Project 2.

DATA MINING

TO cluster questions from Stack Exchange sites, by using only the question text and its title.

(1st December 2013)

Report submitted by
Anuja Banekar
The University of Texas at Arlington
anuja.banekar@mavs.uta.edu

Student ID: 1000994277

Net ID: avb4277

Design:

The design is similar to Project 1. The first project was implemented using C++ language. And for Project 2, Java language is used.

Project1: The code is modified to output the cos similarity of each word. The words in a questions set are broken down and for each word their frequencies are counted depending whether they are present in title or text body. (The output is present in tags_file.txt)

Then for each word present in tags_file.txt their tf-value are counted. (The output is present in vector_file.txt)
Then the tf- values present in vector file are sorted and sorted output in stored sort.txt

When the user inputs arg[1] for number of clusters e.g k, the k values are outputted from sort text file and stored in cluster_values.txt

Project2:
The main logic is implemented in Java.

Clustering method logic used is somewhat similar to K-means.
To choose initial seed for each cluster the top k words are choosen depending on tf-values in descending order.

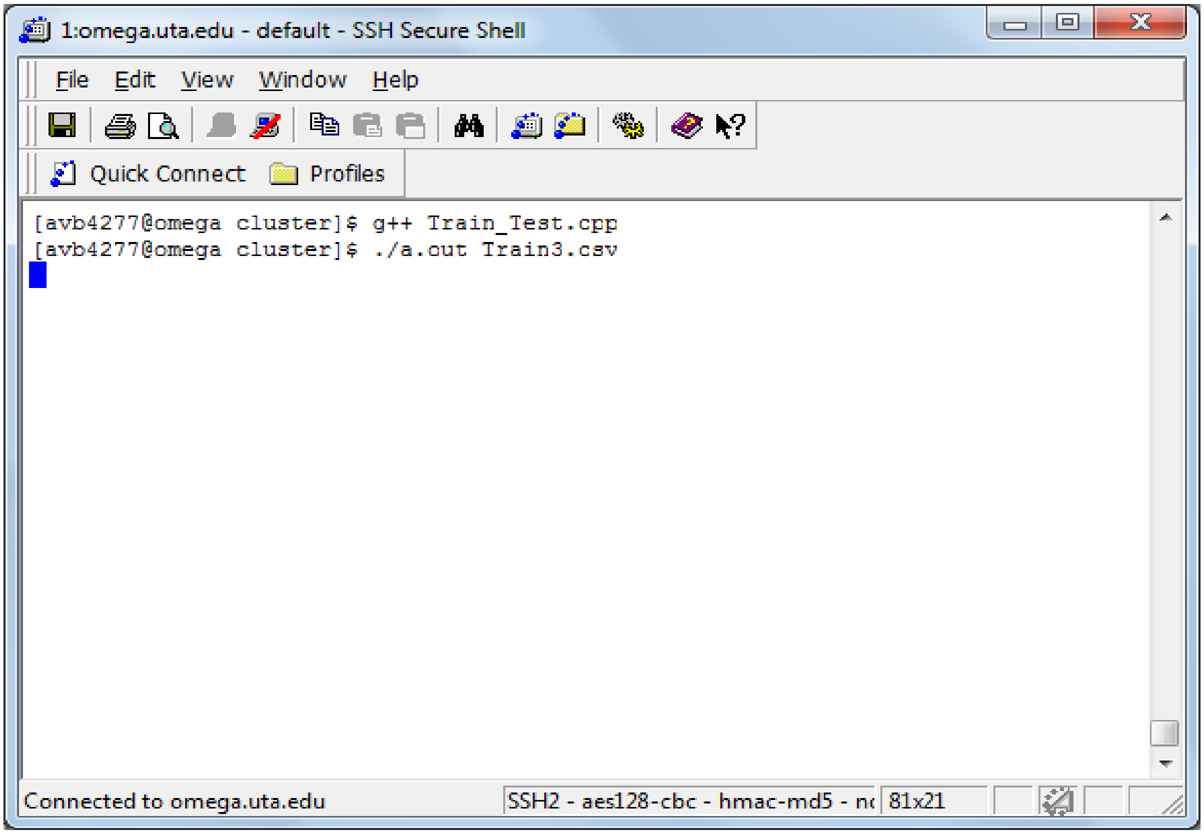
```
while(train_file is not empty)
{
        for(each token in string [] present in next line)
                if(column is 1 i.e Id column)
                        set the id.
                if( column is 2 || column is 3)
                {
                    while(cluster file is not empty)
                        read cluster file
                        search each token in cluster file
                        if token present then set the cluster no to the id associated to token.
                        and output in result.txt file.
                }

}
```

Still many of the questions are skipped during this process of evaluation. Still the output obtain is much more relevant and performance is reliable.
Also to execute use opencsv.jar

opencsv-2.3.jar

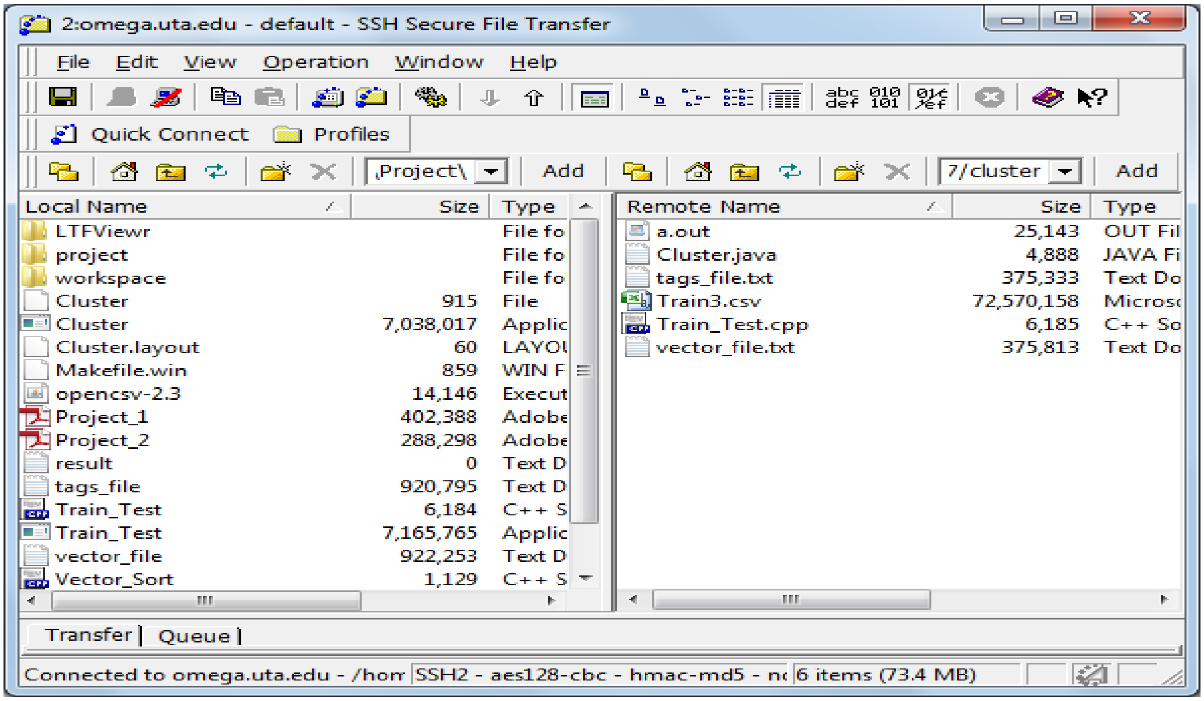
Screen Shots for execution:

Compilation and Execution:


1:omega.uta.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

```
[avb4277@omega cluster]$ g++ Train_Test.cpp
[avb4277@omega cluster]$ ./a.out Train3.csv
```

Connected to omega.uta.edu
SSH2 - aes128-cbc - hmac-md5 - nc
81x21

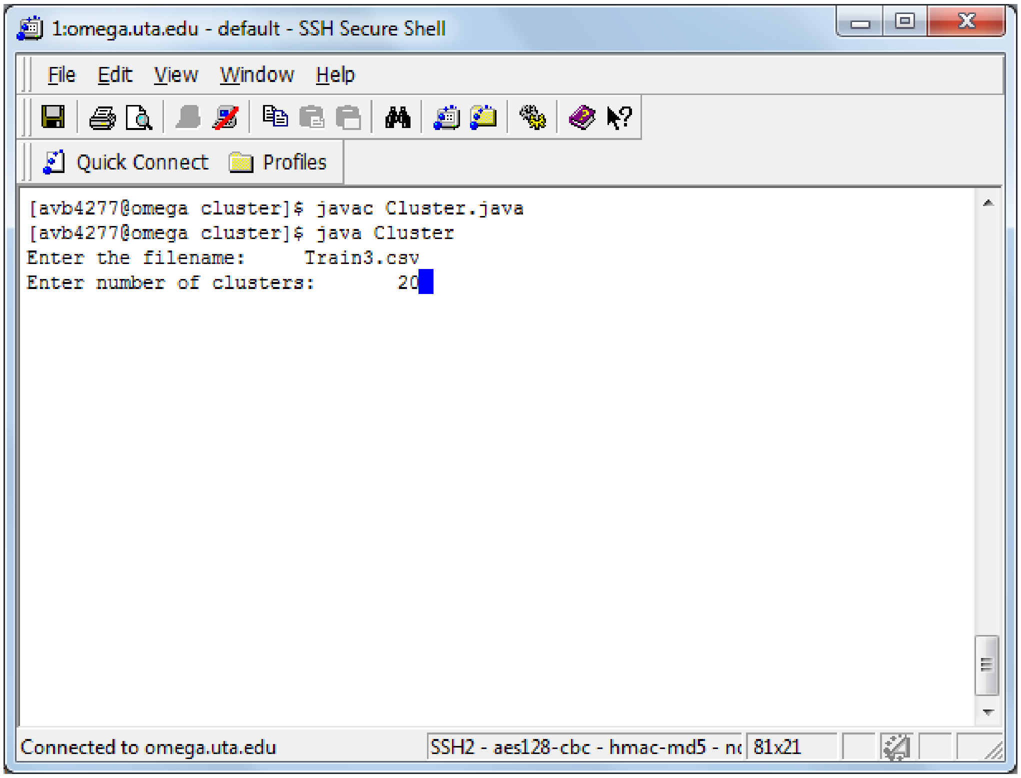


2:omega.uta.edu - default - SSH Secure File Transfer
File Edit View Operation Window Help
Quick Connect Profiles
Project\
Add
7/cluster
Add
Transfer
Queue

Local Name	Size	Type
LTFViewr		File fo
project		File fo
workspace		File fo
Cluster	915	File
Cluster	7,038,017	Applic
Cluster.layout	60	LAYOU
Makefile.win	859	WIN F
opencsv-2.3	14,146	Execut
Project_1	402,388	Adobe
Project_2	288,298	Adobe
result	0	Text D
tags_file	920,795	Text D
Train_Test	6,184	C++ S
Train_Test	7,165,765	Applic
vector_file	922,253	Text D
Vector_Sort	1,129	C++ S

Remote Name	Size	Type
a.out	25,143	OUT Fil
Cluster.java	4,888	JAVA Fi
tags_file.txt	375,333	Text Do
Train3.csv	72,570,158	Micros
Train_Test.cpp	6,185	C++ So
vector_file.txt	375,813	Text Do

