

Data Mining

Project 3 report

The following project has been created using python programming language.

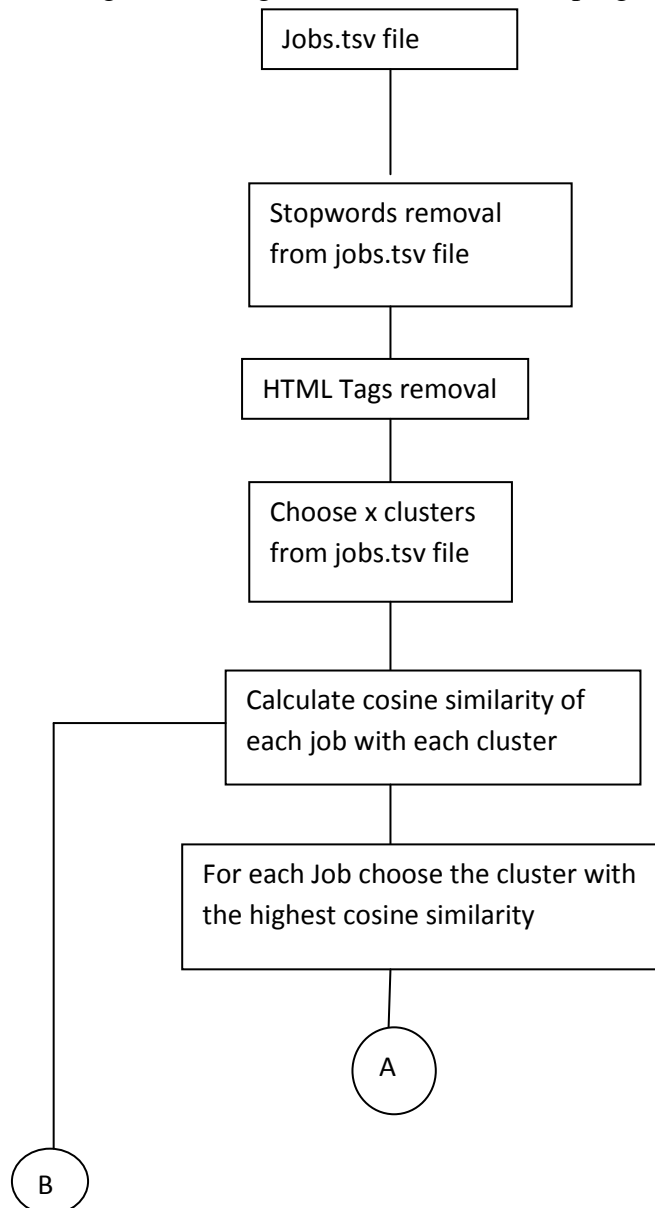
Eclipse Integrated development environment is used for compiling and running the source code.

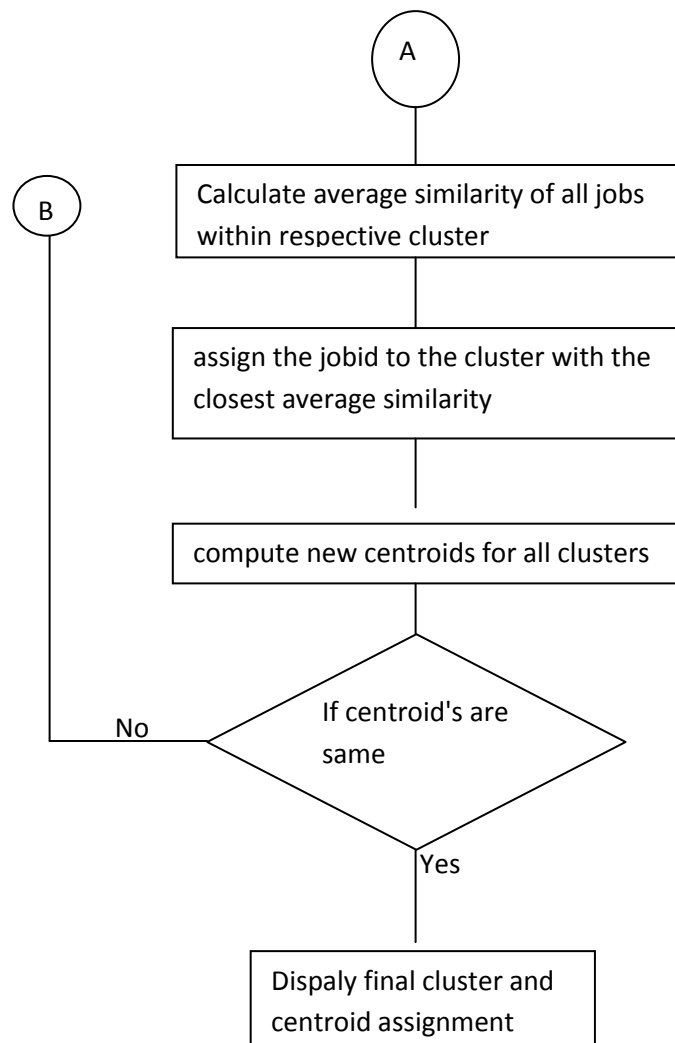
Stopwords list has been taken from following websites.

a) <http://www.ranks.nl/stopwords>

b) <http://www.webconfs.com/stop-words.php>

Design: The following flowchart gives a brief idea on the program logic flow.





Implementation:

- 1) The jobs file which is given as a input to the program is stored as a list .Each record in the file is stored as an object in the list .Each object in the list contains JobID, description and requirements.
- 2)K-means clustering algorithm is used for clustering the jobs .The similarity measure being used is the cosine similarity .
- 3)Using the random function 10 random jobid's are chosen as initial centroid's for k-means algorithm. Each centroid is assigned with a cluster number.
- 4)The cosine similarity between each centroid to each jobid is calculated .The jobid the cluster number and the cosine similarity obtained is stored in a new list .

5)Each job has similarities with a respective centroid. The respective jobid is assigned to the cluster to which it has the highest similarity .All the jobs are assigned similarly to their respective clusters .

6)For the new centroid calculation I consider average similarity of the cluster .Average similarity of each cluster is calculated and Jobid that has the closest average similarity in that cluster is chosen as the new centroid for the respective cluster.

7)The steps 4 ,5 & 6 are repeated until the centroids of last iteration and new iteration become equal. At this point the k-means algorithm is said to converge .

Execution:

1) The stop words file that is being used for removing stop words is included within the source code folder .

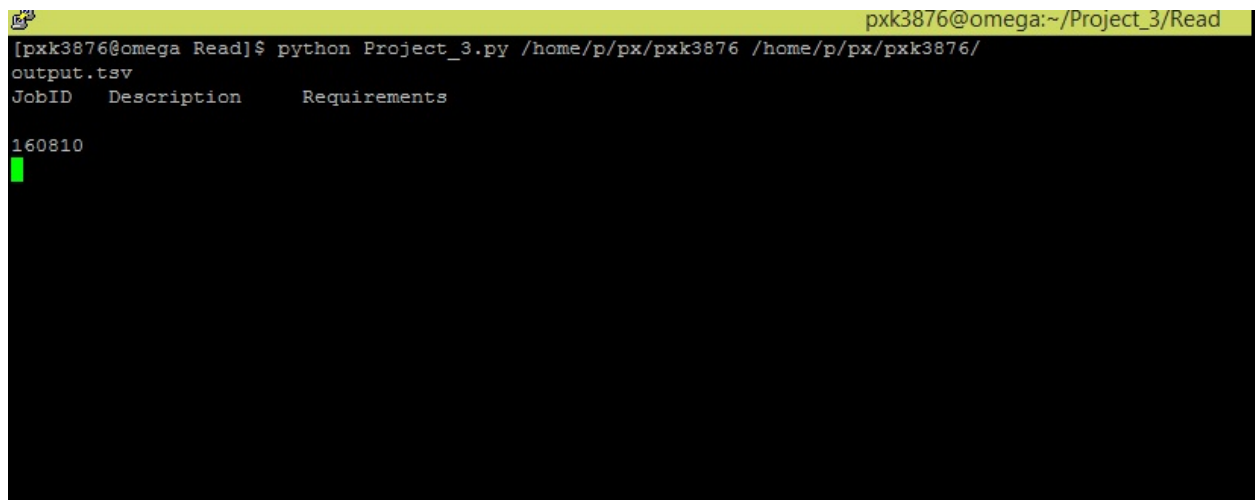
2) The following command line code is used for execution of the program.

```
python Project_3.py /home/p/px/pxk3876 /home/p/px/pxk3876/output.tsv
```

3) The above python command has to be executed after going into the read folder,which is placed inside Project_3 folder .

The output.tsv file is included in the zip file .

Execution screen shots:



```
pxk3876@omega:~/Project_3/Read
[pxk3876@omega Read]$ python Project_3.py /home/p/px/pxk3876 /home/p/px/pxk3876/
output.tsv
JobID  Description      Requirements
160810
```

```
pxk3876@omega:~/Project_3/Read
[pxk3876@omega Read]$ python Project_3.py /home/p/px/pxk3876 /home/p/px/pxk3876/
output.tsv
JobID      Description      Requirements
160810
[pxk3876@omega Read]$
```

C:\Users\Pavan S Koundinya\Desktop\output.tsv - Notepad++

File Edit Search View Encoding Language Settings Macro Run TextFX Plugins Window ?

output.tsv

1	852932	3
2	938157	8
3	393096	8
4	618414	8
5	518889	2
6	92616	8
7	1071882	7
8	61664	10
9	726772	6
10	538875	7
11	754114	4
12	20641	10
13	746954	2
14	40956	8
15	770853	1
16	92682	8
17	767633	7
18	1115709	8
19	673245	3
20	721221	4
21	237376	3
22	991176	8
23	262382	3
24	92643	8
25	1057972	2
26	87856	3
27	816214	3
28	592974	3
29	932465	7
30	280549	8
31	21187	8
32	18060	3
33	171056	2