Assignment 1 Report

BD\_0002\_0275\_0314\_0366

1. Task 1

We are using the Hadoop Streaming API and Python3 to build the mapper and reducer for Task 1.

1.1 Mapper

The mapper first accepts a command line argument for the word to be checked against. We load the dataset into memory, one record at a time, and initially eliminate the record if does not satisfy the predefined conditions (thus making it a “bad record”) in the problem statement. If the record is classified as good, we compare the word in the record with the argument received and retrieve the recognized status of the record. If the status is True, we output from the mapper “RECOGNIZED 1”. Otherwise, we check if the timestamp of the record corresponds to any day of the weekend, and output “UNRECOGNIZED 1”.

1.2 Reducer

The reducer receives the intermediate key-value pairs from the mapper. This input is then split to obtain the recognized flag and the count, and the corresponding counters are incremented with the count. The final counts for both flags are then displayed on separate lines.

2. Task 2

We are using the Hadoop Streaming API and Python3 to build the mapper and reducer for Task 2.

1.1 Mapper

The mapper first accepts a command line argument for the distance metric K and the word to be checked against. We load the dataset into memory, one record at a time, and initially eliminate the record if does not satisfy the predefined conditions (thus making it a “bad record”) in the problem statement. If the record is classified as good, we compare the word in the record with the argument received and retrieve Euclidean distance between the Origin (0, 0) and the first pair coordinates in the first stroke. If the distance is greater than K, we output the country of the record and the count ‘1’.

1.2 Reducer

A dictionary is initialised to store the country and count key-value pairs from the mapper. The reducer receives these intermediate key-value pairs from the mapper and splits them to obtain the country and the count, and the corresponding country key in the dictionary is incremented with the count. This dictionary is then sorted based on the country in lexicographical order and then displayed.