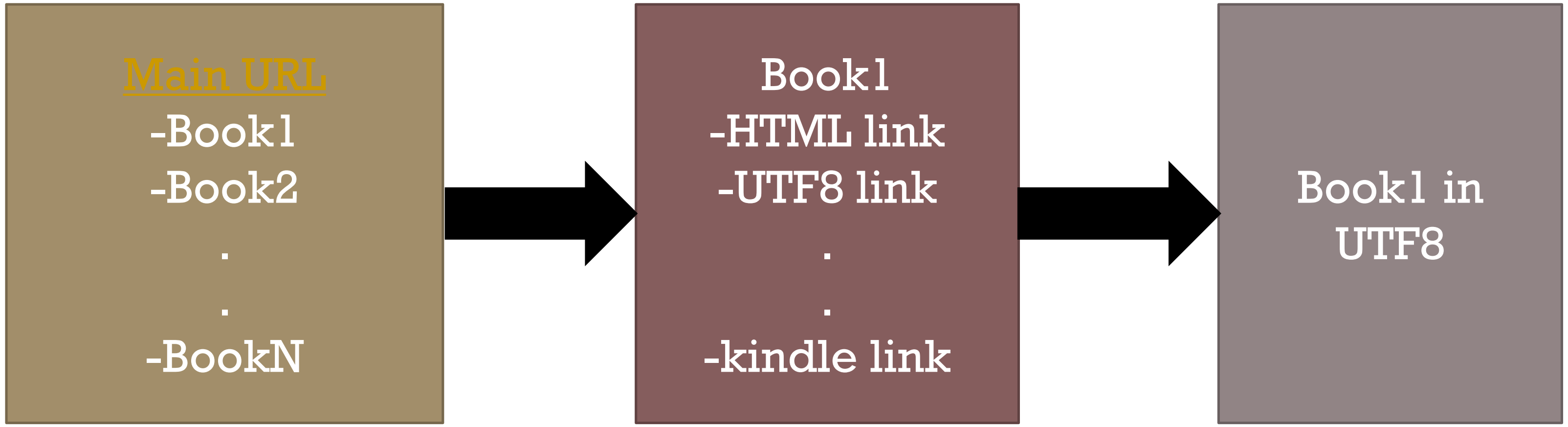
TEXT-ANALYSIS ON PROJECT GUTENBERG

Submitted By:

Tanvi Arora

HOW DO I GET TEXTBOOKS ?

- ABOUT PROJECT GUTENBERG :
  - Project Gutenberg was the first provider of free electronic books, or eBooks
  - It contains more than 59K ebooks completely free of charge in various categories like School Readers, Science & Nature, History , Geography ,etc.
  - Main Web page has links to 104 books in different categories and each book link redirects to another web page which has multiple links for the same book but in different formats

