Project 1.

DATA MINING

TO predict the tags for questions from Stack Exchange sites, given only the question text and its title.
(6th November 2013)

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

Student ID: 1000994277

Net ID: avb4277

Design:

Its a simple probabilistic based design. The probability that a word present in the question text is a tag for that question is directly proportional to the average number of occurrences of the word in the questions present in the training set where it appears as a tag.

Implementation:

For each tag present in the question in the training set, we count the number of times the tag appear in the question. Thus the tag file contain the following format
"tag name","title-count---number of times tag occurs in the title of the question","body-count-----number of times tag occurs in the body of the question","tag-count------number of questions in the training set that contain the tag"
This information is used to define a criteria to be used to decide whether or not a given word is a tag word.

Training Algorithm:

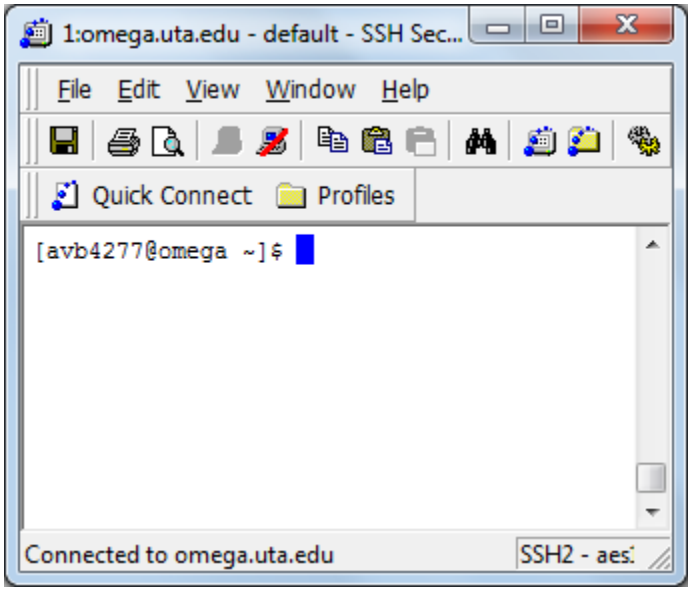
```
for (each question q in the training dataset) {
  for(each tag t present in the question){
       count = number of occurrences of t in q
       store count in the tag_file.txt

  }
}
```

Prediction algorithm:

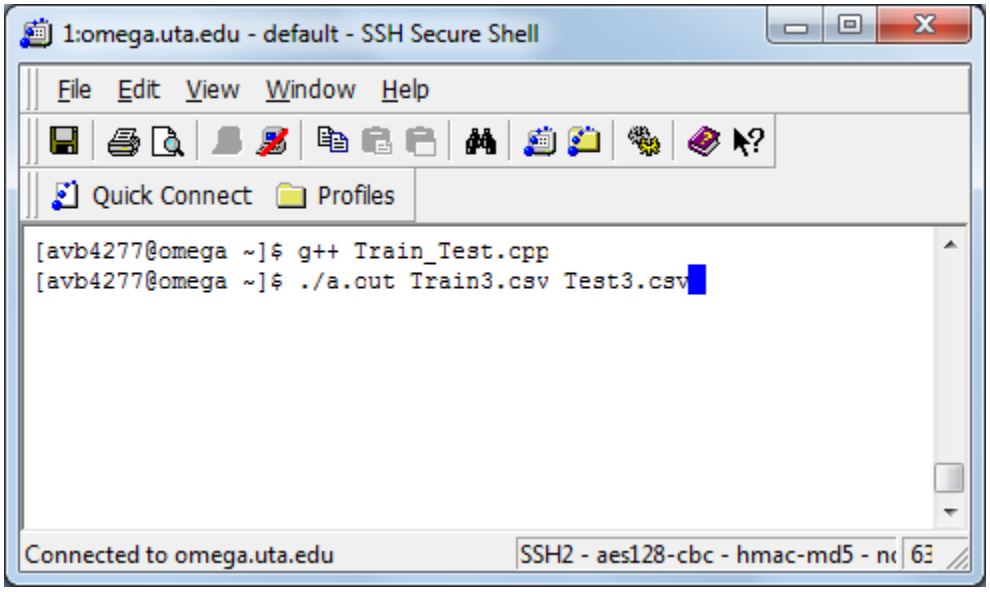
```
for(each question q in the test dataset) {
  for(each word w in q){
    count = number of occurrences of w in q
    if ( w is found in the tags_file) {
       criteria = (title-count + body-count) / tag_count     //refer the underlined words for their meaning
       if (criteria <= count) {
           w is included as a tag for q in results.txt
       }
    }
  }
}
```

