CSE 4334/5334 – Data Mining

Fall 2014 - Project 1 Due: 11:59pm Central Time, Tuesday, Oct 16, 2014

Submitted By: Sarvesh Sadhoo (1000980763)

Design and Implementation:

1. **Part One:** For the first question following where the way in which the program was designed and implemented**.**
    a. A Hash Map <hash_users >was created for the users.tsv file of the format {user_id: [state,country]}
    b. A Hash Map <hash_apps> was created for the apps.tsv file of the format {job_id: {users_state: [job_id]}}. This was a Hash Map of a Hash Map which contained a list of all the job_id
    c. A count operation was used on the user_state list, job_id to count the number of job using Hash Map <hash_apps>.
    d. A max operation as applied to get max value of the number of jobs for a particular state and job_id using Hash Map <hash_apps>.
    e. Finally the top 5 values for a particular job id and state where taken from the complete Hash Map <hash_apps>.

2. **Part One:** For the first question following where the way in which the program was designed and implemented**.**
    a. A Hash Map <hash_apps> was created for the apps.tsv file of the format {job_id: {users_country: [job_id]}}. This was a Hash Map of a Hash Map, which contained a list of all the job_id with country. It was a roll up operation to get to the country level.
    b. A Hash Map <hash_jobs_cube > was created for the jobs.tsv file of the format {job_id: job_title}
    c. A count operation was used on the user_country list, job_id to count the number of job using Hash Map <hash_apps>.
    d. The Hash Map was sliced on country name to get the data for a particular country given as the user input.
    e. A new Hash Table <hash_title > was created to get total count for the particular job title. Format {job_title: total count}
    f. The above hash table was sorted to in a descending order to get the top 5 job title.

Execution Steps:

1. Copy the python file (project1.py) included in the zip folder to the desired location on your machine.

2. Open your terminal/command prompt.

3. To run the project1.py run give the command in the below format. **/p/t** refers to the path of the input file

   Format: python project1.py country_name /p/t/apps.tsv /p/t/users.tsv /p/t/jobs.tsv /p/t/user_history.tsv

   Example: python project1.py 'US' '/Users/srv/Desktop/Data Mining/Project/data/apps.tsv' '/Users/srv/Desktop/Data Mining/Project/data/users.tsv' '/Users/srv/Desktop/Data Mining/Project/data/jobs.tsv' '/Users/srv/Desktop/Data Mining/Project/data/user_history.tsv'

   Note: Make sure you give the path and country variables as a string. Also the input file sequence should also be same as shown in the example

Execution Screen Shot: