MAT 331 Project, Fall 2018
Distinguishing authors

This is a rather open-ended project compared to several of the others and perhaps a little more difficult (but also a little more fun): most of the decisions will be up to you and I am not sure if what I am asking is possible. However, as long as you report your results, even if they fail, you will get full credit.

Can we use the techniques from class to identity the author of a passage? In the `Crypt` folder, I have placed several books by Jane Austin in files (`pride.txt`, `persausion.txt`, `mansfield.txt`, `sense.txt`,) and several by Mark Twain (`sawyer.txt`, `finn.txt`, `innocents.txt`, `roughing.txt`). More books by each author are available from the Project Gutenberg website (they are in several formats, but I have always taken the 'Plain Text' version, which gives an ASCII file).

You may use different authors than I have suggested (e.g., Charles Dickens,...). The Project Gutenberg site has many such books; ones where the copyright has expired, so they tend to be older.

(1) Can you find a way to get `MATLAB` to distinguish Jane Austin from Mark Twain? The idea is to use the books listed above to gather statistics about each author, for example, letter frequencies, how often they use certain common words, the average length of sentence, the number of punctuation marks, or whatever else you can think of.

(2) Write a program that takes in a file and outputs whether it was written by Jane Austin or Mark Twain by comparing it to the statistics you came up with above. Test the program on passages from each author. I suggest taking ten passages of around 10,000 characters for each author (choose a random starting points and take the next 10,000 characters). Test to see how often your code gets the decision correct.

(3) Feel free to refine your ideas to get better results. If you prefer to try different authors than the ones I suggest, that would be fine, but describe what sources you are using.