For the personal big data project, I set out to classify lines from different works based on their authors. I chose this project because I am an English/Computer Science double major, and I love reading and analyzing literary works. This project demonstrated a way for me to analyze literature in a different way and introduced me to the field of the digital humanities, something that one of my English teachers discussed with me previously. I learned about this field more in our Taming Big Data course, and it is something that interests me greatly. I hope to build on this project in the future and strengthen it. Eventually, I want to conduct research into a machine learning project in which a program could classify unknown ancient manuscripts and sort them into genres, authors, and time period. For this project, I got my data from Project Gutenberg, which allowed me to download the plain text files of several books by a handful of major authors. The authors I used in this project are all highly respected authors who have influenced me substantially. While a simple classification project, it challenged me to analyze the inner workings of the classification algorithm because sometimes the size of the data prevented the program from working properly.

I chose six prominent authors (Shakespeare, Edgar Allen Poe, Charles Dickens, Mark Twain, F. Scott Fitzgerald, and Oscar Wilde) and used 20,000 lines from some of their most well-known works as training data. The six authors are the categories into which lines may be classified. I used multiple works from each author to guarantee that the training data is diverse and numerous for each author. I did not exclude any words from the training data because the particular style of each author could affect the frequency of even simple words like “the,” marking that they are all important clues to be taken into account during classification. However, I glanced over each document and excluded the lines at the beginning of the works that listed a table of contents or described the Project Gutenberg disclaimer. I utilized a Fisher Classifier similar to how I used it in the shared big data project. My program runs the classifier on 10 lines from the middle of a literary work by each author. It chooses the best guess for which author it should be based on the training data. My program classified the lines correctly for all six test cases. However, when I increased the lines of test data to 50 lines, it made one error. Increasing the bounds to 100 lines or more returns a mathematical error. My program also prints the probabilities of each work being assigned to each respective author. While each probability is minuscule given the quantity of words trained and tested, there are noticeable and exponential differences between the correct author assignment and the other potential authors. I also ran the program using the Naïve Bayes Classifier, and that returned the correct results as well. However, the probabilities it calculated were extremely smaller than those found by the Fisher Classifier. The table below illustrates this finding. The highest percentage for each piece of literature is bolded; each of these values aligns with the proper author.

