Iteration 2 Summary Beau Carlborg, Clara Richter, Sunny Chiu

First steps: Stripping the books of introductions and project gutenberg info

Following our data collection process from iteration 1, the most obvious next step was to strip the books of all irrelevant data to be able to use them for training data. Due to the crowdsource nature of Project Gutenberg, this posed a significant challenge. There was no consistent way to remove the beginning and end of every book using scripting, so we decided instead to take a more manual and brute force approach to stripping the data. To do this, we divided the nearly 700 books we had present in our project into three equal groups, and each manually stripped the beginnings and ends of each book.

Redefining the problem:

While the brute force approach to stripping the books was time intensive, it provided us the opportunity to explore the data in much more depth. In manually parsing every text to remove the table of contents and other irrelevant info, we each gained a much more in-depth understanding of our data set. In reading through the data, we learned that our data set was much more diverse than we initially imagined. Our data set consisted of many novels and stories, but also, reference books for botany, children's books, the kama sutra, and poetry.

This more in-depth understanding of our data led us to feel more doubtful about our project and also lead us to question the possible outcomes of our project. Due to our doubt, we decided to meet with Matthew Whitehead to gain more intuition for our problems and create a more clear path forward for our project. When we talked to Matthew, he helped give us a better understanding of possible implementations for our project by helping us wrap our heads around the size and scale of our problem. He made the point that with our dataset, we might be lucky to create a valid identifier for 50-year intervals, but even that may be difficult.

After talking with Matthew, our group reconvened on multiple occasions to decide on a path forward. To gain a deeper understanding of our problem and the difficulties we might face moving forward, we decided to do a sort of literature review of existing research in the field of text classification. We read numerous articles and research papers on our topic to gain a deeper understanding of the problem. The pieces we found all focused on the concept of identifying the author of a work as opposed to determining the period a work was published in. Upon reading extensively about the existing literature on author identification, we decided to approach a more straightforward problem before attempting to identify texts based on their author rather than by their publication date.

Machine learning project 2.0:

For our "new and improved" project, we have decided to create a document classifier that works for a small set of authors. We believe this project will serve as a stronger stepping stone for future developments for a few reasons.

1. We believe this project will have a higher initial success rate. This initial higher success rate will allow us to have a stronger starting point for understanding our model. We

believe that it will be easier to find flaws in our model and data formatting if we are working on a simpler problem.

2. We believe this authorship identification problem will generalize to our date identifier well. Both projects fall under the same classification of document identification. It is our understanding that creating a strong identifier based off of stylometric differences among authors works will generalize well to identifying stylometric differences from different eras.
3. This project offers many stepping stones that we can use to quantify our progress as we move forward. We believe that using our current dataset, we will be able to build a classifier that works for increasingly larger numbers of authors.
4. In our current dataset, there are a few authors such as Charles Dickens, T.S. Elliot, Mark Twain and others whom we have an extensive number of books for.

Building a stylometric profile for the Authors:

To efficiently create an author identifier, it is crucial that we identify a series of stylometric features that we can use as columns in our model matrix. To build this stylometric profile, we have created some metrics to quantify the different qualities of an author's work. To choose these metrics, we talked with Matthew Whitehead and referenced a large number of academic articles on the topic of written text identification. From all of these sources, we were able to come up with two general groups of metrics. One based off of n-grams of different words from the books and the parts of speech, and another based off of quantifiable qualities of the work. The metrics we chose to use to quantify the texts are the length of the sentences, the word length, the word richness, the number of function words, as well as the distribution of punctuation and function words.

Of the metrics we are planning to use, we have various options moving forward. For each of these metrics, we can either add columns for the relative frequency of each of the occurrences (for example, we could add a column for the relative frequency of words that are one character long, another for words that are two characters long, and so on). Or, we could choose ways to quantify the distribution to the capture the qualities of the distribution (for example, the standard deviation, and a measure of center for the distribution). These are decisions we have to make for each of the columns.

Evaluating our columns:

One significant question that remains unanswered in our project as of now is the relationship that the number of columns in each of the rows we process has on the model itself. We view ourselves as moving down one of two paths as of right now. There is one situation in which we see ourselves using a model that is trained on data rows with are thousands of columns wide which are based on relative frequency of n-grams. The other situation we envision ourselves in is one in which we are using fewer rows that are based on the metrics we have constructed using NLTK.

Another question that our group is still grappling with is how we can decide on columns that are effective at capturing data and contribute to a strong model. As of now, we are using our knowledge of simple distributions to capture similarities and differences in our metrics across various books and authors. We believe that with our current understanding of machine learning and training this is the most robust predictor we can use to identify useful strong column metrics.

Deciding on a model for moving forward:

In order to move forward, we still need to weigh our options when considering which models to use in order to perform our analysis. In the works we read, we observed that other researchers were using support vector machines, artificial neural networks, decision trees, random forest, and k-nearest neighbor in order to conduct their classification. At this moment, we are strongly learning towards using a support vector machine or Artificial Nueral network to perform our classification. We are planning to perform identification on a small number of authors initially. And then continue to perform classification on a larger number of authors. While we are constructing our model, we are considering a few options for our training. Based on the availability of data, we are weighing our options between using a k-fold cross-evaluation. To decide, we need to conduct more research and possibly attempt running our algorithm using different data training methods. To evaluate our model, we are going to begin by using true and false positive evaluations for our model and then continue to possibly use precision, recall, and f-measure.

Group Member Involvement:

During this iteration the entire group collaborated on many different elements of the project. At the onset of the iteration, all members of the group met and conducted data cleaning on over 700 books which we planned to include in our data set. Upon completing this crucial step, Beau and Clara approached Matthew Whitehead to talk with him and get his perspective on the ideal path forward to construct a matrix which could be used to train our model. Upon talking to Matthew, the group met up again and began constructing scripts using NLTK which could be used to extract word and part of speech N-Grams from the texts. Beau and Clara also used this time to conduct background research on the domain of classification problems. Eventually, after each group member conducted quite a bit more research, the group began creating python scripts which could be used to extract the quantitative stylometric features from the texts which we passed to it. After redefining our project to be author classification based, Beau, Cleaned the new data set, and while Clara and Sunny allowed the data pipeline to accommodate the new dataset. During the final stages of the project, Beau made the presentation and the summary while Clara and Sunny created the Visualizations and finalized the notebooks.

How to run our code:

A description for running each of our notebooks can be found in the top 3 cells of each notebook.