PDC Assignment-4

Hammad Habib

20i-0864

“A”

Question 1

The input data is a csv file containing approx 3.1M articles data. The format is such that each line contains data for 1 article in the following format
Publication year, Journal name, and authors.

The goal of the algorithm is to count the number of articles published in each journal per year.

In the mapper stage, each record is processed one by one. The mapper extracts the publication year and the journal name from each record by tokenizing the input string with the comma delimiter. Then, the mapper concatenates the year and journal name and uses it as the key. The value associated with each key is a constant integer value of 1. The key-value pairs generated by the mapper represent the year and journal name as the key and the value 1 as the associated value.
In the reducer stage, all the key-value pairs generated by the mapper are grouped based on their keys. For each key (which represents a year and journal name), the reducer calculates the sum of all the values associated with that key, which represents the count of articles published in that journal in that year. Finally, the reducer outputs the year, journal name, and the count of articles published in that journal in that year.
Overall, the mapper stage generates key-value pairs where the key represents the year and journal name, and the value is a constant 1. The reducer stage aggregates the values for each key, producing the count of articles published in each journal per year.

The result is in the following form

2018,meltdownattack.com 2

Question 2

The mapper function takes each line of input data as an input record and converts it into a key-value pair. The key of the pair is a Text object, and the value is an IntWritable object.

The key is composed of two author names separated by a comma, and the value is always 1.

The mapper function extracts the author list from the input data and sorts it alphabetically to ensure that the order of authors does not matter.
It then generates all possible pairs of authors from the list using two nested for-loops and emits a key-value pair for each pair.
The output of the mapper function is a set of key-value pairs, where the key is a pair of authors who have published articles together, and the value is always 1. The reducer function then takes these key-value pairs as input and counts the number of times each pair appears in the input data.
The final output is a set of key-value pairs, where the key is a pair of authors who have published articles together, and the value is the number of articles they have published together. The reducer function aggregates the counts of each pair of authors, which are the same keys, and returns the final output.

The result is in the following form

(**Abdel Razik Sebak,Tayeb A. Denidni) 4**