CSE 4334/5334 – Data Mining

Fall 2014 - Project 3
Due: 11:59pm Central Time, Tuesday, Dec 2, 2014
Submitted By: Sarvesh Sadhoo (1000980763)

Implementation Idea:

For the purpose of implementing the project the idea is to initially, remove all stop words, perform stemming and merge specification & requirements together. Similarity is computed between jobs and they are clustered together based on the similarity value.

Design and Implementation:

Following steps where implemented for implementing the above-mentioned idea:

Step 1:
In the first step the jobs.tsv is parsed and stored in a hash table, with key as job_id. The job specification and requirement are merged together

Step 2:

In this, the above-generated hash is feed into a function that cleans the data by removing junk value, stop word and performing stemming. The result of this is a hash table with key as job id and value as list contains job data

Step 3:

In this step frequency is counted for every unique word in the job data and a count frequency matrix is generated.

Step 4:

Now we choose 10 random clusters for the purpose of clustering. After selecting the clusters we compute the similarity between the cluster and jobs and assign the job to cluster with max similarity

Step 5:

The above step is repeated 150 times to get a better centroid and clustering. In every iteration a new centroid is generated and clustering is done. Also a final cluster is selected which has a highest similarity sum. Similarity sum is simply a sum of all individual similarity score for job ids in a particular cluster.

Step 6:

In the final step, the clustering output is written to output.tsv file.

Dependency: Stemming library package is included in the project3 folder. The library used is stemming 1.0 (https://pypi.python.org/pypi/stemming/1.0). The folder also include similarity_compute.py file used to compute similarity and stop_word.py containing list of stop words.

Execution Steps:

The main file to run is project3.py. All other python files are for support.

1. Unzip the folder and copy the **project3** to the desired location on your system.

2. Open your terminal/command prompt.

3. Change directory to project3 folder and run the project3.py file.

4. To run the project3.py run give the command in the below format.

Execution Format:

python your-script-file.py /path/to/data/file/directory/ /path/to/output.tsv

Example:

python project3.py '/Users/srv/Desktop/Code/DataMining3/' 'output.tsv'

/p/t refers to the path of the input file

Note:

1. The main file to run is project3.py

2. Make sure you give the path variables as a string. Also the input file sequence should also be same as shown in the example for the program to execute properly.

3. Make sure any empty output.tsv already exists.

Execution Screen Shot:

```
Sarveshs-MacBook-Pro-2:CSE5334 Project 3 1000980763 srv$ python project3.py '/Us
ers/srv/Desktop/Data Mining/Project 3/' 'test_output.tsv'
Program Execution Started
Program Executed Successfully.
Output File Generated
Time Elapsed: 1011.73660684
Sarveshs-MacBook-Pro-2:CSE5334 Project 3 1000980763 srv$
```