Iteration Two Update Summary: Beau Carlborg, Clara Richter, Sunny Chiu

Improving our data processing:
This iteration, we revisited and expanded upon a lot of the work from our second iteration of this project. We drastically expanded upon our models ability to accomodate large amounts of n-gram based features. We represent this change to our second iteration in our new jupyter notebook titles feature_reduction.ipynb. In this file, we display the functionality of our new feature reduction functions that can reduce the number of columns in our data using concepts of sparsity and of important differences in the summary statistics of our data. In order to decide which of our authors would be the best to process, we chose to graph the total number of bytes in each of the authors books.

Displaying the reduction:
In order to display the data reduction, we show the process of reducing the number of columns in a text when generating the bigrams for two authors works in order to create a model. This general trend carries through most of our data reductions. While we do use our jupyter notebook to display the changes in data, we do not perform all of the transformations of our data set in the jupyter notebook because it simply represents too much computation for a single machine. We constructed the csv's which we used for the graphs in our jupyter notebooks on each of our machines over the course of this iteration.

Files we updated

1. **data_retrieval.ipynb** -- Added two graphs showing the distribution of file sizes for authors.
2. **feature_extraction.ipynb** -- Updated our feature extraction files to be more optimal.
3. **feature_reduction.ipynb** -- Created this file in order to show how we limit the overall size of the data.