Lab 5

The objective of this laboratory is to start playing around with Apache Spark.

1. Problem specification

If you completed Lab 1, you should now have (at least one) files with the word frequencies in the Amazon food reviews, in the format `word\tfreq`, where *freq* is an integer.
A copy of the output of Lab 1 is available in the cluster shared folder:

/share/students/bigdata/Dati/Lab2/OutputFolderLab1

Your task is to write a **Spark** application to filter these results, analyze the filtered data and compute some statistics on them.

Task 1

The first filter you should implement is the following:

- Keep only the lines containing words that start with the prefix “ho”

The returned RDD contains the set of lines (`word\tfreq`) that satisfy the filtering operation.

Print on the standard output the following statistics, based on the content of the RDD returned by the filtering operation:

- The number of selected lines
- The maximum frequency (*maxfreq*) among the ones of the selected lines (i.e., the maximum value of *freq* in the lines obtained by applying the filter).

Task 2

Extend the previous application. Specifically, in the second part of your application, among the lines selected by the first filter, you have to apply another filter to select only the most frequent words. Specifically, your application must select those lines that contain words with a frequency (*freq*) greater than 80% of the maximum frequency (*maxfreq*) computed before.

Hence, implement the following filter:

- Keep only the lines with a frequency *freq* greater than 0.8**maxfreq*.

Finally, perform the following operations on the selected lines (the ones selected by applying both filters):

- Count the number of selected lines and print this number on the standard output
- Save the selected words (without frequency) in an output folder (one word per line)

2. Testing the application

Run your application

1. Create a Jupyter notebook (select PySpark (Local)) and run you application on the input shared folder (*/share/students/bigdata/Dati/Lab2/OutputFolderLab1/*). Set the name of the output folder in your code.
2. Analyze the returned results (the statistics/results printed on the standard output and the content of the output folder)

How to run your application

- Pyspark (Local) notebook - To run your application on the gateway
  - Open a browser and connect to jupyter.polito.it
  - Log in and open a “Pyspark (local)” notebook
  - Write your application in the notebook and run it on the gateway (data are read from and stored on the cluster file system but driver and executors are instantiated on the gateway)

- PySpark (Kubernetes) notebook - To run your application on the nodes of the cluster
  - Open a browser and connect to jupyter.polito.it
  - Log in and open a “Pyspark (Kubernetes)” notebook
  - Write your application in the notebook and run it on the nodes of the cluster (data are read from and stored on cluster file system and driver and executors are instantiated on the nodes/servers of the cluster BigData@Polito)

⚠⚠⚠ Shut down JupyterHub container ⚠⚠⚠

As soon as you complete all the tasks and activities on JupyterHub environment, please remember to shut down the container to let all your colleagues in all the sessions connect on JupyterHub and do all the lab activities.

1. Go into File -> Hub Control Panel menu
2. A new browser tab opens with the “Stop My Server” button. Click on it and wait till it disappears.