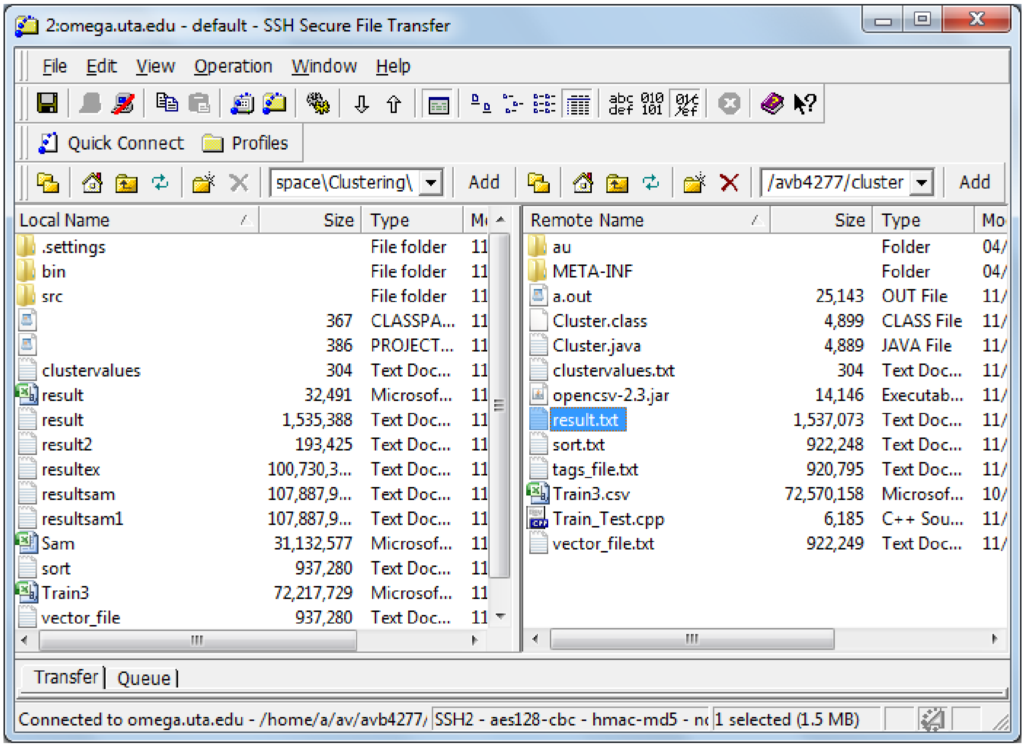
Connected to omega.uta.edu - /hom | SSH2 - aes128-cbc - hmac-md5 - nc | 6 items (73.4 MB)



1:omega.uta.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

```
[avb4277@omega cluster]$ javac Cluster.java
[avb4277@omega cluster]$ java Cluster
Enter the filename:     Train3.csv
Enter number of clusters:       20
```

