DS5001 Exploratory Text Analytics
Final Project - The Harvard Classics
Final Report

Derek Liu (hdl5hz@virginia.edu)

May 11, 2021

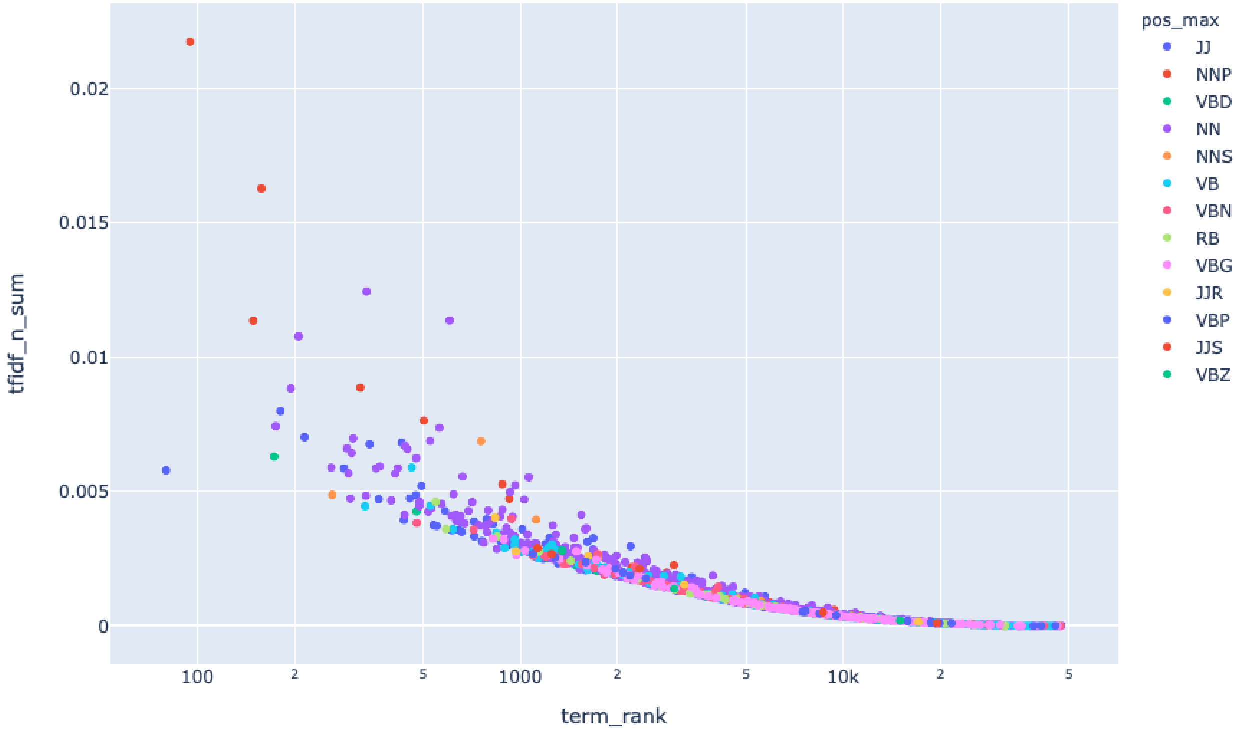
The Harvard Classics is a 50-volume series of classic work from world literature, speeches and historical documents that are compiled by Harvard University President Charles William Eliot. Formerly known as Dr. Eliot's Five-Foot Shelf of Books, this collection was believed to have the benefits of a liberal arts education if carefully read. Archive.org has transcribed the entire collection, however, due to formatting and conversions, the PlainText files are littered with foreign symbols and spelling errors caused by scanning and potentially faulty optical character recognition. Project Gutenberg, however, has the majority of individual books that are within the collection manually transcribed and free to download on their website. This project then aims to use text analytics methods to convert, annotate, and explore the texts that are in the Harvard Classics Collection.

The collection on Project Gutenberg is not based off of each volume, rather than individual works within each volume of the collection. The database's Harvard Classics bookshelf is sorted by popularity in downloads, and while the number of books on this bookshelf covers almost all works in the Harvard Classics, it is difficult to download every book, poem and essay within the collection without large manual hours of navigating the website. For this project, it was decided that the most popular books - those with more than 1,000 downloads, are to be used for the analysis. There are 39 individual texts that fit the criteria, and after further exploration of each text, 22 of the texts remain. The other 17 texts were either in another language, or didn't have a structure that could be documented in the Machine Learning Corpus Format and analyzed in the Standard Text Analytic Data Model. The list of 22 source files are included in the Project Manifest.

