The Word

A textual analysis of the King James Version, American Standard Version, and World English Bible.

Wesley Chioh | wc1728

April 30, 2020

Word Count: 2275 (excl. references and appendix)

Page Count: 17 (Cover Page:1, Report: 10, References: 1, Appendix: 5)

Github Repository: [www.github.com/westerleyy/bible\_study](http://www.github.com/westerleyy/bible_study)

Introduction

The New York Times copyrighted bestseller list changes every week. But on the time scale of centuries and millennia, the longstanding global copyright-free bestseller has not. It is most probably the Bible. The Bible commonly used today is quite different in style and phrasing from that a mere century ago. It undergoes revisions for clarity and ease of comprehension as linguistic norms change. The subject matter and meaning of its verses have not, however. But have they?

The Bible is a “multi-parallel corpora” (Xia and Yarowsky, 2017, p.448) with multiple versions of what is essentially a highly similar corpus. There are 66 books in the Protestant Bible, each of which is distinct from another in terms of authorship, stylometry and topics. Some, such as the four Gospels of Matthew, Mark, Luke, and John are among the most similar in terms of content and topics, despite being written by different authors. Of the four Gospels, Matthew, Mark, and Luke are also known as the Synoptic Gospels (Linmans, 1998; Murai, 2006) with corresponding sections. Murai (2013) argues that from a network analysis perspective, they can be characterized as a series of multiple one-to-many relationships.

Thus, this paper will seek to test the hypothesis that despite controlling for topical similarity, the four Gospel books remain stylistically distinct from another. This distinction is retained over the ages and various versions of the Bible. This paper will use the King James Version (KJV), American Standard Version (ASV), and World English Bible (WEB) as the basis of comparison.

Similarity

First, a spreadsheet of various versions of the Bible was downloaded from BibleHub and the KJV, ASV, and WEB were selected. The preprocessing involved lower-casing and stemming each word, and removing punctuation and numerals. Stopwords were not removed because the English stopwords available through quanteda are contemporaneous with English language norms today but not with the Elizabethan English of the KJV and ASV (Hamlin, 2011). Furthermore, removal of stopwords is likely to result in document-feature matrices (DFMs) with high frequencies of proper nouns in the modern WEB, as well as other stopwords in the older KJV and ASV. This results in an imbalance that might adversely affect a true assessment of their similarities and differences.

From a bag of words perspective, word choice and frequency contain latent information. Cosine similarity thus hints at the relative similarity of the four books in terms of topics, and sentiments.

| Table 1: Cosine Similarity of KJV Gospel Books | | | | |
| --- | --- | --- | --- | --- |
|  | **Matthew** | **Mark** | **Luke** | **John** |
| Matthew | 100.0 | 97.7 | 83.9 | 80.4 |
| Mark | 97.7 | 100.0 | 90.3 | 86.2 |
| Luke | 83.9 | 90.3 | 100.0 | 95.8 |
| John | 80.4 | 86.2 | 95.8 | 100.0 |

| Table 2: Cosine Similarity of ASV Gospel Books | | | | |
| --- | --- | --- | --- | --- |
|  | **Matthew** | **Mark** | **Luke** | **John** |
| Matthew | 100.0 | 92.9 | 91.1 | 94.0 |
| Mark | 92.9 | 100.0 | 98.5 | 98.5 |
| Luke | 91.1 | 98.5 | 100.0 | 97.6 |
| John | 94.0 | 98.5 | 97.6 | 100.0 |

| Table 3: Cosine Similarity of WEB Gospel Books | | | | |
| --- | --- | --- | --- | --- |
|  | **Matthew** | **Mark** | **Luke** | **John** |
| Matthew | 100.0 | 94.8 | 93.1 | 94.8 |
| Mark | 94.8 | 100.0 | 98.3 | 98.6 |
| Luke | 93.1 | 98.3 | 100.0 | 97.2 |
| John | 94.8 | 98.6 | 97.2 | 100.0 |

From the tables above, the average cosine similarity of the four Gospel books across the KJV, ASV, and WEB is 91.79%, 96.58%, 97.1% respectively. This is fairly high, but they are not identically so. The degree of similarity between the books is much higher for the ASV and WEB as compared to the KJV. The ASV and WEB agree on the relative similarity of Matthew, Mark, and Luke, but not the KJV. For example, in the KJV, Matthew is most similar to Mark. But in both the ASV and WEB, Mathew is most similar to John. Furthermore, the ASV and WEB suggest that Mark and Luke are more than 98% similar; but in the KJV, it is Matthew which is almost identical to Mark, not Luke. However, the three versions agree that John is most similar to Mark and Luke.