Connected to omega.uta.edu | SSH2 - aes128-cbc - hmac-md5 - nc | 81x21



2:omega.uta.edu - default - SSH Secure File Transfer

space\Clustering\

/avb4277/cluster

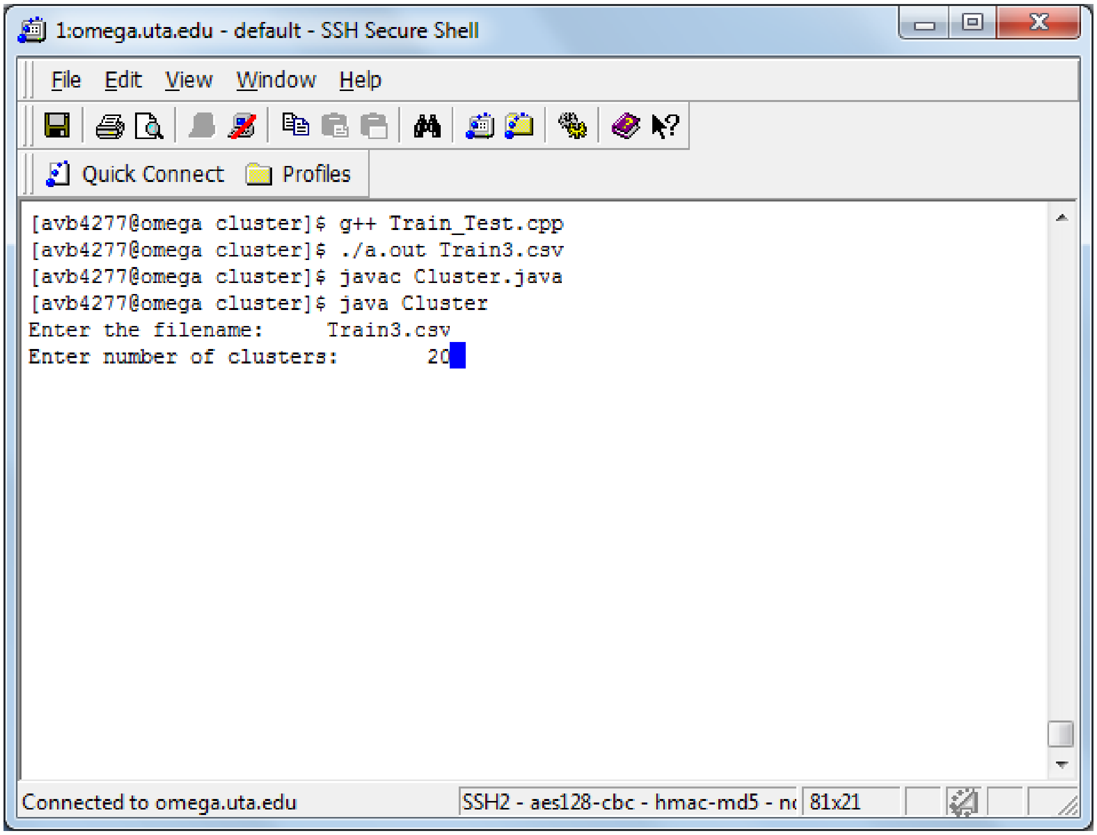
Local Name	Size	Type	M
.settings		File folder	11
bin		File folder	11
src		File folder	11
	367	CLASSPA...	11
	386	PROJECT...	11
clustervalues	304	Text Doc...	11
result	32,491	Microsof...	11
result	1,535,388	Text Doc...	11
result2	193,425	Text Doc...	11
resultex	100,730,3...	Text Doc...	11
resultsam	107,887,9...	Text Doc...	11
resultsam1	107,887,9...	Text Doc...	11
Sam	31,132,577	Microsof...	11
sort	937,280	Text Doc...	11
Train3	72,217,729	Microsof...	11
vector_file	937,280	Text Doc...	11

Remote Name	Size	Type	Mo
au		Folder	04/
META-INF		Folder	04/
a.out	25,143	OUT File	11/
Cluster.class	4,899	CLASS File	11/
Cluster.java	4,889	JAVA File	11/
clustervalues.txt	304	Text Doc...	11/
opencsv-2.3.jar	14,146	Executab...	11/
result.txt	1,537,073	Text Doc...	11/
sort.txt	922,248	Text Doc...	11/
tags_file.txt	920,795	Text Doc...	11/
Train3.csv	72,570,158	Microsof...	10/
Train_Test.cpp	6,185	C++ Sou...	11/
vector_file.txt	922,249	Text Doc...	11/

Connected to omega.uta.edu - /home/a/av/avb4277/ | SSH2 - aes128-cbc - hmac-md5 - n | 1 selected (1.5 MB)

