The Word

Wesley Chioh

April 9, 2020

#### Introduction

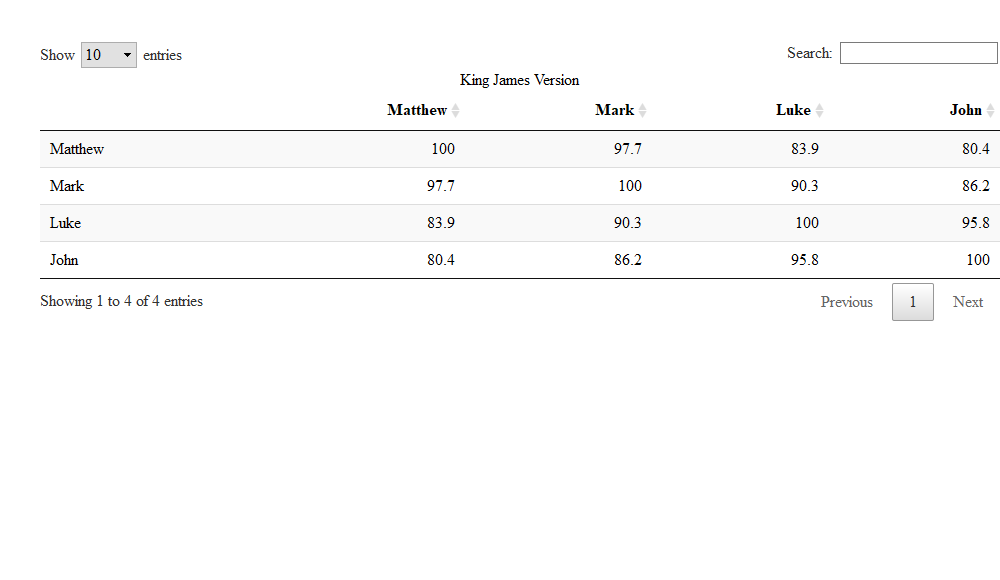
The New York Times bestseller list changes every week. But on the time scale of centuries and millenia, the longstanding global bestseller has not. It is most probably the Bible. The version of the Bible commonly used today is quite different in style and phrasing from that a mere century ago. It undergoes 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?

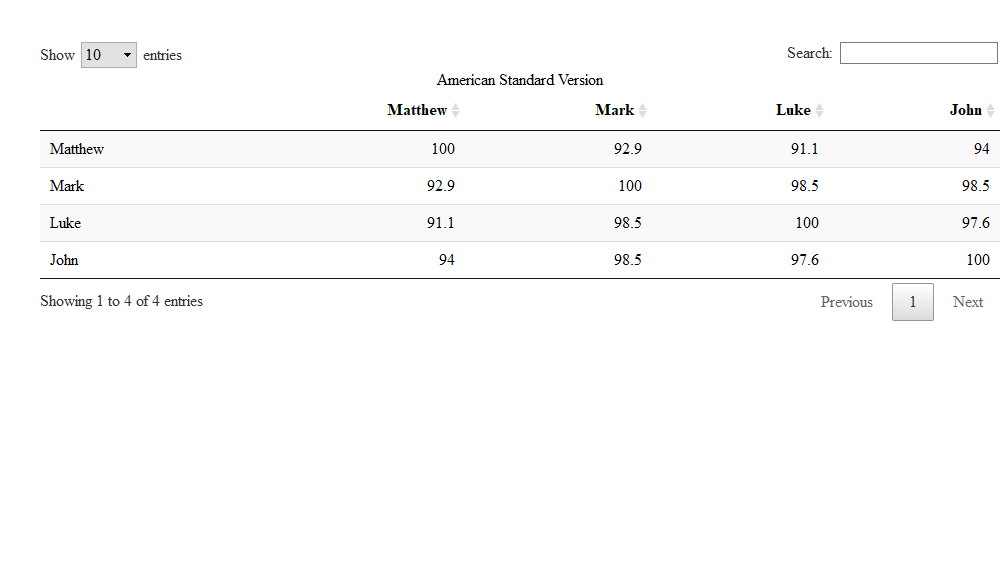
The Bible itself can be seen as a “multi-parallel corpora” (Xia and Yarowsky, 2017, p.448) with multiple versions of what is essentially a highly similar corpus. The Bible used today contains 66 books, each of which is distinct from another in terms of authorship, stylometry and topics. However, the degree to which they differ depends on the books in question. 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. In fact, the Gospels are also known as the Synoptic Gospels (Linmans, 1998; Murai, 2006) with corresponding sections among the four. Murai (2006) argues that from a network analysis perspective, they can be characterized as a series of multiple one-to-many relationships.

