Big Data Analysis Coursework 1

NCDC Weather Data Analysis

Sandor Kanda

2024

Table of Contents

Introduction *3*
Methods *3*
Weather Data 3
Cluster analysis 4
Results *5*
Weather Data Analysis 5
Western Classics Text Analysis 8
Conclusion *10*
References *11*

List of Figures

Figure 1. Pearson Correlation Matrix Visualisation 7
Figure 2. Elbow Graph for Inter-Cluster Density 9

List of Tables

Table 1. Descriptive Statistical Outputs for the Weather Data 5
Table 2. Correlation Matrix using Pearson Correlation 6

List of Equations

Equation 1. Variance Equation 5
Equation 2. Pearson Correlation Equation 6

Introduction

Weather forecasting plays a significant role in different aspects of life, such as operating hydropower plants, renewable energy, flood management, and agriculture. Recently, machine learning techniques have been used for long periods for weather forecasting, as they are more accurate than models based on physical principles. [1]
The selected file consists of the July 2007 weather data, which will be used to process the descriptive statistics on several weather metrics.

Methods

Weather Data

The data file was downloaded from the https://learn.gold.ac.uk domain and uploaded to the Hadoop Apache framework. The mapper reducer and support files were written using the Python programming language in VS Code. Based on the problem requirements and the nature of the computations involved, using a single mapper and a single reducer for this task is a justifiable design choice. In this problem, the input data is partitioned by date, as each line in the dataset represents measurements for a specific date. Therefore, multiple mappers are not required to perform additional data partitioning or filtering. The calculations needed for each date (difference between maximum and minimum wind speed, minimum relative humidity, mean and variance of dew point temperature, and correlation matrix) were performed independently for each date. Combining or aggregating data across dates within the MapReduce job was unnecessary.
Using a single mapper and reducer, the amount of data that needed to be transferred between the map and reduce phases was minimal. The mapper emits key-value pairs where the key is the date, and the value contains the required measurements. The reducer receives these key-value pairs grouped by date, allowing it to process the measurements for each date independently. A single mapper and reducer simplify the overall design and implementation of the MapReduce job.

It avoids additional coordination or communication between multiple mappers or reducers, which would be necessary if we split the computations across various stages.

Even with a single mapper and reducer, the MapReduce framework can still parallelise the computations by assigning different subsets of the input data to different mapper and reducer tasks. This parallelisation allows the job to scale horizontally across multiple nodes in a cluster, ensuring efficient processing of large datasets.

The design adheres to the principles of the MapReduce paradigm, where the mapper extracts and emits key-value pairs, and the reducer aggregates and processes the data based on the keys (dates in this case).

While it is possible to split the computations into multiple MapReduce stages or use numerous mappers and reducers, it would introduce unnecessary complexity and overhead for this specific problem. The one-mapper-one-reducer approach is a simple and efficient solution that meets the requirements while leveraging the MapReduce framework's parallelisation and scalability benefits.

Cluster analysis

The cluster analysis task used Western Classics text data. The data was downloaded and uncompressed before being uploaded to the Hadoop environment. Using Mahout, an open-source project, the raw text was used to create sequence files. Next, the sparse representation of the vectors created term frequency (TF) and term frequency / inverse document frequency (TF*IDF). The text analysis was performed using the K-Means algorithm. In clustering sparse text data, feature weights are used to discover clusters from subspaces of the document vector space and identify keywords that represent the semantics of the clusters. Canopies were generated as an initialisation to help approximate the centroids for the K-Means algorithm. The low inter-cluster density and high intra-cluster density indicate a good solution. The final iteration of the cluster solution was evaluated.

Results

Weather Data Analysis

In this task, the difference between the minimum and maximum Wind Speed Range, the daily minimum Relative Humidity, and the mean and variance of the Dew Point Temp variables were calculated.

The variance was calculated using the formula below:

Equation 1. Variance Equation

$$variance = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)$$

Where $\bar{x}$ is the mean, xi _is the i^{th} observation, and N represents the number of observations.

The descriptive statistic output generated by the MapReduce framework is listed below.

