IST 557 Data Mining Prerequisite Quiz

Instructions and requirements

1. The purpose of this quiz is to test whether you meet the basic prerequisite for this class.
2. You must complete the quiz **independently** without receiving any help from other people. You are allowed to use Internet to seek information in programming (Q1 and Q2).
3. The quiz counts 5% in the final grade.
4. All answers should be included in **a single Word file**.
5. Turn in your answer **by 11:59pm Wednesday August 23rd** to Canvas/Assignments/Prerequisite Quiz.
6. If you have any questions, please ask TA Hua Wei: hzw77 at ist.psu.edu

Q0: Background Information

1. Name
2. Major, Year, PhD or Master
3. Research area (if any)
4. Experience in data mining (if any)

Q1: Data Processing

Given a set of documents, count the frequency of each word in all documents. The words should all be turned into lower case. Output the top 20 words and their frequencies in descending order, separated by a space. For example, the output should be like

```
the 300
an 292
a 240
i 210
```

What to submit. (1) Copy and paste your code in the Word file. You should only use programming language Python. The code should be able to be compiled and executed. (2) Download the test data from `http://bit.ly/2ZnXU3K`. Run your code on this test data. Copy and paste the results in the Word file.

Q2: Data Statistics and Visualization

Given a dataset with 6,040 users rated 1,000,209 for 3,900 movies, write a program to get the statistics about the following fun facts:

1. Give the top 10 movies (name) that get the most number of ratings during August of 2000 (from 08/01/2000 00:00:00 to 09/01/2000 00:00:00). Give the top 10 movies that get the best rating during August of 2000 (from 08/01/2000 00:00:00 to 09/01/2000 00:00:00)?
2. Plot a histogram to show the total number of rating for movies in the genre of "Horror" during different months. The x-axis should be months, and y-axis should be ratings.

The dataset contains three files (detailed description can be found in the README of the dataset):

- Ratings.dat: User id, movie id, rating (1 to 5), time stamp
- Users.dat: gender, age, position, zip code
- Movies.dat: movie id, title, genre

What to submit. (1) Copy and paste your code in the Word file. You may use Python only. (2) Download the test data from `http://bit.ly/2MBd8ME`. Run your code on this test data. Copy and paste the results in the Word file.

Q3: Linear boundary

Given a line in the 2-D plane $y = 0.5x + 10$, where are the following points located? (Answer with on the line, in the half-plane above/below the line)

1. A (10, 15)
2. B (-100, -50)
3. C (0, 0)

Given another line in the 2-D plane $y = ax + b$, what is the necessary and sufficient condition that the point (x_0, y_0) lies in the half-plane above the line?

Q4: Algorithm complexity

```
for i = 1 .. n
    for j = i .. n
            print i * j
```

What is the big-O time complexity in terms of n?

```
function dummyRecursion( depth: Int )
        if depth == 1:
                return
        for i = 1 .. 3
                dummyRecursion( depth - 1 )

dummyRecursion(k)
```

What is the big-O time complexity in terms of k?