Screen Shots for execution:

Compilation:

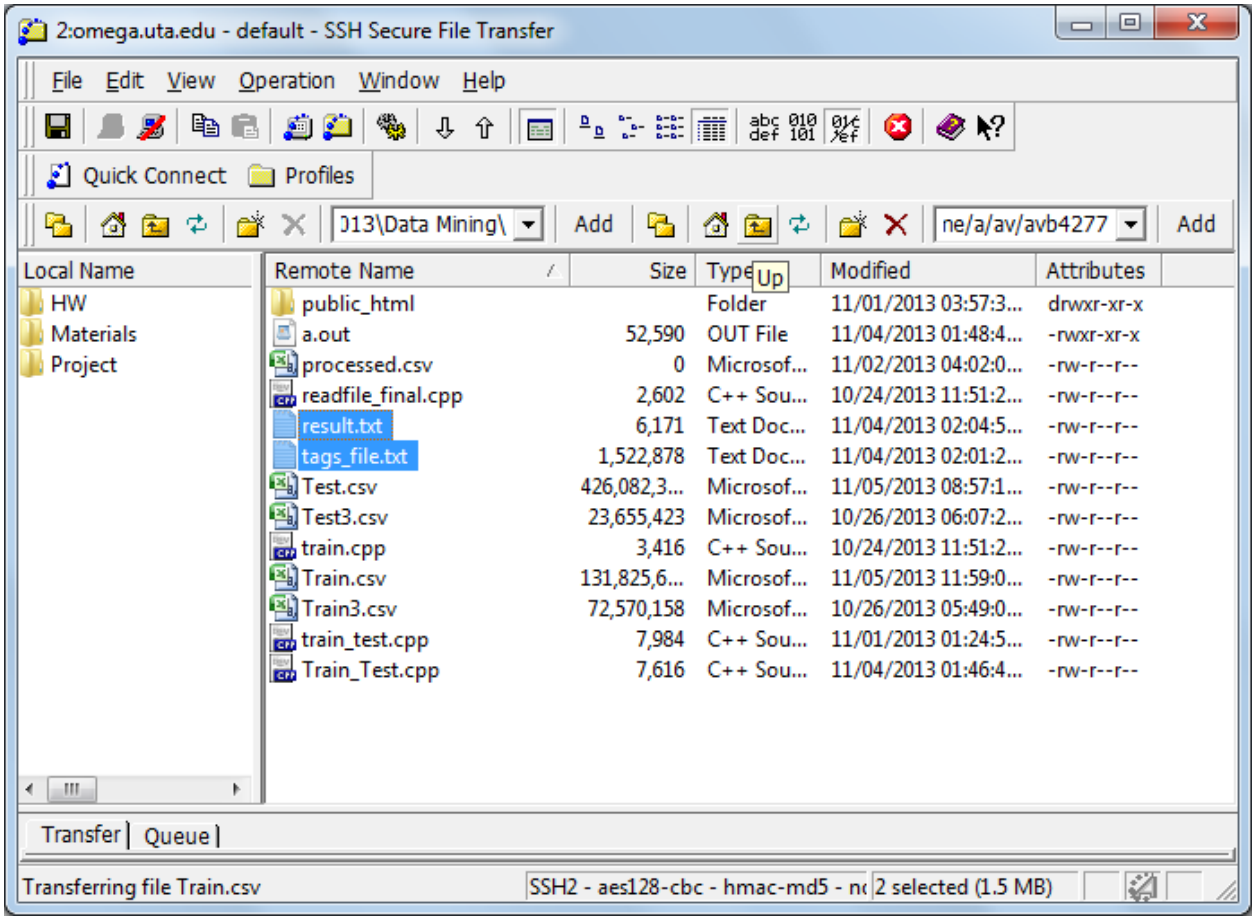


Execution:



It takes around 10 mins to process the file.

And the result text file is generated.



Result file contents: