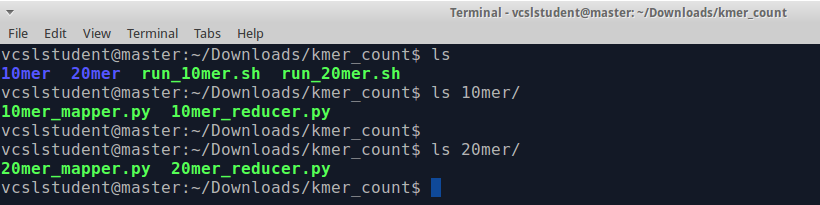
**READ ME**

A K-mer is a substring of length K (K > 0), and counting the occurrences of all such substrings is a central step in many analyses of DNA sequence data. Counting K-mers for a DNA sequence means finding frequencies of K-mers for the entire sequence. Counting the K-mers in a DNA sequence is a very important step in many bioinformatics applications. The following are examples of applications of K-mer counting: i) determining whether a misalignment between reads is a sequencing error or a genuine difference in sequence; ii) detecting repeated sequences, such as transposons, which are an important factor for biological role applications; iii) correcting short-read assembly errors; iv) computing metrics such as relatedness and unique enough (useful in metagenomic applications)

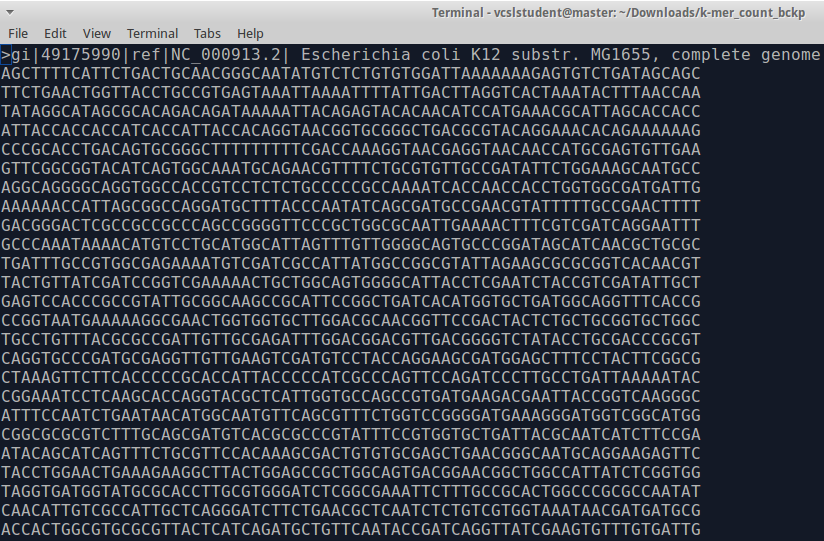
Although simple in principle, K-mer counting is a big data challenge, since a single DNA sample can contain several billion DNA sequences.

**A. FOR EXTRACTING TOP 10-MERS**

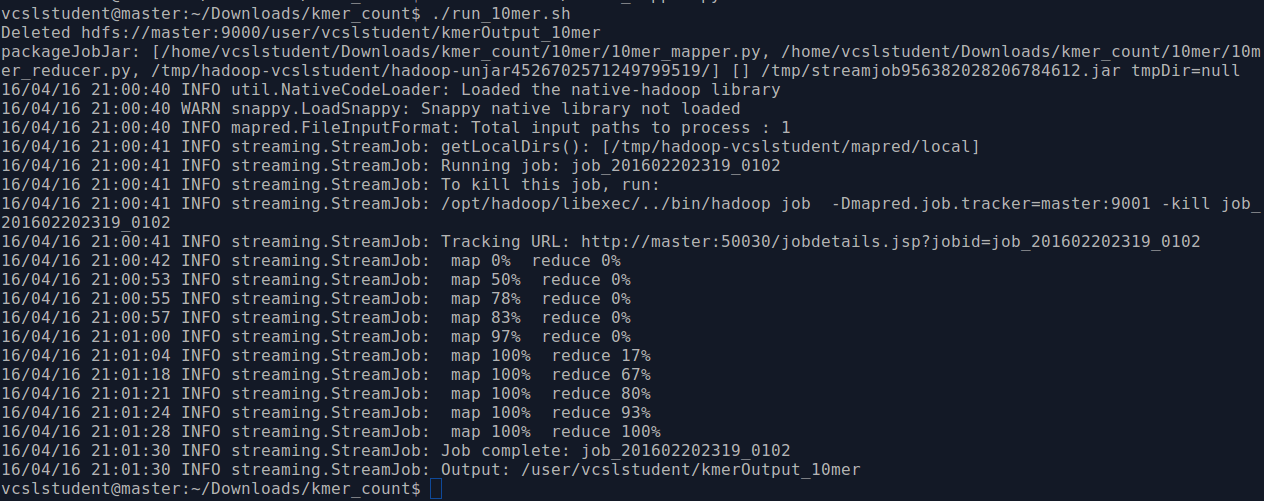
1. There are two input folders one for 10-mers and other for 20-mers as shown below.