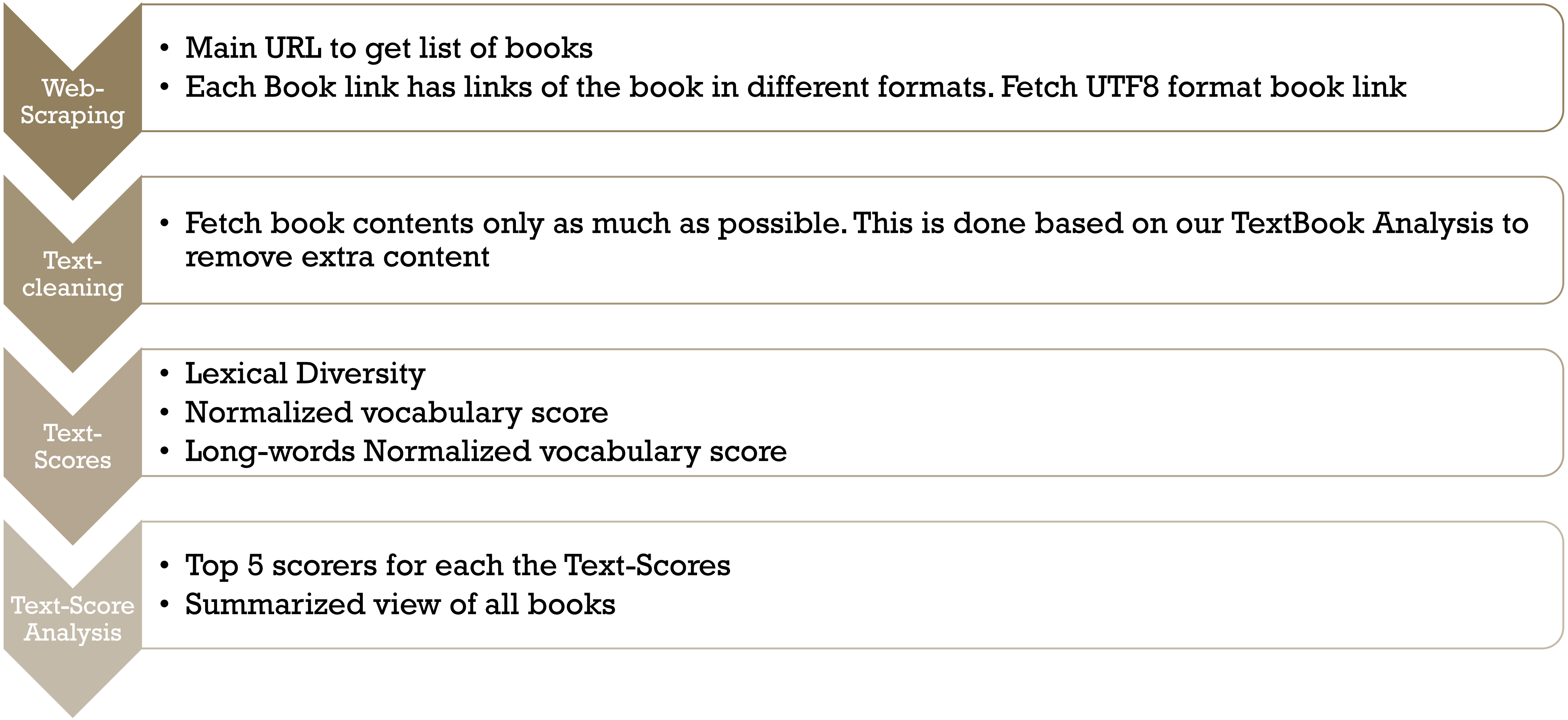


Main URL
-Book1
-Book2
.
.
-BookN

Book1
-HTML link
-UTF8 link
.
.
-kindle link

Book1 in UTF8

TASKS



Web-Scraping

- Main URL to get list of books
- Each Book link has links of the book in different formats. Fetch UTF8 format book link

Text-cleaning

- Fetch book contents only as much as possible. This is done based on our TextBook Analysis to remove extra content

Text-Scores

- Lexical Diversity
- Normalized vocabulary score
- Long-words Normalized vocabulary score

Text-Score Analysis

- Top 5 scorers for each the Text-Scores
- Summarized view of all books

TEXTBOOK ANALYSIS

START Tags

- Gutenberg ebooks contain lot of header information like licensing, book preface, acknowledgements, etc which actually do not contribute to the book content
- START Tags identified :
  - START OF THE PROJECT GUTENBERG , INDEX , PREFACE , CONTENT , ILLUSTRATION
- Starter position of 13000 was identified as maximum acceptable starter position. Maximum value among starter positions of above tags that are at position < 13000 is considered as the start position

Body

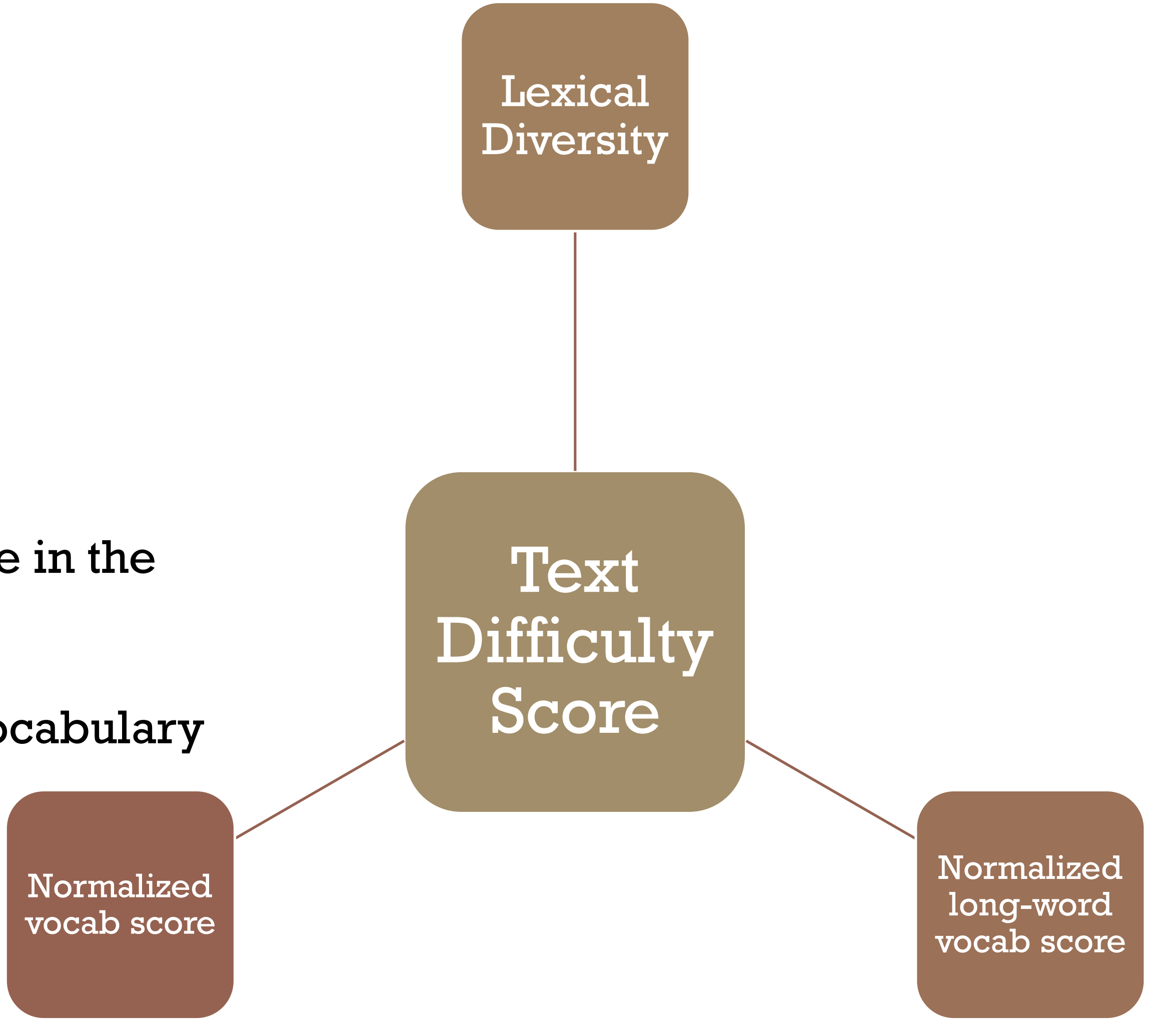
- Books contain lot of data – alphabetical words , numbers, alphanumeric words, punctuations. Which words contribute to vocabulary ? I have considered only alphabetical words. All others are ignored
- To perform case-insensitive analysis, convert all words to lower, that reduces duplicate words

END Tags

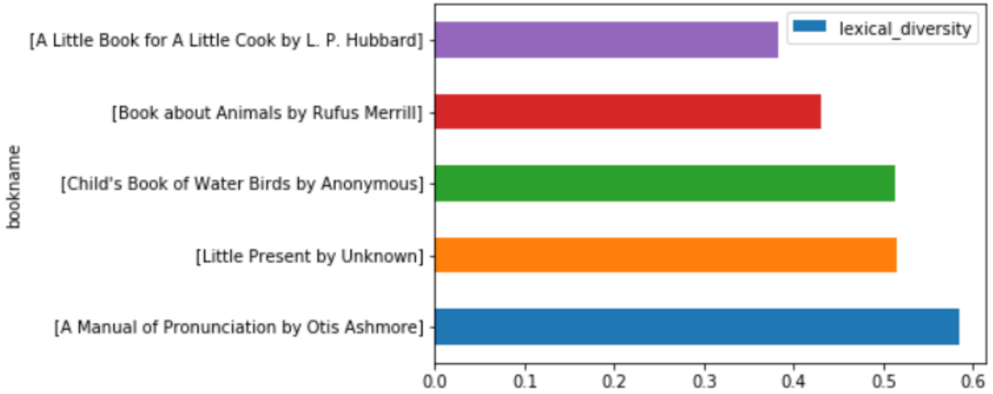
- Gutenberg ebooks also contain license information as footer of the book.
- END Tags identified :
  - End of the Project Gutenberg , END OF THIS PROJECT GUTENBERG
- If above tags are not found then end position is the length of the text i.e. end of the text. Minimum value amongst the end positions of above tags is considered as the end position

TEXT-SCORES

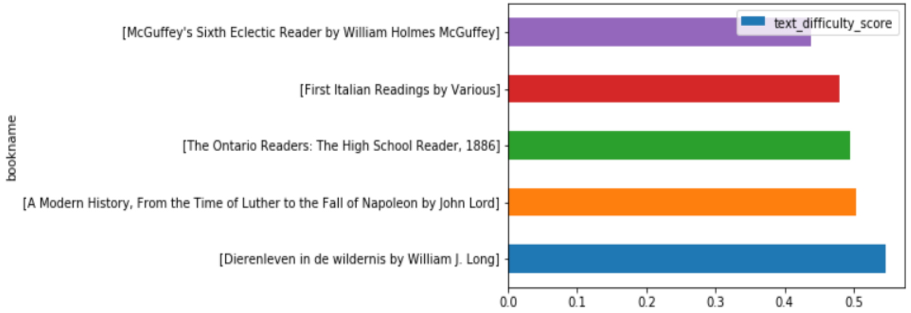
- Vocabulary size = number of distinct words
- Lexical Diversity : vocabulary size/number of words
- Normalized vocabulary score :
  - Raw score : vocabulary size of book/max vocabulary size in the collection
  - Sqrt score : sqrt of Raw score
  - Rank score : average of ranking in ascending order of vocabulary size
  - **(1/2) * (Sqrt score + Rank score)**
- Normalized long-word vocab score :
  - All above scores on the long-word vocabulary size
  - Long-word vocabulary size = number of words > 15 characters



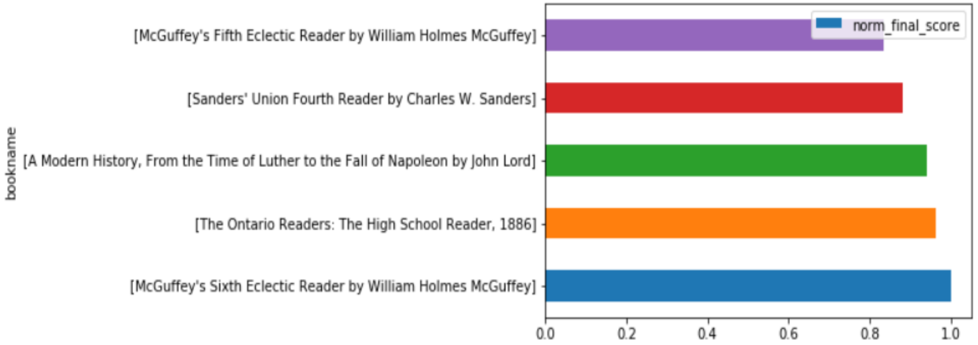
TEXT-SCORE ANALYSIS



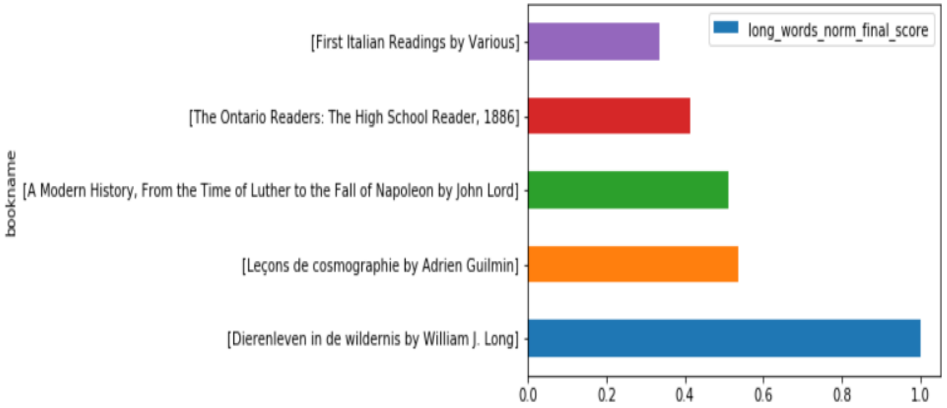
Highest lexical diversity is close to 0.58, looking at top 5 books based on their lexical diversity, they are quiet close



Text difficulty score tend to be biased towards the normalized vocabulary scores for all words plus long-words.None of the books with top lexical diversity are in the list of top 5 books with high difficulty score. As we know from our previous analysis that the 2 normalized scores have ranges , they affect the text difficulty score. But the text difficulty score is in the range similar to lexical diversity i.e. 0 to 0.6 and the top 5 books are not to far away from each other.