The remaining 22 texts are still as diverse as the original 50 volume collection of the Harvard Classics, ranging from *The Odyssey* by Homer to *The School for Scandal* by Richard Brinsley Sheridan. These texts' total download amount is greater than the next 150 texts combined in the Project Gutenberg database, which could imply the popularity of the texts even through the test of time. I believe that the 22 texts could be a good representation of the collection in terms of general theme and topics, but could also have sample bias as it is still a small sample compared to the entire collection.

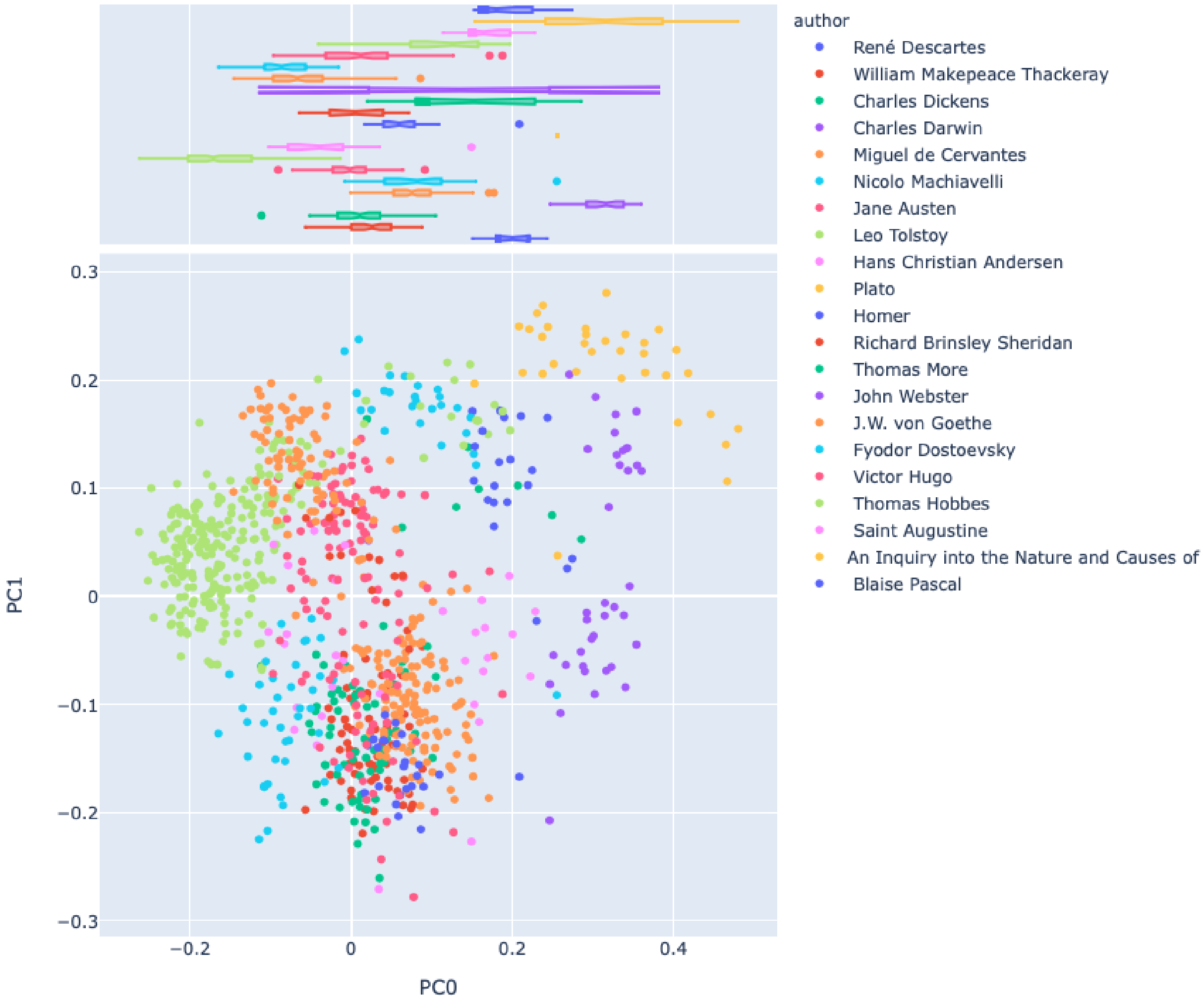
First, the source formats (F0) in PlainText are converted into the Machine Learning Corpus Format (F1) through Python data-wrangling. Next, the corpus was combined and converted into the Standard Text Analytic Data Model (F2). After these two conversions, we arrive at these following tables: LIB, VOCAB, TOKEN. The LIB data frame contains high level data such as title, author and source. The VOCAB table contains chapter, paragraph, sentence and more specific line-by-line data of all the texts combined. The TOKEN table tokenizes the VOCAB table and is used further along the VOCAB table for exploration and analysis.

