Class: CSC-415.02

Name: Arslan Alimov

Student ID: 916612104

GitHub ID: ArslanAlimov

Project: Assignment 4 – Word Blast

File: Alimov_Arslan_HW4_main.c

Description:

The purpose of the code is to read text file and output top 10 most frequent words from the file. The software implements a multithreading with different protections.

Process of creating the code:

I wrote a code in a way where I first decided to create 2 data structures one is for the file and another one for text that appears in the file. File data structure holds information about the file , file size, e.t.c . And data structure for text we fill it up when we find some word that appears frequent and we take a count of it how many times it was used. I realized that we will need a reader function to read from file, I used stat() instead of lseek , because stat fits better for this purpose and it is also was simpler for my usage. It is also much faster than lseek. I also started to realize that we need an add function to add words to our list from which we will count frequency. I created a function to print top 10 , it uses a bubble sort method which pushes the highest value to the top while the lowest values stay on bottom, it was probably not the best solution for this assignment , but it does work. I also came to realization that my threads are working so fast that they are accessing the memory that is unable to be accessed yet due to problem with allocation. This problem I was unable to fix. I was struggling for 4 days straight trying to figure out the problem and was unable to locate it, I was trying not to ask any friends or anyone for that matter. It was the hardest assignment ,so far due to the fact that we need to properly allocate memory, make sure each thread works on their own memory segment e.t.c … I was not able to make this code "bug free".

Analysis:

        When running my code, I realized that the more threads I use the faster my tool completes the task and the faster it finds the words. When I use 1 thread the code can give me 1.0+ time of finding words. I feel like it is based on the way we implement our algorithm of sorting as well as our loop operations. The time varies , because of the fact that the code needs to look for all the words inside our array and compare them as well as give us the top 10 most frequent, I also found out that while increasing threads the total time is being increased , I would guess it is caused because of our thread lock (mutex). I also found out that sometimes one thread performs better than multithreading , though the time is getting lower and lower with a single thread. It is due to the fact that our operation does not really need multithreading only for a calculation of words and organizing them from bottom to top.

Compilation:


```
student@student-VirtualBox: ~/Desktop/Homework/assignment-4-wordblast-ArslanAlimov

File Edit View Search Terminal Help
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$ make
gcc -c -o Alimov_Arslan_HW4_main.o Alimov_Arslan_HW4_main.c -g -I.
gcc -o Alimov_Arslan_HW4_main Alimov_Arslan_HW4_main.o -g -I. -l pthread
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$
```

Running with one thread:

```
student@student-VirtualBox: ~/Desktop/Homework/assignment-4-wordblast-ArslanAlimov
File  Edit  View  Search  Terminal  Help
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$ make run
./Alimov_Arslan_HW4_main WarAndPeace.txt 1

 Word Frequency Count on WarAndPeace.txt with 1 threads
 Printing top 10 words 5 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 thought with a count of 767
Number 10 is CHAPTER with a count of 732
Total Time was 1.922705807 seconds
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$
```
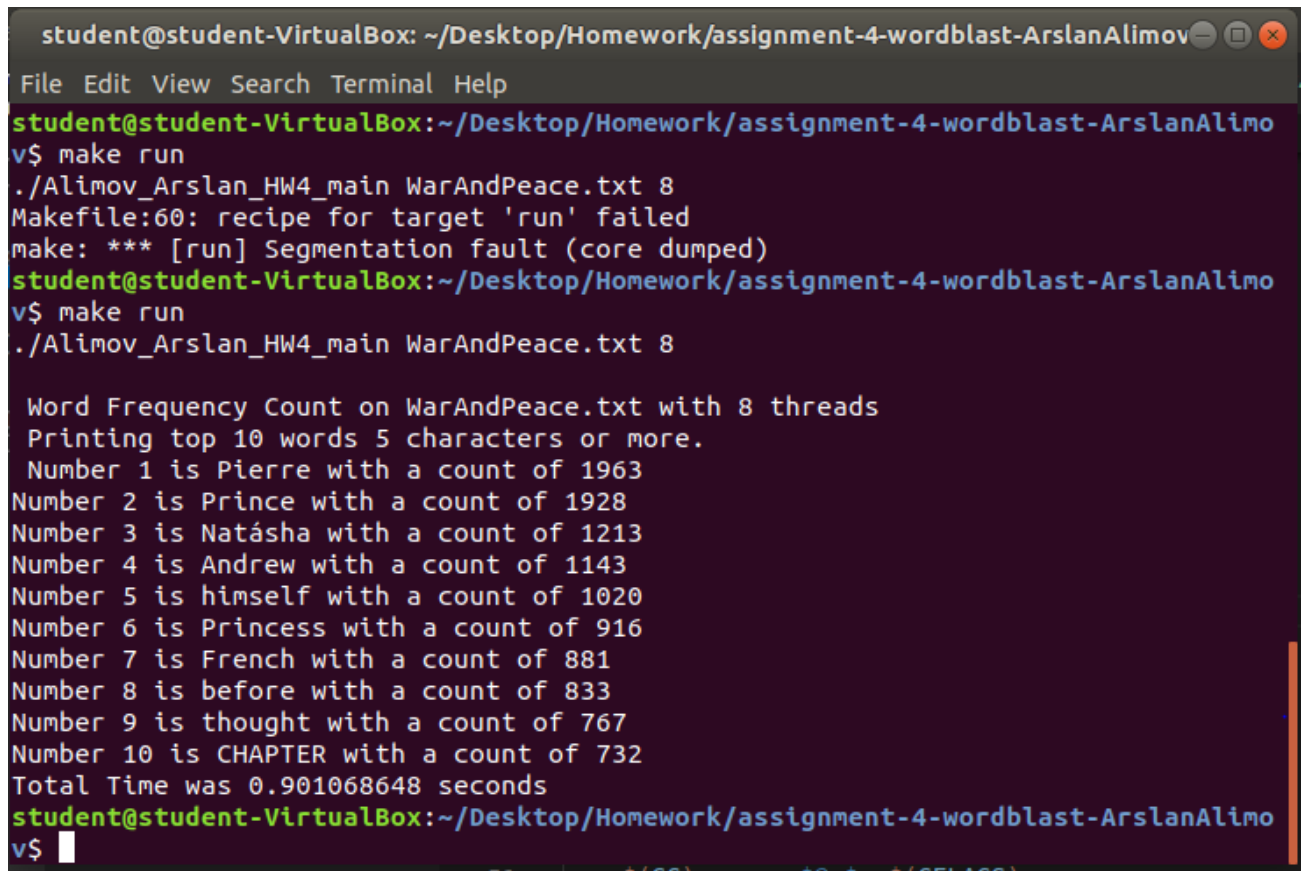
Running with 4 threads:

```
student@student-VirtualBox: ~/Desktop/Homework/assignment-4-wordblast-ArslanAlimov
File  Edit  View  Search  Terminal  Help
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$ make run
./Alimov_Arslan_HW4_main WarAndPeace.txt 4

 Word Frequency Count on WarAndPeace.txt with 4 threads
 Printing top 10 words 5 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 French with a count of 881
Number 7 is before with a count of 833
Number 8 is Rostóv with a count of 776
Number 9 is thought with a count of 767
Number 10 is CHAPTER with a count of 732
Total Time was 0.910822323 seconds
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$
```

Running with 8 threads:


```
student@student-VirtualBox: ~/Desktop/Homework/assignment-4-wordblast-ArslanAlimov
 File  Edit  View  Search  Terminal  Help
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$ make run
./Alimov_Arslan_HW4_main WarAndPeace.txt 8
Makefile:60: recipe for target 'run' failed
make: *** [run] Segmentation fault (core dumped)
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$ make run
./Alimov_Arslan_HW4_main WarAndPeace.txt 8

 Word Frequency Count on WarAndPeace.txt with 8 threads
 Printing top 10 words 5 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 thought with a count of 767
Number 10 is CHAPTER with a count of 732
Total Time was 0.901068648 seconds
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordblast-ArslanAlimo
v$
```

Errors:

I have a problem with my add text where it would access memory faster than it is supposed to causing that error. I was unable to fix it and was not sure how to do it. I decided not to hide this , though the fix to that could be placing another mutex lock in my for loop. But it causes my code to run slower. The software code works perfectly when you run in 1 thread but more than 1 thread it crashes and there was nothing I could really do. (if you can help me fix that error or describe what is happening or how to properly do that I would appreciate it)

Thank you

File Edit View Search Terminal Help

student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordbla
st-ArslanAlimov$ make run
./Alimov_Arslan_HW4_main WarAndPeace.txt 8
Makefile:60: recipe for target 'run' failed
make: *** [run] Segmentation fault (core dumped)
student@student-VirtualBox:~/Desktop/Homework/assignment-4-wordbla
st-ArslanAlimov$