**BSCSE2002 : Tools in Data Science**

**May to August 2022**

**PROJECT 2**

**Released on 11-July-2022**

**Due Date for Submission 20-July-2022 at 23:59 hrs IST**

The Student has been provided with a dataset for this Project. They will have to use the tools and techniques learnt in this course and answer the below questions for their submission.

The Format of submission is as follows:

* A .csv file with Question numbers and answers written..
* Images for specific questions have to be uploaded as a response in the given [form](https://www.google.com/url?q=http://tiny.cc/TDS_Project2&sa=D&source=editors&ust=1657634901037018&usg=AOvVaw2DJZaJbpmRWre6XjGb4pqB)
* Please check the [Sample Submission file](https://www.google.com/url?q=http://tiny.cc/TDS-Project2-Submission&sa=D&source=editors&ust=1657634901037349&usg=AOvVaw2yWIgEpvU6j1L8s_irG0tP)

We will be evaluating both the .csv file and the published links. Any formatting issue in the .csv file will result in zero marks for the relevant question(s)**.**

**FILES:**

* **ratings.csv**contains ratings sorted by time.Ratings go from one to five. Both book IDs and user IDs are contiguous. For books, they are 1-10000, for users, 1-53424.
* **to\_read.csv**provides IDs of the books marked "to read" by each user, as *user\_id,book\_id* pairs, sorted by time.
* **books.csv**has metadata for each book (goodreads IDs, authors, title, average rating, etc.). The metadata has been extracted from goodreads XML files.
* **book\_tags.csv** contains tags/shelves/genres assigned by users to books. Tags in this file are represented by their IDs. Each book\_id  has multiple tag\_id.The field "count" denotes ‘user records’ (the number of users tagged the given tag\_id with the goodreads\_book\_id).

**NOTES :**

* Answering:
* Dataset to be referred will be mentioned in the square brackets for each question.eg [books.csv]
* String answers (“S”)   will be evaluated against entries from the column “title” in [books.csv]
* Data Cleaning:
* For Q1 Find the value from the original csv file [books.csv]. For remaining questions  remove null values in 'original\_title'.
* Ensure for  Q14 null values are removed from  'original\_title' to perform sentiment analysis.

**QUESTIONS :**

1. How many books do not have an original title [books.csv] ? N
2. How many unique books are present in the dataset ? Evaluate based on the 'book\_id' [books.csv]  N
3. How many unique users are present in the dataset [ratings.csv] ? N
4. Which book (title) has the maximum number of ratings based on ‘work\_ratings\_count’  [books.csv] ? S
5. Which tag\_id  is the most frequently used ie. mapped with the highest number of books [book\_tags.csv]  ? (In case of more than one tag, mention the tag id with the least numerical value) N
6. Whichbook (title) has the most number of counts of tags given by the user [book\_tags.csv,books.csv]  ? S
7. Which book (goodreads\_book\_id) is marked as to-read by most users [books.csv,toread.csv] ? N
8. Which is the least used tag, i.e. mapped with the lowest number of books [book\_tags.csv]   ? (In case of more than one tag, mention the tag id with the least numerical value)  N
9. Which book (title) has the minimum ‘average\_rating’  [books.csv] ? S
10. Which book (goodreads\_book\_id) has the least number of count of tags given by the user  [book\_tags.csv,books.csv] ? N
11. How many tags are there in the dataset [book\_tags.csv]  ? N
12. What is the average rating of all the books in the dataset based on ‘average\_rating’  [books.csv]  ? N
13. Find the number of books published in the year ‘2000’ based on the ‘original\_publication\_year’ [books.csv] ? N
14. Predict sentiment using Textblob. How many positive titles (title) are there [books.csv] ? (cut-off >0) N

**(3 marks)**

1. Plot a bar chart in Flourish with top 20 unique tags  in descending order of ‘user records’ (the number of users tagged the given tag\_id with the goodreads\_book\_id) [book\_tags.csv] and share the published link.

**(2 marks)**

1. Bucket the average\_rating of books into 6 buckets [0,1,2,3,4,5] with 0.5 decimal rounding (eg: average\_rating 3.5 to 4.4 will fall in bucket 4). Plot bar graph in Flourish to show total number of books in each rating bucket. [books.csv] and share the published link.

**(2 marks)**