Table 1. Descriptive Statistical Outputs for the Weather Data

YearMonthDay	Wind Speed Range	Min Humidity	Dew Point Mean	Dew Point Variance
20070701	40	0	43.257286525747276	624.9856241110248
20070702	80	0	43.30948373262007	637.0499690870984
20070703	44	0	45.331646617751915	664.8773141527266
20070704	37	0	44.73220316680556	756.1105516220839
20070705	41	0	44.044602236489645	791.8741108147059
20070706	64	0	45.80418118466899	705.6320832256285
20070707	51	0	48.46495029914731	644.2359105394834
20070708	38	0	49.487639543978574	680.9509920131239
20070709	42	0	47.71681375774213	762.4912600521964
20070710	82	0	46.387206058190515	792.5132174761362
20070711	47	0	44.86981633104669	735.0527357708429
20070712	39	0	46.402308424559735	573.0062345569646
20070713	43	0	45.88569690731296	591.7969755086526
20070714	38	0	46.12726309485419	631.5229619800607
20070715	43	0	46.460367526055954	664.4679849444353
20070716	38	0	48.18726186556282	636.1003109512698
20070717	38	0	44.317908874220755	832.8663661073616
20070718	39	0	43.36320811401927	936.4618741792118
20070719	12	0	47.96666666666667	1286.148888888888

Next, the correlation matrix was created, which describes the monthly correlation between the Relative Humidity, Wind Speed and Dry Bulb Temp.

The Pearson Correlation was calculated with the formula below:

Equation 2. Pearson Correlation Equation

$$\text{Pearson Correlation} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left[\left(\sum x^2 - \frac{(\sum x)^2}{N}\right)\left(\sum y^2 - \frac{(\sum y)^2}{N}\right)\right]}}$$

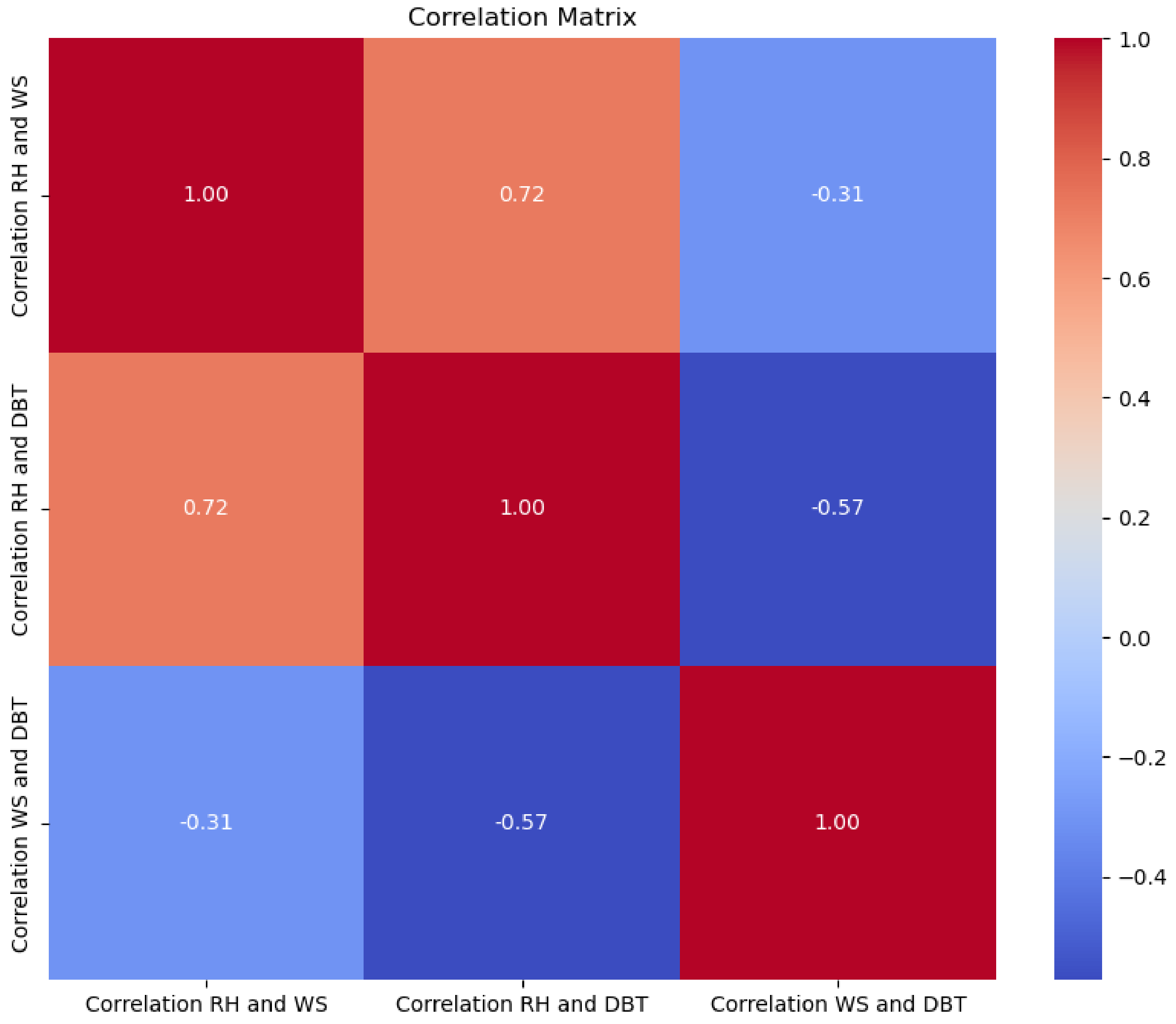
The correlation matrix output is summarised in Table 2 below.

