

CSC-415-01 Fall 2020

Rinay Kumar

Student ID: 913859133

Project: Assignment 4 - Word Blast

Github: rinaykumar

Link: <https://github.com/CSC415-Fall2020/assignment-4-word-blast-rinaykumar>

Description:

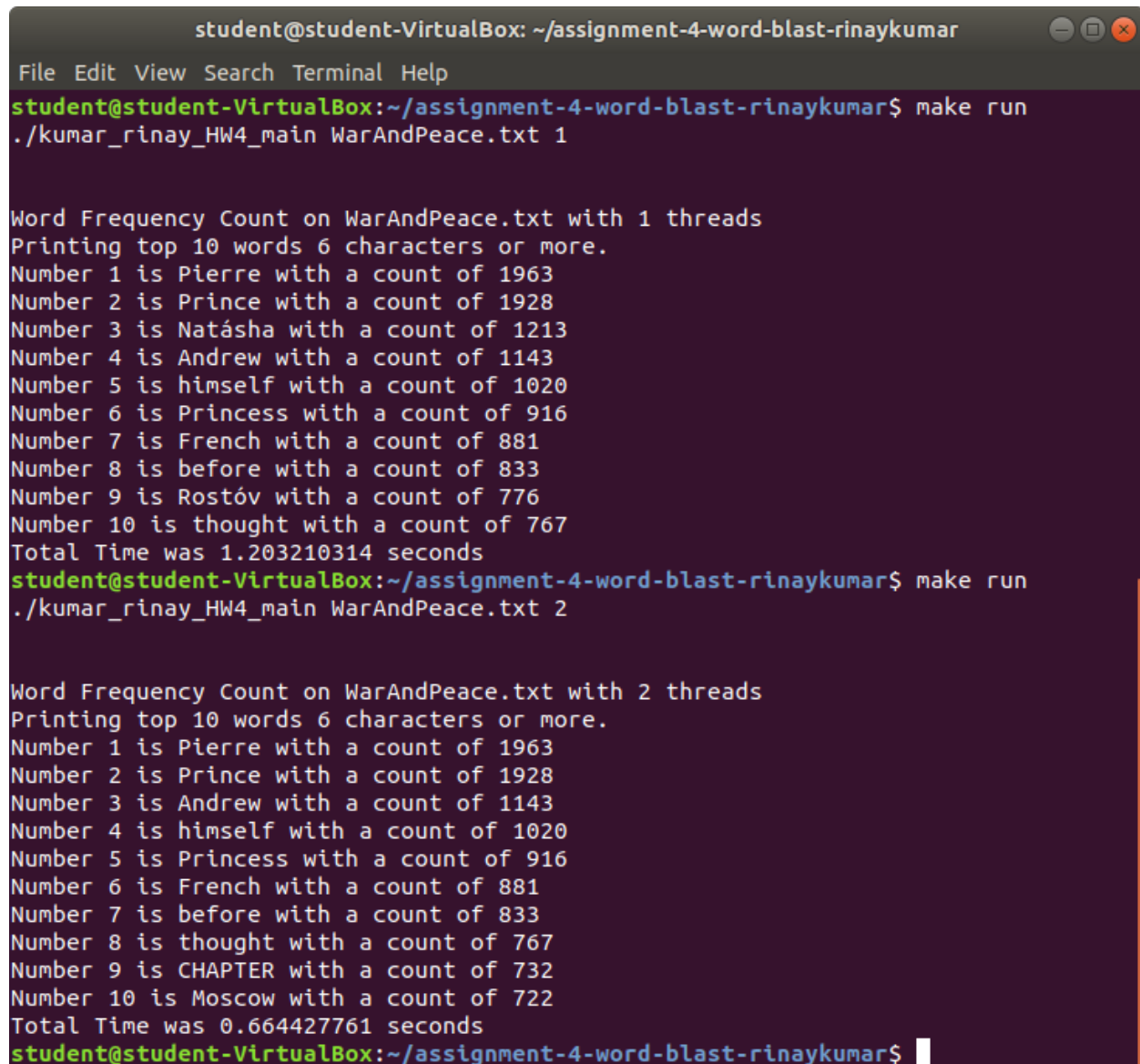
This project takes a text file, parses it, finds words with 6 or more characters, counts their occurrences, then displays the top 10 most used 6+ character words in the file, all implemented using threads. The file is opened, then threads are created, and each thread processes a chunk of the file storing results to a global shared array of structures.

Mutex locks were used to protect the critical sections. At first I placed the entire thread function in locks, which was the only way I could get the program to run correctly. This was due to my implementation using strtok. I then changed my implementation to use strtok_r, and placed a mutex lock only around the code that modified the shared array containing the words and word frequency. To my understanding, the threads run in parallel, and are only locked when accessing the shared array.

The runtimes for the program were affected by the number of threads and by the number of cores assigned to the VirtualBox. With the default of 2 cores on a dual-core machine, the runtime for 1 thread was 1.2x seconds (Fig. 1a), then for 2 threads was 0.6x seconds (Fig. 1a). The runtime halved, which is what would be expected when doubling the thread count from 1 to 2. However, the runtime for 4 threads was also 0.6x seconds (Fig. 1b), and for 8 threads again was 0.6x seconds (Fig. 1b), with slight variations in different runs where 4 thread runs were sometimes slightly faster than 8 threads by hundredths of a second.

At first this seemed incorrect, as per logic the runtime should cut in half with each doubling of the thread count. Tests were then conducted with the core count set to 4 on a multi-core machine. The runtime halved from 1 thread to 2, then halved again from 2 threads to 4, then stayed about the same for 8 threads. Here with 4 cores, the runtimes halved from 1 to 2 to 4 threads, and with 2 cores only halved from 1 to 2 threads, indicating a relationship with the number of cores and the number of threads for runtimes. As an extra aside, tests were run on an AWS EC2 instance with a multi-core setup, where the runtime for 1 thread was 1.7x seconds (Fig. 2a), for 2 threads was 0.9x seconds (Fig. 2a), for 4 threads was 0.4x seconds (Fig. 2b), and for 8 threads was 0.2x seconds (Fig. 2b). Again, the runtimes roughly halved as the number of threads were doubled, as expected with additional cores available.

Sample Output:

A terminal window titled 'student@student-VirtualBox: ~/assignment-4-word-blast-rinaykumar' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows two runs of a program. The first run uses 1 thread and takes 1.203210314 seconds. The second run uses 2 threads and takes 0.664427761 seconds. Both runs print the top 10 words by frequency from 'WarAndPeace.txt'.

```
student@student-VirtualBox: ~/assignment-4-word-blast-rinaykumar
File Edit View Search Terminal Help
student@student-VirtualBox:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 1

Word Frequency Count on WarAndPeace.txt with 1 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Natásha with a count of 1213
Number 4 is Andrew with a count of 1143
Number 5 is himself with a count of 1020
Number 6 is Princess with a count of 916
Number 7 is French with a count of 881
Number 8 is before with a count of 833
Number 9 is Rostóv with a count of 776
Number 10 is thought with a count of 767
Total Time was 1.203210314 seconds
student@student-VirtualBox:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 2

Word Frequency Count on WarAndPeace.txt with 2 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Andrew with a count of 1143
Number 4 is himself with a count of 1020
Number 5 is Princess with a count of 916
Number 6 is French with a count of 881
Number 7 is before with a count of 833
Number 8 is thought with a count of 767
Number 9 is CHAPTER with a count of 732
Number 10 is Moscow with a count of 722
Total Time was 0.664427761 seconds
student@student-VirtualBox:~/assignment-4-word-blast-rinaykumar$
```

Figure 1a: Thread count 1 and 2 on the default 2 core setup

```
student@student-VirtualBox: ~/assignment-4-word-blast-rinaykumar
File Edit View Search Terminal Help
student@student-VirtualBox:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 4

Word Frequency Count on WarAndPeace.txt with 4 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Natásha with a count of 1212
Number 4 is Andrew with a count of 1143
Number 5 is himself with a count of 1020
Number 6 is princess with a count of 916
Number 7 is French with a count of 881
Number 8 is before with a count of 833
Number 9 is Rostóv with a count of 776
Number 10 is thought with a count of 767
Total Time was 0.624500727 seconds
student@student-VirtualBox:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 8

Word Frequency Count on WarAndPeace.txt with 8 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Natásha with a count of 1213
Number 4 is Andrew with a count of 1143
Number 5 is himself with a count of 1020
Number 6 is princess with a count of 916
Number 7 is French with a count of 881
Number 8 is Before with a count of 833
Number 9 is Rostóv with a count of 776
Number 10 is thought with a count of 767
Total Time was 0.646564489 seconds
student@student-VirtualBox:~/assignment-4-word-blast-rinaykumar$
```

Figure 1b: Thread count 4 and 8 on the default 2 core setup

```
ubuntu@ip-172-31-36-137: ~/assignment-4-word-blast-rinaykumar
File Edit View Search Terminal Help
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 1

Word Frequency Count on WarAndPeace.txt with 1 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Natásha with a count of 1213
Number 4 is Andrew with a count of 1143
Number 5 is himself with a count of 1020
Number 6 is Princess with a count of 916
Number 7 is French with a count of 881
Number 8 is before with a count of 833
Number 9 is Rostóv with a count of 776
Number 10 is thought with a count of 767
Total Time was 1.750836724 seconds
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ vim Makefile
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 2

Word Frequency Count on WarAndPeace.txt with 2 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Andrew with a count of 1143
Number 4 is himself with a count of 1020
Number 5 is Princess with a count of 916
Number 6 is French with a count of 881
Number 7 is before with a count of 833
Number 8 is thought with a count of 767
Number 9 is CHAPTER with a count of 732
Number 10 is Moscow with a count of 722
Total Time was 0.918017168 seconds
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$
```

Figure 2a: Thread count 1 and 2 on a multi-core AWS EC2 instance

```
ubuntu@ip-172-31-36-137: ~/assignment-4-word-blast-rinaykumar
File Edit View Search Terminal Help
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ vim Makefile
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 4

Word Frequency Count on WarAndPeace.txt with 4 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Natásha with a count of 1212
Number 4 is Andrew with a count of 1143
Number 5 is himself with a count of 1020
Number 6 is princess with a count of 916
Number 7 is French with a count of 881
Number 8 is before with a count of 833
Number 9 is Rostóv with a count of 776
Number 10 is thought with a count of 767
Total Time was 0.454366643 seconds
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ vim Makefile
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$ make run
./kumar_rinay_HW4_main WarAndPeace.txt 8

Word Frequency Count on WarAndPeace.txt with 8 threads
Printing top 10 words 6 characters or more.
Number 1 is Pierre with a count of 1963
Number 2 is Prince with a count of 1928
Number 3 is Natásha with a count of 1212
Number 4 is Andrew with a count of 1143
Number 5 is himself with a count of 1020
Number 6 is princess with a count of 916
Number 7 is French with a count of 881
Number 8 is before with a count of 833
Number 9 is Rostóv with a count of 776
Number 10 is thought with a count of 767
Total Time was 0.263625594 seconds
ubuntu@ip-172-31-36-137:~/assignment-4-word-blast-rinaykumar$
```

Figure 2b: Thread count 4 and 8 on a multi-core AWS EC2 instance