Floura Angel

Jayashree

Nifasath

**YELP**

**INDEX**

**2.1 INTRODUCTION**

**2.2 PROBLEM STATEMENT**

**2.3 ABSTRACT**

**2.4 DATASET DESCRIPTION**

**2.5 PRE PROCESSING**

**2.6 PROBLEM STATEMENT 1**

**2.6.1 KEY VALUE**

**2.6.2 MAPPER**

**2.6.3 REDUCER**

**2.6.4 CLEANUP**

**2.6.5 SORTBY VALUES**

**2.6.6 SCREENSHOTS**

**2.6.7 OUTPUT**

**2.7 SORTBY VALUES FUNCTION**

**2.8 PROBLEM STATEMENT 2**

**2.8.1 KEY VALUE**

**2.8.2 MAPPER**

**2.8.3 REDUCER**

**2.8.4 CLEANUP**

**2.8.5 SORTBY VALUES**

**2.8.6 SCREENSHOTS**

**2.8.7 OUTPUT**

**2.9 INFERENCE**

**2.10 FUTURE IMPLEMENTATION**

**INTRODUCTION**

Yelp is an American multinational corporation that develops, and markets the online review platform Yelp.com and the Yelp [mobile app](https://en.wikipedia.org/wiki/Mobile_app), which publishes [crowd-sourced](https://en.wikipedia.org/wiki/Crowd-sourced) reviews about local businesses, as well as the an online reservation service called Yelp Reservations and an online food-delivery service Eat24. The company also has a couple of additional services such as training small businesses in how to respond to both negative and positive reviews, hosts a number of social events for reviewers, and provides other insights about the businesses.

Yelp is online platform that provides rating of various local businesses by crowdsourcing user reviews. It also provides other data such as health inspection scores.

Yelp made a massive dataset of available to the public in context of their Yelp Dataset Challenge. These large datasets can be manipulated and maneuvered to generate various insights about the various businesses on the platform. To manage this data we need more advanced data processing methods than just RDBMS.

In recent times data processing has taken enormous strides in terms of management of large volumes of data and to derive practical insights from this data. Hadoop technology can be used effectively to handle big data and convert it into actionable results. The Hadoop MapReduce distributed framework allows processing of large datasets simultaneously on a large cluster of nodes where the dataset is broken into chunks and stored and processed .

**ABSTRACT**

In this project, we intend to design and implement the MapReduce Jobs using Hadoop MapReduce framework on popular Yelp dataset provided by Yelp Inc. The purpose of the project is to analyse the Yelp dataset and extract various facts to help business owners understand their business growth and estimate the impact of users rating over a period of time. In this particular project we will be performing manipulations on the dataset to determine the areas where most businesses are located and find the businesses with the highest average ratings.

**PROBLEM STATEMENT**

Yelp is an American multinational corporation which publish crowd-sourced reviews about local businesses. Given to you is the business and review datasets of Yelp. Come up with a map reduce solution to find out

1. top 10 zip codes where the most businesses are located.

2. top 10 businesses using the average ratings.

**DATASET DESCRIPTION**

1. Business.csv

file contain basic information about local businesses.

Business.csv file contains the following columns

· business\_id: (a unique identifier for the business)

· 'full\_address': (localized address),

· 'categories': [(localized category names)]

2. Review.csv file contains the star rating given by a user to a business. Use user\_id to associate this review with others by the same user. Use business\_id to associate this

review with others of the same business.

review.csv file contains the following columns

· "review\_id","user\_id","business\_id","stars"

· 'review\_id': (a unique identifier for the review)

· 'user\_id': (the identifier of the reviewed business),

· 'business\_id': (the identifier of the authoring user),

· 'stars': (star rating, integer 1-5),the rating given by the user to a business

**PRE PROCESSING**

Data has been cleaned by removing empty string and has been assigned 0 for classes which is not assigned in the dataset. We trimmed extra spaces for example (string.trim()) To get a single column we used awk command.

**PROBLEM STATEMENT1**

1. top 10 zip codes where the most businesses are located.