Table 2. Correlation Matrix using Pearson Correlation

Correlation RH and WS	**Correlation RH and DBT**	**Correlation WS and DBT**
-0.166386323	0.6126492093831369	0.1024520657364338
-0.168006034	0.6217832242877797	0.082021245
-0.170774145	0.6316600826723565	0.029201834
-0.130372868	0.6810410763988378	0.068620177
-0.140032209	0.6917319327461712	0.061365681
-0.197917167	0.6314202307509842	0.09143895
-0.208671147	0.5849786642778098	0.14118903373536498
-0.163721752	0.6316146871759112	0.1682678336037442
-0.168388479	0.6812511648467756	0.039015676
-0.111228864	0.7210120698188767	0.0635959
-0.129571702	0.722947163	0.075695757
-0.19881657	0.6050037834367099	0.070545156
-0.206620103	0.612778789	0.092035494
-0.153491086	0.629277126	0.12906198421633197
-0.185115903	0.6343925648147245	0.065847085
-0.210607436	0.6171669660083516	0.045522560891902146
-0.192886429	0.736544691	0.048879589
-0.158870133	0.7840953762394797	0.078727897
-0.090565128	0.9295436692179059	-0.002676403

The correlation matrix visualisation is visualised below.

Figure 1. Pearson Correlation Matrix Visualisation



The strongest positive correlation (0.72) appears between the Relative Humidity and wind Speed and the Relative Humidity and Dew Point Temperature variables.

Western Classics Text Analysis

The Euclidean, Manhattan and Cosine Distance measures implemented the K-means algorithm. The density measures are summarised below:

K-means algorithm

Inter-Cluster Density: 0.3671243679212697
Intra-Cluster Density: 0.5643600814569034

