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 Bible as we know it today is probably quite diffeent 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?

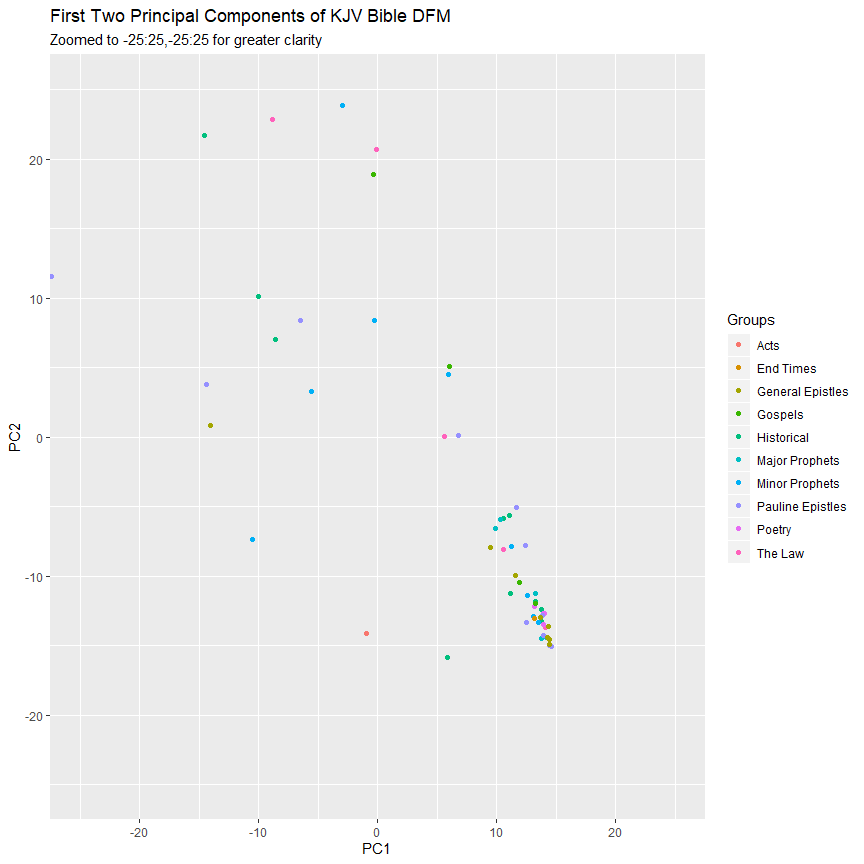
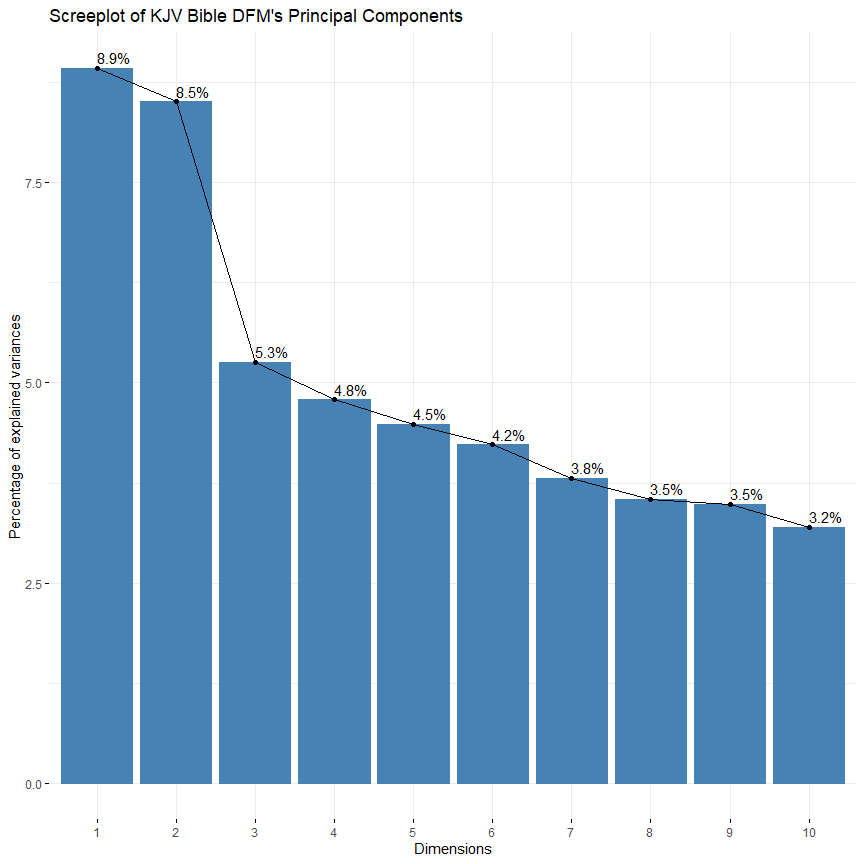
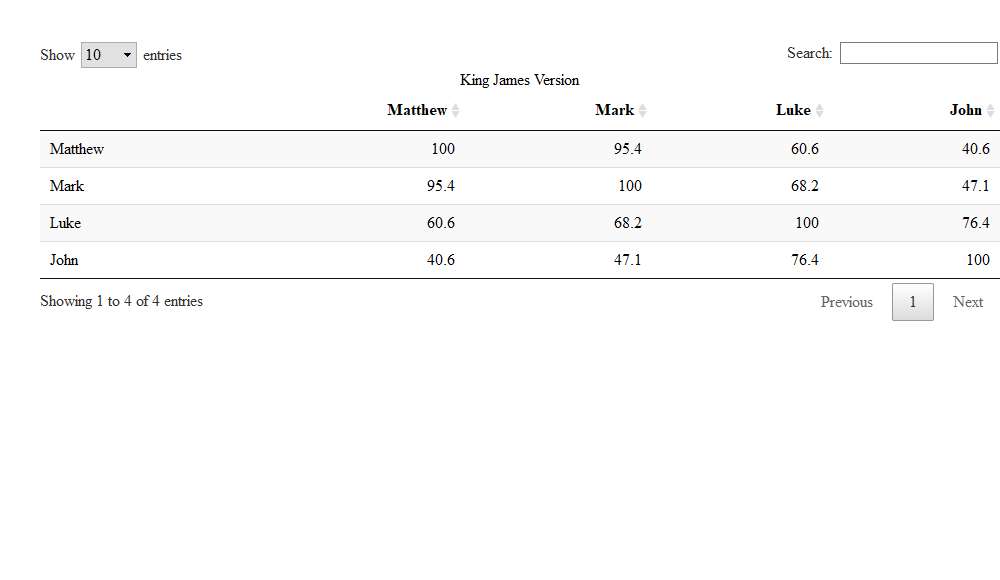
This paper will therefore seek to test the hypothesis despite the linguistic and ease of readability differences, various versions of the Bible maintain a significant degree of similarity in terms of its topics, and sentiments. Furthermore, given that the Bible is effectively an agglomeration of religious treatises by various authors, stylistic differences should be retained even with the revisions.

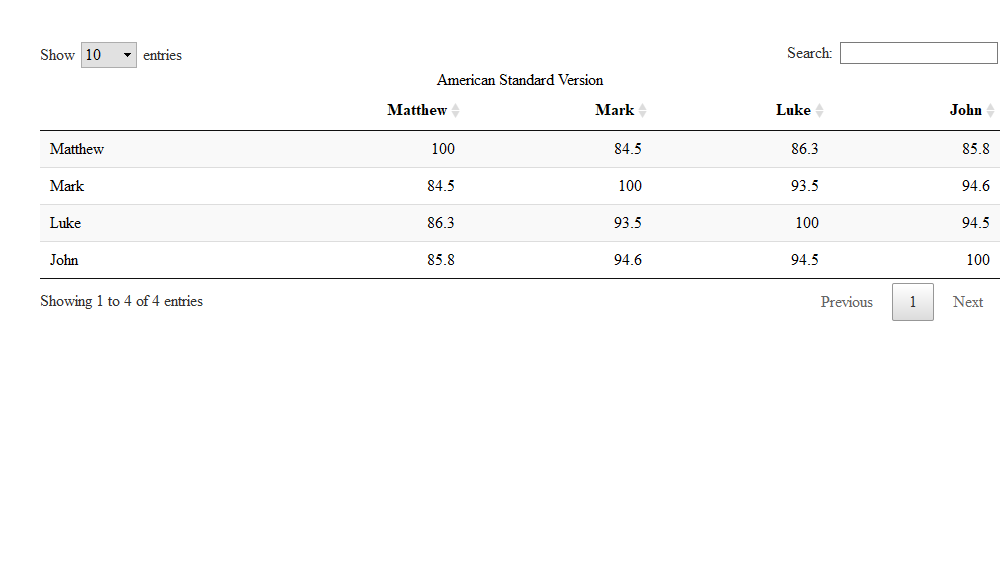
### Methodology and Data Sources

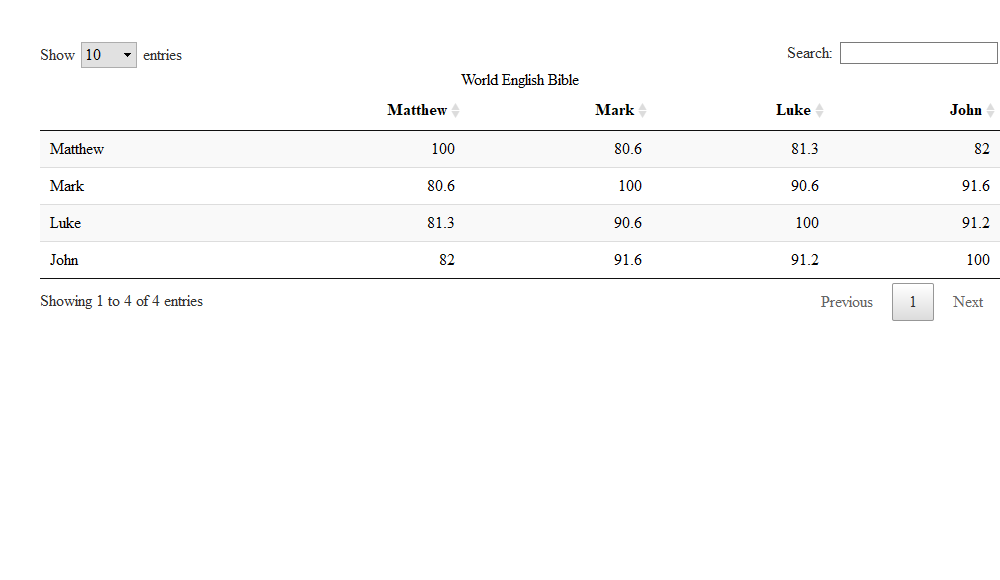
The Bible itself can be seen as a “multi-parallel corpora” (Xia and Yarowsky, 2017, p.448) with multiple versions of what is essentially a highly similar corpus. Stylometric analyis (McDonald, 2014).

First, this paper will seek to validate a commonly held opinion among Christians, and Bible scholars that among the various books within the Bible, the Gospels of Matthew, Mark, Luke, and John are among the most similar. 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.

Using the King James’ Version of the Bible, each verse was collapsed into a row with each row comprising of verses from the same book. A document-feature matrix (DFM) was then created. Stopwords, punctuation, and numbers were removed, every word was lower-cased and then stemmed. To reduce the number of dimensions, Principal Components Analysis (PCA) was performed. The scree plot suggests that each component captured relatively little of the variation between books. Furthermore, the biplot of the first two principal components does not support the hypothesis that the four Gospel books are relatively similar based on variances within the entire Bible’s DFM. Instead, a more distinct clustering of the Pauline and General Epistles can be observed.

Extending the bag of words approach, we can calculate the cosine similarity among the four Gospel books. As cosine similarity is based on the DFM of the entire KJV Bible, 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. The data suggests that Matthew and Mark, and Mark and Luke are the most similar. On the other hand, John is signifcantly different from Matthew and Mark but highly similar to Luke.  


This analysis can be repeated with the American Standard Version of the Bible and the World English Bible to see if a similar relationship is observed. 



This analysis can be repeated with the American Standard Version of the Bible and the World English Bible to see a similar relationship is observed. Based on the tables above, all three versions of the Bible, are relatively uniform in ranking the cosine similarity among the four Gospel books. However, the more antiquated KJV shows a greater divergence among the books. Furthermore, the KJV suggests that John is least similar to Matthew and Mark. With the two newer versions however, that analysis is flipped; John is most similar to both Matthew and Mark.