**KEY VALUE**

The key in the problem statement 1 is the zipcode. The reason why zipcode is taken as the key is to find the particular area in which the business is blooming.

The value is 1 for every business in that zipcode

**MAPPER**

The mapper function splits the data according to the ^ . Retrieves the zipcode that is needed for the problem statement. The extra spaces in the zipcode is trimmed and only the number is obtained. The key value is sent from mapper to the framework.

**REDUCER**

Reducer does the summing up. It takes the list of values from framework and group them.The counter and reducer variable in the reducer function calculates the sum. The sum is then written to the output. This is the job of reducer.

**CLEANUP**

cleanup() is used to clean up any resources that is allocated. There are other uses too, which is useed to flush out any accumulation of aggregate results. setup() and cleanup() are methods that can override they are there to initialize and clean up your map/reduce tasks. The lifecycle of a map/reduce task is (from a programmer’s point of view):

setup -> map -> cleanup

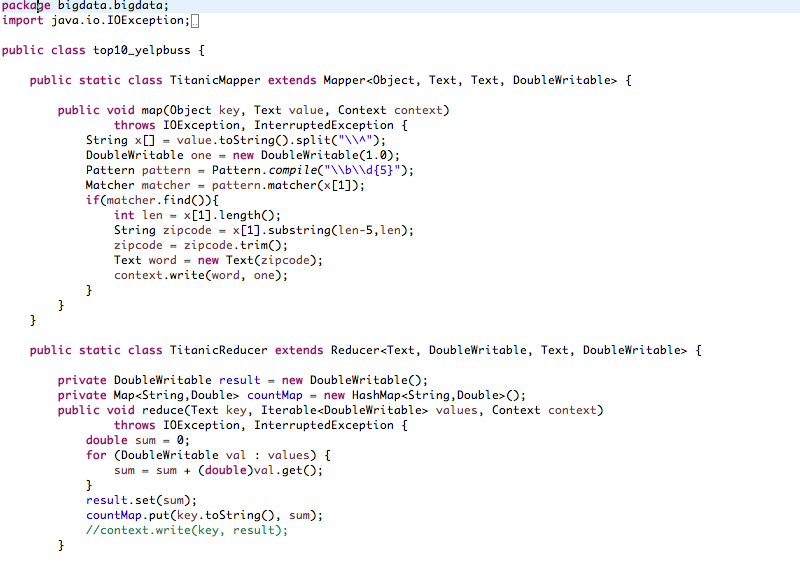
setup -> reduce -> cleanup

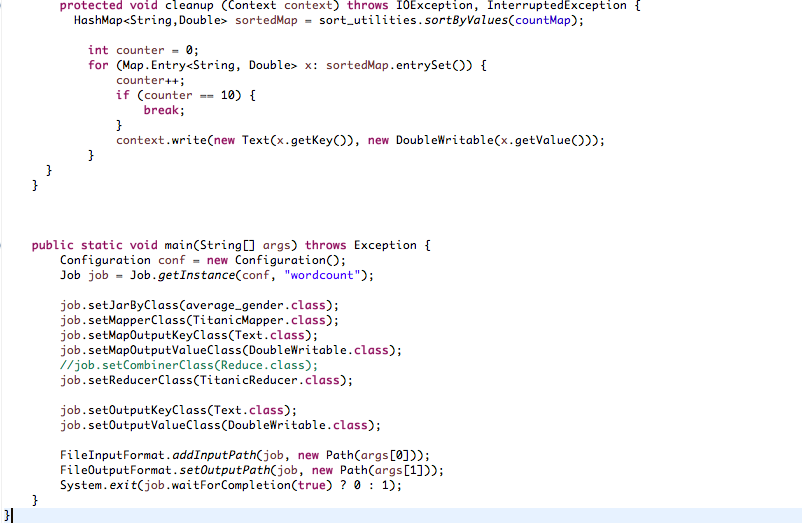
So, for each mapreduce first setup() method is called then map()/reduce() method is called and later cleanup() method is called before exiting the task. With the help of this function we get only the top 10 zipcodes where the most businesses are located.

**SORTBY VALUES**

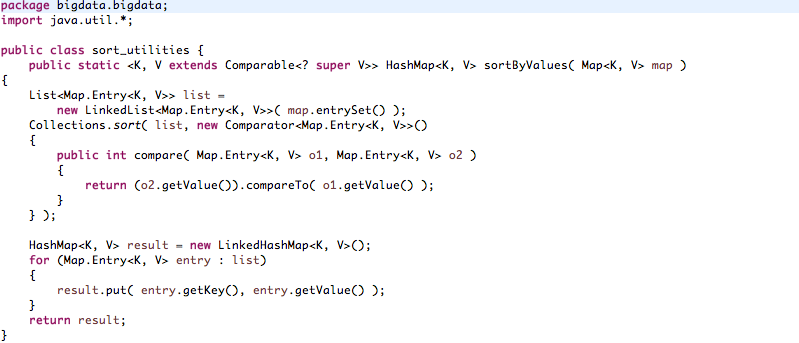
Sortby values uses comparable interface which is a data structure which is used to sort the hash map.It sorts the output of the reducer.

**SOURCE CODE**

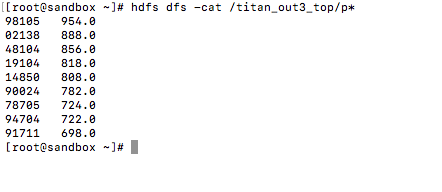
****

****

**SORTBY VALUES FUNCTION**

****

**OUTPUT**

****

**PROBLEM STATEMENT2**

2. top 10 businesses using the average ratings.

**KEY VALUE**

The Key for the problem statement 2 is business ID because we are finding the average rating for the business.

The value is rating for the business.

**MAPPER**

The mapper function splits the data according to the ^ . Retrieves the business ID and ratings that is needed for the problem statement. The extra spaces in the business ID is trimmed and only the number is obtained. The key value is sent from mapper to the framework.

**REDUCER**

Reducer used to find the average ratings in top 10 business. It takes the list of values from framework and group them.The counter and reducer variable in the reducer function calculates the sum and average. The sum and average is then written to the output. This is the job of reducer.

**CLEANUP**

cleanup() is used to clean up any resources that is allocated. There are other uses too, which is useed to flush out any accumulation of aggregate results. setup() and cleanup() are methods that can override they are there to initialize and clean up your map/reduce tasks. The lifecycle of a map/reduce task is (from a programmer’s point of view):

setup -> map -> cleanup

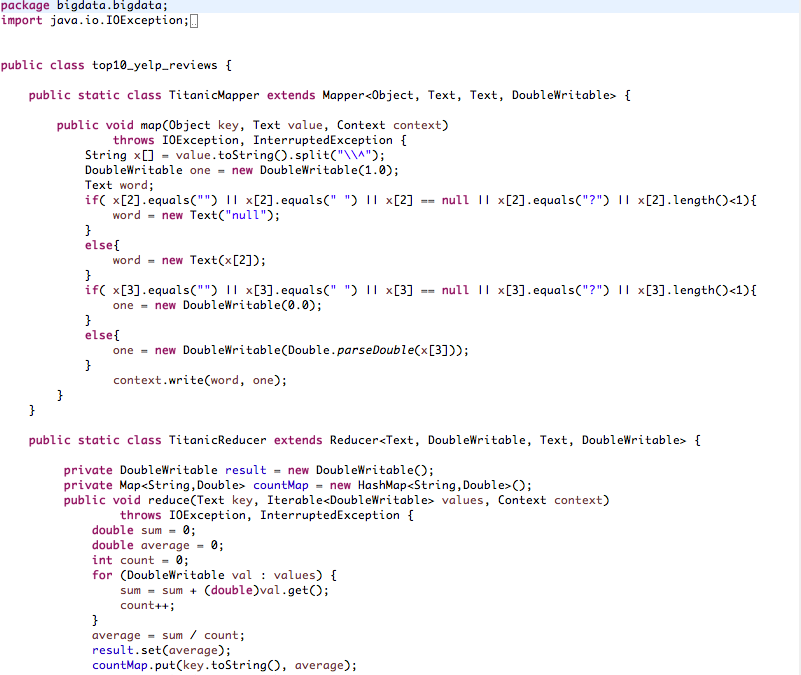
setup -> reduce -> cleanup

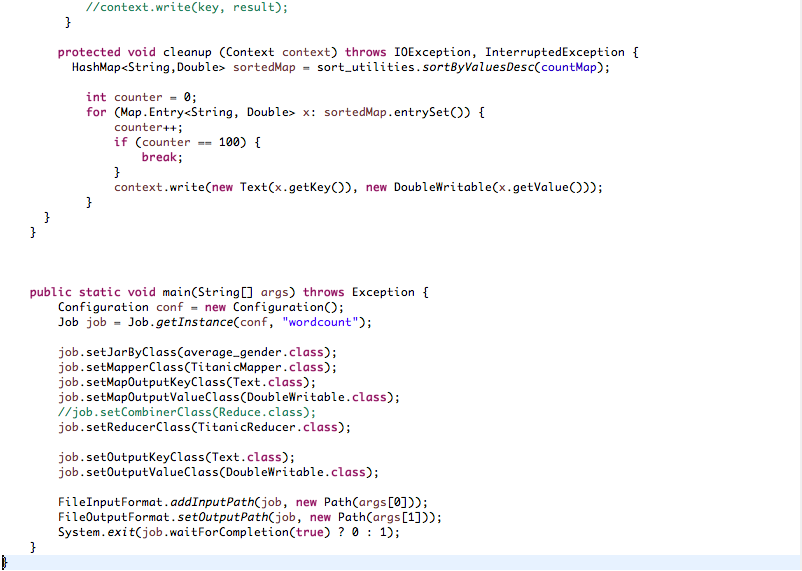
So, for each mapreduce first setup() method is called then map()/reduce() method is called and later cleanup() method is called before exiting the task. With the help of this function we get only the top 10 zipcodes where the most businesses are located.

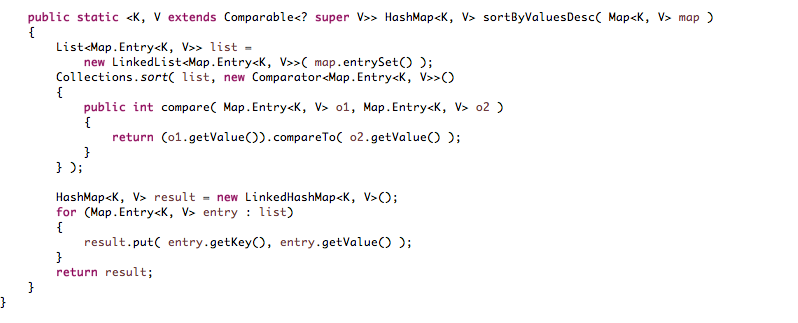
**SORTBY VALUES**

Sortby values uses comparable interface which is a data structure which is used to sort the hash map.It sorts the output of the reducer. Top N-Analysis Algorithm is used to get the top values of the dataset given.

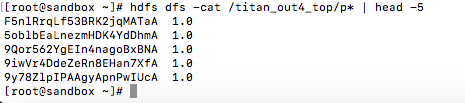
**SOURCE CODE**

****

****

****

**OUTPUT**

****

**INFERENCE**

Using Map-reduce we were able to analyse the enormous dataset to arrive at a suitable output. In the first analyse we were able to determine the top 10 zip codes where the most businesses are located. This information can be used to determine the busiest areas and which areas are more viable for future growth and expansion and those areas which are already saturated.

In the second analysis we determined the top 10 businesses according to their average ratings. With this information we can determine the competition and market leaders and learn which businesses are the most successful.

**FUTURE IMPLEMENTATION**

By analysing the given dataset further a number of conclusions can be made. We can make further analysis on the dynamics of the region in which a particular business is located to determine whether it will be successful or not. Also we can determine which businesses are more successful than others and also maybe determine the reason.

In this project we have used the terminal to implement the instructions. For future works we can develop a GUI to interact with the data and arrive at conclusions.