Next, these tables are manipulated to contain the TFIDF (term frequency - inverse document frequency) data that reflects how important each data point is to the entire corpus (F4).



The table above displays the term rank and TFIDF sum of counts in a scatter plot, separated by part of speech. We gain insights about the entire corpus of source files and see that the words “Sir”, “God”, “Good”, “Money” are all in the top left of the graph, signifying their importance in the corpus. These words are definitely attractive to those who would purchase the collection, and showcases values of Dr. Eliot himself as well. They can be inferred as being well mannered and good hearted with a belief in God is important to those who read this collection, and Money is something that is natural as those associated with Harvard are also usually associated with money.

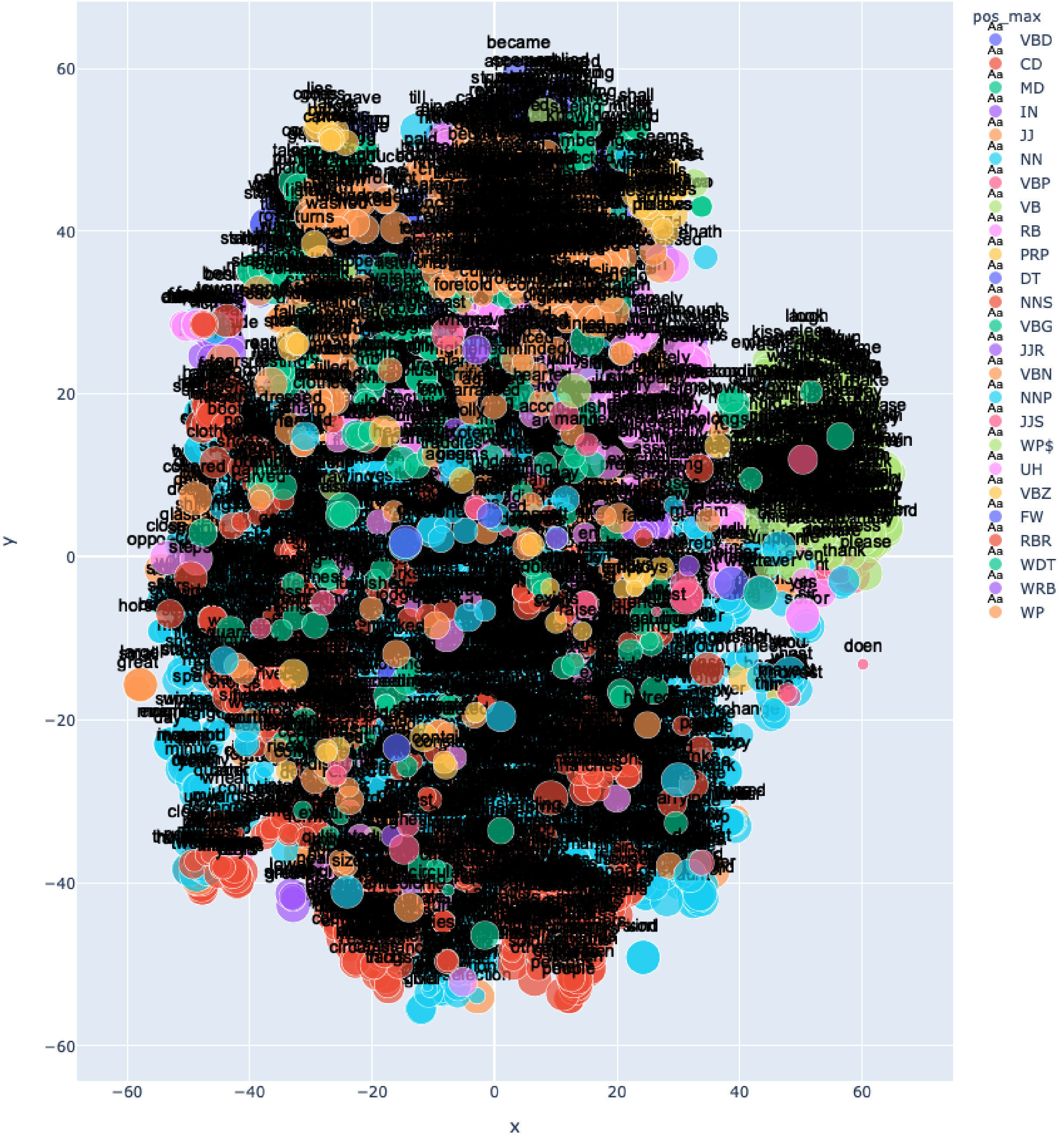
Next, we conducted Principle Component Analysis, which resulted in interesting results of the top authors in the corpus.



