**CEG-7370-01**

**Distributed Computing**

**Fall 2015**

**Project #3**

**Write-up File**

**DHRUVKUMAR NAVINCHANDRA PATEL**

**U00791652**

**Music Meta Data Processing Using Hadoop Map/Reduce**

1. **What is Hadoop – MapReduce**

* Hadoop MapReduce is an Algorithm to processing metadata from large input files.
* It contains two important Task. 1) Map 2) Reduce
* Map function divide the data from the input file and formatted those data into Key, value pair.
* Reducer takes key, value pair from mapper and generate output which is in key, value pair format.
* This algorithm is used for the processing large size of input files into small size files with metadata.

**Following Task to get the metadata according to each task requirement.**

**Task0:**

1. Here I used 3 mapper to generate key and value pair from each file.
2. Take artist name as a key for this task.
3. Song title, artist id and location are the values from each file.
4. Pass these values into one reducer to processing all the key and value pairs and generate key, value pair according to output which mentioned into documentation.
5. Here I checked output manually and it is true. So I store that output in Table0.txt.
6. Output is in following format

**Artistname artistid<SEP>artistname<SEP>location<SEP>songstitle**

If more than one value for location or any other than separated by <I>.

So you can just ignore First Artistname from my file and check remaining string with your output to verify answer.

I couldn’t found the how to remove space between key and value.

**Task1:**

I used Table0.txt file as an input and generate output in following format using Map Reduce functionality.

Store this output in Table1.txt. Which is exactly similar with what you required.

**Task2:**

1. **To find top 5 location with the highest number of artist.**

I used Table1.txt as an input file and make location as a key and store all the artistname of that string into arraylist and put key and value in Hashmap and then proceeds into reducer to check which arrylist connected with location has highest artistname. Then I find top 5 location from key.

1. For second task I used Table0.txt and Table1.txt to generate final string and store that output in Table2.txt.

I Followed all the steps of submissions and put all the files inside zip folder.