Lab 2.

In this lab, you will write by your own a complete Hadoop application. Start by importing the template project available in Lab_Skeleton.zip. Once you have imported the template, modify the content of the classes to implement the application described in the following. The template contains the skeleton of a standard MapReduce application based on three classes: Driver, Mapper, and Reducer. Analyze the problem specification and decide if you really need all classes to solve the assigned problem.

From now on, keep in mind always (even if we do not explicitly ask for it) the complexity of your program. Specifically try to understand, before even submitting a job, what will be the effort you will require to the cluster in terms of time, network and I/0

- How many pairs and bytes will be read from HDFS?
- How many pairs and bytes will be emitted by the mappers and hence how many data will be sent on the network?

You can then check on HUE if you guessed correctly (more or less).

1. Filter an input dataset

If you completed Lab 1, you should now have (at least one) large files with the word frequencies in the amazon food reviews, in the format `word\tnumber`, where number is an int (a copy of the output of Lab 1 is available in the HDFS shared folder /data/students/bigdata-01QYD/Lab2/). You should also have realized that inspecting these results manually is not feasible. Your task is to write a Hadoop application to filter the content of the output of Lab 1 and analyze the filtered data.
The filter you should implement is the following:

- keep only the words that start with “ho”.

How large is the result of this filter? Do you need to filter more?

Modify the application in order to accept the beginning string as a command-line parameter.
Execute the new version of the program to select the words starting with the prefix that you prefer.

Bonus task

If you completed the bonus task of lab 1, try your filter on the 2-grams you have generated.
If you did not complete the bonus task of lab 1, you can use the files available in the HDFS shared folder /data/students/bigdata-01QYD/Lab2BonusTrack/

What is the size of this new input dataset, compared to the simple word counts (1-grams) we used in the previous step? Did you really need the cluster to filter 1-grams? What about 2-grams?

Implement a new application that selects all the 2-grams that contain, at any position, the word “like” (i.e., “like” can be either the first or the second word of the selected 2-grams). What do you think will be, most likely, the other word?