Jae Hyun Lee (jl3fp@virginia.edu)
DS 5001
30 April 2020

Exploratory Text Analytics Final Project

On March 1, 1999, BBC News published an article with a list of the Top 10 Writers voted by readers.[1] I've compiled a collection of works by some of the writers in the list and other internationally renowned authors to analyze their similarity and difference. The set of writings examined in this study mainly explore American (Mark Twain), English (Charles Dickens and Robert Louis Stevenson), Russian (Fyodor Dostoevsky), Spanish (Miguel de Cervantes Saavedra), and French (Gaston Leroux) literature. In the end, this project seeks to answer a question: is there any linguistic or cultural pattern that makes them the greatest writers of all time?

I. Principal Components (PCA)

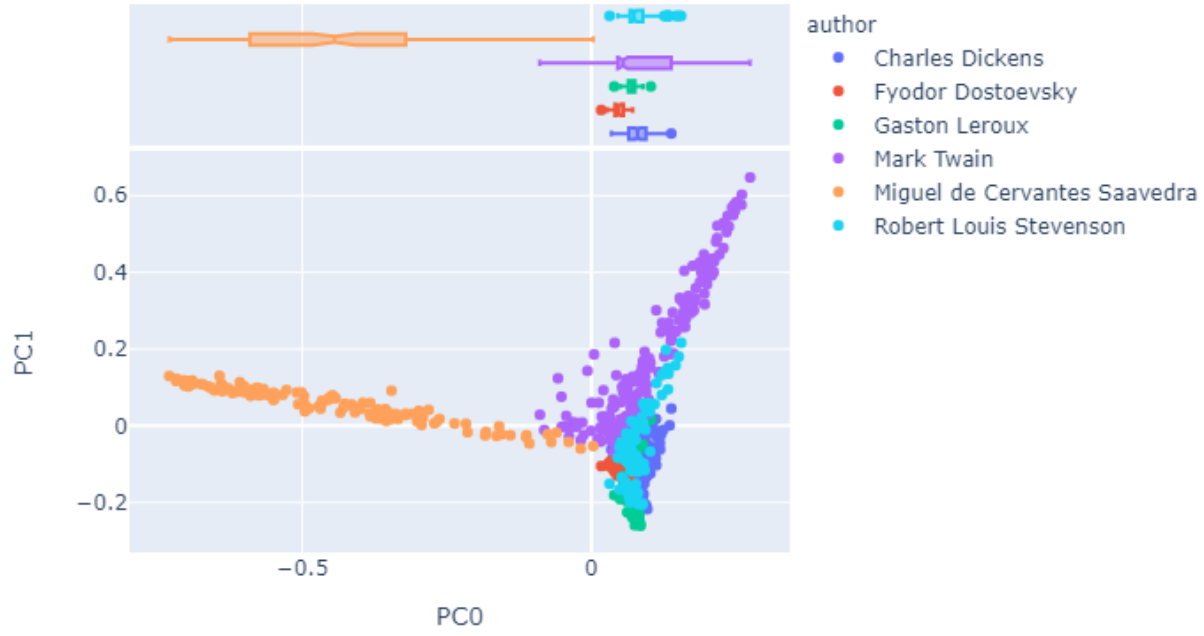

Figure 1.

[1] "Shakespeare voted millennium's best writer" BBC News, 1 May 1999, http://news.bbc.co.uk/2/hi/286082.stm. Accessed 30 April 2020.

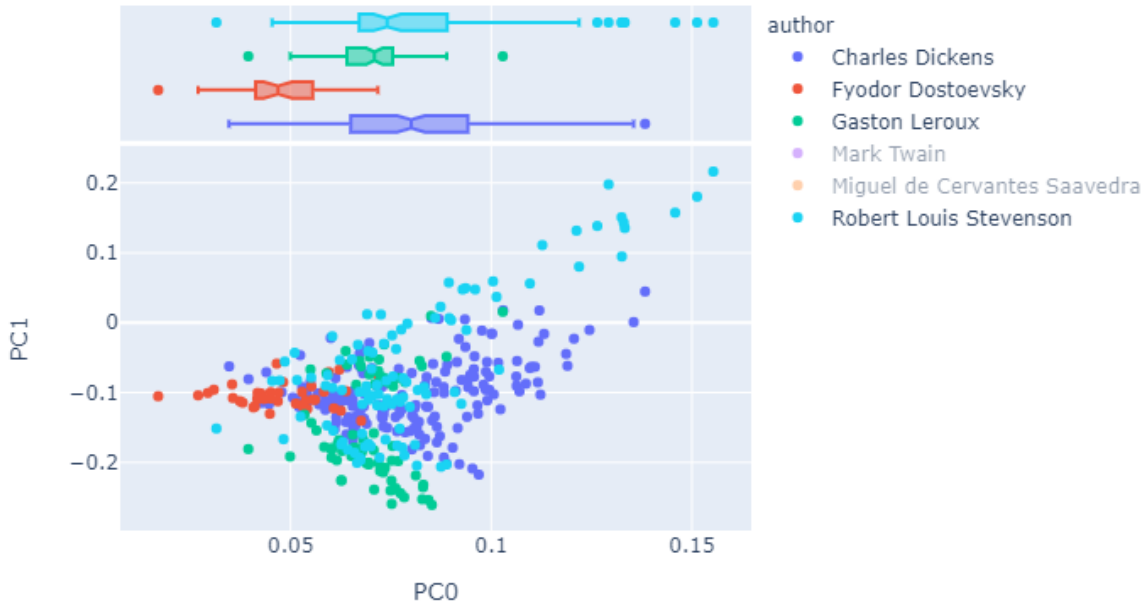

Figure 2.

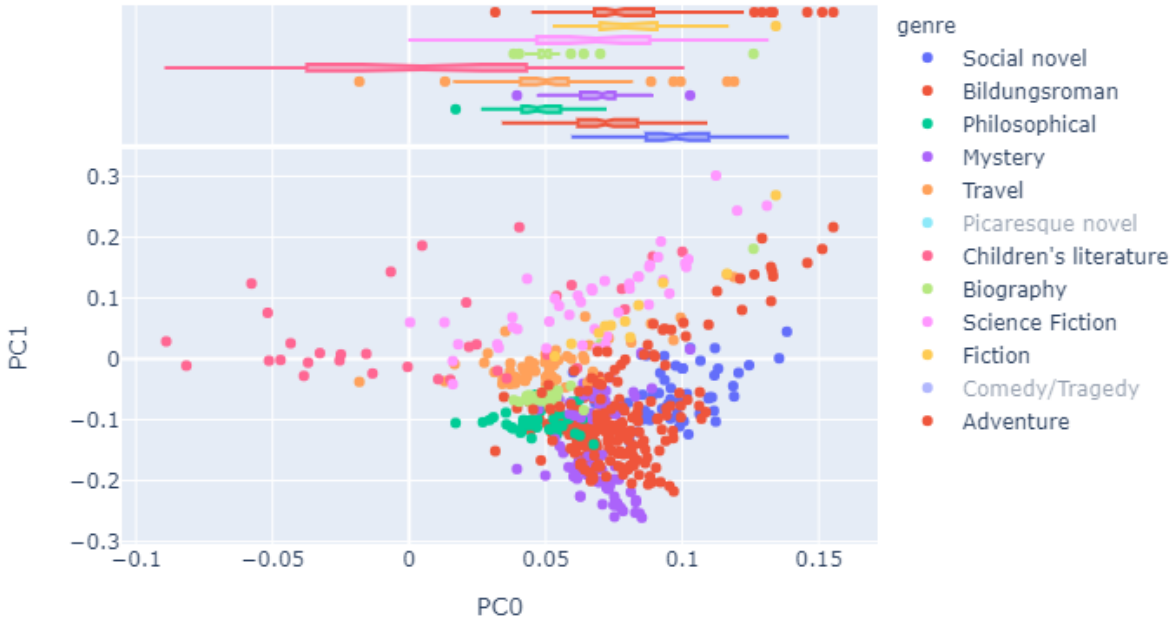

Figure 3.

A visualization comparing PC0 and PC1 is shown above (Figure 1). It looks like a wide v-shaped scatter plot stemming approximately from the origin (0, 0). The two legs are composed of Mark Twain and Miguel de Cervantes Saavedra. This might possibly suggest that these two authors have a wider spectrum of contents, considering that Don Quixote is the only work of Cervantes studied in this collection and Mark Twain's seemingly outlying segment at the top is composed of "Adventures of Huckleberry Finn" and "The Adventures of Tom Sawyer". With the two legs removed, the scatter plot doesn't immediately reveal any noticeable pattern between genres (Figure 2). But if more carefully studied, one can suggest the lighter genres such as adventure, science fiction, and child's literature are located on the top (Figure 3). On the other hand, the relatively more serious genre such as mystery, social, philosophical novels are grouped together near the bottom of the plot.

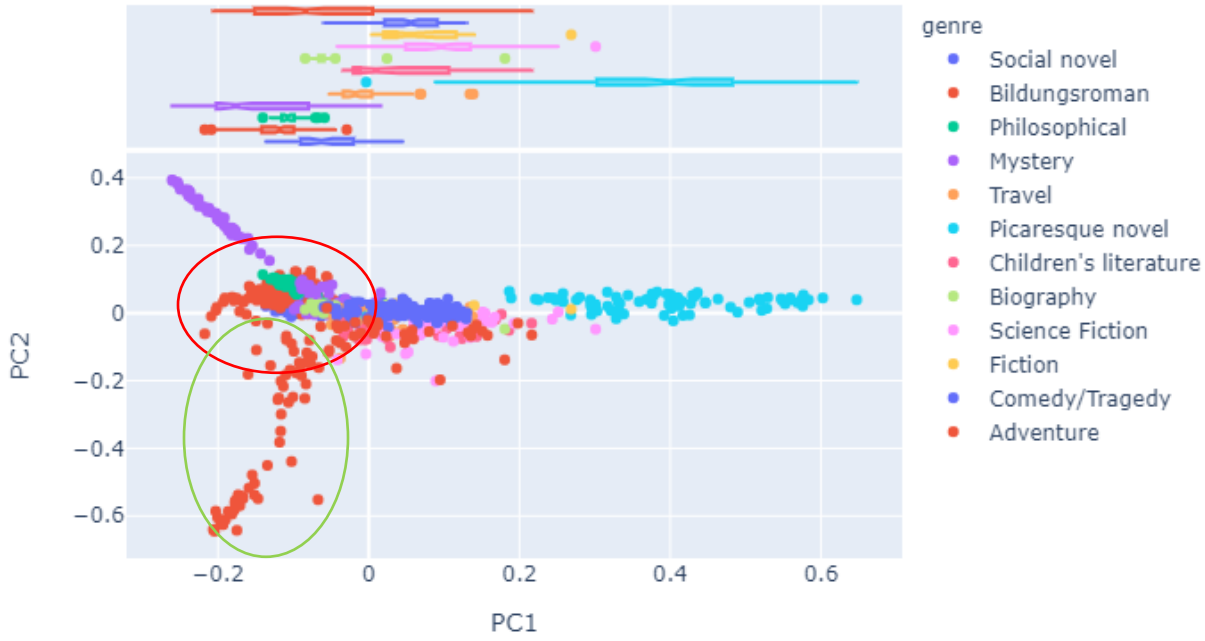

Figure 4.

A comparison of PC1 against PC2 results in three legs, also stemming from the origin (Figure 4). It's interesting to note that Comedy/Tragedy, which is seemingly the most different genre in this collection, is now a part of the main cluster. Instead, mystery is now one of the obvious deviations from the concentrated cluster. Considering the adventure and picaresque novels are extremely similar genre, it's also interesting how the scatter plot was successful in separating

these two. This visualization also shows that Dickens's Bildungsroman novels[2] (red oval) are highly similar to Stevenson's adventure novels (green oval).

Figure 5.

Figure 6.

[2] A literary genre that focuses on the psychological and moral growth of the protagonist from youth to adulthood (coming of age), in which character change is important.

With the two easily distinguishable segments of Spanish and American heritage removed, the remaining authors seem to be grouped by their cultural background as well. The two English writers, Charles Dickens and Robert Louis Stevenson, are overlapping and grouped together (Figure 5). On the other hand, the non-English works of Dostoevsky and Leroux (but Cervantes, which is removed on this plot) are located in the main cluster and are closest to each other (Figure 6).

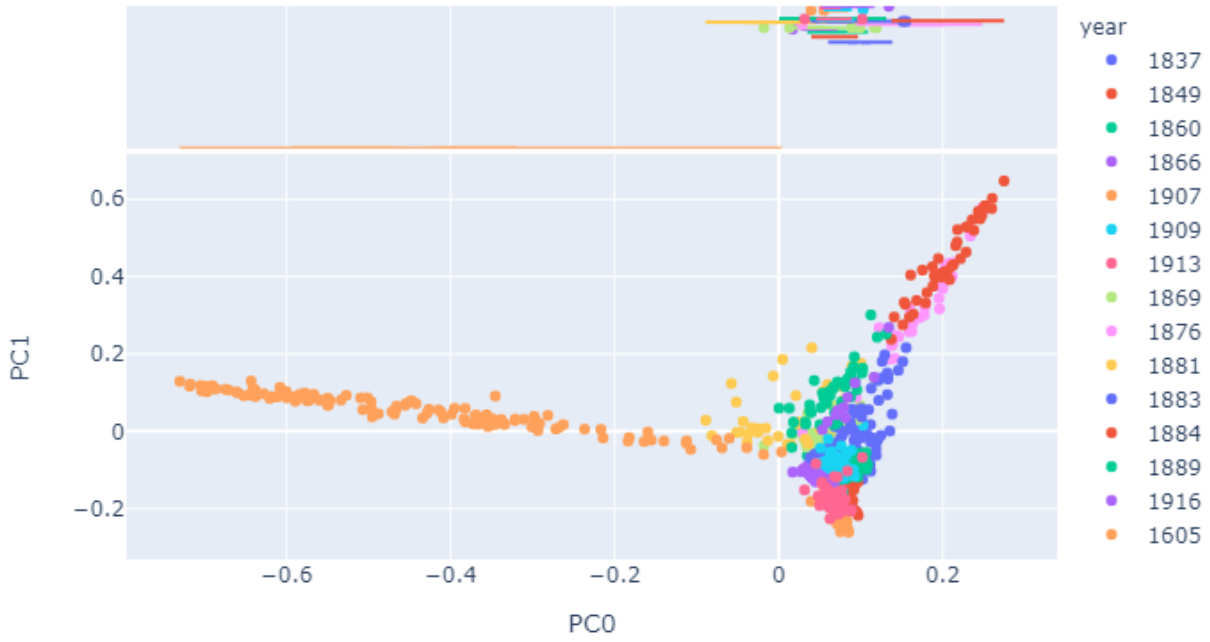

Figure 7.

I created another visualization to determine potential pattern in years the works are published (Figure 7). It seems that except for Don Quixote which was written in 1605, no other notable pattern could be examined based on the year the works are published.

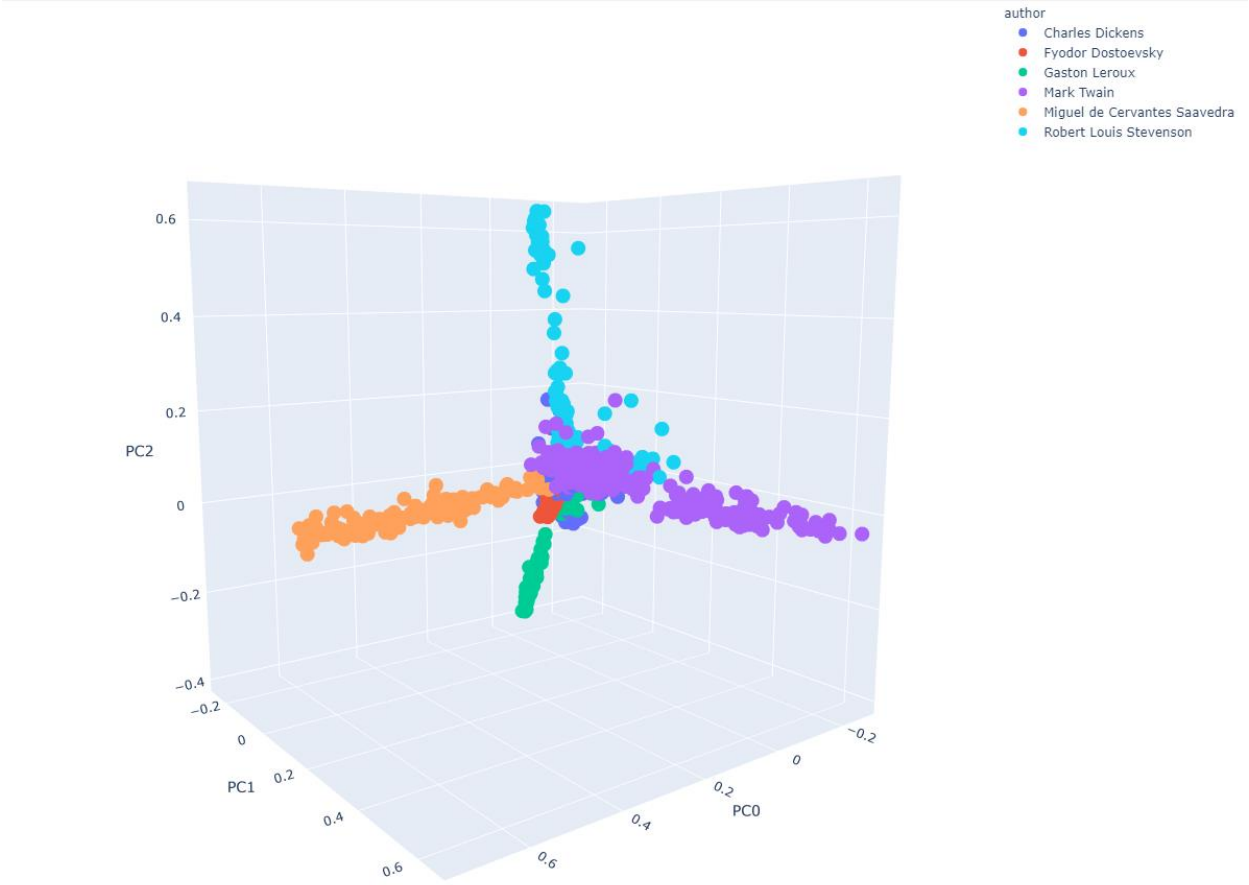

Figure 8.

The 3d scatter plot with scikit-learn shows a four-way segments, each extremely distinctive in its character in terms of a cultural background (Figure 8). To the right (purple, positive PC1) is the American writer Mark Twain. To the left (orange, positive PC0) is the Spanish writer Cervantes. The upper leg (dark and light blue, positive PC2) is composed of English writers Charles Dickens and Robert Louis Stevenson. The last one with all negative PC0, PC1, and PC2 values (red and green) is Fyodor Dostoyevsky and Gaston Leroux, representing Russia and France, respectively.

II. Topic Modeling (LDA)

Top terms for each author in this collection with OHCO level of paragraph is as below:

Charles Dickens	door room moment table hand floor chair face light bed
Fyodor Dostoevsky	door room moment table hand floor chair face light bed
Gaston Leroux	door room moment table hand floor chair face light bed
Mark Twain	time people night town men things day way horses country
Miguel de Cervantes Saavedra	thee thou world knight knights art errant murderer time squire
Robert Louis Stevenson	ye gentleman right man water minute horse bit miles mile

It's immediately noticeable that Dickens, Dostoevsky, and Leroux share the same top terms. I couldn't figure out why and could only suspect a small sample size that might have affected Dostoevsky. Mark Twain has location topics such as "town" and "country," probably due to the adventure or/and traveling content. Cervantes has the most distinctive topics such as the old English ("thee" and "though") and medieval status titles ("knight" and "squire"), given that Don Quixote is written in 1600s, while other works are written in 1800s and early 1900s. Stevenson's top terms are related to adventure or traveling, such as "water", "horse", and "miles."

Top terms for each author in this collection with OHCO level of book is as below:

Charles Dickens	**time** hand **man** head **way** aunt night face house room
Fyodor Dostoevsky	**man time** room day door **way** face eyes moment woman
Gaston Leroux	room voice door **time** ghost eyes **man** evening chamber night
Mark Twain	**time man** day thing **way** people men place years night
Miguel de Cervantes Saavedra	**time** master thou worship **way** thee **man** world life knight
Robert Louis Stevenson	**man** ye **time** men hand house captain **way** face sea

Almost all of the authors share the topic "time", "man", and "way". It would be interesting to conduct a further research and determine whether these topics are central to any novels or unique to these renowned authors. As a comparison, the top terms for Austen and Melville were "**time man** day sister room thing father mother house **way**" and "**man time** men ship day **way** deck sailors sea night" respectively; these also contain the three topics in question.

This time, Dostoevsky and Leroux both have top terms more unique to itself. It seems as if Dostoevsky has a focus on facial expressions ("face" and "eyes") while Leroux is creating a mood for mystery and horror ("ghost", "chamber", and "night"). Cervantes still has the same old-English terms; the set not includes a broader topic such as "world" and "life." Stevenson's top terms now include "captain" and "sea"; they are probably related to Treasure Island, his one of the most well-known novels.

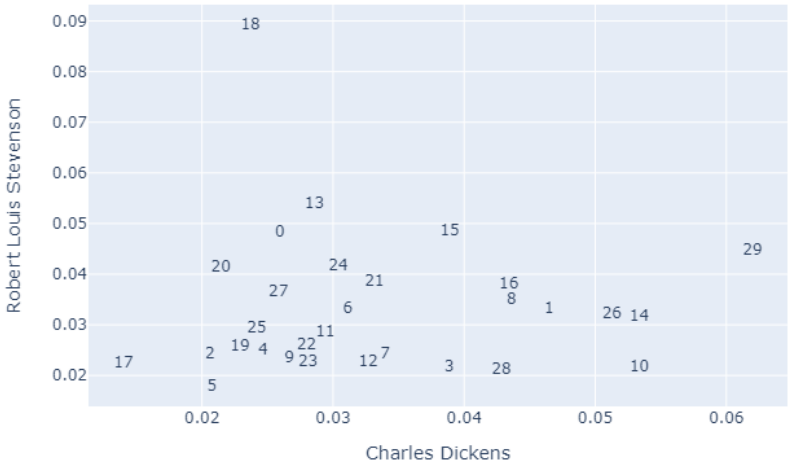

Figure 9.

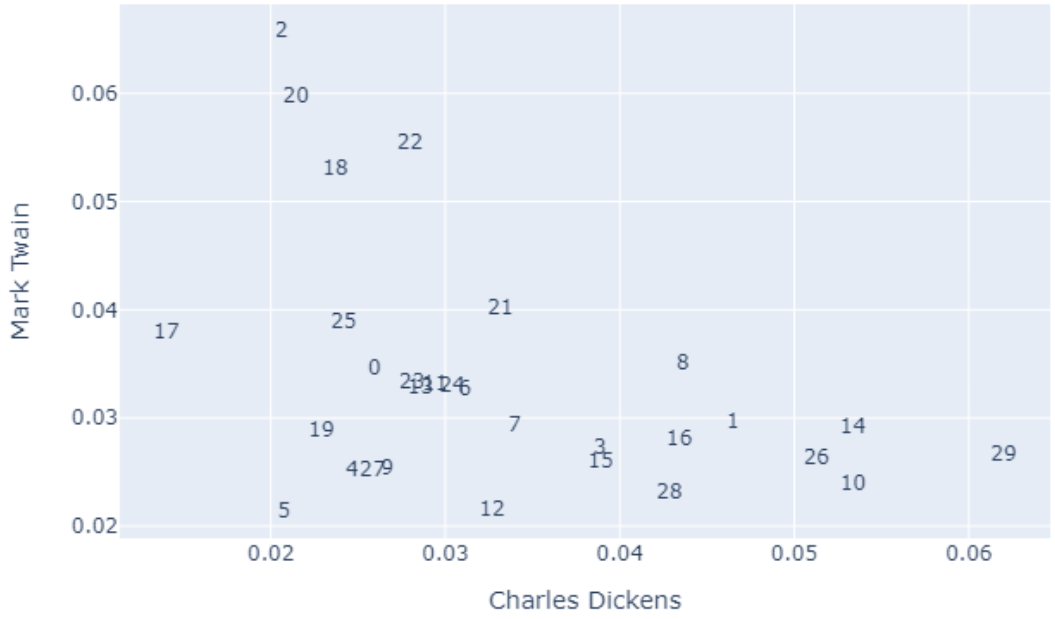

Figure 10.

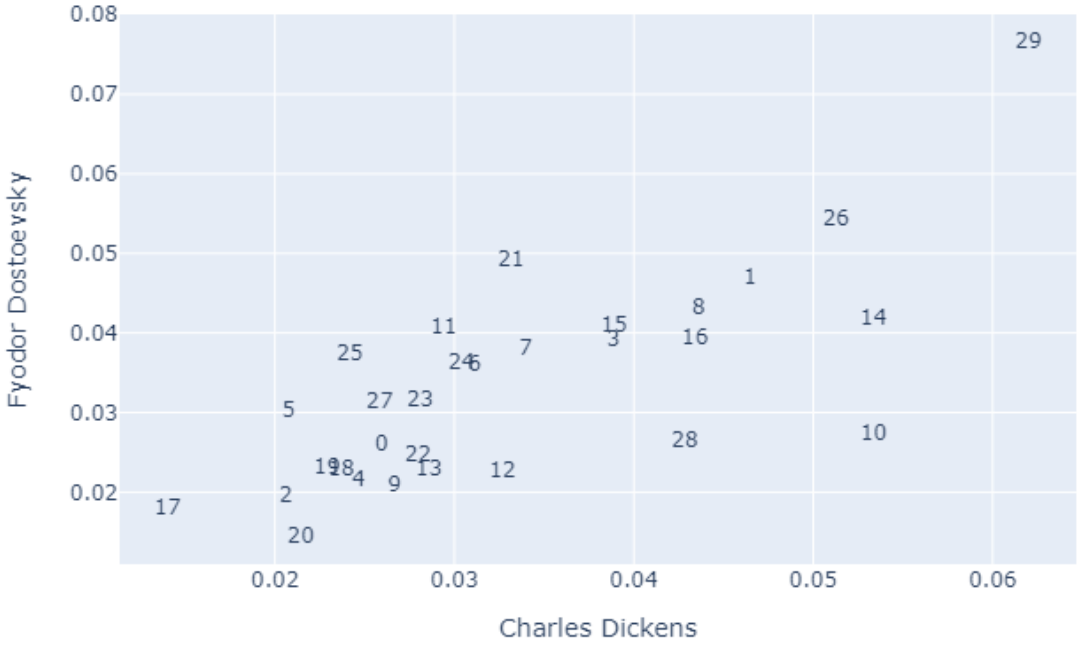

Figure 11.

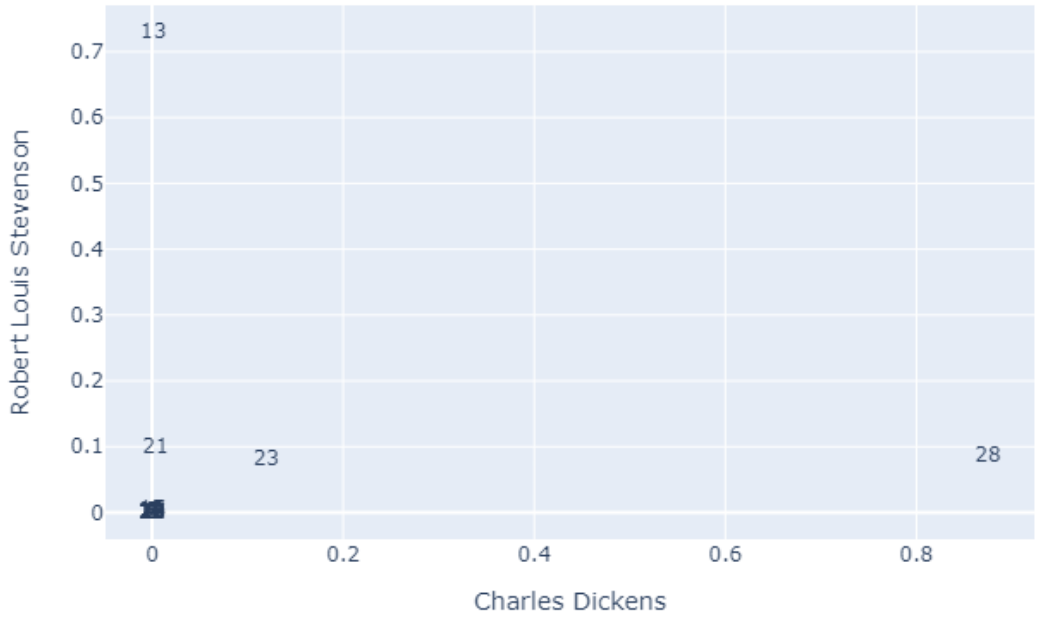

Figure 12.

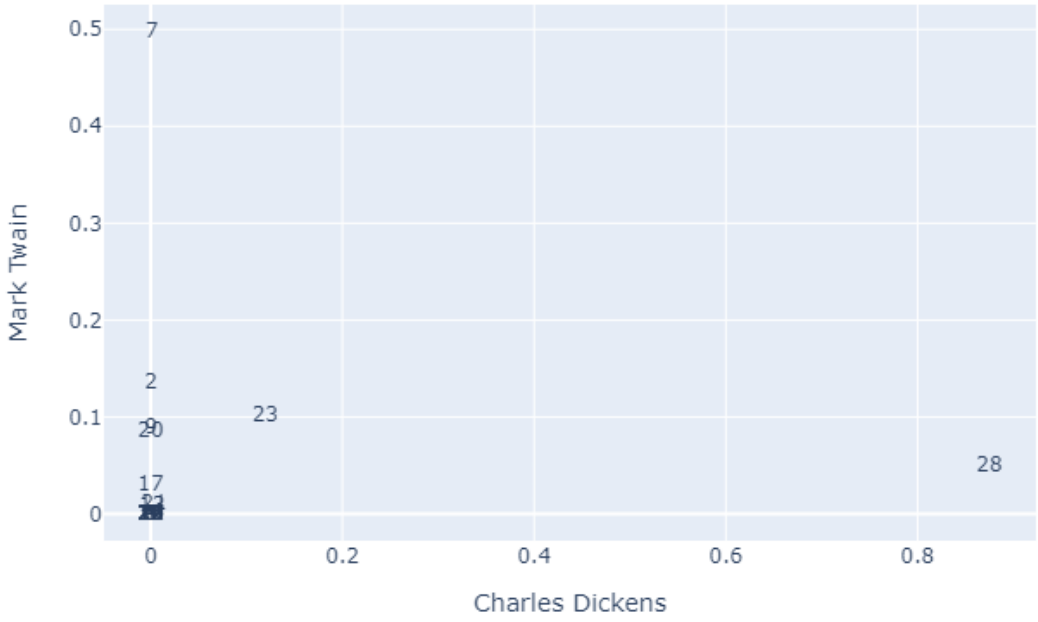

Figure 13.

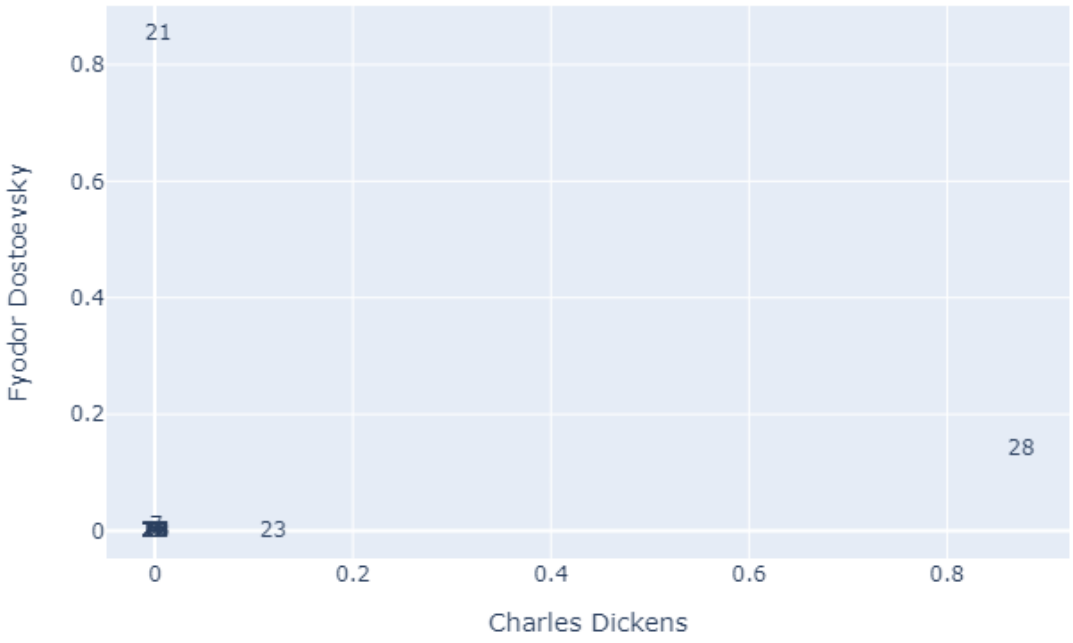

Figure 14.

Figures 9, 10, and 11 explore relationships of topics in a paragraph level in Dickens's works against that of Stevenson, Twain, and Dostoevsky, respectively. Figures 12, 13, and 14 study the same but in a book level, which show significantly less relationship, or even lack thereof. This might suggest that the authors share a similar writing style in terms of plot progression or word choices at a lower level, while the main theme or storyline is unique to each author.

III. Word Embedding (word2vec)

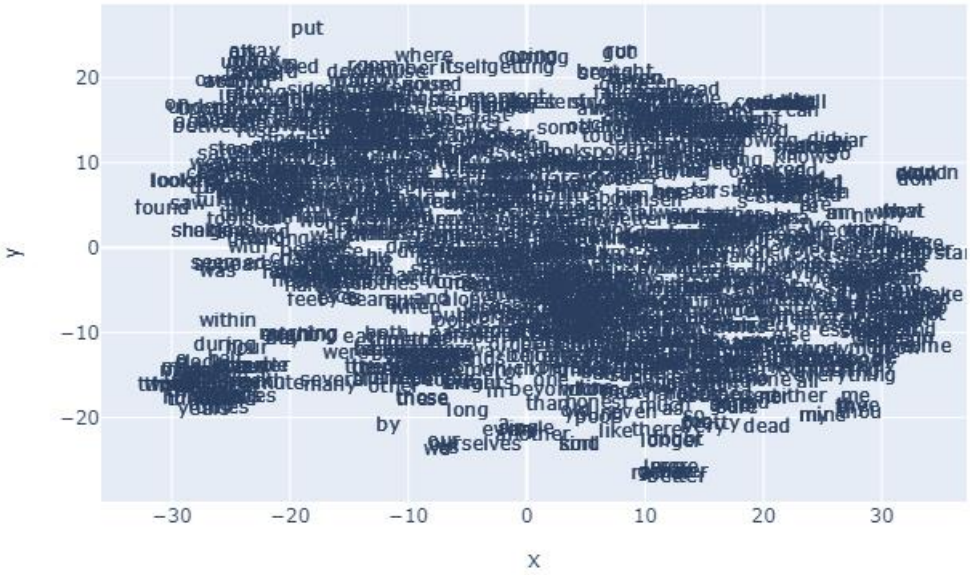

Figure 15.

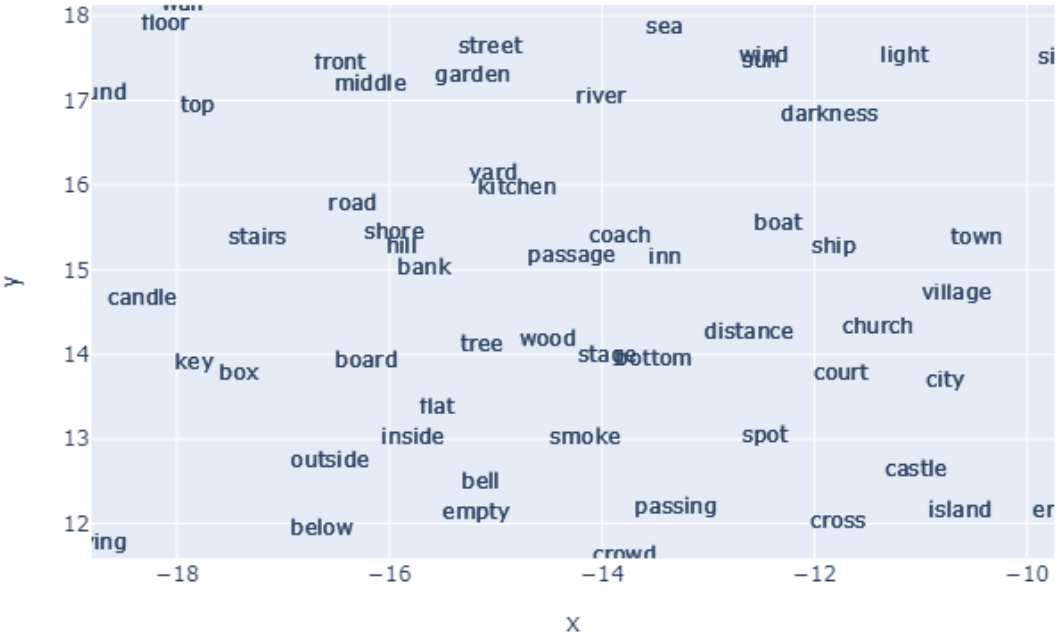

Figure 16.

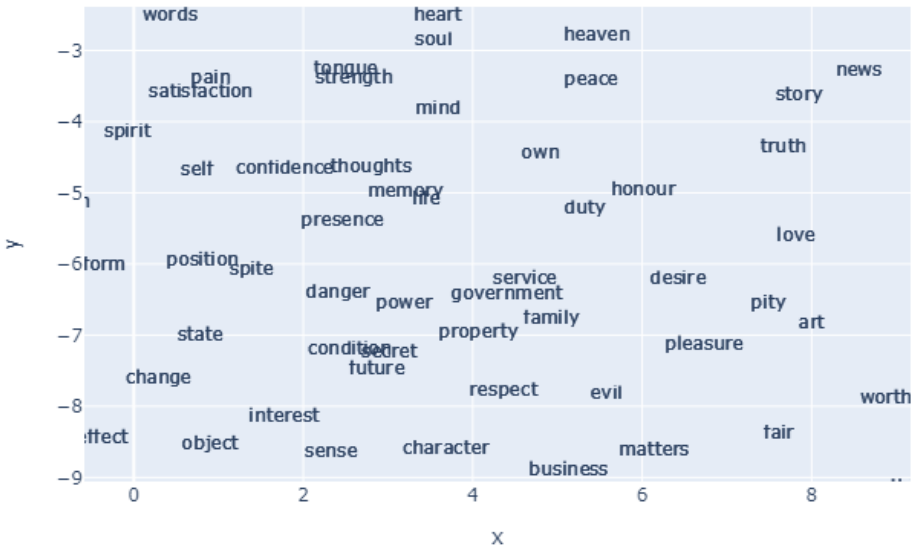

Figure 17.

There are many word clusters in this corpus but only a few noteworthy ones, excluding numbers, body parts, family members, etc. (Figure 15). The one I found interesting was a set of locations: shore, hill, bank, inn, boat, road, kitchen, church, village, court, city, town, street, etc. (Figure 16). Another cluster that seem important includes more conceptual words such as confidence, thoughts, memory, life, duty, honour, peace, soul, heart, pain, satisfaction, etc. (Figure 17). This cluster seems to be composed of both positive and negative words, such as danger, power, evil, pleasure, pity, love, etc. What I didn't expect in this cluster was the word "family," which is usually associated with other family member terms. This made me think that this cluster could be heavily associated with the central motifs of the novels.

IV. Sentiment Analysis

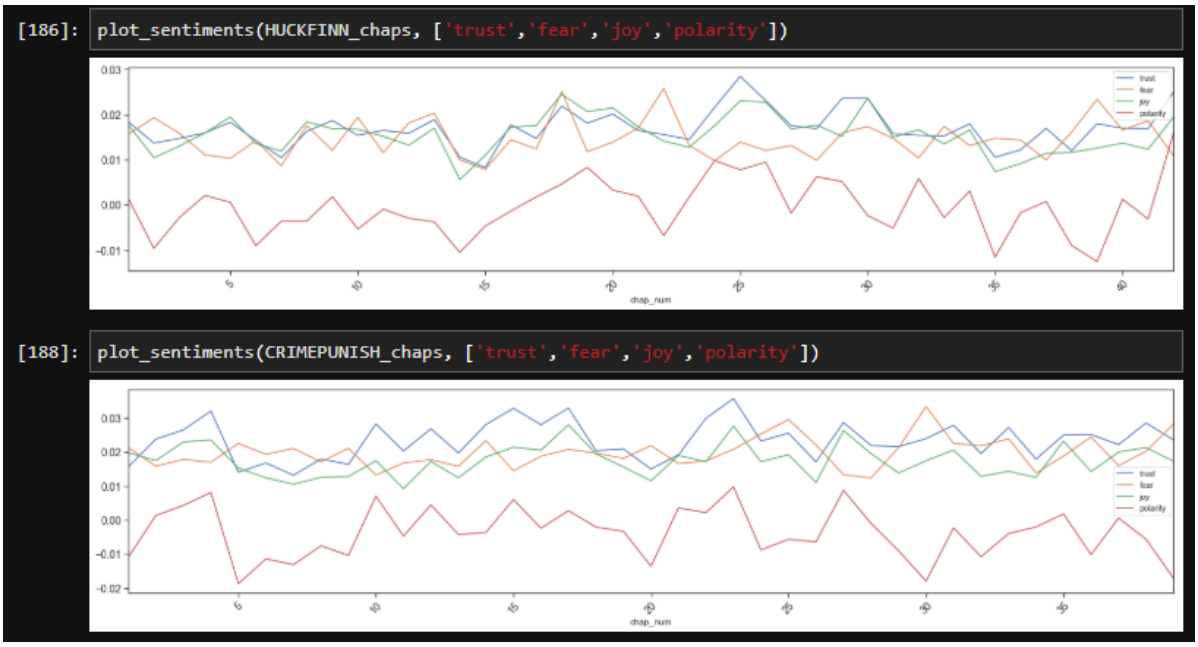

Figure 18.

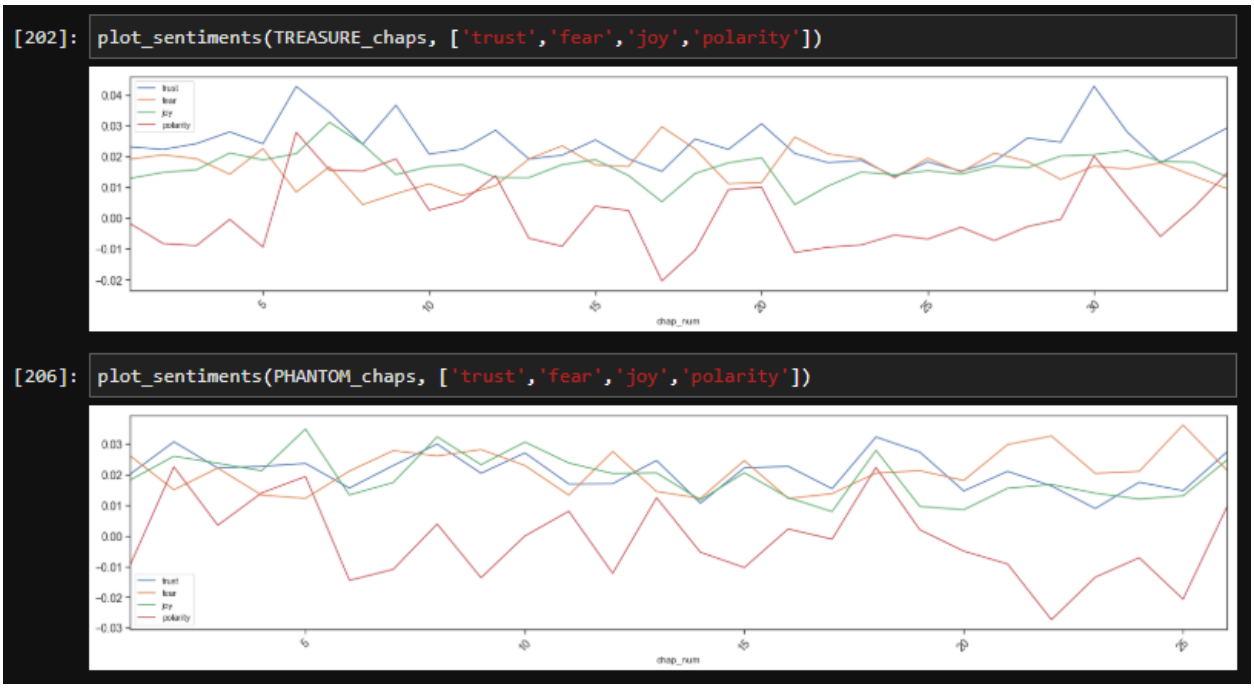

Figure 19.

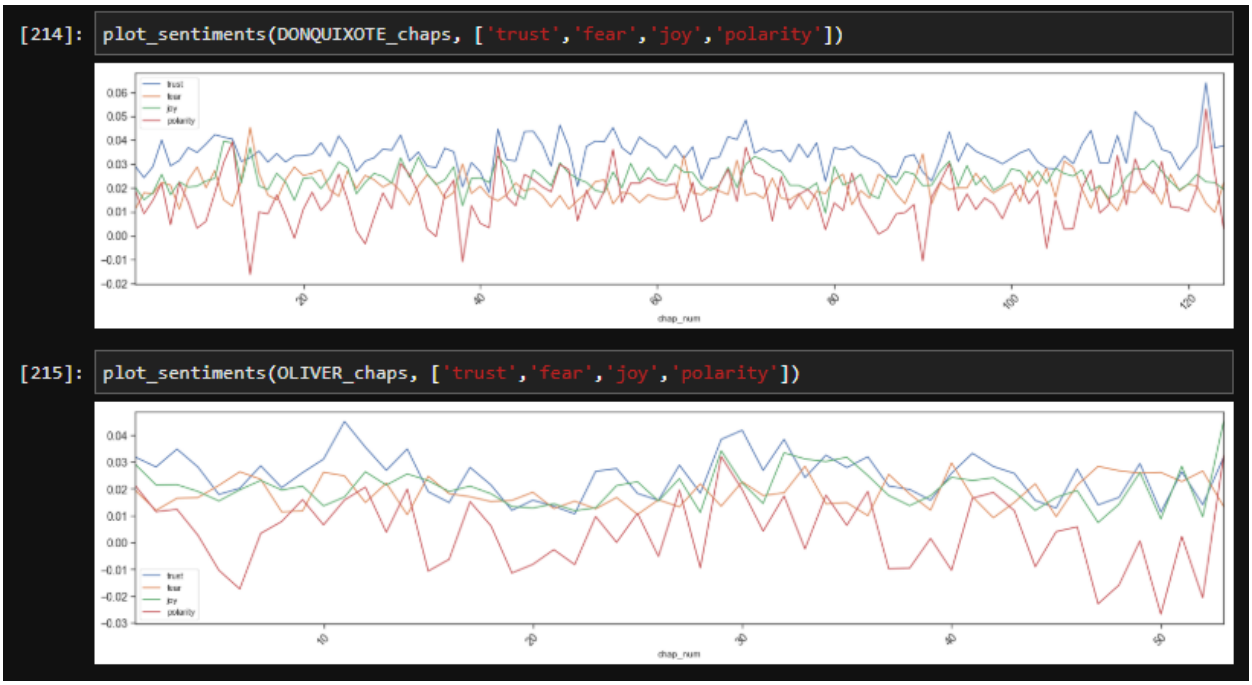

Figure 20.

