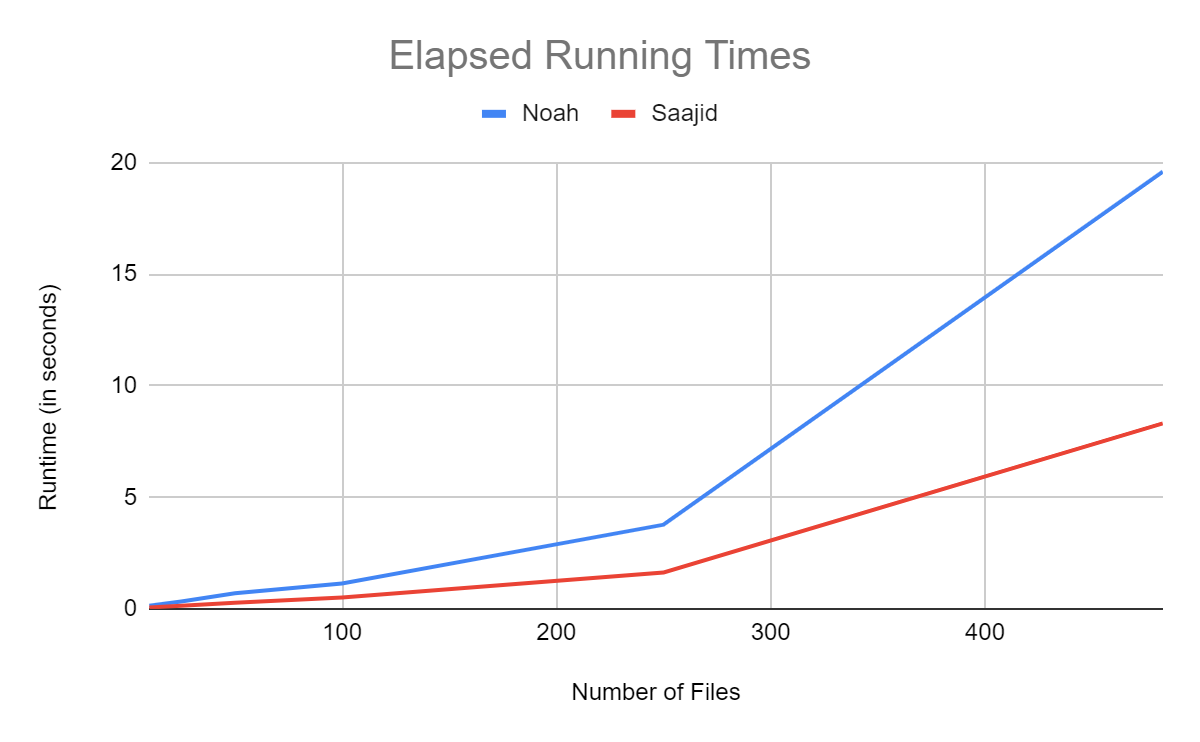
Project 1 Report

How I executed program:

* I used Python
* First off, I included a number of libraries to help this program function:
  + os, bs4 (beautiful soup), re, csv, operator
* Using the os library I would input html files (the given ones) into my program in a separate folder containing only those files
* I had a loop that would open files in that folder one at a time and perform certain actions on that file
  + It would first pass the html to beautiful soup to decode and then I would use a bs4 function to take away all the html code and leave only the content
  + Then I would strip all the unnecessary punctuation so that only words would be left
  + I then took all that was left and wrote that file into another directory as a text file
  + Finally, I took the striped content and looped through it to take each individual word and add it to a global dictionary of word counts (counting words of all files). If said word was already there, I would increase count otherwise I would hash a new key with value 1
* Finally, my program had one last thing to do and that was to output the dictionary in two different ways:
  + One is by order of values, in other words, the highest count words would be first and the rest follow in decreasing order. I sorted dictionary and wrote it into a csv
  + The other one was sorted alphabetically. So, I had to sort by keys and use the operator library in order to alphabetize these strings. The operator library would allow me to compare the strings. After sorting this way, I would output this dictionary into another csv.
  + Link to github: <https://github.com/Saajidf/CMSC476-Proj1>

Runtime:

|  |  |  |
| --- | --- | --- |
|  | Noah | Saajid |
| 10 | 0.1399755 | 0.066193 |
| 25 | 0.3299164 | 0.143665 |
| 50 | 0.6977398 | 0.278305 |
| 100 | 1.1344264 | 0.512824 |
| 250 | 3.7756374 | 1.626152 |
| 483 | 19.5988492 | 8.3144444 |



My program ended up faster, but I think it might have had to do with hardware/software in use because we had similar thought process in how we coded our programs though we did use different tactics for making dictionary and sorting it.

Results:

Our outputs were very similar except he and I had a few different words in dictionary. He had the key ‘s’ in his dictionary, and I didn’t. That may be because of how we striped punctuation differently.

Errors:

* Some files were not working with my program, so I excluded them and put them in another folder. I think my program was not properly identifying their encoding. These are the files: