ScriptScript - A Comparative Text Analysis Library

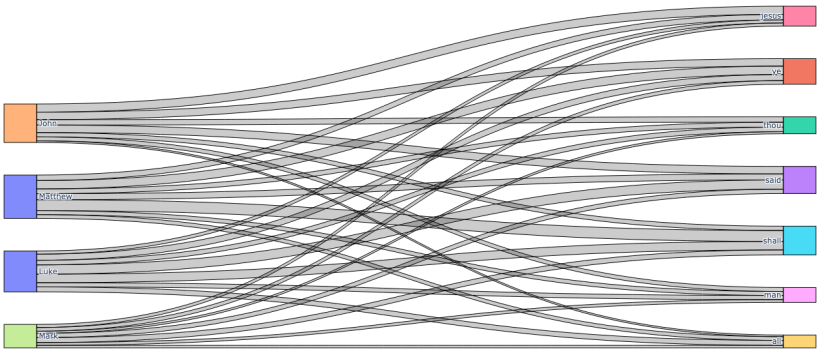
Benjamin Pierce | Personal Repo

Data Sources

To highlight my library's proficiency in extracting and visualizing information from a host of text files, I looked to use a set of texts that were similar enough to allow comparison, but different enough to pose a meaningful reason for evaluation. The King James Bible provides a plethora of text files to analyze in this regard, as there are many different books and writers all focused on a few general themes. Analyzing each book, however, would be well out of the scope of this assignment, as 1) there are over 60 books in the Bible, and 2) using each book of the Bible would hinder the library's ability to visualize meaningful comparisons. Thus, I needed a smaller subset of books to compare. The most logical subset I came upon was the grouping of the four Gospels; Matthew, Mark, Luke, and John. Each Gospel highlights the life and teachings of Jesus, but all address the audience and portray Jesus in slightly different regards. By comparing these four texts, we see how the differences in linguistics and sentiment feed into the varied audiences and interpretations of Jesus.

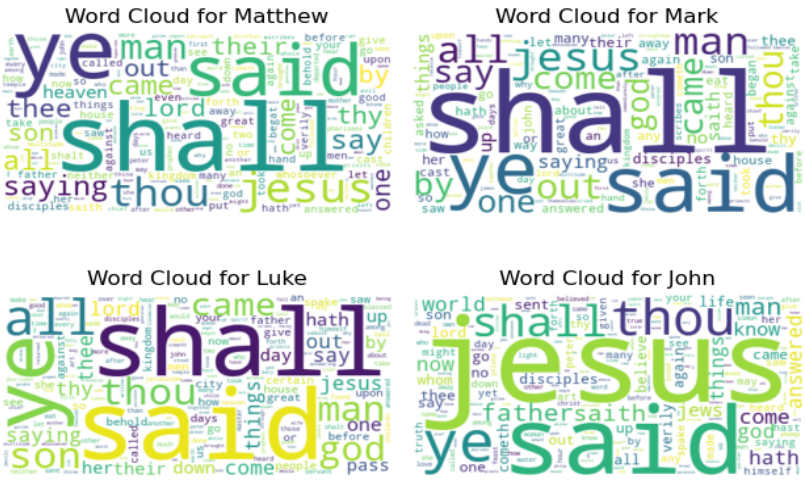
Visualizations

1) Text-to-Word Sankey Diagram



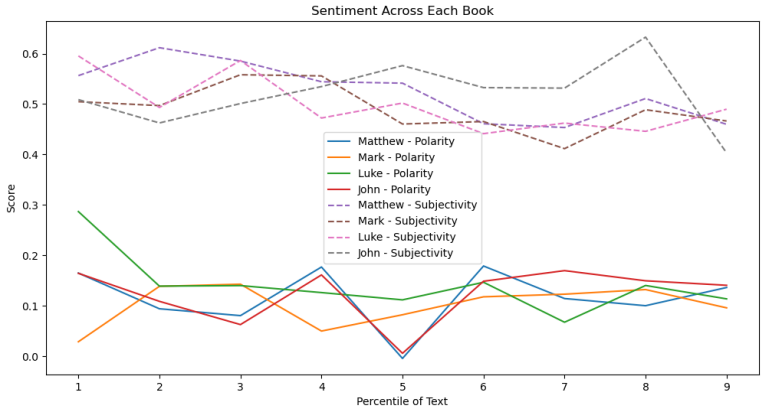
Here is the Text-to-Word Sankey Diagram produced by my ScriptScript library. This Sankey diagram is produced using the k most common words across each text file. My library supports the functionality of passing a user-defined list of words here as well but for a general analysis, the default functionality of the k most common words provides good insights. Through this visualization, we can see the interconnectedness of the different texts to their most commonly used words, with the thickness of the connection representing the word count in each text. This visualization is very useful in understanding the interrelation across text files of the k most common words (or a user-specified set of words).

2) WordCloud Sub-plots



Here is my second visualization, which contains sub-plots of WordClouds for each text file. In this user-oriented visualization, the size of each word in a given text is represented by the frequency of the words, the most frequent of which are then formatted into an attractive visualization. This visualization conveys somewhat similar information to the Sankey diagram but presents it in a more user-friendly, eye-catching manner. While this visualization is less useful for drawing raw insights, it is very useful as a starter visualization for orienting oneself with the texts. While the Sankey diagram provides more specifics on word counts and interconnectedness, it can be intimidating for someone who may be unfamiliar with Sankey diagrams and/or the texts. Thus, when implementing this library on a set of texts for the first time, this simple, yet attractive visualization is incredibly useful in familiarizing oneself with the most frequently used words in each text.

3) Comparative Sentiment Analysis



Here is my third visualization, overlaying the sentiment information of each text onto a single visualization. In this visualization, we track the overall sentiment of each text over its course by dividing the text into percentiles for even comparison. The subjectivity and polarity are

then calculated through each of these percentiles. This visualization is incredibly useful in understanding the sentiment over the course of the texts, as well as comparing the sentiments across the course of the texts. While subjectivity and polarity are on different scales, the scales are similar enough that both can be displayed in the same visualization. Through this visualization, we can track how negative/positive (polarity) and how opinionated (subjectivity) the text fares over its course. We can analyze trends and associate them with certain regions of the text, as well as associate trends across the different texts.

Results (source)

The Gospel of Matthew was primarily written for Jewish Christians and emphasizes Jesus as the fulfillment of the Old Testament. Since he is addressing a grounded and educated audience, he must do so appropriately. We can see from our results that 'ye' and 'shall' are his most frequently used words, commanding assertion. Next, we can see that Matthew's subjectivity starts relatively high, but gradually declines through the text. This likely has to do with his audience, an established group with strong opinions he needs to captivate the attention of. In terms of polarity, Matthew's positive and negative sentiments can flip-flop quite a bit, reaching the lowest of each of the texts, offering a varied portrayal of Jesus that an experienced audience will understand.

The Gospel of Mark was written largely for new converts & common Greeks and emphasized the immediacy of Jesus's actions. From our results, we can see that 'shall' once again is a commanding word, which is logical for this audience. Mark is trying to address new converts and call them into action. In terms of subjectivity, Mark remains relatively consistent, likely avoiding a higher frequency of opinionated statements to captivate a newer audience. For polarity, Mark starts and ends with the lowest, but does not achieve the overall lowest. This is likely reflective of a neutral stance and portrayal of Jesus.

The Gospel of Luke was written mainly for Gentile Christians and educated Greeks, emphasizing Jesus's compassion and social righteousness. We can see from our visualizations that Luke commonly uses the words 'ye', 'all', 'shall', and 'said'. This vocabulary emphasizes Jesus's communal and instructional figure, appealing to the audience. Luke starts out the most subjective of the texts but gradually declines, likely due to the initial emphasis on Jesus's moral figure. Luke's text also has the highest polarity across the texts, plausibly due to the emphasis on Jesus's goodness.

The Gospel of John was written for a wide range of audiences, ranging from new converts to established Christians. John looks to portray Jesus as a philosopher, emphasizing his divine nature. From our results, we can see that unlike the other three texts, John uses the word 'Jesus' the most. This is likely due to his diverse audience, as he is not really catering to any specific portrayal of Jesus. We can further see that John's subjectivity starts relatively lower, but reaches the highest subjectivity out of any text towards the end. This is likely intentional, as it is hard to captivate a wide range of audiences by starting heavily opinionated. John's polarity is very similar to Matthew's polarity, offering a varied interpretation of Jesus to appeal to a broad audience.