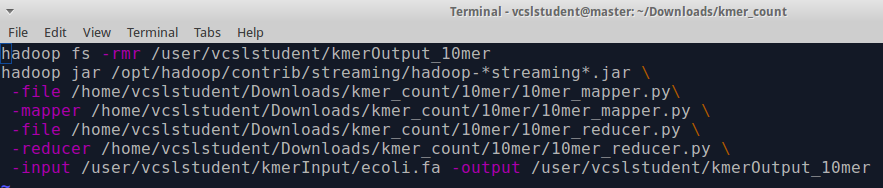
1. The input file is as shown below



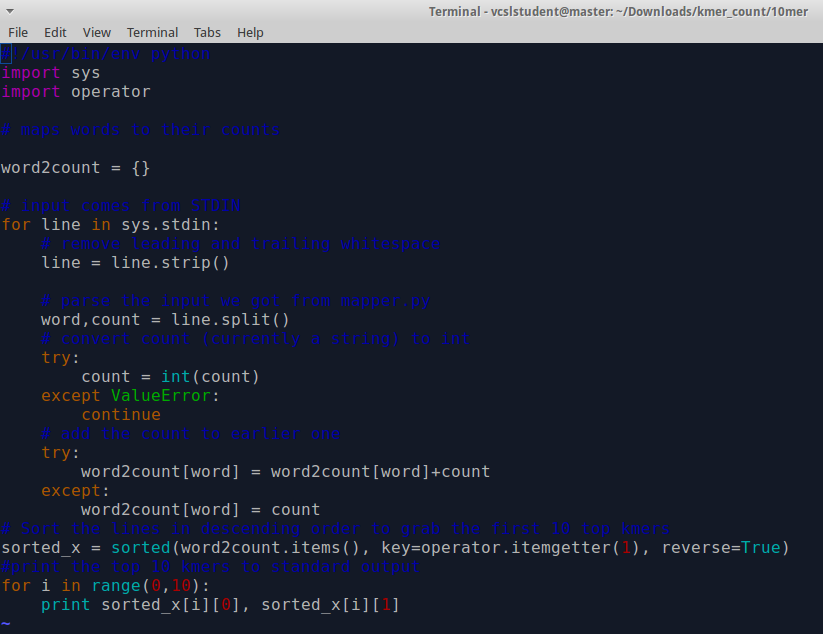
1. The code can be run as follows



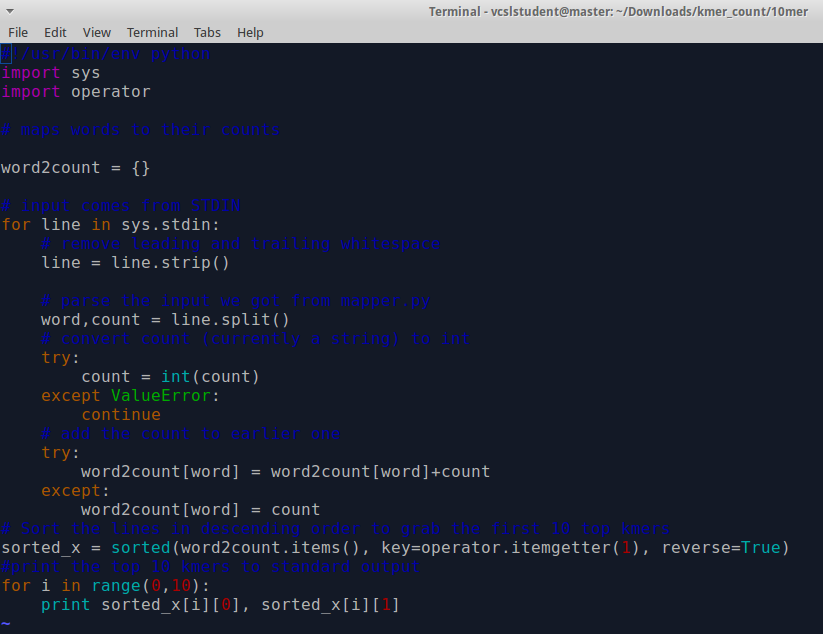
1. The sh file is run to trigger the Hadoop file system to run the mapper and reducer. The first line is to delete the output folder if it exists already. The contents of sh file are as follows-



