**Data Analysis Using Map/Reduce**

Submitted by: Team 01 – Sameer Moses Murala (1001556494)

– Sameena (1001679154)

Under Supervision of: Prof. Sharma Chakravarthy

Department of Computer Science,

University of Texas at Arlington

**Contents**

**Overall Status**

Map Implementation

Reduce Implementation

Analysis Results

**File Description**

**Division of Labor**

**Logical Errors Faced**

**Overall Status**

We first understood the Map Reduce Concepts provided in blackboard and in the course notes. With the help of Professor and TA (Teaching Assistant), we got understood the concept of a Map Reduce and how we could implement a simple Map Reduce framework.

We have successfully implemented the following methods:

* **Map Implementation**

In this method we take each line in the input file provided as a single string. We then split the string according to ‘;’ character into string tokens. From the input data we know that the token 4 is contains the data corresponding to the year of the movie and the token 5 contains the genre list data. Before creating the key value pairs we check if the year token corresponds to an year on or after 2000 and doesn’t contain ‘\N’ character and the genre list token does not contain ‘\N’ string. We take the genre list token and again split them into tokens of each genre in that list. We then create a key containing both the year and genre information. We then create a key value pair of key and value of each key as one for the count.

* **Reduce Implementation**

In this method we take the key and value list. Here the value list the number of 1’s of each key that is the year and genre of the movies released in that year. For each value in that list we increase the count by 1. We create a list of key value pairs in which the key is the movie year and genre and the value is the number of movies of that genre in that particular year.

**Analysis Results**

**Most Popular Genre (2000-2020)**

We can see that the most popular genre in the period of 2000-2011 remains to be drama. It has the most number of movies released in that period, with a steady increase in the number of movies released each year.

But there is change in the years from 2012-2017 where comedy films are in demand and more movies of such genre are released in that period with a consistent growth up to 2016 after which there is a little fall in the numbers in 2017 and after that the drama genre again prevails as the most released genre from 2018 to 2020.

**Least Popular Genre (2000-2020)**

It is quite evident from the above graph that the movie genre western has the least number of movies released of such type in the period from 2000 to 2017.There is no consistent increase or decrease in the number of movies of such genre released in that period. It reaches the highest number in the year 2017 with 355 movies of that genre being released in that year.

From the period of 2018 to 2020 there is a new least released genre that is Talk Show television having lowest numbers in those years.

**Total Number of Movies of Each Genre (2000-2020)**

The above pie chart gives us an idea of the total number of movies of each genre released in the period of 2000-2020 and how much percentage does each genre constitute.

Drama is the most the genre that has maximum number of movies released in that period an constitutes a major 15% of the total number of movies released, followed by the genre comedy that is the second most watched and released genre in that period with about 14% of the total number of movies released. The genre western and war are the not so preferred genre chosen for movies and have the least numbers.

**Growing Genre (2000-2010)**

From the above graph we can observe the average increase of each genre in the number of movies released in the period of 2000-2010. From the above data it is evident that genre of Reality TV is the most growth in the numbers in this period. From period of 2000 to 2010 the number of titles of genre Reality TV has become the more preferred TV genre and has seen the more dramatic increase in the number of such titles released with dramatic average increase of 33% every year.

On the other hand the TV genre of Game Show has been the most consistent or the one with least increase in numbers over the years with a mere average increase of 5% each year.

**Total Number of Movies Released Each Year (2000-2017)**

The above analysis gives us an understanding of the total number of movies that were released each year in the period from 2000-2017. We can see that there has been tremendous increase in the numbers over years with an average increase of 15% which is huge number.

In the period of 2000-2008 the increase had been steady but there are dramatic increase in the period from 2008-2015 with highest number of movies being released in the year 2016. The numbers are consistent with not much increase in the 2013 to 2016 period with a little dip in the numbers in the year 2017.

**File Description**

No additional files were created for our project. All the files we have used is present in the skeleton code given.

**Division of Labor**

In the first few weeks, we setup the Ubuntu OS using a Virtual Machine and installed Hadoop 2.9 in the single cluster mode using the pdf given. We met a couple of times to understand the Map reduce framework and studied the information and Map Reduce Concepts given in the pdf files. Figured out the use of each function using the example word count program and found out how to use those functions and variables in our implementation .Then we decided on how to distribute our work.

We followed the pdf given in class started working on the code. We have worked on both the Map and Reduce implementations together.

We created the Analysis Report using histograms.

Each of us spent around 15 hours each for the implementation of the project.

**Logical Errors**

1. Installation of Hadoop was difficult as we were unfamiliar with Hadoop. The instructions provided were not clear enough so went through some tutorials to figure out the Hadoop installation.
2. During the implementation of the Map method we had initially spilt both the input string and the genre list according to the ‘;’ character. This caused array out of bound exception as it did not split the genre list.
3. During the implementation of the Map method initially the information regarding the ‘\N’ information present in the input data. Whether such data should be included in the final output or not was not clear. After the update from TA we proceeded as instructed.
4. During the implementation of the Map method initially we created a key with only the year information. In the reduce method we were unable to get the genre information. Later we created a key with both the year and genre information.