Project Gutenberg Books Sensitivity Analysis

Sujata Verma

9/28/2020

This project aims to predict the "feel good" factor or the proportion of positive sentiments over negative sentiments in books across multiple genres using the Project Gutenberg free e-books dataset utilizing Sensitivity Analysis or Emotional Artificial Intelligence.

Introduction

In times of the pandemic, there has been an increase in the overall stress and anxiety among people and one of the ways of relaxing while sheltering at home has been reading books. Project Gutenberg is an online library of free e-books established by Michael Hart in 1971 to "encourage the creation and distribution of eBooks".There are over 50,000 works in this library, in many languages and across many genres.

The central **objective** of this report is to identify the best books containing the highest proportion of words reflecting a positive sentiment, using the AFINN Sentiment Lexicon, in order to help readers select "feel good" books to read. AFINN Sentiment Lexicon is a list of words assigned numerical integer values for sentiments and range from -5 (negative sentiment) to +5(positive sentiment), manually compiled by Finn Årup Nielsen.

Obtaining the data

In this section, the data is imported from the Gutenberg Library and the summary statistics are obtained to get an overview of the data.

Data Exploration

The data set contains more than 50,000 books, documents etc. GutenbergID assigns a number to each of the works. Each work is specified with title and author name and assigned author ID number. Other variables are Language, Bookshelf (Genre), Rights (pertaining to copyright) and whether the work contains text.

Descriptive statistics

Nest, the distribution of non-binary numeric variables is explored. The table given below includes the mean, standard deviation, median, minimum value, maximum value, the range and skewness of the numeric variables.

```
## # A tibble: 51,997 x 8
##    gutenberg_id title author gutenberg_autho~ language gutenberg_books~
```

```
rights
##           <int> <chr> <chr>             <int> <chr>    <chr>
<chr>
##  1            0  <NA> <NA>                NA en       <NA>
Publi~
##  2            1 "The~ Jeffe~           1638 en       United States L~
Publi~
##  3            2 "The~ Unite~              1 en       American Revolu~
Publi~
##  4            3 "Joh~ Kenne~           1666 en       <NA>
Publi~
##  5            4 "Lin~ Linco~              3 en       US Civil War
Publi~
##  6            5 "The~ Unite~              1 en       American Revolu~
Publi~
##  7            6 "Giv~ Henry~              4 en       American Revolu~
Publi~
##  8            7 "The~ <NA>                NA en       <NA>
Publi~
##  9            8 "Abr~ Linco~              3 en       US Civil War
Publi~
## 10            9 "Abr~ Linco~              3 en       US Civil War
Publi~
## # ... with 51,987 more rows, and 1 more variable: has_text <lgl>

##   gutenberg_id       title              author
gutenberg_author_id
##  Min.   :     0   Length:51997       Length:51997       Min.   :    1
##  1st Qu.: 12999   Class :character   Class :character   1st Qu.:  641
##  Median : 25998   Mode  :character   Mode  :character   Median : 3612
##  Mean   : 26016                                         Mean   :12930
##  3rd Qu.: 38997                                         3rd Qu.:31606
##  Max.   :999999                                         Max.   :46625
##                                                         NA's   :3457
##    language         gutenberg_bookshelf    rights           has_text
##  Length:51997       Length:51997        Length:51997       Mode :logical
##  Class :character   Class :character    Class :character   FALSE:1367
##  Mode  :character   Mode  :character    Mode  :character   TRUE :50630
##
##
##
##
```

Method and Analysis

In the following section, the metadata containing 51,997 rows is reduced to about 1000 observations for speed up calculations. This means that around 1000 books will be downloaded and each word in each book will be examined and assigned a sensitivity score by comparing it against the AFINN Lexicon.

The following criteria are used to subset the data. Only english language books that contain text and fall under 'Public Domain in the U.S.' are considered. Stop-words like 'if', 'but', 'he' etc. are eliminated as are the numbers such as chapter numbers. The metadata defines numerous 'bookshelves' or sub-categories from 'Animals' to 'Zoology'. In this project, two or three bookshelf sub-categories are grouped together to define a genre and five different genres are defined from 16 sub-categories. The different genres are:

1. *Lyric* consting of Operas, Playes and Poetry categories.
2. *Classic* consisting of Harvard Classics, Historical Fiction,Classical and Antiquity categories.
3. *Mystery* consisting of Detective Fiction, Crime Fiction, Gothic Fiction, and Horror categories.
4. *Fun* consisting of Humor, Adventure and Fantasy categories.
5. *Great* consisting of Best Books Ever Listings and Biographies categories.

Next, sentiment analysis is applied to the books to predict the proportion of positive sentiment using the AFINN lexicon. The mean sentiment score, defined as the difference between positive over negative sentiments is then calculated for each book in the genre and aggregated for each genre.

The top 10 books in each genre according to the mean sentiment score are tabulated.

Genre 1: Lyric

title	author	gutenberg_bookshelf	mean_sentiment
Chapters of Opera Being historical and critical observations and records concerning the lyric drama in New York from its earliest days down to the present time	Krehbiel, Henry Edward	Opera	0.553
Poems 1817	Keats, John	Poetry	0.632
A Little Book of Western Verse	Field, Eugene	Poetry	0.517
Life And Letters Of John Gay (1685-1732), Author of "The Beggar's Opera"	Melville, Lewis	Opera	0.535
The Standard Operas (12th edition)			

Their Plots, Their Music, and Their Composers	Upton, George P. (George Putnam)	Opera	0.535
Old Scores and New Readings: Discussions on Music & Certain Musicians	Runciman, John F.	Opera	0.521
The Opera A Sketch of the Development of Opera. With full Descriptions of all Works in the Modern Repertory.	Streatfeild, R. A. (Richard Alexander)	Opera	0.530
The Vision of Sir Launfal And Other Poems by James Russell Lowell; Edited with an Introduction and Notes by Julian W. Abernethy, Ph.D.	Lowell, James Russell	Poetry	0.563
A Dark Month From Swinburne's Collected Poetical Works Vol. V	Swinburne, Algernon Charles	Poetry	0.565
Some Forerunners of Italian Opera	Henderson, W. J. (William James)	Opera	0.659

The top ten works in the Lyric genre are dominated by Operas, demonstrating the uplifting power of Operas.

'Some Forerunners of Italian Opera' by W.J. Henderson is the highest positively ranked work in this genre with a score of 0.659. Project Gutenberg website introduces the book as follows "The purpose of this volume is to offer to the English reader a short study of the lyric drama in Italy prior to the birth of opera, and to note in its history the growth of the artistic elements and influences which finally led the Florentine reformers to resort to the ancient drama in their search for a simplified medium of expression."

Genre 2: Classic

Let us see how do Classics fare in terms of the sentiment analysis.

title	author	gutenberg_bookshelf	mean_sentiment
New Atlantis	Bacon, Francis	Harvard Classics	0.816
Theocritus, Bion and Moschus, Rendered into	NA	Classical Antiquity	0.571

English Prose			
An Egyptian Princess — Volume 04	Ebers, Georg	Historical Fiction	0.765
The Sisters — Volume 2	Ebers, Georg	Historical Fiction	0.558
Barbara Blomberg — Volume 03	Ebers, Georg	Historical Fiction	0.626
Barbara Blomberg — Volume 04	Ebers, Georg	Historical Fiction	0.700
The Parisians — Volume 01	Lytton, Edward Bulwer Lytton, Baron	Historical Fiction	0.507
The Parisians — Volume 04	Lytton, Edward Bulwer Lytton, Baron	Historical Fiction	0.668
The Bible, King James version, Book 47: 2 Corinthians	Anonymous	Harvard Classics	0.655
Cicero's Brutus or History of Famous Orators; also His Orator, or Accomplished Speaker.	Cicero, Marcus Tullius	Classical Antiquity	0.503

In the sub-category, The Harvard Classics,the book titled 'The New Atlantis', by author Francis Bacon is ranked at the top with the positivity score of 0.816. An introductory note about the book on the Guternberg website states "Bacon's literary executor, Dr. Rowley, published"The New Atlantis" in 1627, the year after the author's death. It seems to have been written about 1623, during that period of literary activity which followed Bacon's political fall. None of Bacon's writings gives in short apace so vivid a picture of his tastes and aspirations as this fragment of the plan of an ideal commonwealth. The generosity and enlightenment, the dignity and splendor, the piety and public spirit, of the inhabitants of Bensalem represent the ideal qualities which Bacon the statesman desired rather than hoped to see characteristic of his own country; and in Solomon's House we have Bacon the scientist indulging without restriction his prophetic vision of the future of human knowledge."

It is not surprising that The Bible finds a place in the top ten list as there is a surge of positive emotions when reading religious texts.

Genre 3. Mystery

Next we analyze the sentiments of books in the mystery genre.

title	author	gutenberg_bookshelf	mean_sentiment
Northanger Abbey	Austen, Jane	Gothic Fiction	0.251
Mosses from an Old Manse, and	Hawthorne,	Gothic Fiction	0.190

Other Stories	Nathaniel		
His Last Bow: An Epilogue of Sherlock Holmes	Doyle, Arthur Conan	Detective Fiction	-0.017
The Man	Stoker, Bram	Horror	0.102
The Lady of the Shroud	Stoker, Bram	Horror	0.331
Nightmare Abbey	Peacock, Thomas Love	Gothic Fiction	0.113
No Hero	Hornung, E. W. (Ernest William)	Crime Fiction	0.202
The Lost Stradivarius	Falkner, John Meade	Horror	0.070
The Passenger from Calais	Griffiths, Arthur	Detective Fiction	-0.020
R. Holmes & Co. Being the Remarkable Adventures of Raffles Holmes, Esq., Detective and Amateur Cracksman by Birth	Bangs, John Kendrick	Detective Fiction	0.190

Even though a number of readers like reading mystery books for the thrill, they do not score very highly on the Sentiment Score. The top book in this genre, The lady of the Shroud, is authored by Bram Stoker, and belongs to the Horror category.It scores only 0.331 on the sensitivity score.

Bram Stoker, the famous author of Dracula, has written a supernatural fantasy in The Lady of the Shroud, with themes of a vampire tale, Balkan politics and a love story.

Genre 4. Fun

We next turn to the genre 'Fun' and analyze the sentiment associated with books in this genre.

title	author	gutenberg_bookshelf	mean_sentiment
Old French Romances, Done into English	Morris, William	Fantasy	0.586
Samantha among the Brethren — Volume 1	Holley, Marietta	Humor	0.428
Samantha at the St. Louis Exposition	Holley, Marietta	Humor	0.448
The Ten Pleasures of Marriage and the Second Part, The Confession of the New Married Couple	Marsh, A.	Humor	0.485

The Rubaiyat of Ohow Dryyam

With Apologies to Omar	Duff, J. L.	Humor	0.714
‘Oh, Well, You Know How Women Are!’	NA	Humor	0.436
M. P.’s in Session: From Mr. Punch’s Parliamentary Portrait Gallery	Furniss, Harry	Humor	0.619
Mr. Punch Afloat: The Humours of Boating and Sailing	NA	Humor	0.541
Mr. Punch’s Book of Love: Being the Humours of Courtship and Matrimony	NA	Humor	0.573
Mr. Punch’s Book of Sport			

The Humour of Cricket, Football, Tennis, Polo, Croquet, Hockey, Racing, &c |Various |Humor | 0.487|

Most of the top books in this genre belong to the humor category, with J.L. Huff’s “The Rubaiyat of Ohow Dryyam, With Apologies to Omar” taking the top Honors with a score of 0.714.The book is a parody of the famous The Rubaiyat by Omar Khayyam and it is about the prohibition-era.

Most titles in the top 10 books in this genre are intriguing and provocative.

Genre 5. Great

Lastly, we analyze the books belonging to the ‘Great’ genre.

title	author	gutenberg_bookshelf	mean_sentiment
Herland	Gilman, Charlotte Perkins	Best Books Ever Listings	0.640
John Jacob Astor	Hubbard, Elbert	Biographies	0.504
Knights of Art: Stories of the Italian Painters	Steedman, Amy	Biographies	0.641
Little Masterpieces of Autobiography: Actors	Iles, George	Biographies	0.419
A Vindication of the Rights of Woman			
With Strictures on Political and Moral Subjects	Wollstonecraft, Mary	Best Books Ever Listings	0.392
Famous Affinities of History: The Romance of Devotion. Volume 1	Orr, Lyndon	Biographies	0.414

Eugene Field, a Study in Heredity and Contradictions — Volume 1	Thompson, Slason	Biographies	0.479
Eugene Field, a Study in Heredity and Contradictions — Volume 2	Thompson, Slason	Biographies	0.515
George Du Maurier, the Satirist of the Victorians	Wood, T. Martin	Biographies	0.504
Translations of Shakuntala and Other Works	Kalidasa	Best Books Ever Listings	0.391

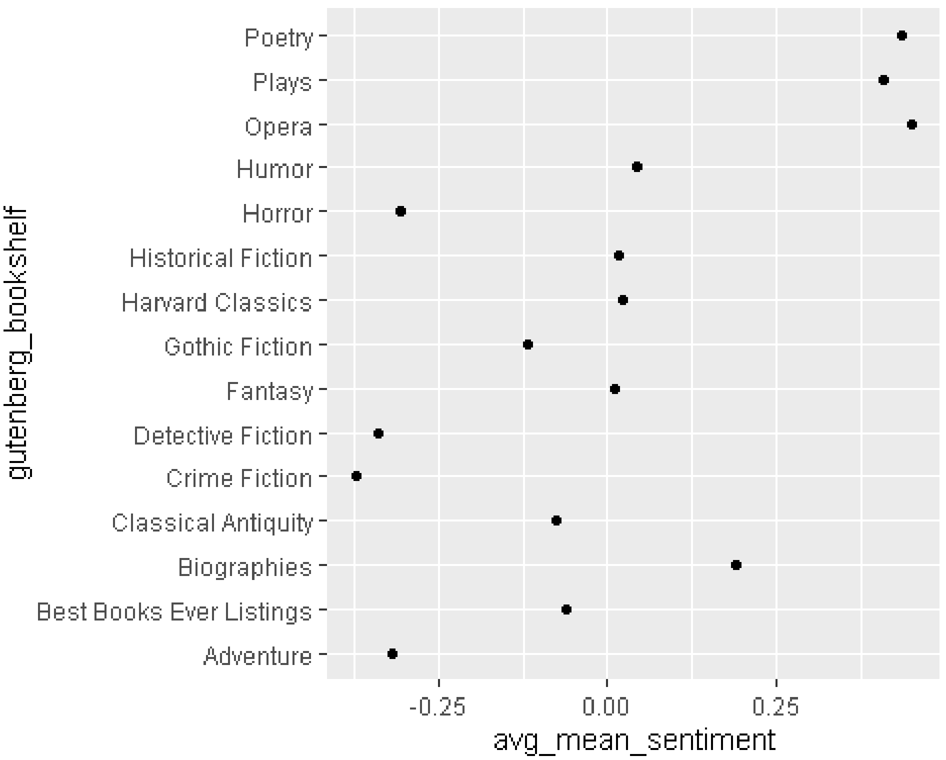
The top book in this genre is titled ‘Herland’ by Charlotte Perkins Gilman and the Sensitivity Score associated with it is 0.64. Herland is an utopian novel about an all-female inhabited imaginary place, free of war and conflict.

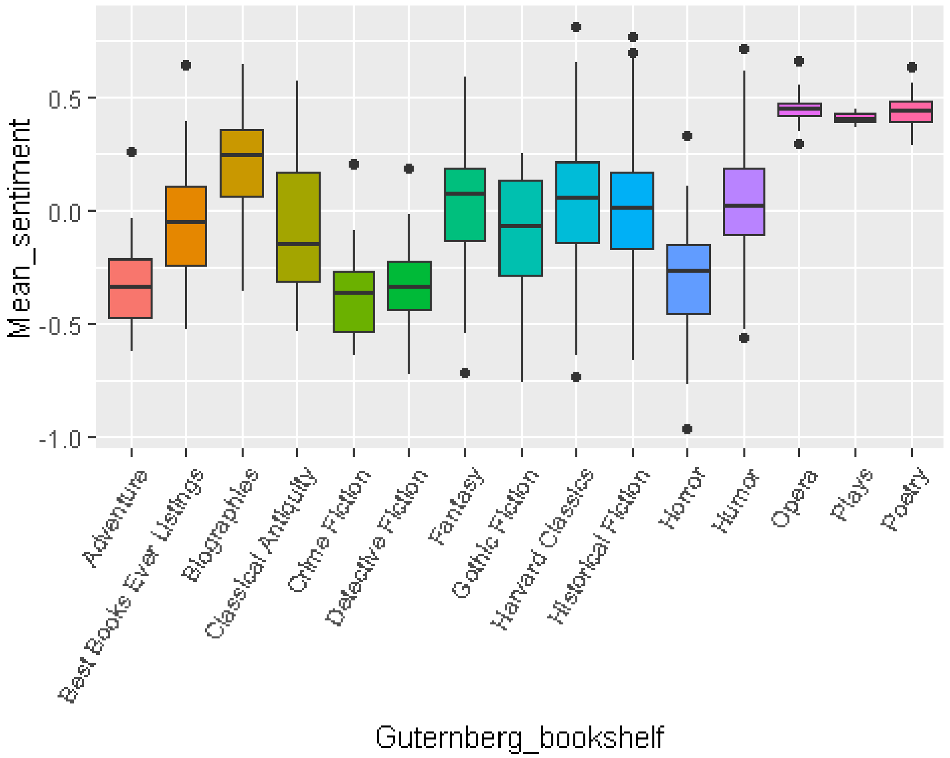
Next the Sensitivity mean scores in the five genre are compared in the table below. Lyric scores the highest followed by Great and Classic. Interestingly, Mystery and Fun(containing adventure books) do not score as high on positive sentiment scores. We need to go into the details at the sub-category level to find out the mean sentiment associated with each sub-category in each genre.

classic	lyric	fun	mystery	great
0.007	0.44	-0.01	-0.326	0.035

In this subsection, all the 934 books are combined into one large data set. Grouping by subcategories of genres, the average mean sentiment for all sub-categories is computed. It is easy to see that Poetry, Operas and Plays score very high on positive sentiment as do Biographies. For a number of sub-categories the average score is close to neutral (0 sentiment score). For Detective and Crime Fiction as well as Adventure Books, negative sentiment dominates.

