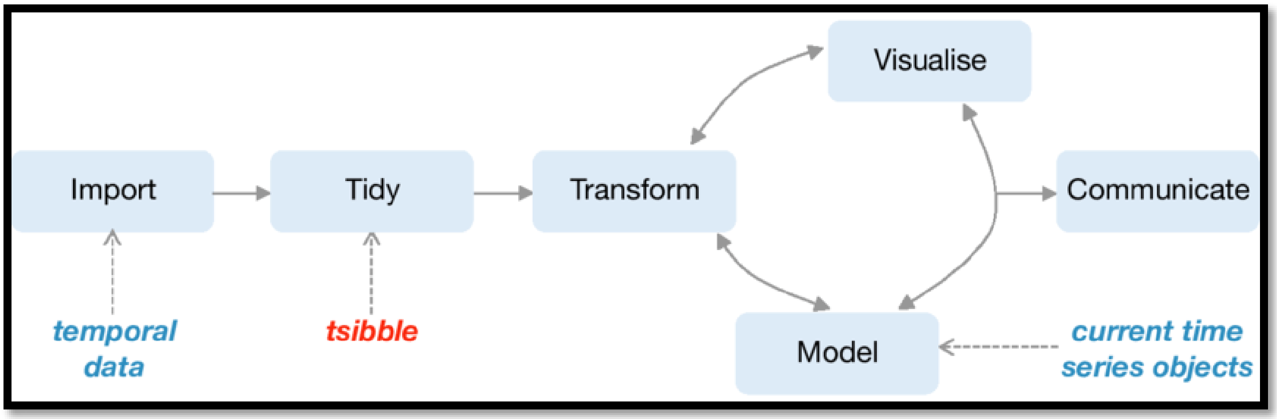
wordcloud	it helps to analyze text and visualize keywords/text	
scales	gives tools to override default breaks, labels and transformations	

B. TECHNIQUES USED:

STEP 1. DATA WRANGLING/ PRE-PROCESSING

The objective of this step is to ensure data quality and usefulness of data, data accuracy. In this step the extracted raw data from Gutenberg package is transformed to meaningful data by un-nesting the words, removing punctuations, breaking text into sentences, lines, grouping by using the following methods from using different libraries.

1) Stop word removal (common English stop words which are irrelevant to sentiment lexicons are removed for example the, that, that's, it, a etc.)
2) Word frequency
3) Stemming and Lemmatisation
4) Tokenization and Negation handling
5) Strip white spaces and sparse terms, remove numbers



*The below picture represents the abstraction of data wrangling process, to prepare data for meaningful insights/to extract patterns from data it is important to remove unnecessary text.

STEP 2. FEATURE EXTRACTION:

In this phase we identify features such as words, sentences, a paragraph, lines or chapters, n-grams are taken from the text for sentiments, common and uncommon words are taken for statistical analysis and distribution of data.

1) Text representation
2) N grams
3) Parts of speech (Pos) tagging
4) Negations
5) Sentence segmentation
6) Co-reference resolution
7) Dependency parsing

STEP 3. SENTIMENT ANALYSIS:

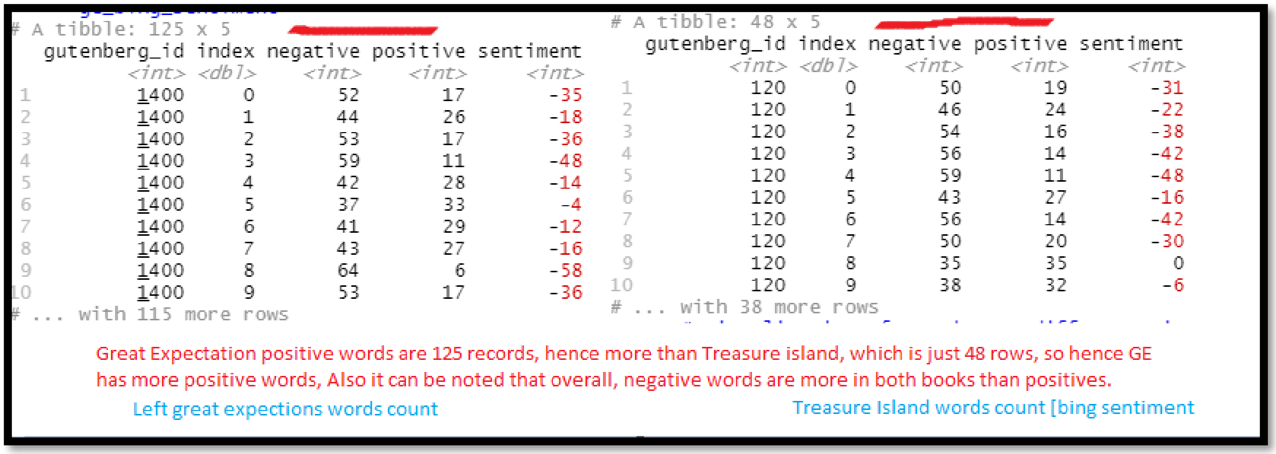
Different tests have been performed and dplyr library to get sentiments from the contents and pairing them and testing them with built in sentiments libraries such as "NRC, Bing, Afinn" to analyze the words with sentiment lexicons. Also, compared and contrast the features of the sentiments.

Registration # 2101142

Two statistical tests have been performed such as Kolmogorov-smirnov and T-test tests. T-test has been performed to find out which book has more positive/negative sentiments than others. Kolmogorov-smirnov is performed to find the fit of the data, to check whether books data is normally distributed or not. Furthermore, n-grams functions are used to clean the data by removing common stop words and to find out how closely data/words are relevant to each other.

Hypothesis are mentioned below which are under investigation for which the analysis is performed.

1) Child list book "Treasure Island' will have more positive sentiment expressions as compared to the adult list book 'Great Expectations'.



```
# A tibble: 125 x 5
  gutenberg_id index negative positive sentiment
         <int> <dbl>    <int>    <int>     <int>
 1        1400     0       52       17       -35
 2        1400     1       44       26       -18
 3        1400     2       53       17       -36
 4        1400     3       59       11       -48
 5        1400     4       42       28       -14
 6        1400     5       37       33        -4
 7        1400     6       41       29       -12
 8        1400     7       43       27       -16
 9        1400     8       64        6       -58
10        1400     9       53       17       -36
# ... with 115 more rows
```

```
# A tibble: 48 x 5
  gutenberg_id index negative positive sentiment
         <int> <dbl>    <int>    <int>     <int>
 1         120     0       50       19       -31
 2         120     1       46       24       -22
 3         120     2       54       16       -38
 4         120     3       56       14       -42
 5         120     4       59       11       -48
 6         120     5       43       27       -16
 7         120     6       56       14       -42
 8         120     7       50       20       -30
 9         120     8       35       35         0
10         120     9       38       32        -6
# ... with 38 more rows
```

Great Expectation positive words are 125 records, hence more than Treasure island, which is just 48 rows, so hence GE has more positive words, Also it can be noted that overall, negative words are more in both books than positives.

Left great expections words count

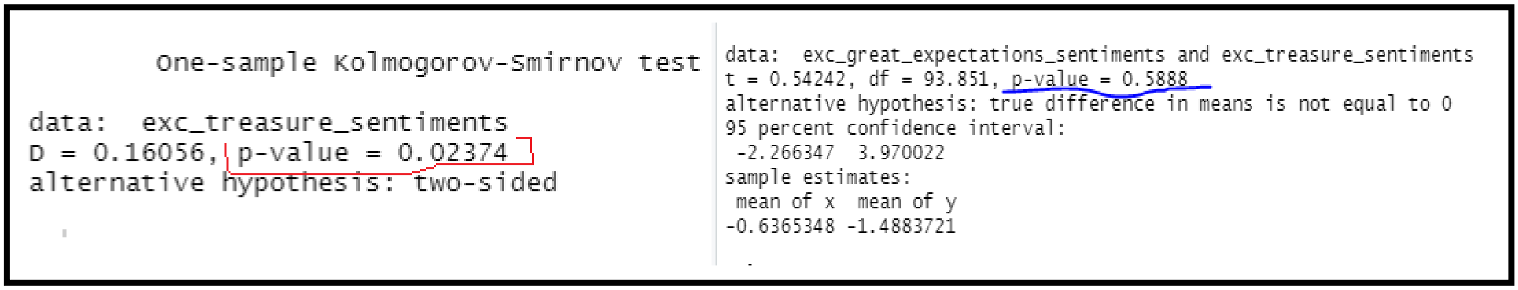
Treasure Island words count [bing sentiment

*It can be seen in the statistics below that great expectations has more positive words than treasure Island.

2) Great Expectations book is for adults which makes up the assumption that it will be lengthy as compared to child book.
   a) This is also true as great expectation is lengthy as compared to treasure island.
   b) After cleaning treasure island data frame has: A tibble: 22,526 x 2 words(repeating =NO)
   c) While great expectation data frame has: A tibble: 56,504 x 2 words (repeating = yes)

3) Finding the difference of distribution of data for two books to find if the data fits perfectly or not.

As can be seen in the below two screenshots, the Kolmogorov-Smirnov test p-value is very low, which states that distribution differs and fit is not perfect.