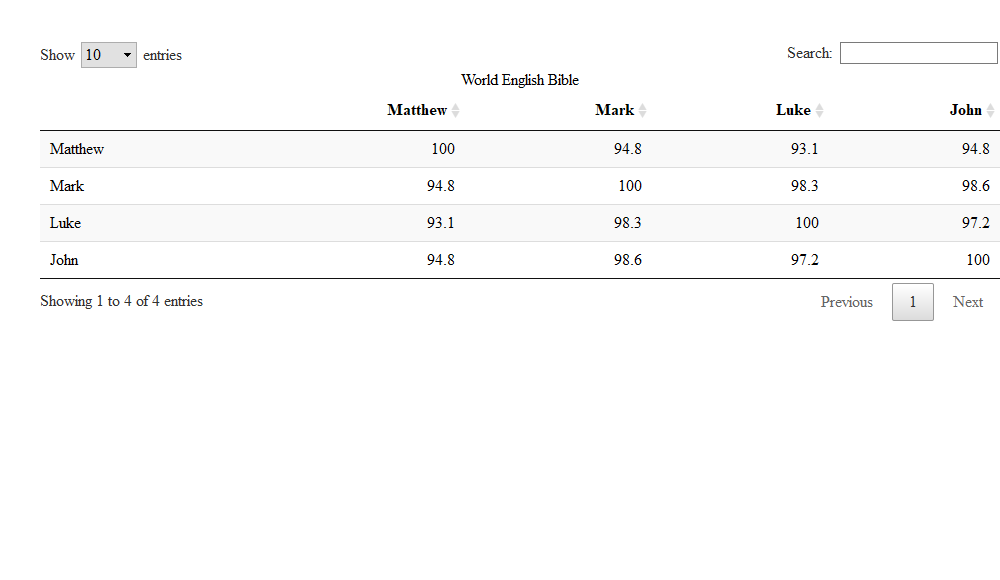
Thus, this paper will seek to test the hypothesis that despite controlling for topical similarity, the four Gospel books remain stylometrically distinct from another. This distinction is retained over the ages and various versions of the Bible. This paper will use the King James Version (KJV), American 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 times when the KJV and ASV were published. 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.

Using a bag of words approach, we can calculate the cosine similarity among the four Gospel books. In so far as word choice and frequency contains latent information, it hints at the relative similarity of the four books in terms of topics, and sentiments. 



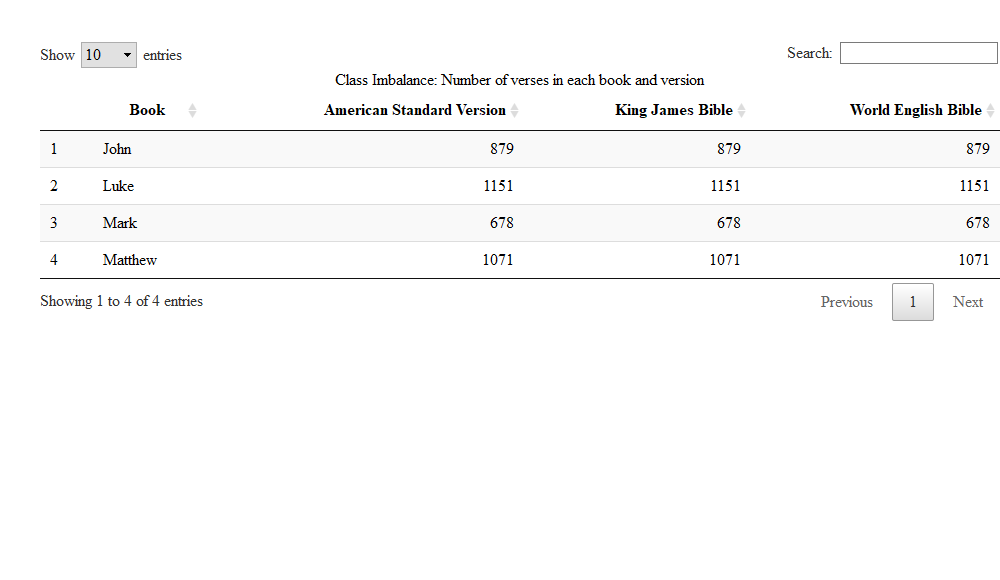


From the tables above, the average cosine similarity of the four Gospel books across the KJV, ASV,and WEB is 89.1%, 95.4%, 96.1% respectively. This is fairly high, but they are not identically so. The degree of similarity between the books is much higher for the ASV and WEB as compared to the KJV as proven by the mean cosine similarity score. 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 takes into account the slight class imbalance shown below 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



With 12 classes to predict, the out-of-sample accuracy is 0.196. 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 table suggests that verses from the ASV and KJV are highly likely to be misclassified as the other. Of the four books, it appears that verses from John is the least likely to be misattributed to others whereas verses from Luke and Mark are especially prone to misattribution. This analysis be further broken down into the versions and the books with fewer classes in each case.

## Precision Recall  
## Class: American Standard Version 0.3710590 0.4051192  
## Class: King James Bible 0.4542652 0.3901147  
## Class: World English Bible 0.7670311 0.8049426

In the previous model, it was observed that the poor performance of the classifier could be attributed to errors in distinguishing books and versions with seemingly more errors in classification stemming from separating the versions. Training another random forest model to just classify the verses based on its bible version, the out-of-sample accuracy is 0.196is much higher but it is hardly very accurate if taken as a whole. Further, we confirm that the precision and recall rates are especially poor for the ASV and KJV as compared to the WEB.

##### Book  
ids\_train3 <- createDataPartition(gospels\_reduced\_class$Book, p = 0.7, list = F, times = 1)  
train\_x3 <- gospels\_dfm[ids\_train3,] %>%  
 as.data.frame()  
train\_y3 <- gospels\_reduced\_class$Book[ids\_train3] %>%  
 as.factor()  
test\_x3 <- gospels\_dfm[-ids\_train3,] %>%  
 as.data.frame()  
test\_y3 <- gospels\_reduced\_class$Book[-ids\_train3] %>%  
 as.factor()  
  
## tuning random forest  
# mtry <- sqrt(ncol(train\_x3))   
# ntree <- 51  
# trainControl <- trainControl(method = "cv", number = 5)  
# metric <- "Accuracy"  
# tunegrid <- expand.grid(.mtry = mtry)  
# cl <- makePSOCKcluster(5)  
# registerDoParallel(cl)  
  
# # running rf  
# gospels\_rf\_book <- train(x = train\_x3, y = train\_y3,  
# method = "rf", metric = metric, tuneGrid = tunegrid, trControl = trainControl, ntree = ntree, doParallel = T)  
# stopCluster(cl)  
  
# save RDS and then load to speed things up  
# saveRDS(gospels\_rf\_book, "gospels\_rf\_book.RDS")  
gospels\_rf\_book <- readRDS("gospels\_rf\_book.RDS")  
  
# print model  
# print(gospels\_rf\_book)  
  
# predict and confusion matrix  
gospels\_rf\_book\_predict <- predict(gospels\_rf\_book, newdata = test\_x3)  
gospels\_rf\_book\_cm <- confusionMatrix(gospels\_rf\_book\_predict, reference = test\_y3, mode = "prec\_recall")  
gospels\_rf\_book\_cm$table

## Reference  
## Prediction John Luke Mark Matthew  
## John 766 9 6 6  
## Luke 15 988 47 59  
## Mark 0 11 525 17  
## Matthew 10 27 32 881

