Project Report

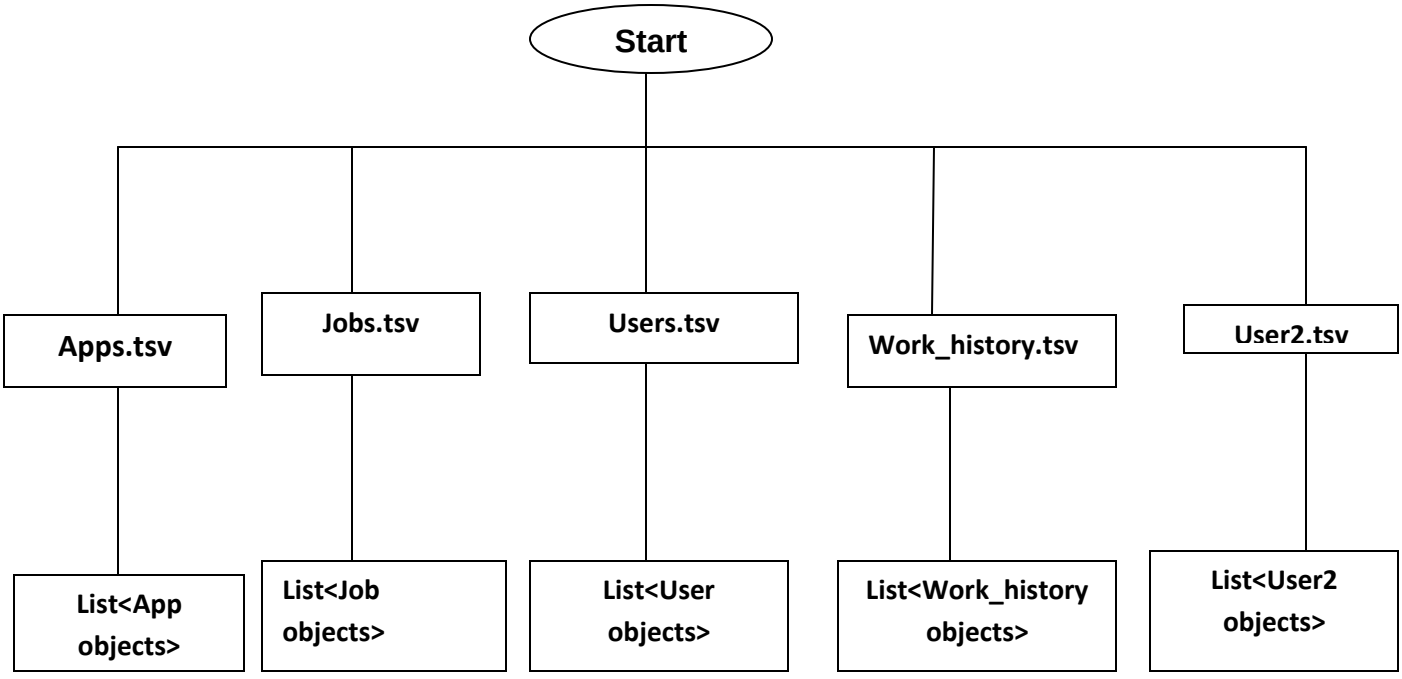
Data mining assignment 2

The following project has been implemented using Python programming language.

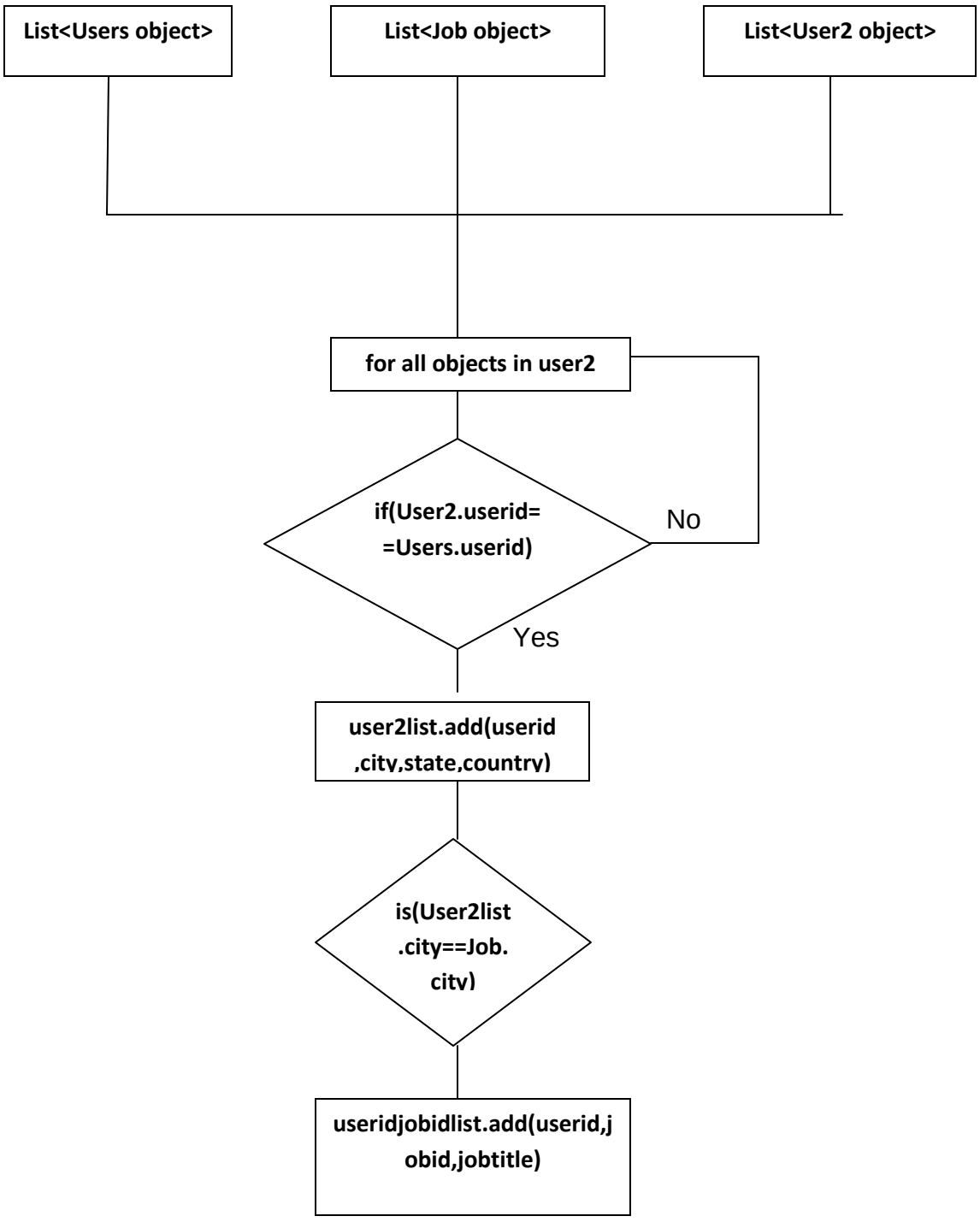
Eclipse Integrated environment is used for creating the source code.

Design:

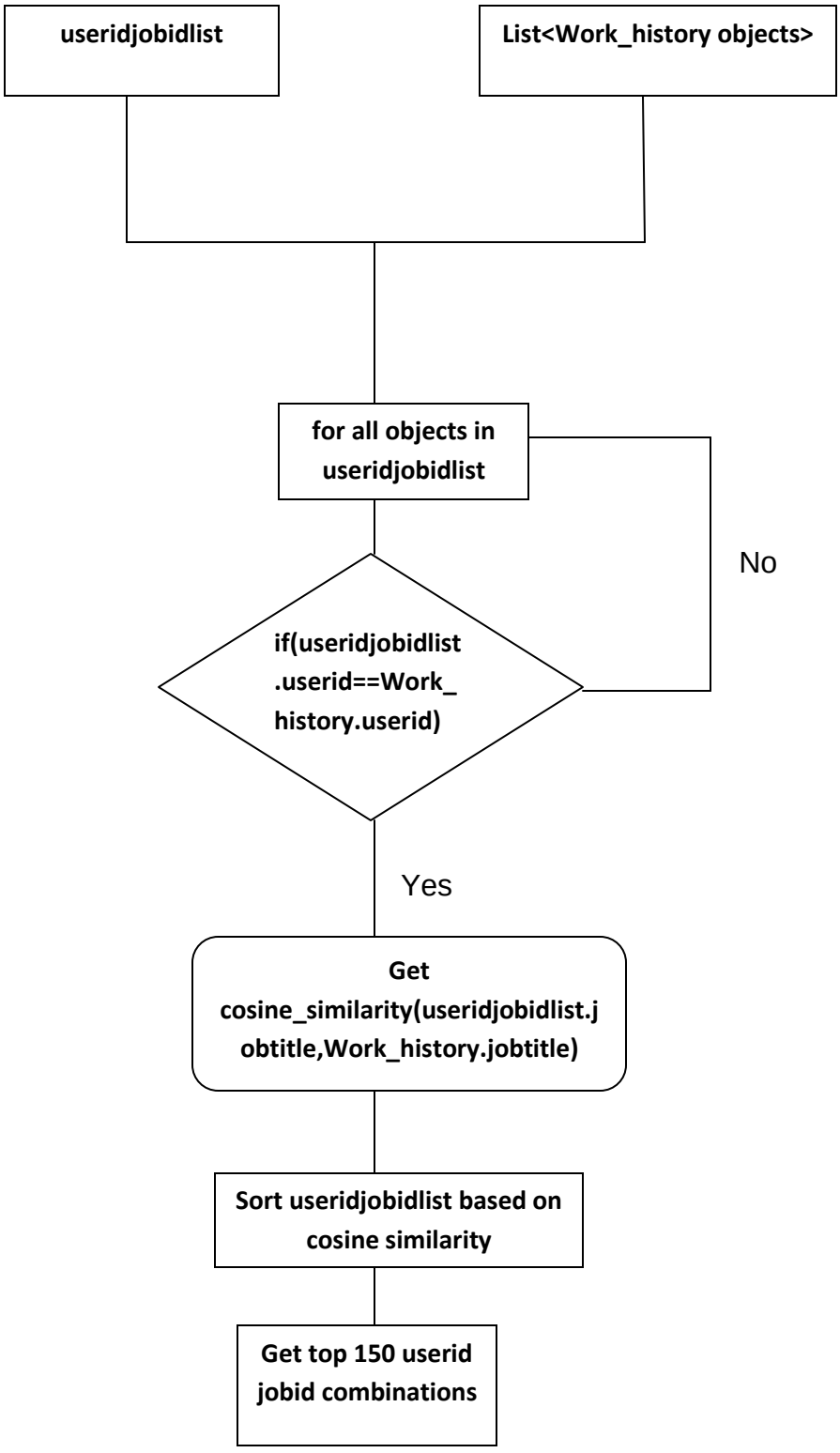
The following flowcharts gives a brief idea on the program logic flow.



Part 1:


List<Users object>
List<Job object>
List<User2 object>
for all objects in user2
if(User2.userid==Users.userid)
No
Yes
user2list.add(userid,city,state,country)
is(User2list.city==Job.citv)
useridjobidlist.add(userid,jobid,jobtitle)

Part2:


useridjobidlist
List<Work_history objects>
for all objects in useridjobidlist
if(useridjobidlist .userid==Work_ history.userid)
No
Yes
Get cosine_similarity(useridjobidlist.j obtitle,Work_history.jobtitle)
Sort useridjobidlist based on cosine similarity
Get top 150 userid jobid combinations

Implementation:

1) Considering all the 5 input files given each file is converted to list .Each record in a file is assigned to a object and all the objects are stored in their respective lists.

2)Every object in the list contains the attribute values related to each record.(for eg apps object contains userid,applicationdatejobid)

3) When the file path is passed as an argument it is stored in a file object ,which is passed as a parameter to the readlineobject. Using the delimiter each row is split and each attribute value is stored into an array . The array is used as an input to load values in the object

4) Now to obtain the location of users in user2 file ,each userid is cheked for equality in users file .If userid in user2 and users match then the corresponding city satte and country is appended to the userid and a new list termed "user2list" is created which contains user2id and location.

5)For each userid in user2list if the city location matches with the city location in job file , then corresponding jobid is assigned to the userid . A new list containing of all such matches is created called "useridjobidlist".

6)From work_history file each userid is compared to uesrid in useridjobidlist ,If a match is found then job title from work_history and jobtitle from useridjobidlist is used for computing cosine similarity .The greater the cosine similarity between the job titles the greater is the chances of the user applying to that job.

7) The userid jobid combination is sorted based on cosine similarity and the top 150 values are retrieved and stored in output.txt file .

Execution:

The following code is used for execution of the source code

python project2.py /home/g/gx/gxz2070/5334 /tmp/pavan

The Omega server has Python 2.4 version which does not support certain methods implemented in the source code .The python version that i'm using is python 2.7 .

I tried running source code in omega but it is producing errors as those methods are not implemented in python 2.4 version .

The output file is attached with in the zip file provided .