Project 2: Clustering Algorithms

Demo time & hard copy report due: October 30 2018
Code submission due: 10:30am October 30 2018
Please bring your hard copy report and UB Card to demo.

Please clearly state the UB Person numbers and UB IT names for all the group members on the cover of the report.

Two gene datasets (*cho* and *iyer*) can be found on Piazza. Please check the README file first for a short description of the two datasets.

Complete the following tasks:

1. Implement three clustering algorithms to find clusters of genes that exhibit similar expression profiles: K-means, Hierarchical Agglomerative clustering with Single Link (Min), and one from (density-based, mixture model, spectral). Compare these three methods and discuss their pros and cons.

For each of the above tasks, you are required to validate your clustering results using the following methods:

- Choose an external index (Rand Index or Jaccard Coefficient) and compare the clustering results from different clustering algorithms. The ground truth clusters are provided in the datasets.
- Visualize data sets and clustering results by Principal Component Analysis (PCA). You can use the PCA you implemented in Project 1 or use any existing implementation or package.

2. Set up a single-node Hadoop cluster on your own machine and implement MapReduce K-means. Compare with non-parallel K-means on the given datasets. Try to improve the running time. To set up single-node Hadoop, we provide an instruction file *Hadoop_setup.pdf* that can be found on Piazza.

Your final submission should include the following:

- Codes: A folder named *Code*, that contains three clustering algorithms, and MapReduce K-means algorithm and a *README* that shows how to run your code.
- Report: A pdf file named *Cluster_report .pdf*. Describe your implementation details about all the algorithms. Compare the performance of these approaches using visualization and external index on the two given data sets. State the pros and cons of each algorithm and any findings you get from the experiments.

Project Submission:

1. Your final submission should be a zip file named as *project2.zip*. In the zip file, you need to include aforementioned folder *Code* and folder *Report*.
2. Log in any CSE department server and submit your zip file as follows:
   >> **submit_cse601 project2.zip**

The details about Demo will be released **no later than Oct.28 through Piazza**. Please note:

- Two new data sets will be given to check your implemented clustering algorithms and validation measures. The data format is the same with the project data sets we provide.
- During the demo, the specific parameter setting will be changed and therefore you need to keep your code flexible to different parameters.

Note that copying code/results/report from another group or source is not allowed and may result in an F in the grades of all the team members. Academic integrity policy can be found at https://engineering.buffalo.edu/computer-science-engineering/graduate/resources-for-current-students/academic-integrity.html