OSP Assignment 1

Github repository: <https://github.com/ZainBangash/OSPAssignment1.git>

Task1:

For task 1, firstly I created a bash file to clean the data given in the wlist\_all folder. Initially I started with the 12th file (wlist\_match12.txt) and worked my way up to the 1st file (wlist\_match1.txt) and experimented with the different ways to sort the data. First, I wanted to remove certain characters from the words that had characters other than letter in them. For example, the word ………apple would be filtered to apple. I decided to remove all the characters that were alphabets with the exception of the apostrophe ( ‘ ). That lead to words that wouldn’t be classified as proper words. After trying different methods to make this work, eventually I came to the realization that the best way to filter it would be to remove all words that had any illegal characters, rather than just removing the illegal characters. This did not have a very major impact on the number of words left as there were still more than a million words left after this filter. A simple grep -E "^[a-zA-Z]\*$" command was used to remove any words with characters other than the characters of the alphabet. Secondly, all words whose length was either less than 3 or greater than 15 were removed using the command grep -E '^.{3,15}$'. Unfortunately, an issue arose of words that had fulfilled the criteria of the previous filters but still wouldn’t be classified as proper words. An example could be ‘aaaaand’, it can be deduced that the word intended is ‘and’ however there are too many a’s at the beginning. There were more than few occurrences of such words, and there all words with more than one occurrence of the same alphabet consecutively were removed using the command  grep -v '\(.\)\1' . Words like ‘aaand’, ‘boooth’ and ‘dooog’ were removed however as a result words such as bareroot were also removed. There were still more than a million words left to sort and do the next tasks with so this filtering approach was kept. Duplicates were also checked for by using the following command uniq -cd command. The words were also sorted on their third letter, using the following command -k1.3 clean.txt. All in all the following coreutils were used shuf, grep, sort, uniq, wc and time (to check how long each program took).

For the Task1Filter.cpp file, both the clean and dirty file paths were given in as arguments, both files are opened and each line from the dirtyfile is checked, if it passes the filtering criteria then it is added to the clean file. Methods were made to check if the word only contained alphabets and if the word contained more than one character consecutively.

Limitations:

After all the filtering there were several words that can be said to be of the English dictionary, however as this was the best and more efficient way to do the filtering, this method was kept. It also left a large set of data to be sorted.

Outcomes:

When we run the bash file, we get the following:

Text

Description automatically generated

When we compile and run Task1 we get the following:

Text

Description automatically generated

The program opened both the wlist\_match1.txt and finalList.txt files and wrote the filtered data from the dirty file into the cleaned file, and closed the file in 1.948s. This is an improvement from the 7.440s taken by the bash file. The bash file also sorted the words, however as we will see the c++ programs achieve the same in a much faster time.

Task2:

For Task2 I had to use forks to sort the data into 13 files, a file per word length and then perform a 13 to 1 merge sort on the files. 13 forks were made and assigned a word length it had to sort. It would take the words of the word length allotted to it from the globalArray and sort it on the 3rd letter and then write the sort list to a .txt file. The reduce method would then open all 13 files, get one word from each file, sort the list, remove the first word from the sorted list, write it to the clean file, get the next word from the file that contains the same word length as the current removed word. If the file to get words from is empty then it will remove the word from the sorted list and sort the next words and the finished file will never again be opened as there is no word of the file’s word length present to trigger a getline() from that file.

Limitations:

A limitation or error or failure of logic initially was that a file was being opened and closed every time a new word was needed to be inputted to the list, to be sorted. As there are around a million words, the whole program was taking around 15 minutes, however once the files were left open outside of the loop the program was significantly faster. Forks were only used with Map3 and not reduce as the instructions provided indicated forks were to be used for sorting the words of a certain length and outputting them to a file for its word length.

Outcomes:

As a result of sorting from the 3rd letter onwards, words that contain ‘a’ as their 1st letter rather than their 3rd letter will not be first on the list, rather words that contain ‘a’ as their 3rd letter will come before words that contain ‘a’ as the 1st but not 3rd letter

When Task2 was compiled and ran, it produced the following results:

