**Coursework Description**

For this coursework, you will solve the given problems using MapReduce computational model and Mahout on Hadoop cluster. This coursework carries 30% weightage of total mark for the module.

**Q1) Find the descriptive statistics for temperature of each day of a given month for the year 2007**.

We use weather data from [NCDC](http://www.ncdc.noaa.gov/). You can access hourly weather data from ‘Data Sets’ folder under tab on the course’s VLE. We have chosen the hourly records of April, May, June and July from year 2007. A month is represented per file. You may select any one of the four months (files) for analysis.

You find the weather data from different weather stations (wban - first column). Using the hourly data across all weather stations, find

* The difference between the maximum and the minimum “Wind Speed” from all the weather stations for each day in the month
* The daily minimum “Relative Humidity” from all the weather stations
* the daily mean and variance of “Dew Point Temp” from all the weather stations
* the correlation matrix that describes the monthly correlation among “Relative Humidity”, “Wind Speed” and “Dry Bulb Temp” from all the weather stations.

You are NOT going to use any package that gives the statistics. You MUST use MapReduce framework. Write the pseudo code for mapper and reducer functions for the above four tasks and implement them in Python. Note that while using mapper and reducer it is helpful to consider the following formulae for variance and correlation:

, where is the mean, is the observation, and N represents the number of observations.

Pearson Correlation = .

**Q2) Cluster Analysis using Apache Mahout**.

For this question, you can use the data (a set of text files that are placed in a folder) provided with Topic 4 for the k-means algorithm. You are welcome to use your own dataset for this question. If you choose to do so, please provide a link to the data in your report.

As we discussed in text clustering (Topic 4), the terms of the documents are considered as *features* in text clustering. The vector space model is an algebraic model that maps the terms in a document into n-dimensional linear space. However, we need to represent textual information (terms) as a numerical representation and create feature vectors using the numerical values to evaluate the similarity between data points.

Use **Apache Mahout** and perform the standard steps for the cluster analysis, 1) create *sequence files* from the raw text, 2) create a sparse (efficient) representation of the vectors, initialisation approximate centroids for K-Means, 3) run the K-Means algorithm, 4) get the final iteration’s clustering solution, and 5) evaluate the final solution

You need to consider the following points in the analysis:

* Implement the K-Means clustering algorithm with Euclidean and Manhattan Distance Measures.
* Find the optimum number (K) of clusters for the K-mean clustering for the above distance measures.
* Implement K-mean clustering algorithm with Cosine Distance Measure and verify the relation between the average distance to the centroid and the K value.
* Plot the elbow graph for K-mean clustering with Cosine Measure. Try to smooth graph so that you can explain the value for K as the best.
* Compare the different clusters you obtained with different distance measures and discuss what is the best setting for K-means clustering for this dataset.

**You need to include the following in your coursework submission:**

1. For Q1, submit the pseudo code and Python code for the mappers and reducers implementation for all the descriptive statistics along with some comments so that a layperson can follow. Anyone should be able to run your code and reproduce your results with the instructions that you have provided.
2. For Q2, write a brief summary on the impact of parameter changes on the performance of k-means algorithm. For example, you may do 1) compare different distance measures in K-Means algorithm discuss the merits and demerits and 2) present a table that shows the performance of K-means algorithm for different K values.
3. Submit a report on the experiments. This report will be a detailed explanation (Max 1500 words, excluding code and references) of what you explored, the results you obtained, and some discussion points on the limitations of MapReduce methodology and Hadoop’s MapReduce computing engine.

Credit will be given for:

* the depth and breadth of your investigation.
* the technical skills you demonstrate in your write-up.
* good use of the Hadoop cluster.
* critical evaluation of your work.

**The grades in this coursework are allocated approximately as follows:**

|  |  |
| --- | --- |
| Activity | Mark |
| * Mappers/reducers and the results for Q2 * K-means clustering with Euclidean and Manhattan Distance Measures cosine * Studying the impact of K on the precision * Considering Cosine Distance Measure distance measure for comparison * Elbow graph * Discussion on MapReduce methodology and K-means algorithms * Overall quality of the report | 12  2  3  2  3  4  4 |