**Group HW 3**

**Required: Must do using HBase**

**General Overview**:

**Individual Part 1**

Download “[movies.txt.gzLinks to an external site.](https://snap.stanford.edu/data/movies.txt.gz)” From the link [https://snap.stanford.edu/data/web-Movies.htmlLinks to an external site.](https://snap.stanford.edu/data/web-Movies.html).

Insert this data into the HBase:

Keep two column families:  
(i) User (contains columns ‘userId’ and ‘profileName)  
(ii) Product (contains remaining columns, ‘productId’, ‘helpfulness’,’score’,’time’,’summary’,’text’)

Create unique row keys for each entry (see which column, OR, concatenation of which two columns can be considered as row key)

After creating the table and inserting data, run 'describe' command to show that the table has been created perfectly and then run 'scan' & 'limit' command to show 5 rows.

You can decide the required datatype of each column.

**Individual Part 2**

Alter ‘Product’ column family to support 3 versions. Now, take a random row to put additional 2 different ‘text’. Then show the all 3 ‘text’ for this row.

Find the 'summary' that have the word ‘touching’.

Find the ‘text' that have any characters apart form alphanumerical characters (use regex).

**Individual Part 3**

Find how many reviews have 'score’ equal to 0, 3, and 5.

Find the average 'score' in the dataset.

Find the average 'helpfulness' in the dataset.

[Note:

\*\*\* It is a good idea to first take a small subset of the provided input file or create

a small dataset by yourself and then complete the programming. So that you can

test the correctness of your code. Then you can run your program for the whole

file and submit the output.

\*\*\* Each student in a group needs to do one separate part. Rest are collaborative

efforts. However, submit the code as one whole project].