**Applying Text Mining Methodologies to Identify Text Characteristics of Bible**

Zhaofeng Tong, Qi Ying Yap, Wilson Ye

December 21, 2018

**Abstract**

Through statistical programming and text mining techniques, we analyzed six of the selected books in Bible: Genesis, Proverbs, Psalms, Matthew, John, and Romans. Aspects investigated include descriptive statistics, vocabulary usages, letter frequencies, and word associations. Our research result provides quantitative evidences to further facilitate religious studies. Furthermore, we investigated potential “markers” for each book, which work as a potential tool for machine learning techniques.

**Introduction**

* 1. **Background**

Text mining is one of the most emerging statistical techniques in the World. Through discovering and analyzing text patterns, usages, and associations, researchers may explain many social and economic aspects of great literatures throughout the history with statistical methods.

Text mining has been widely used in religious studies. Holy texts are oftentimes compiled with works throughout different time periods with different authors. Researchers investigate many aspects, such as word usages, word associations, syntax and diction to discover relationships between Holy texts and their historical context. Such methods apply to non-English languages as well. For example, researchers have investigated vocabulary usages in Arabic *Quran* (Sharaf 2009).

Bible is the most significant holy text in Christianity. It is not only a holy book for religion, but also an important history book for readers to learn about the cultural, social, and religious dynamic more than two thousand years ago. The original text of Bible was written in multiple languages and was translated into English after hundreds of years. Understanding the word use and syntax of holy text is crucial to develop an in-depth understanding of the context of Bible. In this research, we incorporated statistical knowledge to analyze six of the most popular Bible books: Genesis, Exodus, Psalms, Matthew, John, Romans; three of which come from the Old Testament, while three come from the New Testament. The result of this research provided us a snapshot of the textual features of books written in different times.

* 1. **Book Selection**

We select Bible as our primary source, not only because of its historical meaning and religious significance, but also because of its variety of word choices and formats.

Original Bible was written in Hebrew and Greek. Although text mining may be able to analyze texts in these languages, it still requires researchers possess knowledge in these languages. Therefore, we decided to use the English version as a proxy. While the English version is not the original piece, it still contains the original format and structure.

We select King James English version (KJV) as our primary source. It was published in 1611 and is in public domain. KJV is one of the most popular versions of Bible, and it contains variation of English wording since it was published over 400 years ago Using the KJV version also provides us a snapshot of how English usages have changed in the past few centuries[[1]](#endnote-1).

Because Bible has 66 books, we determined that it would be more feasible to analyze a few selected books. We select Genesis, Exodus, Psalms, Matthew, John, and Romans as our targets, because they are well-known books which contain unique historical information. We received this information from Bible Gateway, a well-known Bible resource center[[2]](#endnote-2).

* 1. **Introduction to Books**

Genesis is the starting chapter of Bible, one of the most well-known books. It describes the history of God’s creation of earth and humankind, as well as human’s early interactions with God.

Proverbs is a well-known old testament book which talks about Biblical applications in human life, including human character, ethics, and behaviors.

Psalms is an old testament book which uses the format of Psalms (poetry style) to worship and praise God.

Matthew is the first new testament book, which describes the life of Jesus Christ, his teaching, and his miracles. The book is significant because it is the first book talking about the story of Jesus.

John is a new testament book, which emphasizes the preaching of Jesus and his significance in representation of God.

Romans is a new testament book, which explains how Jesus Christ save human beings (salvation), as well as Paul’s (one of disciples of Jesus) opinions of faith.

* 1. **Summary Statistics**

The table below presents the summary statistics of six books. The first column identifies the number of verses in each book, ranging from Romans with the fewest verses (432) to Psalms with the largest number of verses (2460). To proceed, the length of books, represented by No. of words in the second column, also tells the same story with Psalms being numerically the largest chapter. Nevertheless, the average word length per verse shows a different story, with Genesis having the highest average number of words per verse, whereas Proverbs and Psalms have the least average number of words per verse.

|  |  |  |  |
| --- | --- | --- | --- |
| Book | No. of verses | No. of words | Avg. words per verse |
| Genesis | 1532 | 38243 | 24.96 |
| Proverbs | 914 | 15020 | 16.43 |
| Psalms | 2460 | 42671 | 17.34 |
| Matthew | 1070 | 23658 | 22.11 |
| John | 878 | 19058 | 21.71 |
| Romans | 432 | 9410 | 21.78 |

Furthermore, we examined the letter frequency for each book.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Book | Max Letter | No. of Max Letter | Min Letter | No. of Min Letter |
| Genesis | e | 11866 | q | 17 |
| Proverbs | e | 5387 | z | 1 |
| Psalms | e | 14696 | z | 49 |
| Matthew | e | 8202 | q | 20 |
| John | e | 6275 | q | 9 |
| Romans | e | 3199 | z | 3 |

**Data Cleaning**

We took a series of steps to clean our original text in order to prepare for our analysis. Steps include cleaning of the Bible text and transformation of the text into a corpus.

* 1. **Text Location**

We take the following steps to conduct text cleaning:

1. We read in the entire bible text and remove the leading white space in front of the verses.
2. We filter out book names from all 66 books using “*grep*” function and remove the number in front of book names, so that the format of the text is clearer, which allows us to apply more text data cleaning skills to the file.
3. We attach content of each of the 66 books to their respective book names using a “*for*” loop. Hence, instead of a giant Holy Bible object, we have 66 small objects. By typing out book name such as Genesis or Proverbs, we can easily find the content behind a book.
   1. **Corpus**

We take the following steps to create a corpus to the text we analyze. Corpus is an important step in text mining, as it creates a complex documental matrix for all the words in a specific document, making it easier for us to analyze the text patterns:

1. Use “*Quanteda*” package to create a corpus for each book.
2. Before further processing, we recognized that “God” and “god” have different meanings in Bible. However, Corpus does not recognize the differences between upper and lower cases. Therefore, we manually change all “God” to “cgod” to differentiate.
3. For consistency, we change all characters to lower case.
4. Remove all punctuations.
5. Remove stop words in English, such as “is”, “are”, “will”, etc. The statistical package includes a pre-set group of stop words. It is important to note that KJV is an older version of Bible, therefore it may contain old English stop words not recognized by modern statistical packages. We manually removed some common old English stop words. (i.e., unto, thy, thee, said, thou).

In general, all of these steps can be done in R’s text mining package under “*tm\_map*” function. After all these pre-possessing, we transformed the latest text into a corpus that is ready to apply more text mining skills.

**Initial Results**

We analyzed six books based on the following topics:

1. Single-word frequency
2. Bigram frequency
3. Letter Frequency
4. Notable Wording Combinations and Associations
   1. **Single-word Frequency**

Our initial results show that different books have different top words. For detailed results, we demonstrate top 20 words for each book in Appendix 1.1.

Genesis frequently mentions “God”, “lord”, “land”, and “came”. This may due to its content is closely related to God’s creation of heaven and earth.

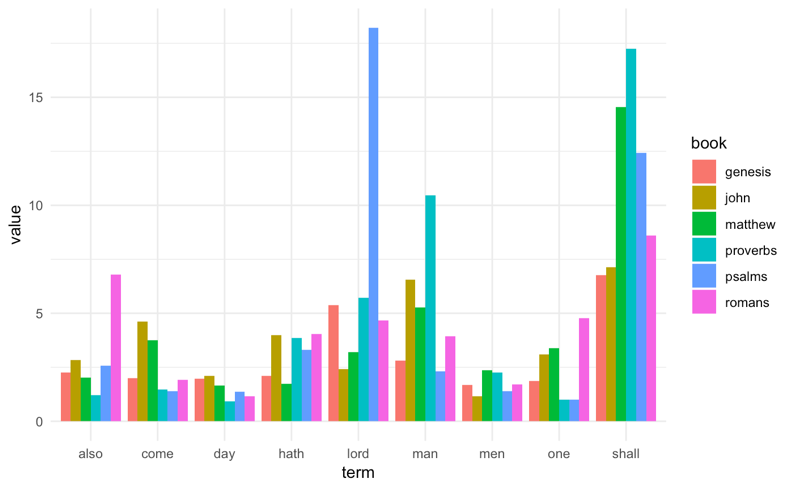
Proverbs frequently mentions “man”, “wicked”, “lord”, and “heart”. This may due to its content is closely related to human ethics. Since Bible places strong emphasis in one’s heart and one’s desire to follow God, it makes sense that these words appear frequently.

Psalms frequently mentions “lord”, “God”, “let”, “upon”, and “hast”. Many of these terms are often times found in poetry. It is also interesting that Psalms mentions “lord” 777 times, much more than other old Testament books.

All New Testament books frequently mention “Jesus”, along with “law”, “saying/say”, which are aligned with the primary topic of New Testament, which is the story and teaching of Jesus Christ. Old Testament mentioned Jesus zero times.

We recognized that since different books have different lengths, therefore, in order to compare across books, standardization is needed. We divide the frequency of a particular word by the total number of words in that book, and multiply by 1,000. Therefore, the standardized number means how many times does this word appear per 1,000 words in a book.

We also cross-analyzed word frequency across books, including: standardized top words across Old Testament (Appendix 1.2) and standardized top words across New Testament (Appendix 1.3).

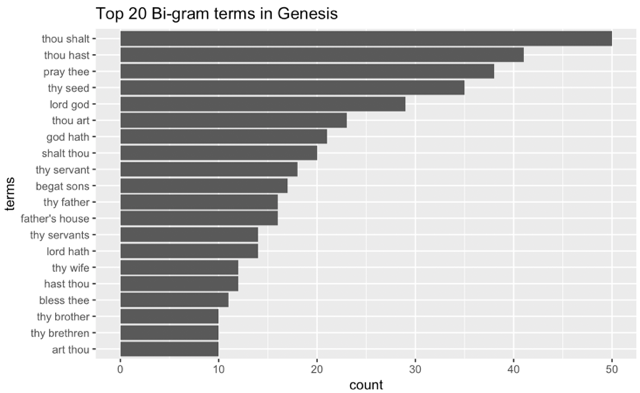


Graph 3.1.1. Standardized top word frequencies for 6 books.

* 1. **Double-word Frequency**

Following the single word frequency, we explore the popular double word frequency (“bi-gram”) in the 6 books. The concept is similar except that it is for two words. This analysis only counts for frequency, and it does not estimate the like-hood of association, as we will discuss in chapter 3.3.

Below we attach a graph of the top 20 bi-gram terms appeared in Genesis as an example of our analysis. From the graph, we can see “thou shall” and “thou hast” are the two most frequent terms in Genesis (The order of two terms doesn’t matter in this case). This result corresponds to our previous analysis on the single word frequency graph because thou and shall are also the most popular words in that scenario. Some of the interesting terms here include “father’s house” and “begat sons”.

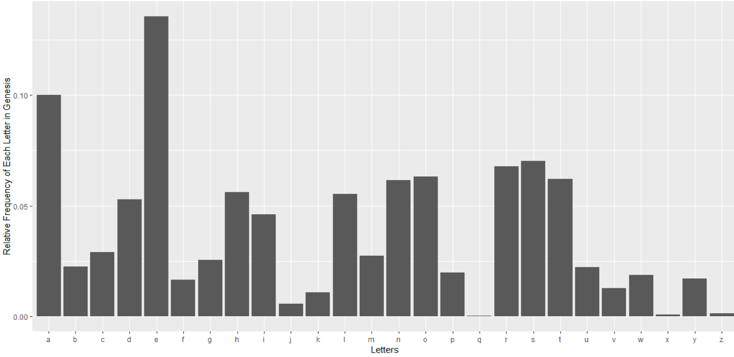


Graph 3.2.1. Double-word frequency for Genesis.

* 1. **Letter Frequency**

Another area of interest that we explored was the relative frequency of letters in the six books. Specifically, we sought to not only observe the differences of the proportions comparing between the different books in the Bible, but also compare these with the global relative letter frequency of the English language. Subsequently, we also ran proportion tests to examine the significance of these letter deviations from the global frequency.

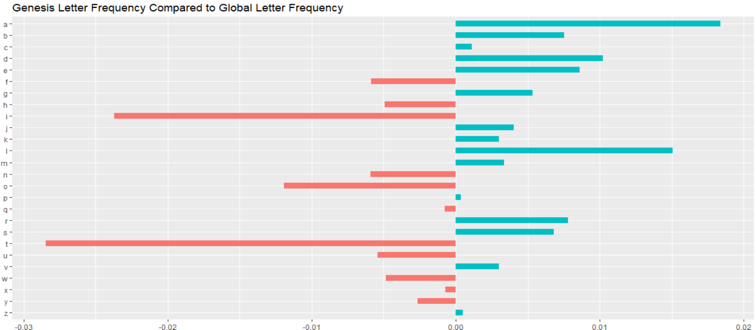
To conduct our tests, further data cleaning in R was necessary. In order to reflect the true letter frequency, we used the original text, instead of cleaned corpus, to conduct the analysis. As results are displayed, we manually filtered out unwanted symbols and punctuation, further splitting the words of each individual book into individual letters and forming a table with the relative frequency of each letter in one book in the Bible. A basic plot of the relative letter frequency of the book of Genesis has been provided below (Graph 3.3.1). For example, the letter ‘a’ was used a total of 8752 times while there was a total of 87481 letters used in the book of Genesis, resulting in a relative frequency of 10.004% (5 s.f.).



Graph 3.3.1. Letter frequency for Genesis

To obtain a more comprehensive understanding of how the letter profiles of the six Bible books compare to a global average, we used an online site for the global relative frequency of letters, a source also used by Agyepong, Buchanan and Jones in a 2018 paper. By comparing a vector of global letter relative frequencies to the relative frequencies of the letters in our paper, we were able to obtain a plot of how far each letter in a certain book of the Bible deviates from the global relative frequency. In the graph below (Figure 3.3.2), bars marked in blue represent letters of Genesis that appeared in greater relative frequencies than the relative frequencies of those same letters globally, and bars marked in red represent the opposite. Once again, we take the letter ‘a’ as an example and can observe that the difference between the relative frequency of ‘a’ in Genesis letter and the relative frequency of ‘a’ globally was 1.8375%. In fact, ‘a’ was the most overrepresented letter in the book of Genesis.

Finally, we conducted a two-proportion z-test to see if the deviations of these relative letter frequencies were large enough to be considered significant. In our code, we elect to use the total number of letters of each book of the Bible as our total counts of trials for both the global and actual letter relative frequencies, as well as the expected (based on global proportions) and actual number of a certain letter for our counts of successes. We first observe that by using a significance level of 0.05. A large majority of the relative frequencies of letters in the Bible differed significantly from the global average. For example, only 2/26 letters in Genesis had a p-value greater than 0.05.



Graph 3.3.2. Letter frequency for Genesis against global letter frequency. Red means underrepresented in Genesis, while Blue meaning overrepresented