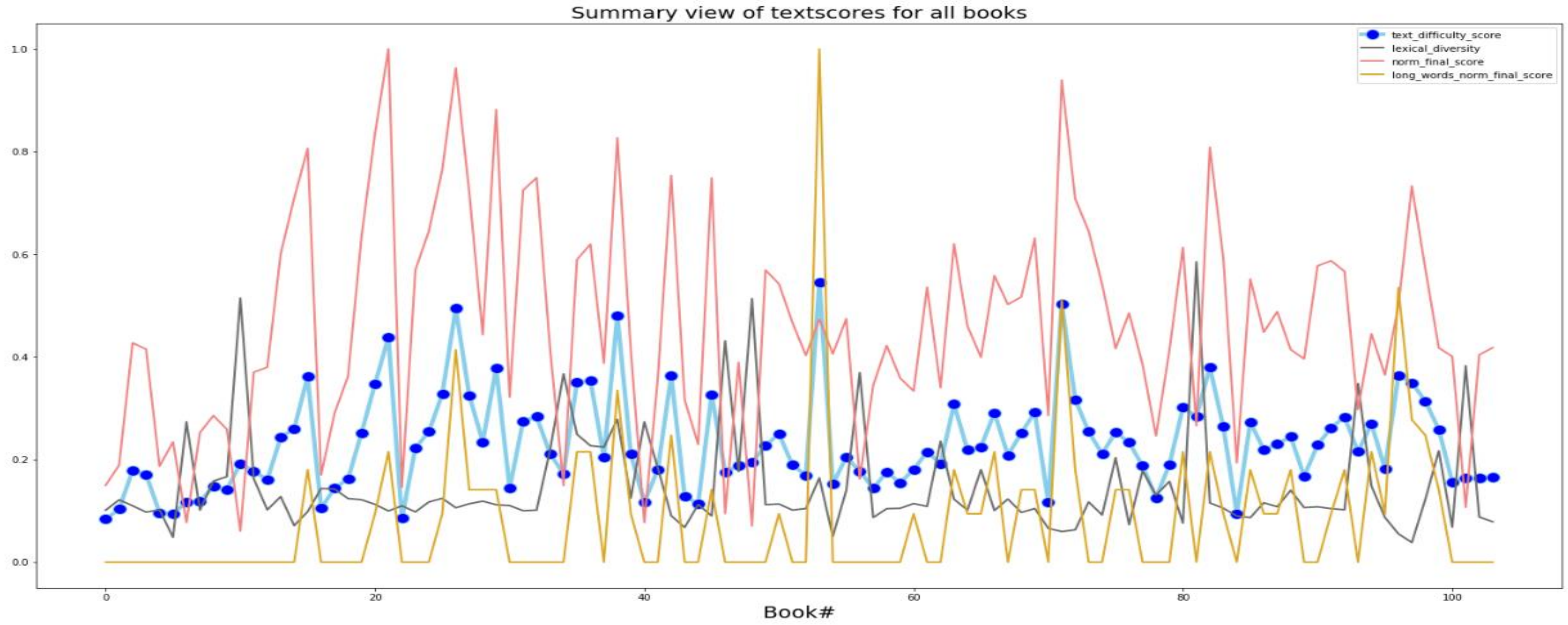


Normalized score is between 0 and 1 , and is based on the list of books being considered, so highest value obtained is 1 definitely. Looking at the top 5 books based on normalized vocabulary score, they are quiet close to each other. However interesting to note is the books with top 5 normalized vocabulary size are different than the top 5 books based on the lexical diversity



We have considered 15 char length as the lower limit of long-words . This analysis may change if the minimum length criteris is changed, more for the lower value than the higher. Normalized score is between 0 and 1 , and is based on the list of books being considered, so highest value obtained is 1 definitely. Looking at the top 5 books based on long-words normalized vocabulary score, it is a skewed plot with 1 book having the maximum number of long-words and second book to top it has 40% less long words. Also interesting to note is the books with top 5 long-words normalized vocabulary size are different than the top 5 books based on the lexical diversity but these have some common traits with the top 5 books of normalized vocabulary score

CONCLUSION



- For most of the books, normalized vocabulary score is high as compared to their lexical diversity. There are few exceptions though where lexical diversity is higher than the normalized vocabulary score. These are the books that are rich in their lexical diversity individually but when compared with other books on the Gutenberg project, these are not as rich in vocabulary.
- The book with highest text difficulty score has lower lexical diversity but a very high long words raw score, i.e. uses a lot of long words (words with length > 15). This increases its text difficulty level. Similarly there are books with 0 long-words , per our pre-assumed long-word length as 15 characters. This brings down its text difficulty score when compared to other books. This can change if our pre-assumed long-word length changes.
- The book with highest number of long-words is a Dutch book. We have not considered any translations for this analysis and for this book vocabulary is in Dutch. Language plays a factor in this case as Dutch language itself may have more long word as compared to English or other language

THANK - YOU