The Principle components graphical output shows the most important Principle Components and the relationships among authors. This graph proves the diversity of authors and their works, and that the entire corpus isn't a mere collection of similar books. Something interesting here to note is that Adam Smith's Wealth of Nations is represented as the yellow dots on the graph, which deviates the most from the other clusters. This could suggest that the majority of the corpus contains fiction rather than educational non-fiction such as the Wealth of Nations.

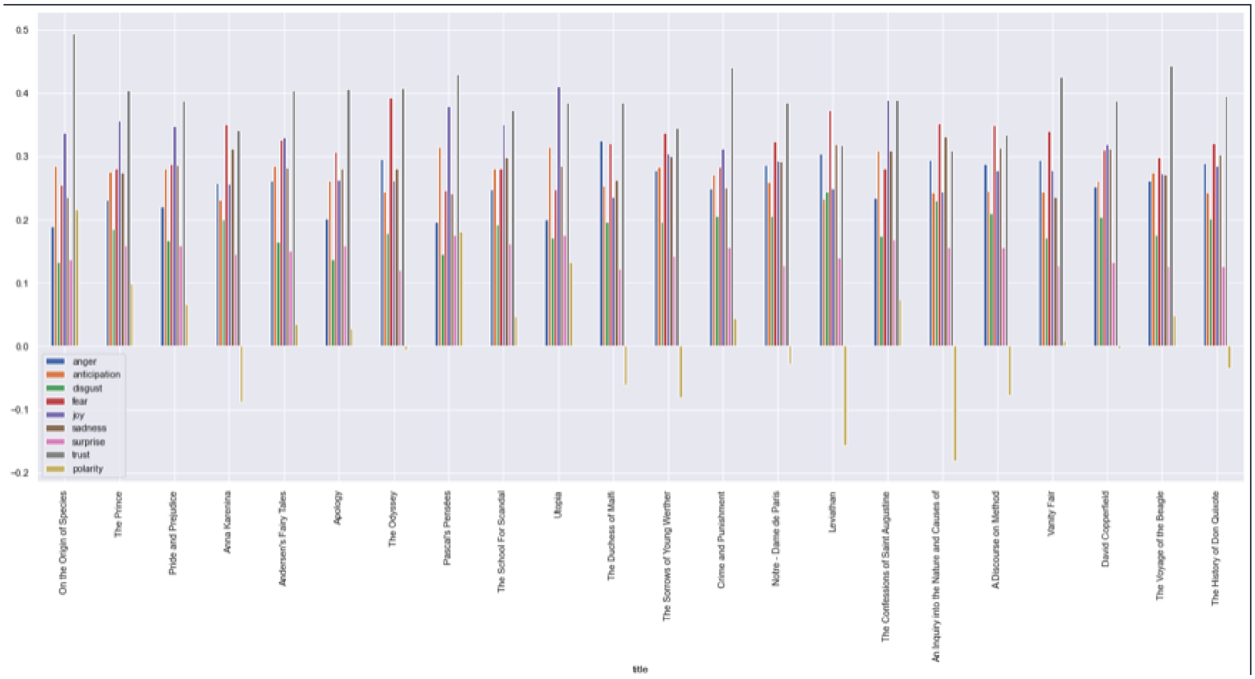
Next, we explored the topics of the corpus through topic modelling. The top topics included words such as man, father, day, house, wife, life, friends, master, knight, etc. These words are family and honor oriented, and can be interpreted as the corpus being centered heavily at family topics. This is not surprising, as we see from the PCA that the majority of the works clustered together are fiction.

Our third exploration focuses on word embeddings, the following graph shows the word embeddings by part of speech.



It is quite difficult to navigate through the graph as it is not interactive here, however, it is an interactive graph within the Jupyter Notebook that is attached with this report. There are a lot of clusters that the analysis captured, and we can recognize that there are a diverse number of word clusters, which makes sense since the collection is very diverse in nature. Something very interesting about the cluster is the small cluster on the right of the graph, colored in green. This cluster seems to be an outlier and contains words such as run, defend, save, tend, etc. This could imply that the genre of combat and escape is not as prevalent and connected to the majority of the collection.

The last exploration we focused on was sentiment analysis. By using the SALEX lexicon, we analyzed each title within the corpus and their corresponding sentiment. As we can denote from the graphical summary, trust and joy are the two major sentiments that are prevalent throughout this corpus. If the model is accurate, it could mean that the entire collection is of a trustworthy and generally happy sentiment, which is something a book buyer back in the early 1900s would presumably want from a highly touted collection from the best University in the world.



In conclusion, this subset of the Harvard Collection gives us a glimpse into the entire collection and the vision of Charles William Eliot himself. Being a good person, believing in God, being family oriented, trustworthy and joyful are all depicted throughout analysis, and if this subset encapsulates the entire collection, then it is indeed a good source of information and education for an individual wanting to become a better person.