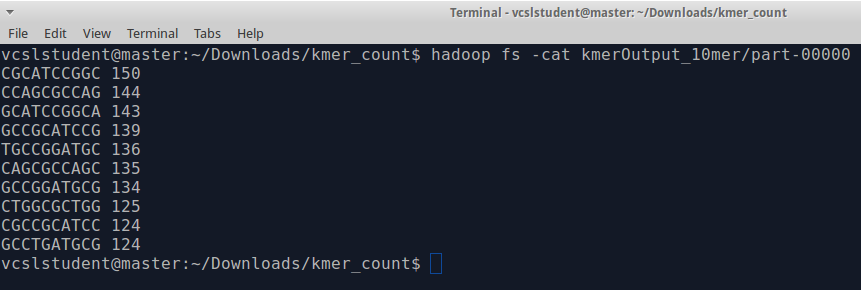
1. The input mapper file is as shown below



1. The output of the mapper file is given to the reducer as input. Below is the code for the reducer.

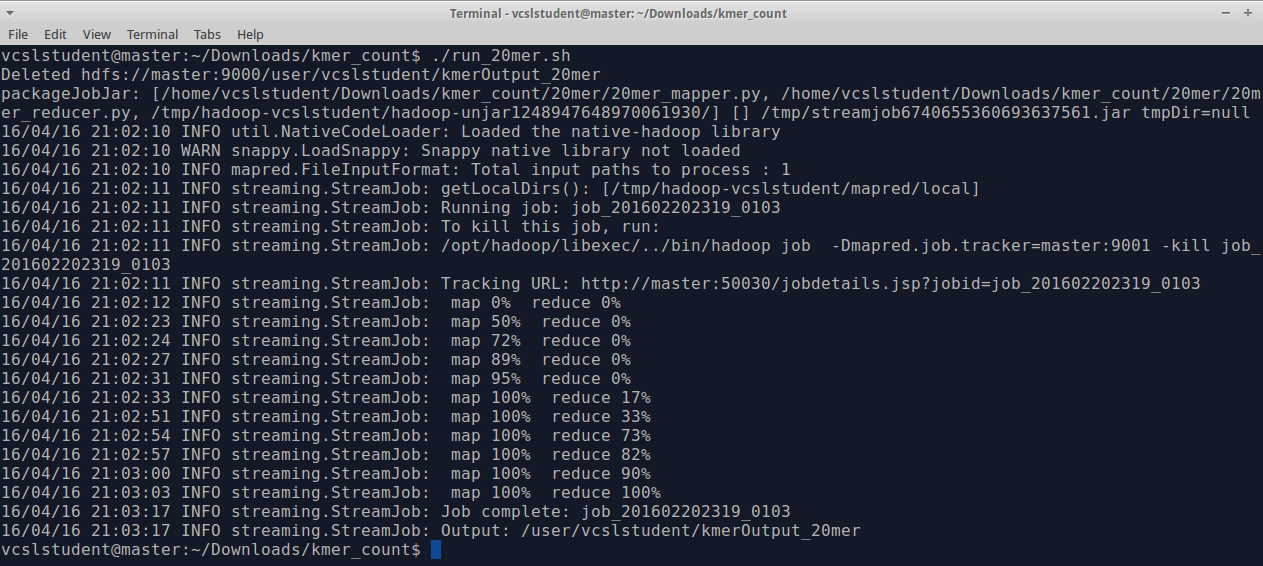


1. The results obtained are as follows

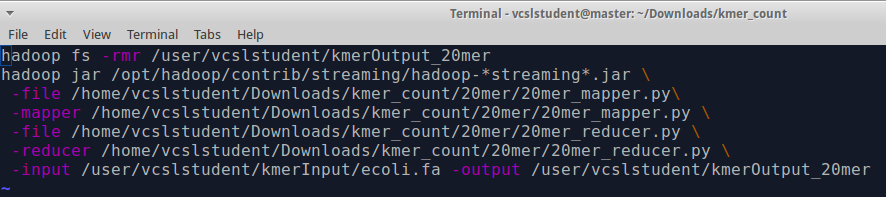


**B. FOR EXTRACTING TOP 10 20-MERS**

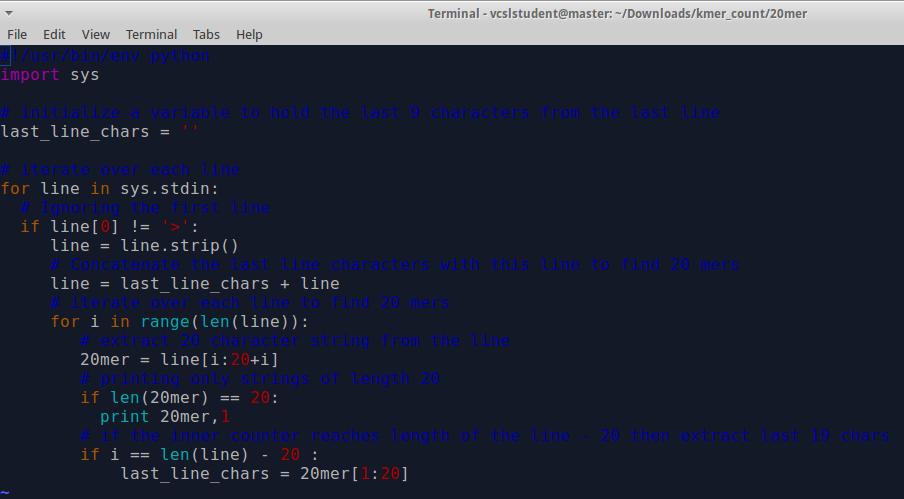
1. The code can be run as follows