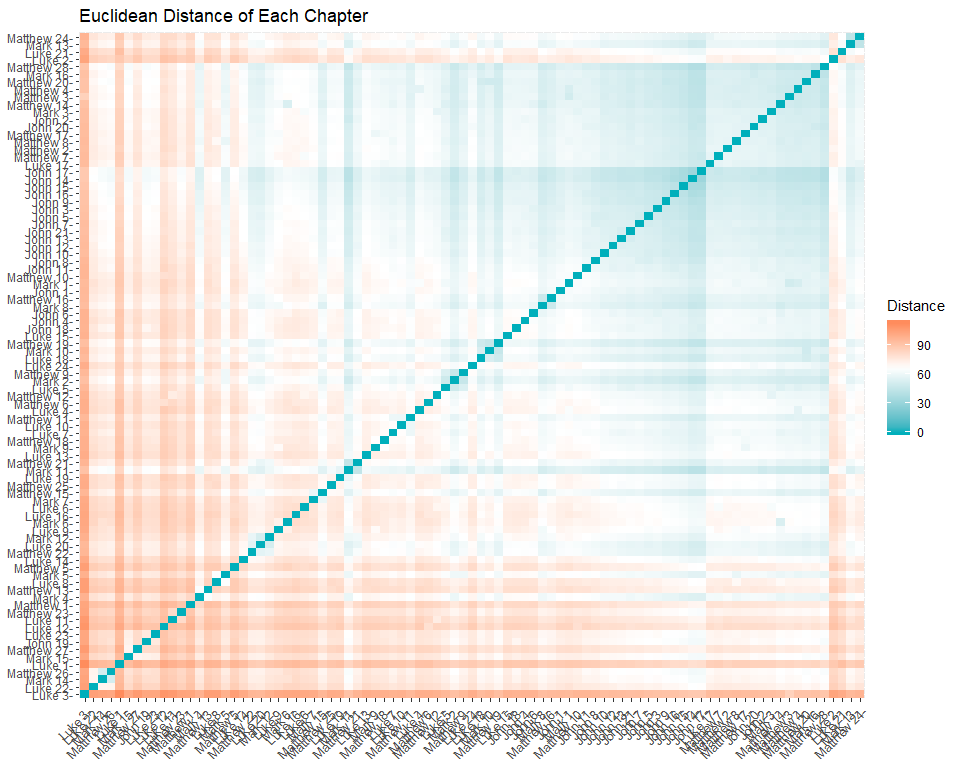
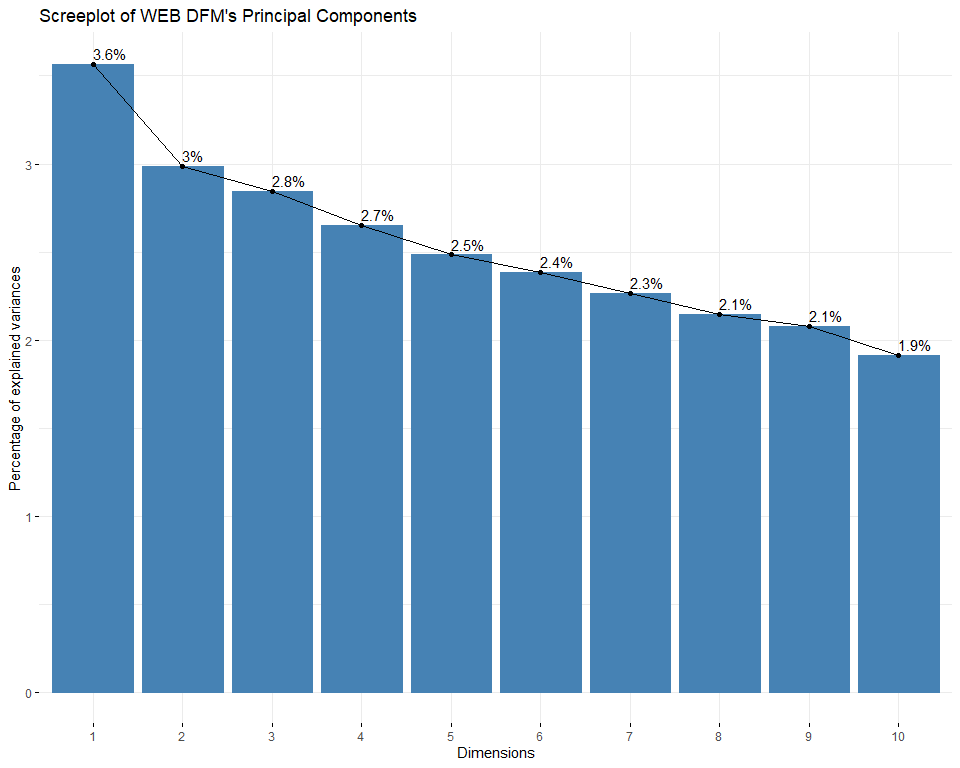
gospels\_rf\_book\_cm$byClass[,5:7]

## Precision Recall F1  
## Class: John 0.9733164 0.9683944 0.9708492  
## Class: Luke 0.8908927 0.9545894 0.9216418  
## Class: Mark 0.9493671 0.8606557 0.9028375  
## Class: Matthew 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 a uncharacteristically low recall score of 0.8606557. This suggests that verses in Mark tend to be misattributed to other books. 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).

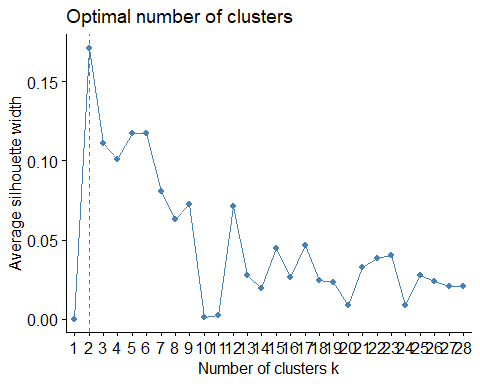
#### Clusters

Murai (2006) argued that the Gospels contain verses that are echoed in one another in syncoptic fashion. Thus, various chapters across the four Gospel books should exhibit high degrees of similarity after controlling for Bible version. This is especially so given the argument of Marcan Priority where Luke and Matthew are derived from the book of Mark.

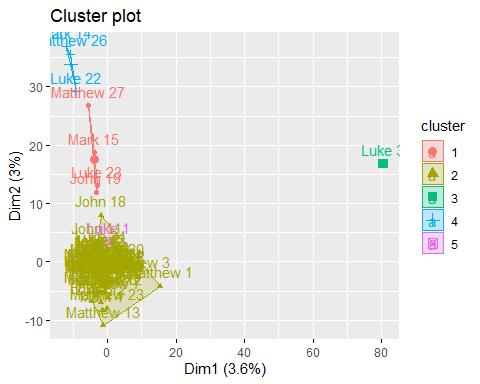
The WEB was used to test this expectation. The WEB is the most recently translated copy of the three. Removal of contemporary stopwords can thus be performed with relative confidence that textual meanings are not severely redacted. This leaves behind proper nouns and other less commonly used words that will aid in the clustering at the chapter level. Verses can be relatively short and easily taken out of context when stopwords are removed. Collapsing them into chapters can provide a better approximation of chapter meaning.  


As a baseline of comparison, the Euclidean distance between each book can be calculated and plotted. For the most part, books and chapters are distinct from one another with the exception of 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 now split between two chapters. The screeplot above suggests that dimensionality has not been significantly reduced. The first two dimensions account for less than 7% of the variability. Thus, principal components analysis is not particularly useful in this case as the trade-off between dimensionality and variance is extremely steep.

# KMeans  
set.seed(1728)  
fviz\_nbclust(web\_chapter\_mat\_scale, kmeans, method = "silhouette", k.max = 28)



k5 <- kmeans(web\_chapter\_mat\_scale, 5, nstart = 25)  
fviz\_cluster(k5, data = web\_chapter\_mat\_scale)



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 2 contains chapters on the betrayal of Jesus by Judas and the Last Supper.  
3. Cluster 3 and 4 are outliers.  
4. Cluster 5 is most probably an amalgamation of dissimilar chapters, given that most chapters are more different than they are similar, as the Euclidean distance plot suggests.