Separability

The second part of this paper builds on the understanding that although the three versions of the Bible and the four Gospels might very well share similar content, they are ultimately distinct. This can be attributed to stylistic differences among the different authors, and translations and word choice contemporaneous to the day and age of each version.

To test this hypothesis, random forest classifiers were fitted on a sample of verses from each of the four books across all three versions. The training and test data split take into account the slight class imbalance (Table A1) by ensuring that every class is sampled according to its distribution in the population of verses. Three random forest classifiers were fitted:  
1. Classify verses based on the version and the book  
2. Classify verses based on the version  
3. Classify verses based on the book

| Table 4: Model (Book and Version) Results | | | | |
| --- | --- | --- | --- | --- |
|  | **Balanced Accuracy** | **Precision** | **Recall** | **F1** |
| Class: A\_Joh | 0.5339414 | 0.1366906 | 0.1444867 | 0.1404806 |
| Class: A\_Luk | 0.5277186 | 0.1396896 | 0.1826087 | 0.1582915 |
| Class: A\_Mar | 0.5120476 | 0.0862069 | 0.0738916 | 0.0795756 |
| Class: A\_Mat | 0.5023220 | 0.0983146 | 0.1090343 | 0.1033973 |
| Class: K\_Joh | 0.5347952 | 0.1538462 | 0.1292776 | 0.1404959 |
| Class: K\_Luk | 0.5393798 | 0.1675676 | 0.1797101 | 0.1734266 |
| Class: K\_Mar | 0.5046291 | 0.0743802 | 0.0443350 | 0.0555556 |
| Class: K\_Mat | 0.5285875 | 0.1445783 | 0.1495327 | 0.1470138 |
| Class: W\_Joh | 0.6875946 | 0.4698276 | 0.4144487 | 0.4404040 |
| Class: W\_Luk | 0.6442193 | 0.3052392 | 0.3884058 | 0.3418367 |
| Class: W\_Mar | 0.5311812 | 0.1764706 | 0.0886700 | 0.1180328 |
| Class: W\_Mat | 0.6165504 | 0.3062500 | 0.3052960 | 0.3057722 |

With 12 classes to predict, the out-of-sample accuracy is 0.195. It is a slight improvement over its base rate of 0.102, which is obtained by simply predicting the most commonly occurring class in the dataset. The confusion matrix (Table A2) suggests that verses from the ASV and KJV are highly likely to be misclassified as the other. Verses from John are least likely to be misattributed to others whereas verses from Luke and Mark are especially prone to misattribution. This analysis can be further broken down into the versions and the books with fewer classes in each case.

| Table 5: Model (Version) Results | | | | |
| --- | --- | --- | --- | --- |
|  | **Balanced Accuracy** | **Precision** | **Recall** | **F1** |
| Class: American Standard Version | 0.5308914 | 0.3710590 | 0.4051192 | 0.3873418 |
| Class: King James Bible | 0.5778906 | 0.4542652 | 0.3901147 | 0.4197531 |
| Class: World English Bible | 0.8413504 | 0.7670311 | 0.8049426 | 0.7855297 |

In the previous model, it was observed that the poor performance of the classifier could be attributed to errors in separating the versions. Training another random forest model to just classify the verses based on its version, the out-of-sample accuracy is 0.195 is similar to the previous model. Further, the precision and recall rates are especially poor for the ASV and KJV as compared to the WEB. This confirms the notion that the KJV and ASV are highly similar and difficult to classify.

| Table 6: Model (Book) Results | | | | |
| --- | --- | --- | --- | --- |
|  | **Balanced Accuracy** | **Precision** | **Recall** | **F1** |
| Class: John | 0.9801711 | 0.9733164 | 0.9683944 | 0.9708492 |
| Class: Luke | 0.9517025 | 0.8908927 | 0.9545894 | 0.9216418 |
| Class: Mark | 0.9253081 | 0.9493671 | 0.8606557 | 0.9028375 |
| Class: Matthew | 0.9432622 | 0.9273684 | 0.9148494 | 0.9210664 |

The out of sample accuracy for predicting the book given a verse is 0.93. Of the three classifiers trained and tested thus far, this shows the highest degree of accuracy. However, given that there is a class imbalance, the F1 score as shown above might be more appropriate instead. The high degree of precision in book classification suggests that distinctions between books are relatively strong and retained over the various versions. However, the book of Mark has an uncharacteristically low recall score of 0.861. This suggests that verses in Mark tend to be misattributed to other books.

Clusters

To test this further, kmeans clustering was performed to test the ability of the algorithm to separate the corpus into four distinct clusters, each corresponding to a book. Collapsing verses into chapters can provide a better approximation of chapter meaning as they can be relatively short and easily taken out of context. Furthermore, stylometric signatures should be stronger at the chapter level.

To reduce dimensionality, principal components analysis was performed. The first two dimensions account for less than 7% of the variability. Thus, principal components analysis is not particularly useful as the trade-off between dimensionality and variance is extremely steep.

As a baseline of comparison, the Cosine similarity between each book was calculated and plotted. For the most part, books and chapters are about 60 to 70% similar except certain chapters that echo each other content-wise such as Mark 11 and Matthew 21. Curiously, Luke 19 and 20 are dissimilar, possibly because the same content is split between two chapters. It also appears that Luke 3, Matthew 1, and John 13 through 17 are relatively distinct from the rest.

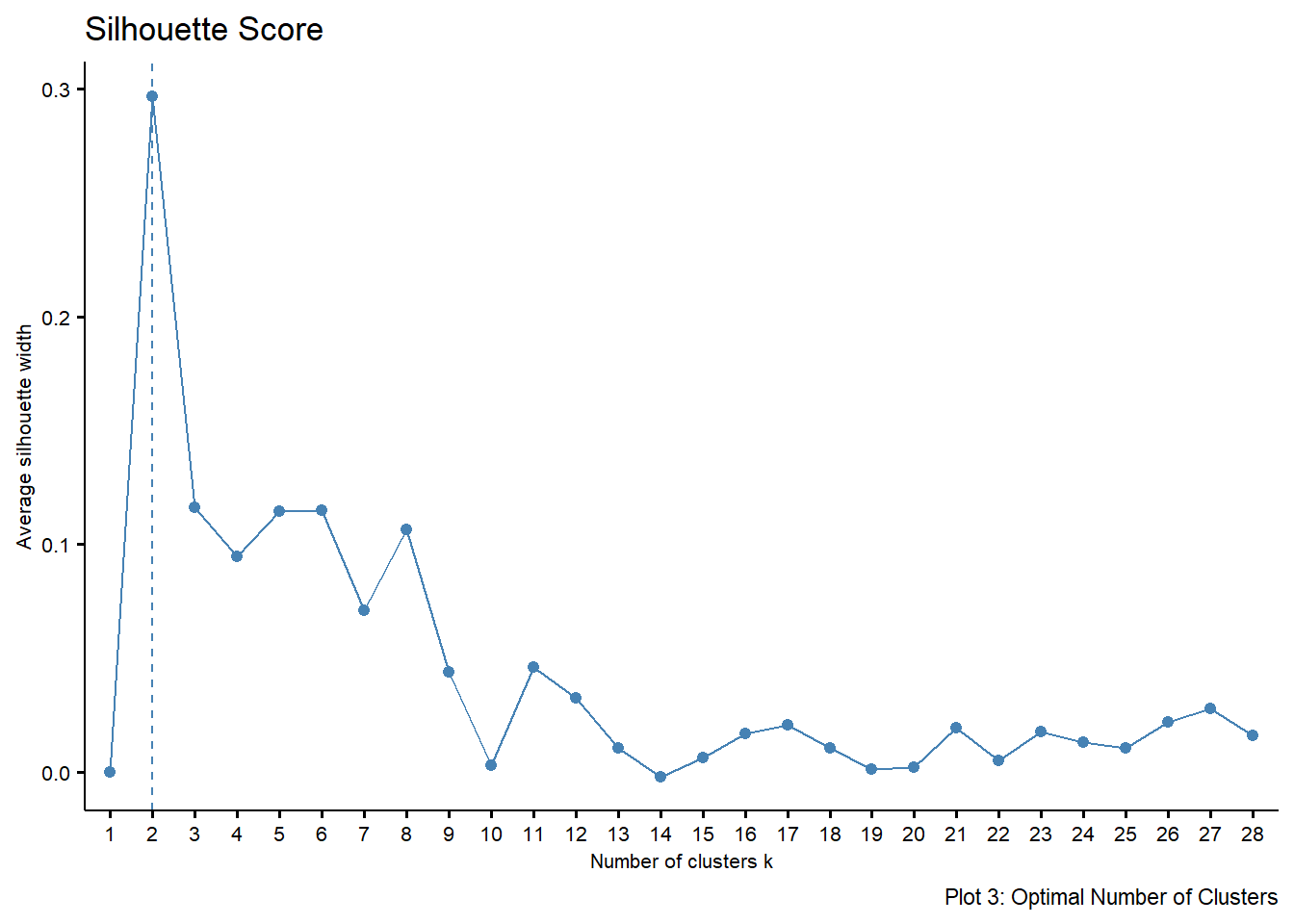


Figure 7: Silhouette score and optimal number of clusters

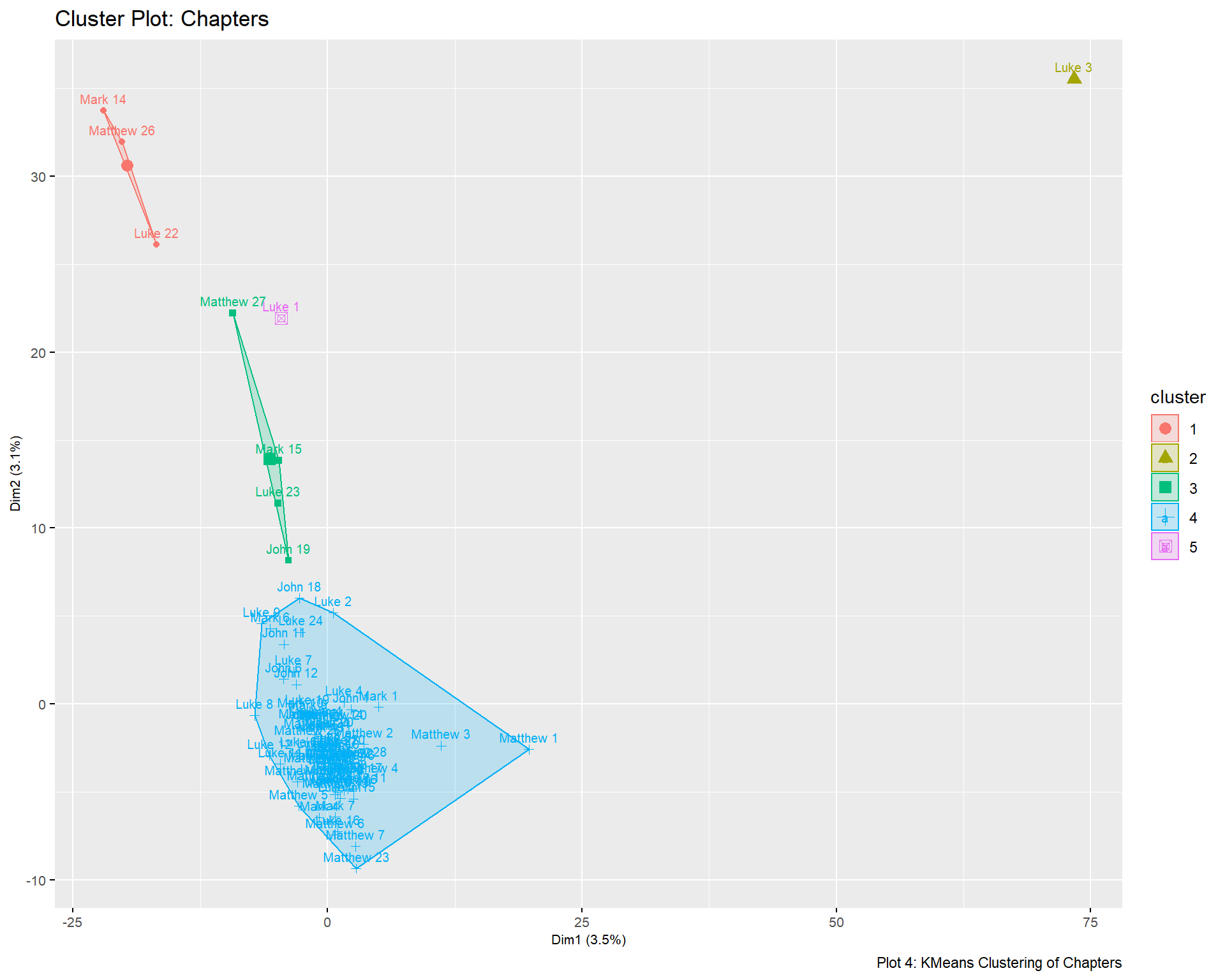
k

Figure 8: KMeans clustering of chapters

KMeans clustering was then performed to see whether books can be identified from one another, in which case there ought to be four clusters. Alternatively, the books can also be clustered by content similarity, in which case there should be anywhere between 16 and 28 clusters given that shortest book has 16 chapters, and the longest has 28. This assumes that content order and chapter delimitation is highly similar. The silhouette score plot suggests that the ideal number of clusters is 5. Plotting the five clusters using the first two principal component, the following observations about the clusters can be made:  
1. Cluster 1 contains chapters on the crucifixion of Jesus.  
2. Cluster 3 contains chapters on the betrayal of Jesus by Judas and the Last Supper.  
3. Cluster 2 is an outlier. Luke 3 was identified as one of the least cosine similar chapters relative to the rest because its content on John the Baptist is not found elsewhere in the Gospels as a standalone chapter.

4. Cluster 5 is distinct. Like Luke 3, the contents of Luke 1, on John the Baptist’s birth, are not found in any of the other three books.

5. Cluster 4 is most probably an amalgamation of other chapters, given that most chapters are somewhat similar, as the cosine similarity matrix suggests.

Discussion

Having performed the analyses detailed above, it is possible to make the argument that using a bag-of-words approach, the three versions are highly similar in terms of word choice, and by extension, meaning. This can be seen from the difficulty that the random forest models have in distinguishing between the KJV and the ASV even though they were published three centuries apart. This is because the “ASV maintained close ties to Elizabethan English” (Hamlin, 2011, pp.165-166), which came to be known as “formal equivalence” (p.166). On the other hand, even though the WEB is an update of the ASV, it is written in modern-day informal English which sets it apart from the KJV and ASV. This implies that the English language has evolved enough over the years for a supervised machine learning technique to set it apart.

On the issue of authorship, the ability of a random forest model to separate the four books across versions imply that verses in all four books retain unique stylometric traces of authorship that have been preserved over the years. In the case of the WEB, of the verses which were misclassified as those of other versions, 87% of them were attributed to the correct book albeit the wrong version.

However, the book of Mark is often misattributed to Luke and Matthew. This lends some support to the theological argument of Marcan Priority in which the Gospels of Matthew and Luke are based on Mark (Goodacre, 2000; Murai, 2006), and therefore similar. To extend this argument a little further, one might expect to find four distinct clusters when performing kmeans analysis on verses collapsed into clusters, with some chapters from Mark, Luke, and Matthew mixed together.

However, this expectation was not supported. Instead, some clusters were topical in nature. Murai & Tokosumi (2006) argued that the Gospels contain verses that are echoed in one another in syncoptic fashion. Thus, various chapters across the four Gospel books could exhibit high degrees of similarity after controlling for Bible version. Furthermore, as the KMeans clustering algorithm minimizes distance across the dimensions, individually similar observations such as the syncoptic chapters are clustered together. This is why the clusters identified mirror to a certain extent, groups of highly similar cells in the cosine similarity matrix. On the other hand, a random forest model might not necessarily be looking to minimize the distance between the observations. Furthermore, the random forest model was trained on verses, and not chapters as the KMeans cluster plot was. There might be broader patterns that the random forest model picked up, which are not observable individually or when it is aggregated.

It is also curious that this was not the case for all topics discussed in the Gospel Books, evident from the Euclidean distance plot, and the optimal number of KMeans clusters. This could be because of chapter delimitation where content covered in one chapter in a book, can be covered in two in another. This reduces the similarity of each chapter as a document.Further, of the two outliers, Luke 1 is uncharacteristically similar to the other outlier, Luke 3, but otherwise typical of other chapters as calculated by cosine similarity. Thus in a high-dimensional space, this sets the two books apart from the others, resulting in their separate classification even though when decomposed into the first two principal components, Luke 1 appears relatively close to the other clusters.

Conclusion

Thus, what is observed here is an inability to satisfactorily explain why a random forest model is better able to classify and predict than a KMeans model. Neither can the hypothesis that the four Gospel are stylistically distinct be conclusively supported. An extension to this work will ideally include some form of synoptic identification and similarity testing using embeddings. Further, if there are common topics and proper nouns among the books, then burstiness can be used to explore intensity. Normalized for length, and relative to one another, burstiness should be identical as the books cover the same issues over the same timeline, which is the life of Jesus Christ. In the final summary, it appears fitting for a text that has been extensively studied for millennia to retain inexplicable features that continue to capture future collective interest.

References

Goodacre, M. (2000). A Monopoly on Marcan Priority? Fallacies at the Heart of Q. In Society of Biblical Literature Seminar Papers 2000 (pp. 538-622).

Hamlin, H. (2011). The King James Bible after 400 years: literary, linguistic, and cultural influences. Cambridge Univ. Press.

Linmans, A. J. M. (1998). Correspondence analysis of the Synoptic Gospels. Literary and linguistic computing, 13(1), 1-13.

McDonald, D. (2014). A text mining analysis of religious texts. The Journal of Business Inquiry, 13(1), 27-47.

Murai, H., & Tokosumi, A. (2006). Synoptic Network Analysis of the Four Gospels. In SCIS & ISIS SCIS & ISIS 2006 (pp. 1590-1595). Japan Society for Fuzzy Theory and Intelligent Informatics.

Murai, H. (2013). Exegetical Science for the Interpretation of the Bible: Algorithms and Software for Quantitative Analysis of Christian Documents. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (pp. 67-86). Springer, Heidelberg.

Xia, P., & Yarowsky, D. (2017, November). Deriving Consensus for Multi-Parallel Corpora: an English Bible Study. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 448-453).

Appendix

| Table A1: Class Imbalance - Number of verses in each book and version | | | |
| --- | --- | --- | --- |
| **Book** | **American Standard Version** | **King James Bible** | **World English Bible** |
| John | 879 | 879 | 879 |
| Luke | 1151 | 1151 | 1151 |
| Mark | 678 | 678 | 678 |
| Matthew | 1071 | 1071 | 1071 |

