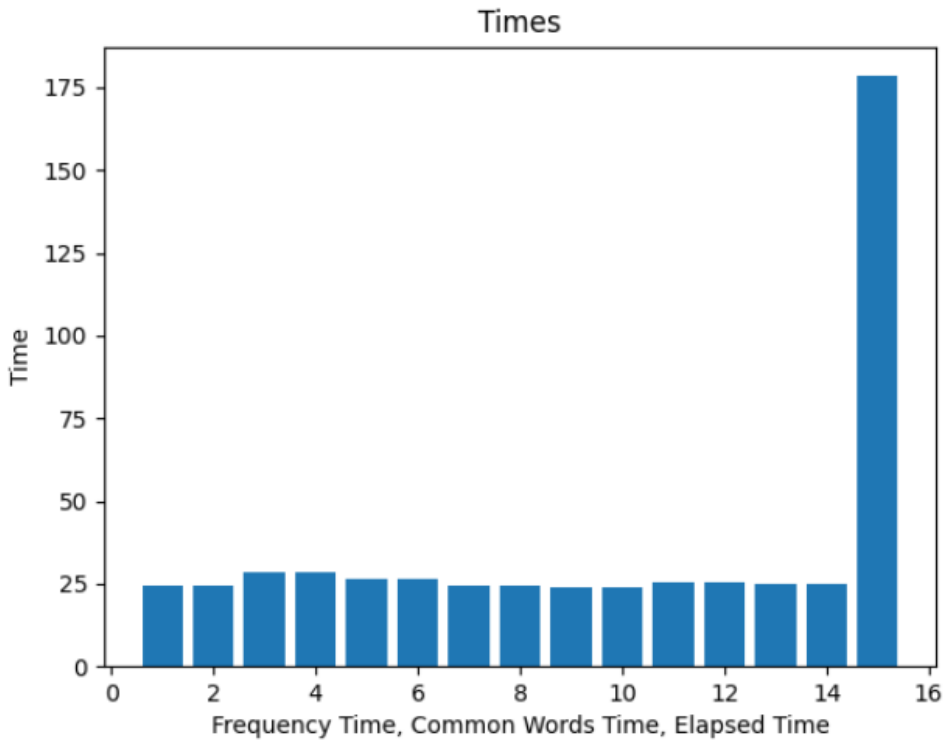
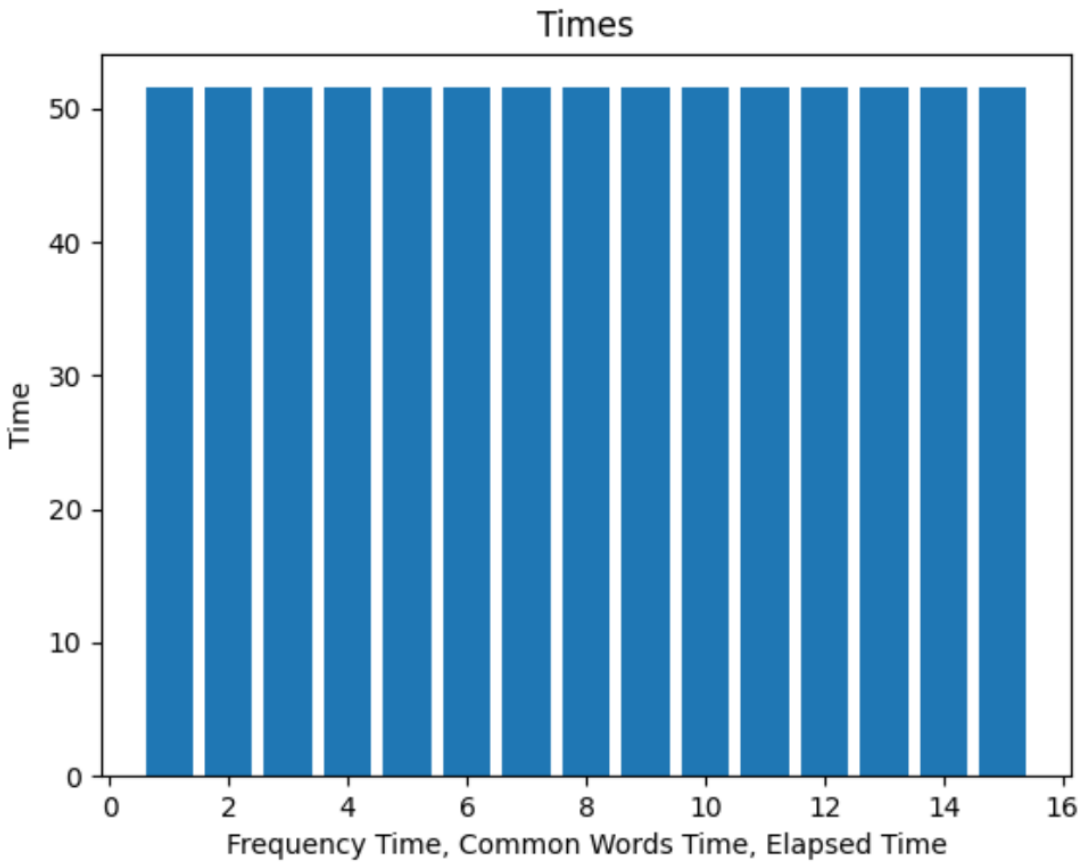
Junnan Shimizu
Project 5: Multitasking - Report

Abstract:
In this project, our goal was to implement any type of multitasking to speed up our reddit comment file projects we did in CS231 where we had to count the words in eight years of reddit comment files and determine the most frequent words as well as the total word count and frequency. Additionally, our goal was to learn multiprocessing and multithreading in Python and learn to implement it.

Results/Discussion:
There are 8 cores on my computer, and I utilized 8 threads within my program.



Every pair of bars represents the time it takes for the program to calculate the frequency of a word and the top 10 most common words respectively. The first two bars are for the first reddit comment file, the second two bars are the second reddit comment file, etc. The last bar shows the TOTAL elapsed time, so in the case of a single threaded approach to this program, as you can see, it took a substantial amount of time to go through all the reddit files.



For this graph on the other hand, the time it takes to calculate the frequency of a word and the top 10 most common words may take longer than the first program, the elapsed time to calculate it for all 8 files is the same as it takes for one because we are distributing the work among 8 threads, so overall it is a lot more efficient to spread the work among the 8 threads.

I saw the process of going through all the reddit files as one large task, and so I spread each file as a smaller task that went to their separate thread. So the processes all finished at very similar times. By running the code, the terminal will also give specific numbers on how long each part took (as the graphs are not very specific), as well as show the top 10 words for each file, total word count, and the frequency of the word “hello” or any word for that matter. Not the entire output is presented below, just snippets.

```
reddit_comments_2009.txt
Frequency of the word: hello, 2.6631570354104626e-05
Time to find word frequency 24.44939241601969
('the', 1670340)
('to', 1057215)
('a', 955408)
('and', 802979)
('i', 790210)
('of', 787500)
('that', 624317)
('is', 602698)
('you', 573452)
('it', 550222)
Time to find 10 most common words: 24.513756791013293
Word Count: 39051396
reddit_comments_2010.txt
Frequency of the word: hello, 2.4667140225970775e-05
Time to find word frequency 28.629104624997126
('the', 1856800)
('to', 1226717)
('a', 1136212)
('i', 1032855)
('and', 949731)
('of', 852227)
('you', 722344)
('that', 686363)
('it', 659529)
('is', 656958)
Time to find 10 most common words: 28.709268749982584
Word Count: 45526153
reddit_comments_2011.txt
Frequency of the word: hello, 2.673959902221168e-05
Time to find word frequency 26.348618500021985
('the', 1671276)
```

```
----------MULTITASKING IMPLEMENTED----------
The number of cores in the system is 8
Frequency of the word: hello, 3.2942170654955496e-05
Time to find word frequency 43.60989212500863
reddit_comments_2013.txt
('the', 1472315)
('to', 983600)
('a', 939100)
('i', 899868)
('and', 794846)
('of', 650920)
('you', 584993)
('it', 550475)
('that', 532495)
('is', 517877)
Time to find 10 most common words: 43.740673916006926
Word Count: 37763146
Frequency of the word: hello, 2.8363194074859203e-05
Time to find word frequency 44.91898458299693
Frequency of the word: hello, 2.6631570354104626e-05
Time to find word frequency 45.01457812500303
reddit_comments_2012.txt
('the', 1502772)
('to', 1023818)
('a', 967683)
('i', 946725)
('and', 815535)
```

Extensions:

I did not do any extensions for this project.

References/Acknowledgements

I did not go to office hours or TA hours for this project. I did get a little help from my brother for this project in order to determine how to tackle multitasking in this project.