Combining all the genres and then estimating the ten highest positive books gives us the following results.





title	author	gutenberg_bookshelf	mean_sentiment
The Rubaiyat of Ohow Dryyam			

With Apologies to Omar	Duff, J. L.	Humor	0.714
Poems 1817	Keats, John	Poetry	0.632
Some Forerunners of Italian Opera	Henderson, W. J. (William James)	Opera	0.659
Herland	Gilman, Charlotte Perkins	Best Books Ever Listings	0.640
Knights of Art: Stories of the Italian Painters	Steedman, Amy	Biographies	0.641
New Atlantis	Bacon, Francis	Harvard Classics	0.816
An Egyptian Princess — Volume 04	Ebers, Georg	Historical Fiction	0.765
Barbara Blomberg — Volume 04	Ebers, Georg	Historical Fiction	0.700
The Parisians — Volume 04	Lytton, Edward Bulwer Lytton, Baron	Historical Fiction	0.668
The Bible, King James version, Book 47: 2 Corinthians	Anonymous	Harvard Classics	0.655

A Deep Dive into the Relationship between Emotions and Books

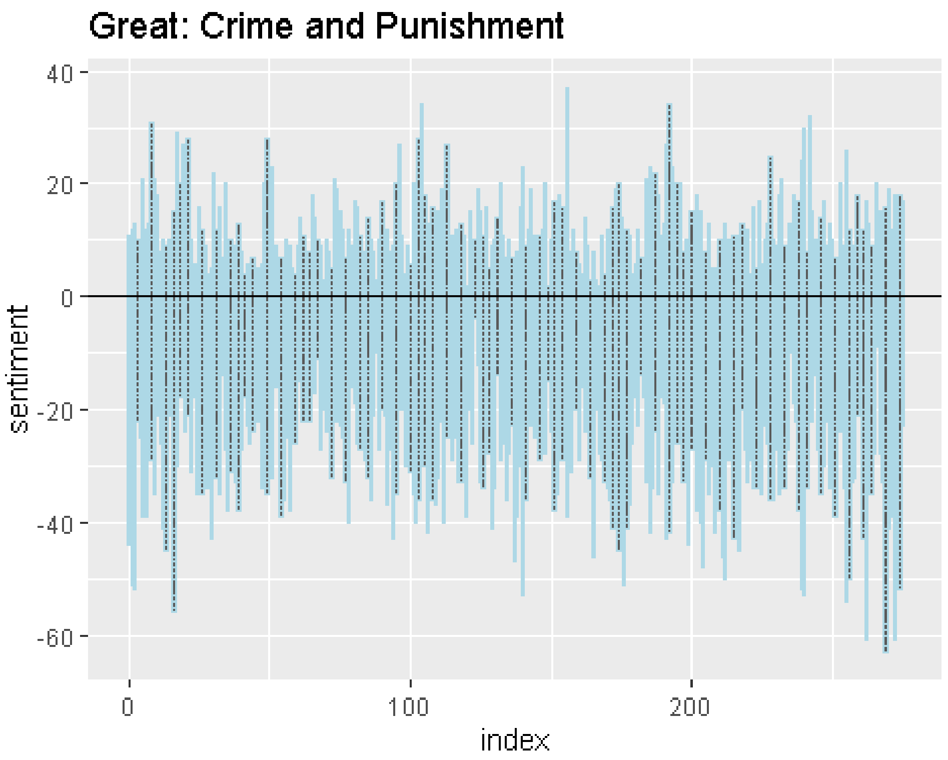
In this section, we will look at a sample of books, one from each of the five genres, and explore how books take us through a roller-coaster of positive and negative emotions as the narrative progresses. The horizontal axis in the charts represents the progression of the book and the vertical axis shows the positive and negative sentiments generated, using the BING lexicon.

BING Sentiment Lexicon was developed by Bing Liu and Collaborators and charateizes words in a binary positive or negative sentiments. Unlike the AFINN lexicon it doesn't assign a numerical sentiment value to words.

The Five books selected are as follows:

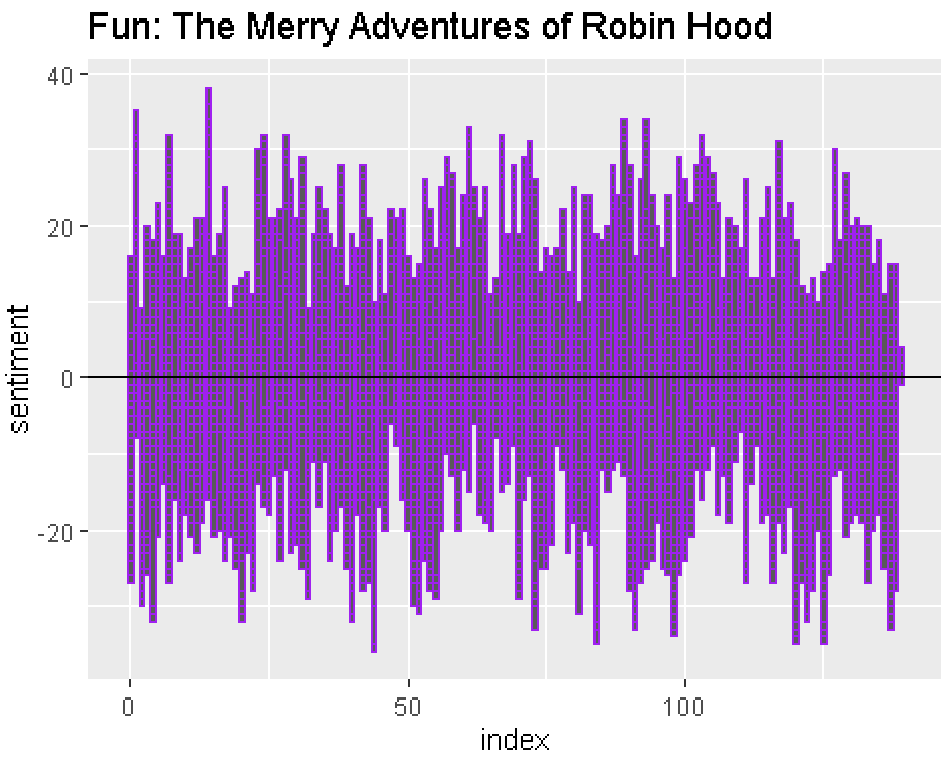
Book 1:Crime and Punishment: Genre: Great

```
## # A tibble: 1 x 8
##   gutenberg_id title author gutenberg_autho~ language gutenberg_books~
rights
##          <int> <chr> <chr>             <int> <chr>    <chr>
<chr>
## 1         2554 Crim~ Dosto~              314 en       Best Books Ever~
Publi~
## # ... with 1 more variable: has_text <lgl>
```



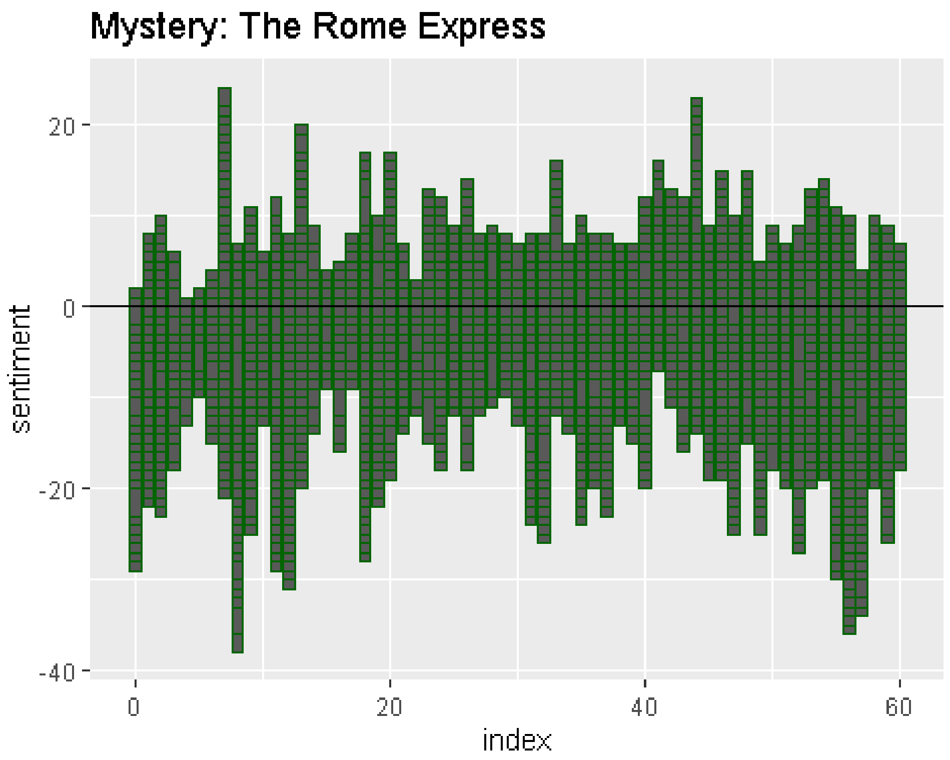
Book 2: The Merry Adventures of Robin Hood: Genre: Fun

```
## # A tibble: 1 x 8
##   gutenberg_id title author gutenberg_autho~ language gutenberg_books~
rights
##          <int> <chr> <chr>             <int> <chr>    <chr>
<chr>
## 1          964 The ~ Pyle,~              491 en       Fantasy
Publi~
## # ... with 1 more variable: has_text <lgl>
```



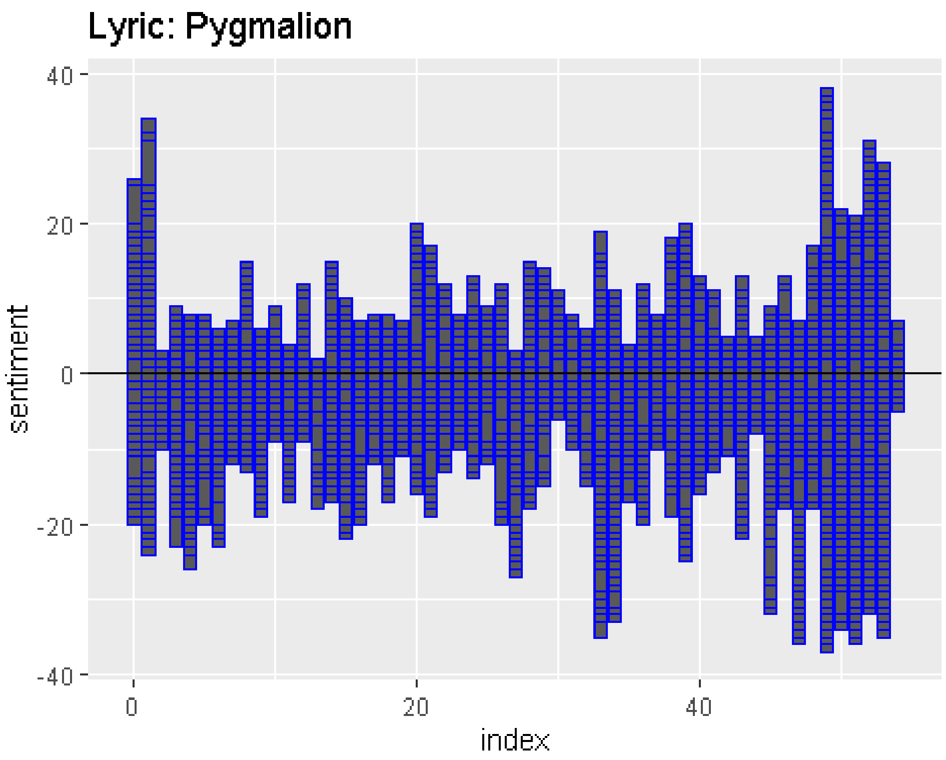
Book 3:Mystery : The Rome Express

```
## # A tibble: 1 x 8
##   gutenberg_id title author gutenberg_autho~ language gutenberg_books~
rights
##          <int> <chr> <chr>             <int> <chr>    <chr>
<chr>
## 1        11451 The ~ Griff~             3987 en       Detective Ficti~
Publi~
## # ... with 1 more variable: has_text <lgl>
```

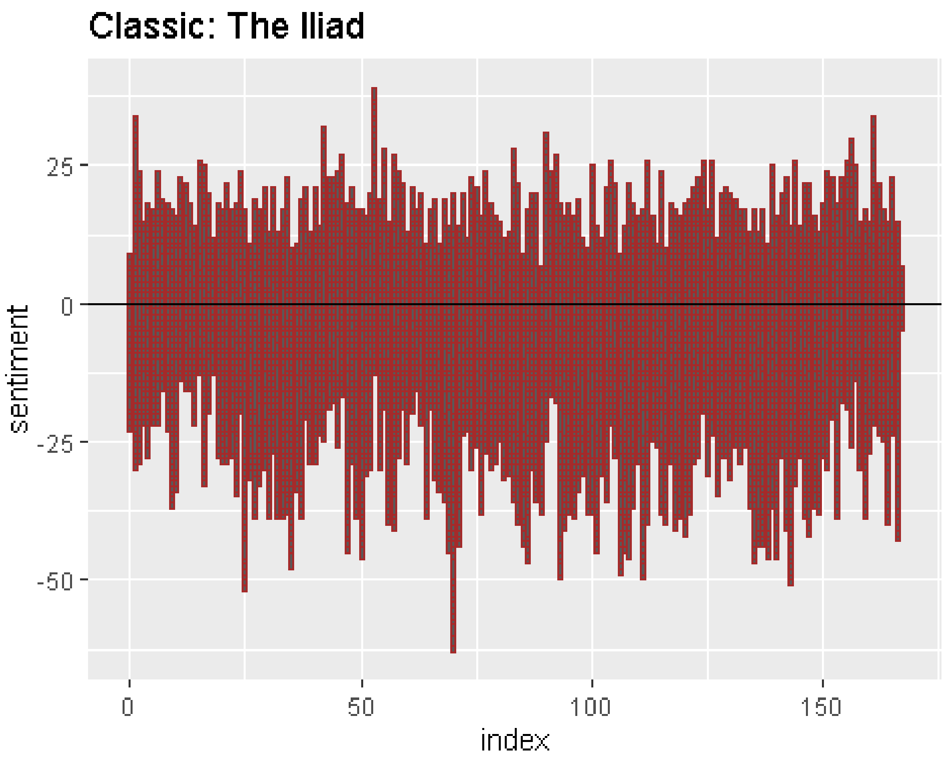


Book 4: Genre: Lyric: Plays “Pygmalion”

```
## # A tibble: 1 x 8
##   gutenberg_id title author gutenberg_autho~ language gutenberg_books~
rights
##          <int> <chr> <chr>             <int> <chr>    <chr>
<chr>
## 1         3825 Pygm~ Shaw,~              467 en       Plays
Publi~
## # ... with 1 more variable: has_text <lgl>
```

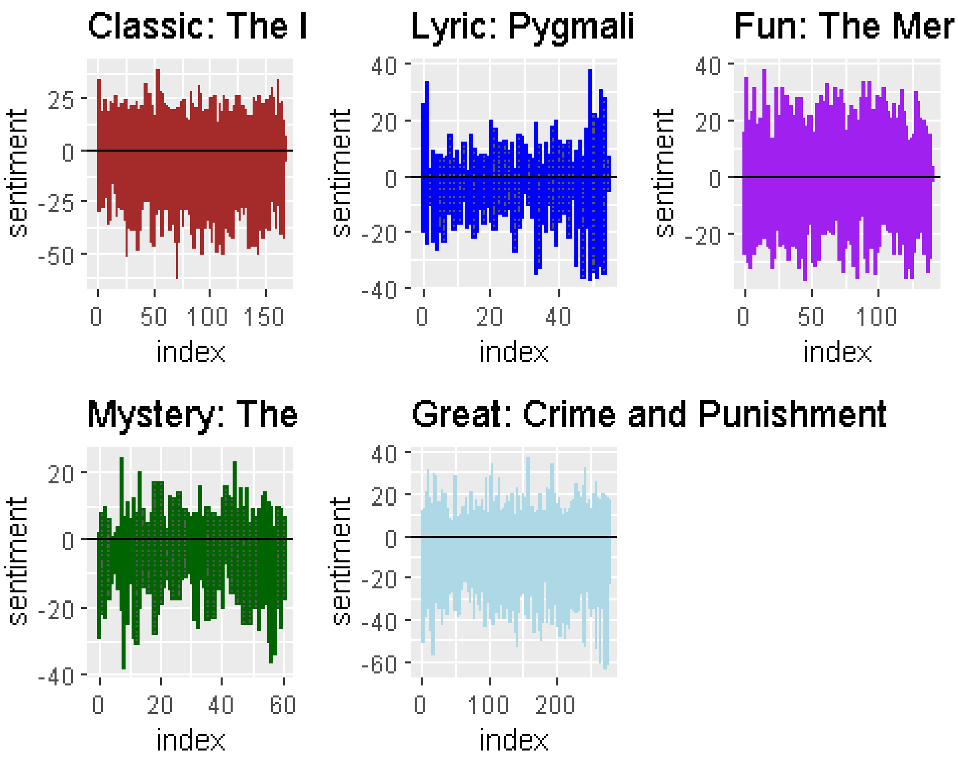


```
## # A tibble: 1 x 8
##   gutenberg_id title author gutenberg_autho~ language gutenberg_books~
rights
##          <int> <chr> <chr>             <int> <chr>    <chr>
<chr>
## 1         2199 The ~ Homer               705 en       Classical Antiq~
Publi~
## # ... with 1 more variable: has_text <lgl>
```



Combining the results below, we see that the Merry Adventures of Robin Hood has the most positive emotions while The Iliad and Crime and Punishment, both books touched the depths of negative sentiments.

The results table shows the proportion of positive sentiments over negative sentiments and the means are pretty close for all the five books. The results also support the charts.



```
## # A tibble: 5 x 2
##   method                                   mean
##   <chr>                                   <dbl>
## 1 Great:Crime and Punishment              0.298
## 2 Fun:The Merry Adventures of Robin Hood 0.476
## 3 Mystery:The Rome Express                0.341
## 4 Lyric: Pygmalion                        0.395
## 5 Classic:The Iliad                       0.374
```

Conclusion

The central **objective** of this report was to identify the best books to read for evoking positive emotions in the reader, based on Sensitivity Analysis, for ebooks freely available via Project Gutenberg. This report examined books from different genres to predict the positive sensitivity score, refrenced by the AFINN sentiment lexicon. Sixteen bookshelf categories were combined to create the five genres, Lyric, Classic, Great, Fun and Mystery.Top ten most uplifting books, based on the sensitivity scores, were selected within each genre as well as for the entire data set containing almost 1000 books.

In a deep dive exercise, a sample book was analyzed from each genre and the relationship between emotions and progression of narrative in these books was explored. It was found that while each book takes the reader on an emotional ride, some books, for example, The Merry Adventures of Robin Hood, have more highs than lows.

Limitations

The Gutenberg Project consists of over 50,000 works. Only a small subset of 934 books was explored in this project. The analysis could be expanded to more works as well as more free ebook libraries for example, Google Books and zlib.

References

- https://rafalab.github.io/dsbook/
- Project Gutenberg. (n.d.). Retrieved September 21, 2020, from www.gutenberg.org
- Finn Årup Nielsen, "A new ANEW: evaluation of a word list for sentiment analysis in microblogs", Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. Volume 718 in CEUR Workshop Proceedings: 93-98. 2011 May. Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, Mariann Hardey (editors)
- https://www.tidytextmining.com/sentiment.html