Sentiment Analysis by chapter of a select novel for each author is shown in Figure 18, 19, and 20. I could immediately note that Cervantes's Don Quixote had the most synchronous sentiments and polarity. The polarity being the difference between positive and negative sentiments, it's almost always inversely proportional to "fear." The adventure and coming-of-age novels tend to have a pattern of negative polarity recovering toward positive near ending. This is probably because most of the stories in this genre have a protagonist overcoming a major obstacle or conflict. In my opinion, sentiment analysis are more closely associated with genres than cultural backgrounds and therefore I tried to avoid generating a false sense of interpretation of the results.

V. Clustering and Similarity

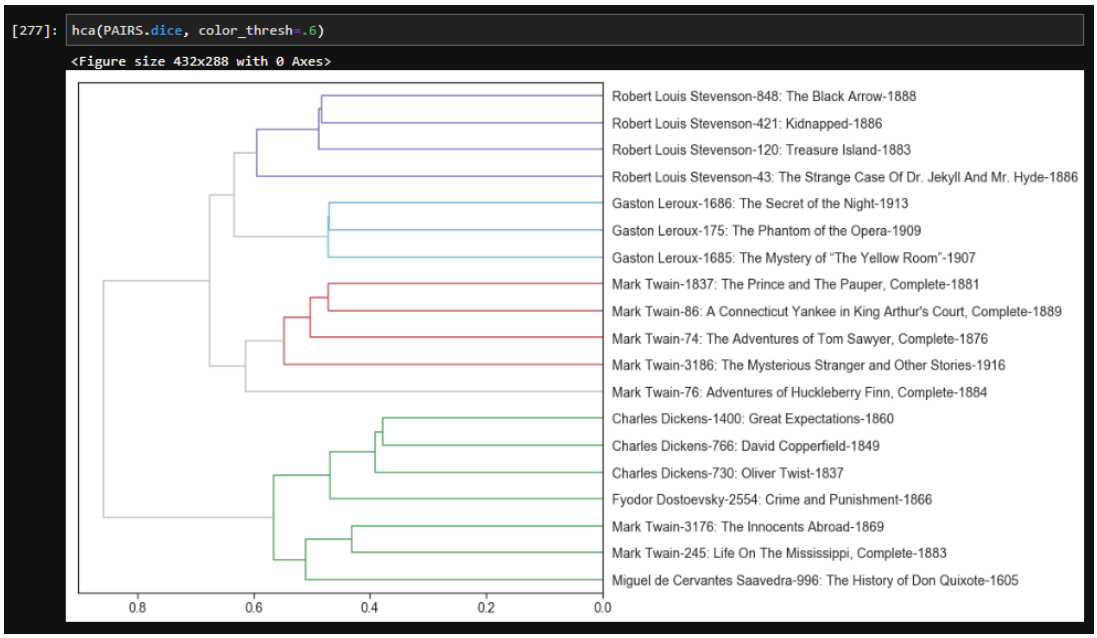

Figure 21.

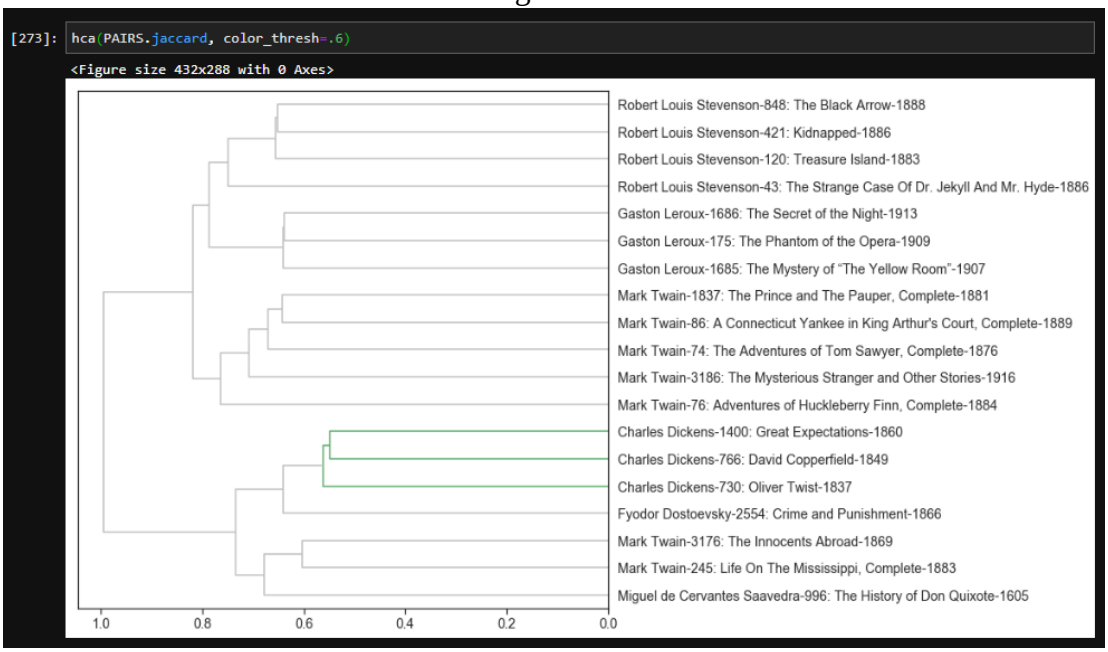

Figure 22.

Most of the novels were grouped together by authors when using jaccard or dice as the linkage method (Figure 21 and 22). There are two Mark Twain writings that aren't grouped together with the rest of his works: "The Innocents Abroad" and "Life On The Mississippi." These are Mark Twain's only two non-fictions in this collection. The fact that these two writings are more

closely associated with Don Quixote might suggest that the writing style for fictional novels back in 1600s more closely resembled that of a non-fiction in modern era.

Conclusion

The writings of the authors studied in this analysis seem to share similar styles at a lower (i.e. paragraph, chapter) level, regardless of genre. However, they have unique characteristics in terms of topics and central themes at a book level. They also seem to connote distinct cultural backgrounds innately. In other words, these books can be varying in topics and themes based on the authors' cultural backgrounds or genres. For example, there is a clear distinction between American, English, European, and Spanish writings based on the results of this analysis. However, there's a similar linguistic pattern in these novels which is what perhaps makes the universal readers to collectively comprehend these pieces as well-written and timeless.