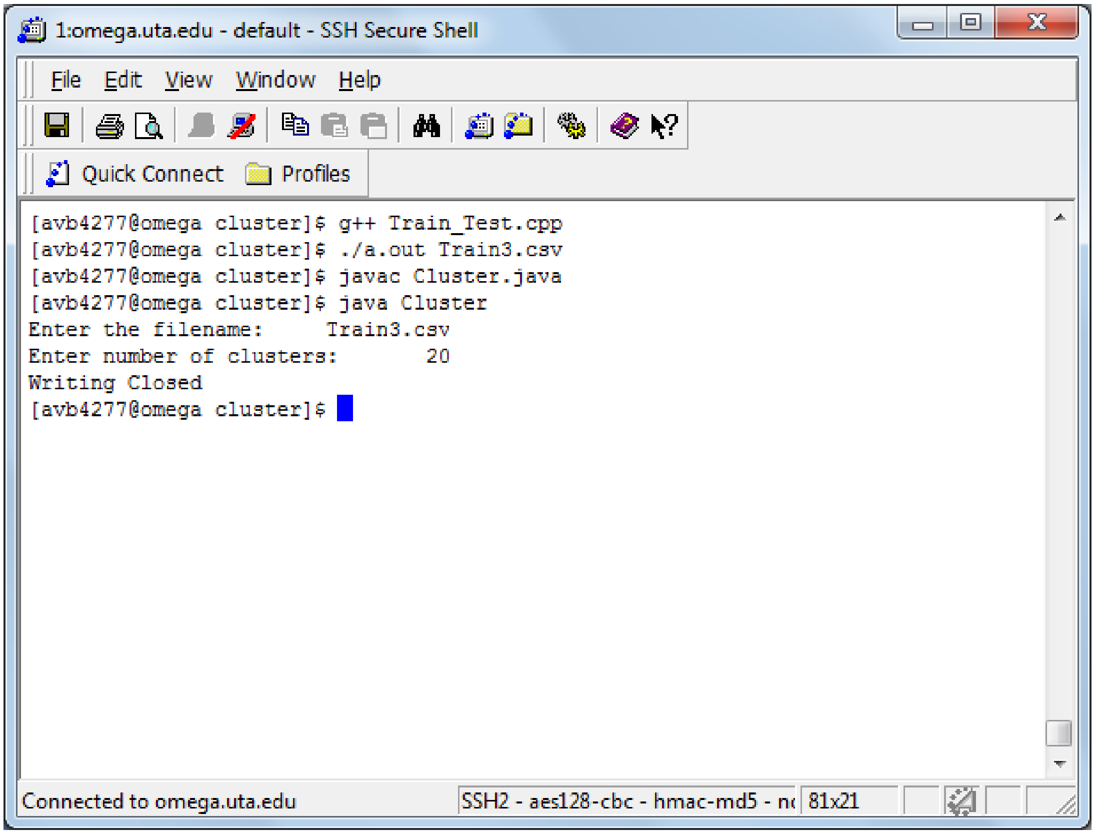
It takes around approx one hour to process the file.

And the result text file is generated.


1:omega.uta.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

```
[avb4277@omega cluster]$ g++ Train_Test.cpp
[avb4277@omega cluster]$ ./a.out Train3.csv
[avb4277@omega cluster]$ javac Cluster.java
[avb4277@omega cluster]$ java Cluster
Enter the filename:     Train3.csv
Enter number of clusters:       20
```