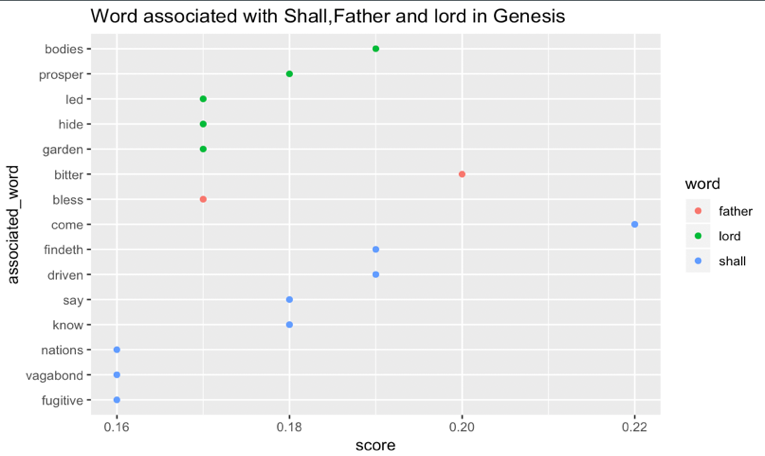
To verify our result, we converted the Z scores through logarithm and the result is consistent.

* 1. **Notable Associations**

We explore the associations between words in six books. Association answers an important question: “*Do we oftentimes see some particular words appear together? If so, how often*”?

We used R’s “*findAssocs*” function to find the association score between two words in a document term matrix. Typically, the score ranges from 0 to 1. 0 means no association (not possible to see these words appear together), while 1 means 100% association (always appear together).

In the case of six books we selected, the score tends to be lower, hovering around 0.1-0.2. In order to better facilitate our research, we select an example of Genesis. We picked three words: father, lord and shall, because these three words are the top words in Genesis, as we discussed in chapter 3.1. Our objective is to find the association score of these three words with respect to other words in Genesis. From the graph below, the word “shall” has the highest association score of 0.22 with the word “come”, which means shall and come are more often to be associated with each other. “nation”, “vagabond” and “fugitive” all have the lowest association score of 0.16.



Graph 3.4.1: Word Association with Shall, Father, and Lord in Genesis. For example, the association score between Father and Bless is 0.17, as shown in red dot.

**Discussions**

After our primary analysis of six books, we are able to find interesting patterns, which help us develop an in-depth understanding of the role of text mining strategy in statistical analysis.

**4.1 Potential Markers**

Markers are identifiers used to train machine to make prediction. In Bible, markers can be used to track what particular passage belongs to what book. Throughout our analysis, we identified some potential markers, including:

1. Particular word frequencies, for example, the frequency of Jesus.
2. Sentence length, for example, Psalms have significantly shorter sentence lengths since it is in poetry format.
3. Letter Frequency, since different books have different letter frequencies. We may compare the passage’s letter frequency profile against the existing profiles for each book.
4. Word Associations. Although our analysis is only primary, being able to find interesting word associations in a particular chapter can make it a useful marker.

We use Psalms as an example:

Through our primary results, we identify the following facts about Psalms, which may be used as markers:

1. On average, Psalms has fewer words per sentence, since it is largely in poetic format.
2. Psalms does not contain words found exclusively in New Testament, such as Jesus.
3. Psalms distinguishes itself from other books in letter frequency.
4. A few unique word associations can only be found in Psalms.

The program will evaluate the statistics of a piece of sample work against these markers and decide the likelihood of matching.

However, we recognize that markers have flaws which may prevent the program from matching correctly. For example, multiple books may have similar letter frequencies. In this case, such marker may be less significant. If a book contains huge variety of sentence structures, it will add confusion to the program.

**4.2 Future Study**

Although our study is informative and intellectually rewarding, there are many areas for future study.

1. Analyze more books in addition to six books. Bible is such a significant masterpiece, and different books may have different historical contexts which help us construct individualized markers.
2. Incorporate more historical context. One clear purpose of text mining is to help humans understand social context better. By incorporate historical background into analysis (for example, if we know Paul wrote these books, is his diction & word choice consistent?). We were able to go over primary screens over these historical contexts (such as patterns between Old and New testament), but more research is necessary to explore the relationships between text usage and historical backgrounds.
3. Analyze Bible in different languages. The original Bible is written in multiple languages, including Hebrew and Greek, but certainly not English. Our research would be more significant if we are able to obtain the original piece of Hebrew. Currently our statistical packages have very limited support over Hebrew.
4. Markers can be investigated further. For example, although Psalms has a few strong markers as we identified in chapter 4.2, we may further break it down into specific syntax and sentence structures, such as classifying sentences based on characteristics and types, and apply different markers to different types of sentences, thus, accuracy of prediction can be increased.