Manhattan Distance

Inter-Cluster Density: 0.5052167024866535

Intra-Cluster Density: 0.5640387657590269

Euclidean Distance

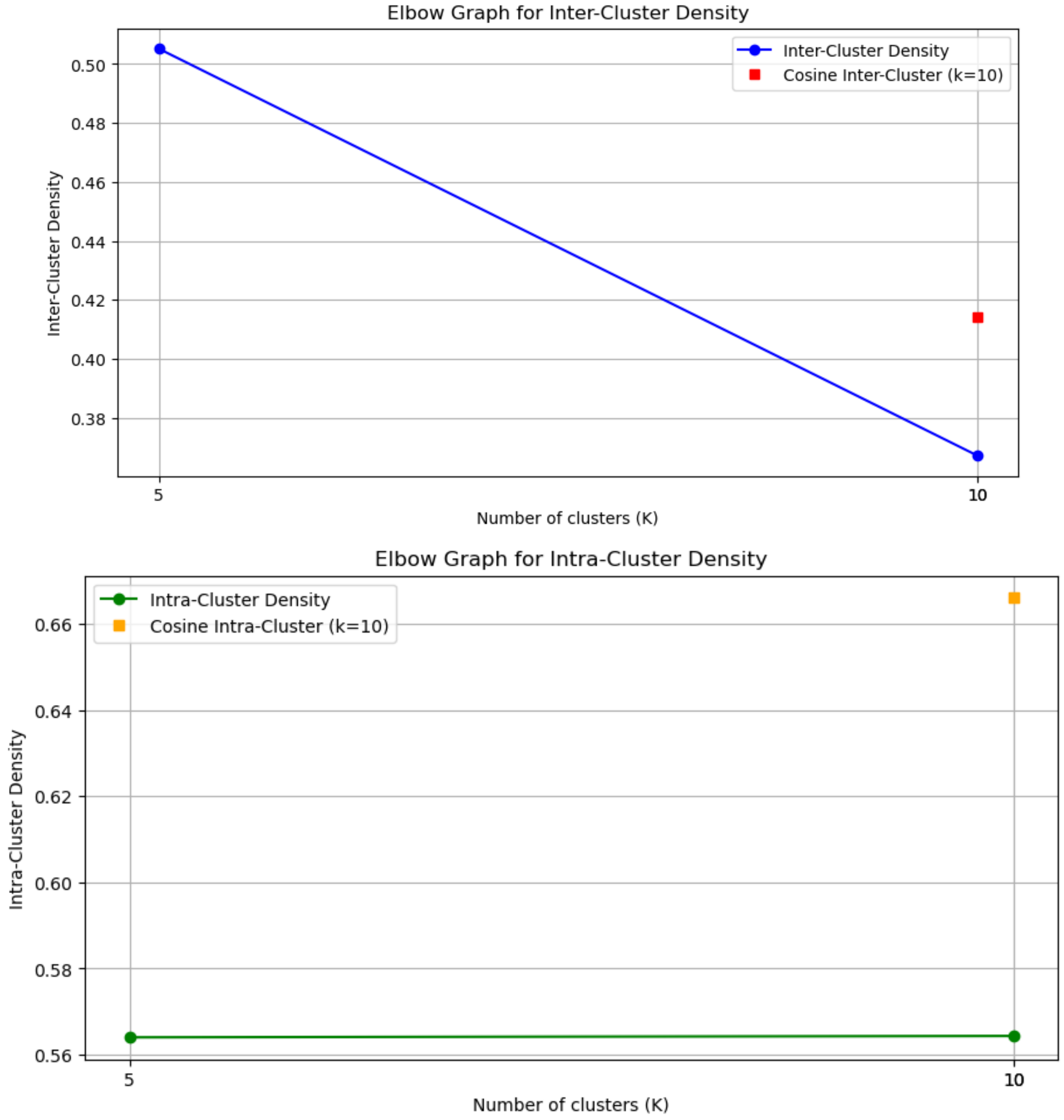
Inter-Cluster Density: 0.5052167024866535

Intra-Cluster Density: 0.5640387657590269

Cosine Distance
Inter-Cluster Density: 0.41435833538955114
Intra-Cluster Density: 0.6660760856365529

The K-Means algorithm produced the lowest inter-cluster density (0.367), and the highest intra-cluster density (0.667) was the Cosine Distance. Implementing the K-mean clustering algorithm with Cosine Distance Measure and verifying the relation between the average distance to the centroid and the K values was not successfully calculated. The elbow graph visualisation for the K-mean clustering with the Cosine Measure is plotted in Figure 2 below.

Figure 2. Elbow Graph for Inter-Cluster Density



Conclusion

The Big Data Analysis concluded the first task of descriptive weather data analysis using the MapReduce framework. The pseudo-code for the mapper and reducer functions are the pseudo_mappers.txt and pseudo_reducers.txt files. The Task was submitted to the Hadoop system using the script:

```
hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \
 -file mappers.py -mapper 'python3 mappers.py' -file reducers.py -reducer 'python3 reducers.py' -input 200707hourly.txt -output output2110.
```

The analysis produced no output beyond the 19th day of the month. With further work, this could be investigated. The limitations of the MapReduce methodology are that it requires real-time processing and that the data must be in key-value pairs. Debugging the script was time-consuming, and it was hard to simulate the distributed environment on a local machine. The text cluster analysis aimed to create sequence files from raw text, make a sparse representation of vectors, and run the K-mean algorithm with Euclidean and Manhattan Distance Measures. The bash_script_Q2_Cluster.sh file contains the script that aims to find the optimum number of clusters (between 2 and 10). However, the script did not produce a valuable output. Due to the time limitation, it was impossible to debug the script satisfactorily. There is also a limitation in the reproducibility of the results, as the server-side script is not in full representation in the submission of this work.

In conclusion, the valuable insights from the weather and text analysis have served a great purpose for the big data framework. However, comparing the newer data processing framework was not a viable option.

References

[1] I. Gad and D. Hosahalli, ‘A comparative study of prediction and classification models on NCDC weather data’, *International Journal of Computers and Applications*, vol. 44, no. 5, pp. 414–425, May 2022, doi: 10.1080/1206212X.2020.1766769.