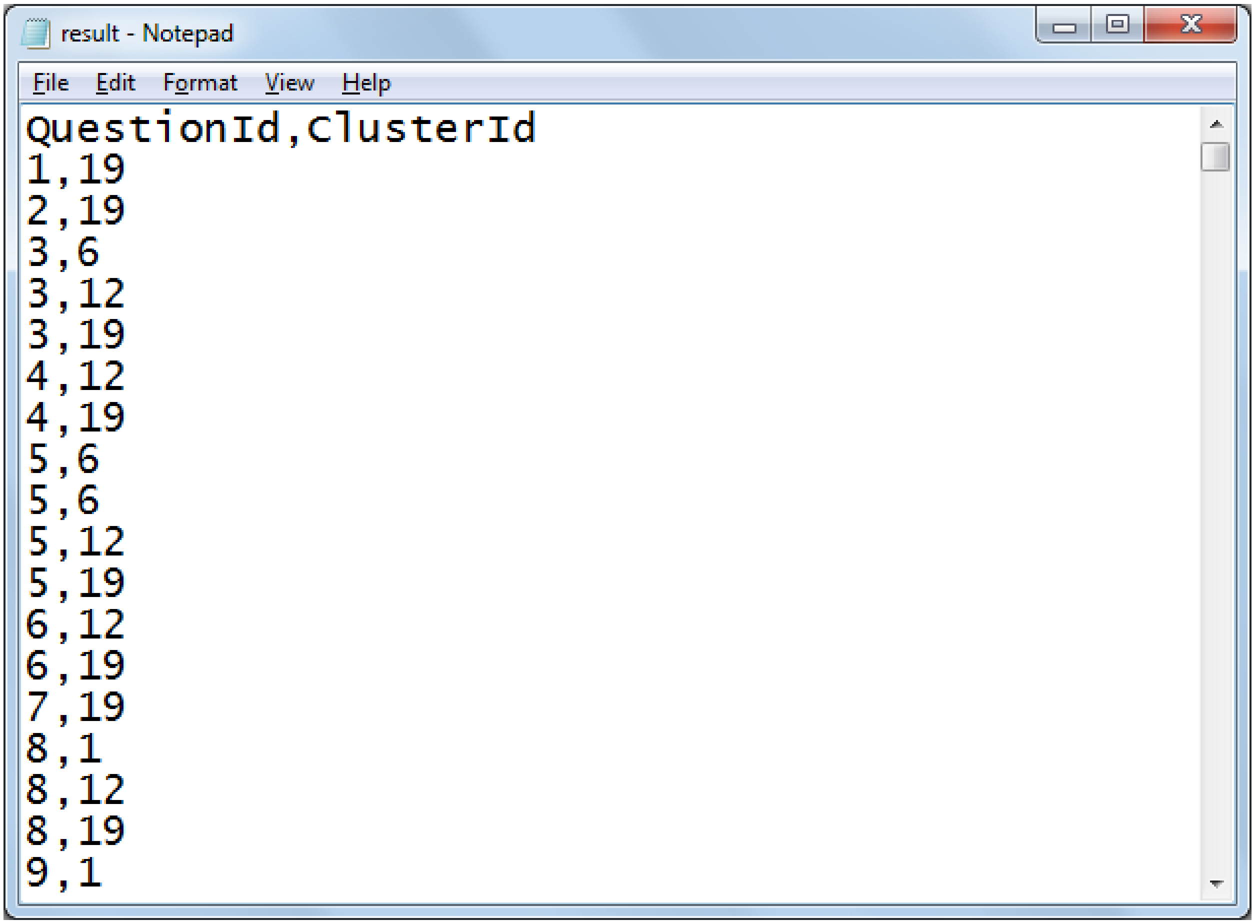
Connected to omega.uta.edu
SSH2 - aes128-cbc - hmac-md5 - nc
81x21


1:omega.uta.edu - default - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

```
[avb4277@omega cluster]$ g++ Train_Test.cpp
[avb4277@omega cluster]$ ./a.out Train3.csv
[avb4277@omega cluster]$ javac Cluster.java
[avb4277@omega cluster]$ java Cluster
Enter the filename:     Train3.csv
Enter number of clusters:       20
Writing Closed
[avb4277@omega cluster]$
```

Connected to omega.uta.edu
SSH2 - aes128-cbc - hmac-md5 - nc
81x21

Result file contents:



```
QuestionId,ClusterId
1,19
2,19
3,6
3,12
3,19
4,12
4,19
5,6
5,6
5,12
5,19
6,12
6,19
7,19
8,1
8,12
8,19
9,1
```