**4.3 Acknowledgements**

We would like to thank Professor Akram Almohawas from UCLA Department of Statistics, who generously offered us research guidance and technical instructions.

**Appendix**

Appendix 1.1. Single-word frequency (not standardized)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Genesis | | Proverbs | | Psalms | | Matthews | | John | | Romans | |
|  | term | freq | term | freq | term | freq | term | freq | term | freq | term | freq |
| 1 | shall | 259 | shall | 259 | lord | 777 | shall | 344 | jesus | 255 | cgod | 162 |
| 2 | cgod | 230 | man | 157 | shall | 530 | jesus | 172 | shall | 136 | shall | 81 |
| 3 | lord | 206 | wicked | 89 | cgod | 424 | man | 125 | father | 126 | law | 78 |
| 4 | land | 187 | lord | 86 | let | 238 | saying | 120 | man | 125 | christ | 66 |
| 5 | came | 176 | heart | 81 | upon | 194 | say | 117 | saith | 113 | also | 64 |
| 6 | father | 169 | wise | 66 | hast | 189 | came | 107 | come | 88 | one | 45 |
| 7 | jacob | 166 | hath | 58 | praise | 156 | come | 89 | cgod | 82 | sin | 45 |
| 8 | sons | 158 | way | 57 | ever | 146 | one | 80 | world | 79 | lord | 44 |
| 9 | son | 148 | thine | 55 | earth | 141 | son | 78 | answered | 77 | faith | 39 |
| 10 | upon | 141 | understanding | 54 | hath | 141 | lord | 76 | therefore | 77 | righteousness | 39 |
| 11 | joseph | 137 | wisdom | 54 | mine | 135 | heaven | 76 | hath | 76 | hath | 38 |
| 12 | earth | 121 | righteous | 54 | soul | 132 | disciples | 71 | now | 75 | things | 38 |
| 13 | abraham | 121 | mouth | 51 | people | 130 | went | 70 | know | 71 | jesus | 37 |
| 14 | behold | 118 | evil | 50 | heart | 123 | behold | 60 | things | 69 | man | 37 |
| 15 | years | 112 | son | 44 | also | 110 | father | 60 | jews | 68 | now | 35 |
| 16 | went | 110 | good | 43 | like | 104 | things | 56 | disciples | 65 | spirit | 29 |
| 17 | man | 107 | knowledge | 42 | mercy | 100 | men | 56 | came | 63 | might | 27 |
| 18 | name | 101 | lips | 42 | name | 100 | kingdom | 56 | son | 63 | therefore | 27 |
| 19 | wife | 101 | words | 40 | man | 99 | cgod | 55 | one | 59 | good | 27 |
| 20 | called | 98 | life | 38 | hand | 97 | therefore | 49 | sent | 57 | flesh | 26 |

Appendix 1.2. Top words in Old Testament (words appeared as top 100 words across all of 3 Old Testament Books)

|  |  |  |  |
| --- | --- | --- | --- |
| term | genesis | proverbs | psalms |
| shall | 6.77283544 | 17.2436751 | 12.4206135 |
| lord | 5.38688842 | 5.72569907 | 18.2090881 |
| upon | 3.68714207 | 1.99733688 | 4.54641325 |
| earth | 3.1641432 | 1.06524634 | 3.3043519 |
| man | 2.79804398 | 10.4527297 | 2.32007687 |
| let | 2.43194477 | 2.13049268 | 5.57755853 |
| hand | 2.30119505 | 1.6644474 | 2.27320663 |
| also | 2.24889516 | 1.19840213 | 2.57786319 |
| house | 2.24889516 | 1.86418109 | 1.12488575 |
| hath | 2.0919955 | 3.86151798 | 3.3043519 |
| come | 1.98739573 | 1.46471372 | 1.40610719 |
| day | 1.96124578 | 0.93209055 | 1.35923695 |
| shalt | 1.93509584 | 1.26498003 | 1.26549647 |
| one | 1.85664601 | 0.99866844 | 1.00771015 |
| give | 1.75204623 | 1.26498003 | 1.68732863 |
| men | 1.6735964 | 2.26364847 | 1.40610719 |
| make | 1.2551973 | 0.86551265 | 1.61702327 |
| good | 1.17674747 | 2.86284953 | 1.47641255 |
| forth | 1.12444758 | 1.06524634 | 1.10145063 |
| away | 1.01984781 | 1.53129161 | 1.03114527 |
| eyes | 1.01984781 | 1.99733688 | 1.03114527 |
| way | 0.94139798 | 3.79494008 | 1.24206135 |

Appendix 1.3. Top words in New Testament (words appeared as top 100 words across all of 3 New Testament Books)

|  |  |  |  |
| --- | --- | --- | --- |
| term | matthew | john | romans |
| also | 2.028912 | 2.83345577 | 6.80127524 |
| among | 0.887649 | 1.20684227 | 1.06269926 |
| called | 1.521684 | 1.15437087 | 1.16896918 |
| cgod | 2.324795 | 4.30265505 | 17.2157279 |
| come | 3.76194099 | 4.61748347 | 1.91285866 |
| day | 1.648491 | 2.09885612 | 1.16896918 |
| even | 1.225801 | 1.31178508 | 2.65674814 |
| father | 2.53613999 | 6.61139679 | 1.27523911 |
| hath | 1.733029 | 3.98782663 | 4.03825717 |
| jesus | 7.27026799 | 13.3802078 | 3.93198725 |
| know | 0.887649 | 3.72546962 | 1.27523911 |
| lord | 3.21244399 | 2.41368454 | 4.67587673 |
| made | 1.056725 | 1.57414209 | 2.55047821 |
| man | 5.28362499 | 6.55892539 | 3.93198725 |
| many | 1.648491 | 1.52167069 | 1.70031881 |
| men | 2.367064 | 1.15437087 | 1.70031881 |
| might | 0.803111 | 1.94144191 | 2.86928799 |
| neither | 1.394877 | 0.83954245 | 0.95642933 |
| now | 1.394877 | 3.93535523 | 3.7194474 |
| one | 3.38151999 | 3.09581278 | 4.78214665 |
| people | 1.014456 | 1.04942806 | 0.95642933 |
| saith | 1.986643 | 5.92926855 | 2.01912859 |
| say | 4.94547299 | 2.72851296 | 2.23166844 |
| shall | 14.540536 | 7.13611082 | 8.60786397 |
| son | 3.29698199 | 3.30569839 | 0.8501594 |
| therefore | 2.071181 | 4.04029804 | 2.86928799 |
| things | 2.367064 | 3.62052681 | 4.03825717 |
| yet | 0.803111 | 1.94144191 | 1.27523911 |

**References**

Kasper Welbers, Wouter Van Atteveldt & Kenneth Benoit (2017) Text Analysis in R, Communication Methods and Measures, 11:4, 245-265, DOI: [10.1080/19312458.2017.1387238](https://doi.org/10.1080/19312458.2017.1387238)

<http://en.algoritmy.net/article/40379/Letter-frequency-English> (other website) to get the global letter relative frequency (was quoted in a 2018 study)

<https://www.researchgate.net/figure/Relative-Frequency-of-Letters-in-the-English-Language_fig2_325714929> (that study)

<http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Diverging%20Bars> for idea of how to create code to compare to a single point

1. References from Bible Gateway, https://www.biblegateway.com/versions/King-James-Version-KJV-Bible/ [↑](#endnote-ref-1)
2. Results drawn from Bible Gateway, https://www.biblegateway.com/blog/2014/04/the-10-most-popular-books-of-the-bible-and-why/ [↑](#endnote-ref-2)