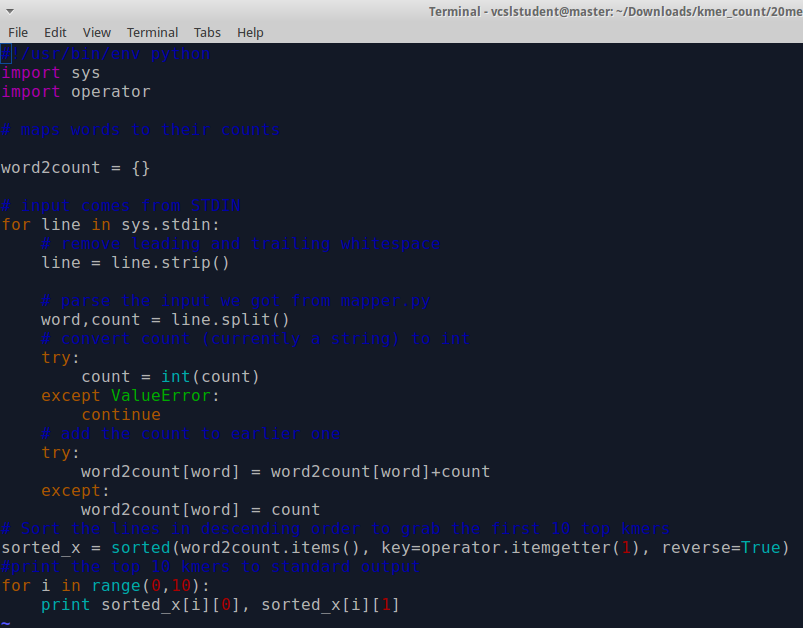
1. The sh file is run to trigger the Hadoop file system to run the mapper and reducer. The first line is to delete the output folder if it exists already. The contents of sh file are as follows-



1. The input mapper file is as shown below



1. The output of the mapper file is given to the reducer as input. Below is the code for the reducer.



1. The results obtained are as follows