| Table A2: Predicting Book and Version with Random Forests | Prediction(V), Reference(H) | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A\_Joh** | **A\_Luk** | **A\_Mar** | **A\_Mat** | **K\_Joh** | **K\_Luk** | **K\_Mar** | **K\_Mat** | **W\_Joh** | **W\_Luk** | **W\_Mar** | **W\_Mat** |
| **ASV** | | | | | | | | | | | | |
| A\_Joh | 38 | 9 | 13 | 7 | 157 | 4 | 3 | 2 | 43 | 0 | 1 | 1 |
| A\_Luk | 18 | 63 | 27 | 39 | 13 | 179 | 11 | 15 | 2 | 72 | 5 | 7 |
| A\_Mar | 3 | 11 | 15 | 17 | 0 | 6 | 78 | 2 | 0 | 1 | 38 | 3 |
| A\_Mat | 8 | 21 | 22 | 35 | 6 | 12 | 16 | 164 | 5 | 4 | 3 | 60 |
| **KJV** | | | | | | | | | | | | |
| K\_Joh | 142 | 0 | 2 | 3 | 34 | 9 | 2 | 9 | 19 | 1 | 0 | 0 |
| K\_Luk | 0 | 172 | 10 | 16 | 12 | 62 | 27 | 40 | 2 | 26 | 2 | 1 |
| K\_Mar | 1 | 0 | 73 | 0 | 3 | 7 | 9 | 16 | 0 | 0 | 12 | 0 |
| K\_Mat | 3 | 3 | 8 | 153 | 16 | 33 | 38 | 48 | 0 | 3 | 2 | 25 |
| **WEB** | | | | | | | | | | | | |
| W\_Joh | 48 | 1 | 0 | 0 | 18 | 2 | 0 | 0 | 109 | 23 | 12 | 19 |
| W\_Luk | 1 | 59 | 8 | 4 | 1 | 28 | 3 | 3 | 52 | 134 | 58 | 88 |
| W\_Mar | 0 | 0 | 22 | 3 | 1 | 0 | 14 | 2 | 3 | 20 | 18 | 19 |
| W\_Mat | 1 | 6 | 3 | 44 | 2 | 3 | 2 | 20 | 28 | 61 | 52 | 98 |

| Table A3: Predicting Version with Random Forests | Prediction(V), Reference(H) | | | |
| --- | --- | --- | --- |
|  | **American Standard Version** | **King James Bible** | **World English Bible** |
| American Standard Version | 459 | 611 | 167 |
| King James Bible | 477 | 442 | 54 |
| World English Bible | 197 | 80 | 912 |

| Table A4: Predicting Book with Random Forests | Prediction(V), Reference(H) | | | | |
| --- | --- | --- | --- | --- |
|  | **John** | **Luke** | **Mark** | **Matthew** |
| John | 766 | 9 | 6 | 6 |
| Luke | 15 | 988 | 47 | 59 |
| Mark | 0 | 11 | 525 | 17 |
| Matthew | 10 | 27 | 32 | 881 |

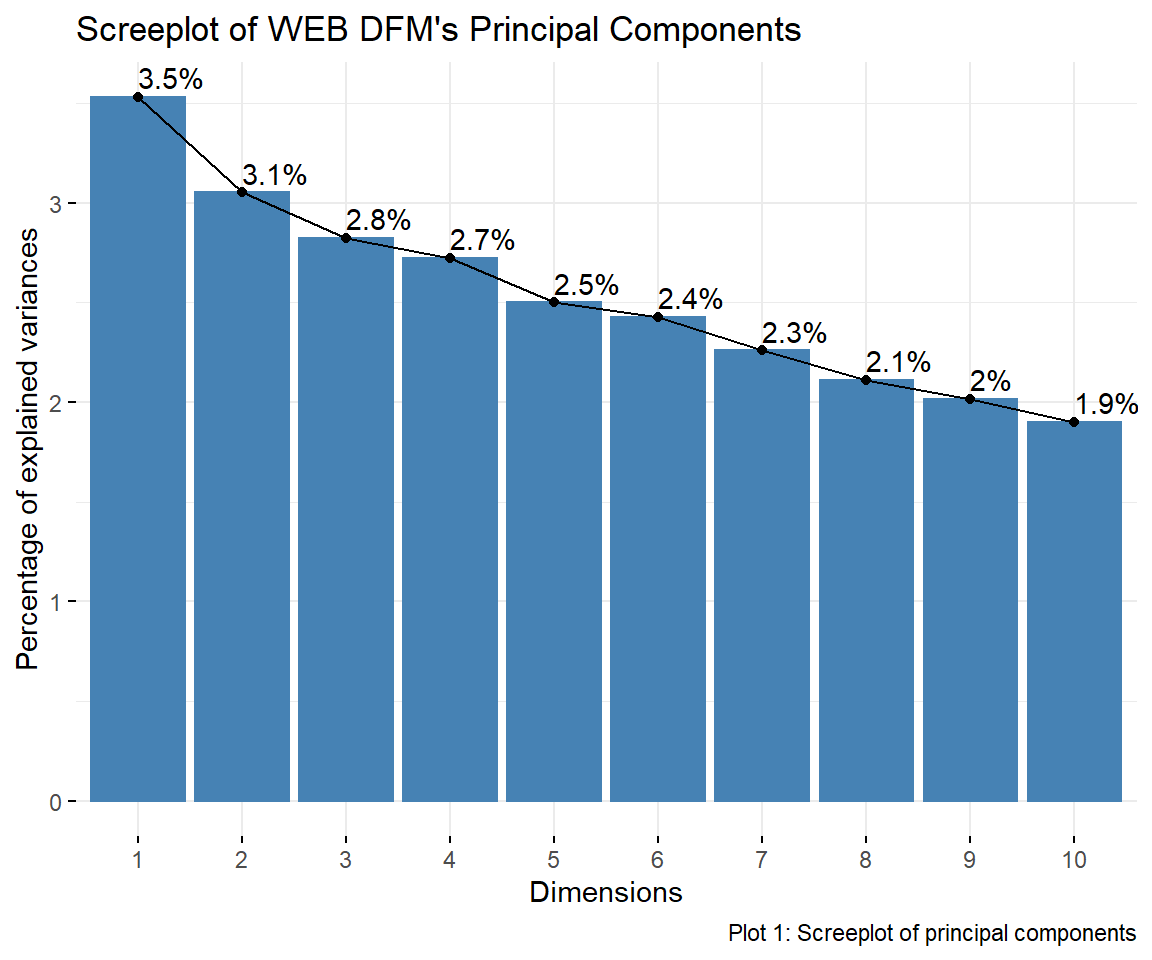


Figure A5: Screeplot of WEB Gospels Chapters’ Principal Components

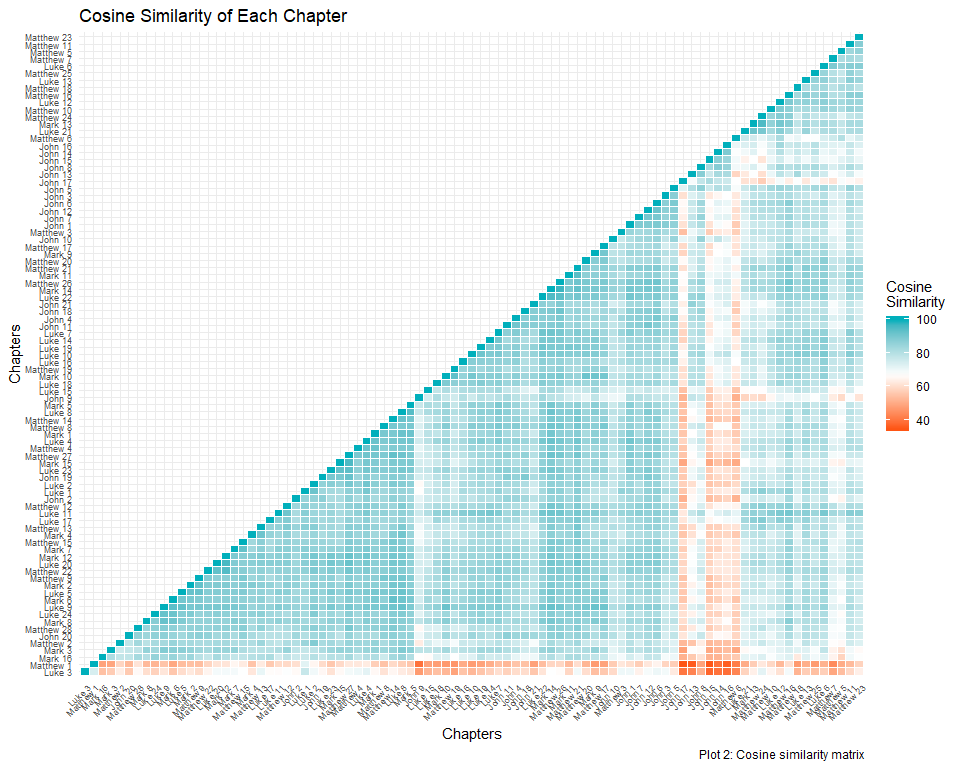


Figure A6: Cosine similarity of each chapter