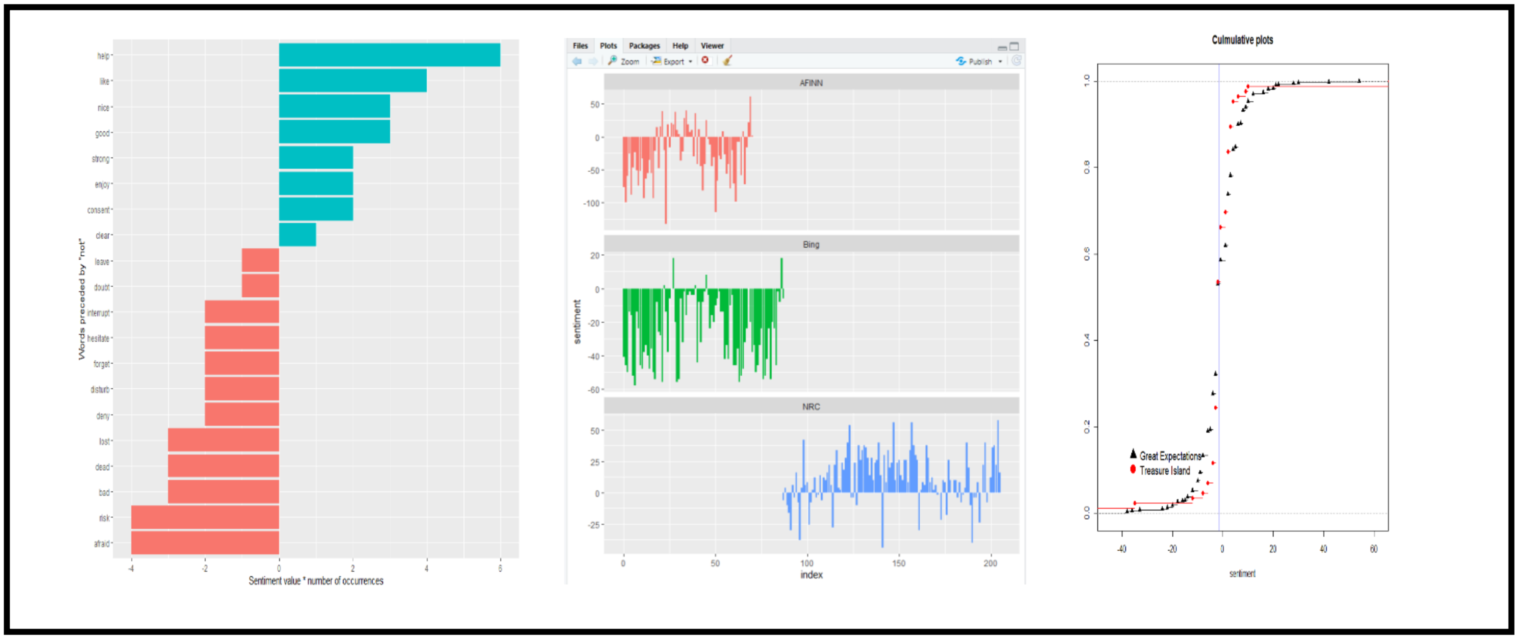
```
        One-sample Kolmogorov-Smirnov test

data:  exc_treasure_sentiments
D = 0.16056, p-value = 0.02374
alternative hypothesis: two-sided
```

```
data:  exc_great_expectations_sentiments and exc_treasure_sentiments
t = 0.54242, df = 93.851, p-value = 0.5888
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.266347  3.970022
sample estimates:
 mean of x  mean of y
-0.6365348 -1.4883721
```

For T-test p-value is more than 0.5 which shows difference of means test for two distributions

4) To check the result of the three dictionaries whether similar or dissimilar



*From the below second screenshot, it can be observed that Afinn and Bing words are more negative while nrc dictionary words are more positive. It can also be observed that Bing has more negative lexicons than nrc lexicons. Hence, we can get clearer picture when using Bing lexicon

Results/conclusion:

Hence from the above report, statistical analysis and sentiment analysis, it can be concluded, that sentiments of different stake holders captured from text analysis can be very useful for understanding the real intent/sentiments of the user for decisions making.

There are certain challenges which can be main blockers in analysing sentiments such as Tone, Sarcasm, Emojis and Idioms problems, biasness of users in different contexts and negations. However, with further research

Text analysis can be further improved, by performing some more statistical analysis and applying different statistical tests and working more on data to clean it more precisely. Some more work can be done by applying machine learning models to train and test the model to achieve better accuracy and to test the sentiments properly which can be very useful for different stakeholders.

References: https://www.google.com/

https://www.researchgate.net/figure/Illustration-of-the-data-science-workflow-drawn-from-Wickham-Grolemund-2016-showing_fig1_330726479

https://www.sparknotes.com/lit/greatex/themes/