Update Description Beau Carlborg, Clara Richter, Sunny Chiu

Updated files

In order to maintain continuity in our project notebooks, we chose to reflect the update to the first project notebook in two seperate jupyter notebooks. In the first of our notebooks titled **data_retrieval_preprocessing.ipynb** we show our updated scraping and parsing scripts which can be used to retrieve works from relevant authors on Project Gutenberg. In the Second of our jupyter notebooks titled, **feature_extraction.ipynb** we show our new method of quantifying the number of books per author we were able to retrieve.

Simplifying and refocusing our project

During this iteration, we gained deeper insight into the general nature of classification problems and the many intricacies that are involved in constructing a robust classifier. Among many other details about classifier problems, we learned about the specific nature of using stylometric features to create an identifier and also the general relationship between bucket size and the number of data points in any given bucket. To our understanding, having fewer buckets, and more data points per bucket enables the model to perform at a higher level allowing more accurate classifications. Due to our more in-depth understanding of stylometric features and classification problems in general, we decided to create a model which identifies the most probable author of a text based on stylometric features.

As can be seen in our updated project_1 notebook (represented in the two files mentioned above), using authors as classes for our model enables us to have more data points per author and also allows us to have stronger similarities among our data points. During our process of data retrieval, we learned that our data was more strongly varied than we initially expected. The variations among our data came as a result of the limiting nature of using documents for which we could find original publication dates for. After beginning our data cleaning process, we learned that our dataset reflects a much more extensive variety of genres and styles than we could have initially guessed. Because of this variance of genres in our data set, we chose to maintain a reasonable number data points per bucket by classifying authors as opposed classifying works by their publication dates.