Spectral Tech AI Hiring Challenge

This document is having **two problem statements** for our Data Science round. Both the problem statements are compulsory to solve. The dataset of respective problems can be found in the folder. You can use the **Jupyter Notebook** for creating a detailed report about your solutions and techniques.
Know more about the problem statement details below.

Problem Statements

1. Can you predict the missing grade?

- ❑ **Introduction**

  The CBSE Class 12 examination, is taken by Indian high school students at the end of K-12 school examination. The scores or grades in this examination form the basis of their entry to the College or University system, for an undergraduate program. At the K-12 level, students appear for examinations in five subjects. These five subjects generally include one language; three elective subjects oriented towards Science, Commerce, or Humanities; and any elective of their choice as a fifth subject.

- ❑ **The Challenge**

  This challenge is based on real school data of the CBSE Class 12 examination conducted in the year 2013. You are given the grades obtained by students with specific but popular combinations of subjects (and all these students had opted for Mathematics). Their grades in four subjects are known to you. However, their grade in Mathematics (i.e, the fifth subject) is hidden. The records provided to you are the grades obtained by students who had opted for the following combinations of subjects or courses and obtained a passing grade in each subject. The individual subjects in the data are:

English, Physics, Chemistry, Mathematics, Computer Science, Biology, Physical Education, Economics, Accountancy and Business Studies. \

The most dominant subject combinations, account for approximately 99% of the data are:

```
English, Physics, Chemistry, Mathematics, Computer Science
English, Physics, Chemistry, Mathematics, Physical Edu.
English, Physics, Chemistry, Mathematics, Economics
English, Physics, Chemistry, Mathematics, Biology
English, Economics, Accountancy, Mathematics, Business Std.
```

The grades of students in four subjects (other than Mathematics) are provided to you. Can you predict what grade they had obtained in Mathematics ?

To help you build a prediction engine, we will provide you with a training f file, containing the grade points obtained by students with the above subject combinations, in all five subjects.

Note about the Grading System

The student is first assessed on a scale of 100. (S)He needs a score of at least 33% to pass in the subject.

Among those who pass: Grade 1 to Grade 8 assign.

If more than 1 student share the same score and lie in the margin, they share the higher grade.

❑ Input Format

The first line will be an integer N. N lines follow each line being a valid JSON object. The following fields of raw data are given in json.

```
SerialNumber (Numeric): The identifier of the student
English (numeric): The grade (between 1 and 8)
Three more numeric fields from among the rest of the
```

```
subjects.
```

The input for each record has the grade for all subjects opted by a student, other than *Mathematics* which you have to predict as the answer.

❑ **Constraints**

$1 <= N <= 10^5$
The SerialNumber field will contain a unique numeric identifier such that
$1 <=$ SerialNumber $<= 5 * 10^5$
All other fields in the JSON fragment will represent the grades obtained on four subjects and will be populated by numeric values between 1 and 8 , both inclusive.

❑ **Output Format**

For each student record that is given as a JSON object, containing the grade obtained in four subjects, output the predicted grade in Mathematics (this will be numeral between 1 and 8, both inclusive) in a newline.

❑ **Training File and Sample Tests**

The *training file with sample test data* is available here. The three files in this package are:
Training.json
sample-test.in.json
sample-test.out.json
Training data as well as sample test cases have been provided in the above file for offline training and to help you build your prediction model. Feel free to include file data inside your cde.

❑ Sample Input

```
12345
{"SerialNumber":1, "English":1, "Physics":5,...}
jsonobject
.
.
.
.

jsonobject
```

❑ Sample Output

```
1
3
4
7
6
5
```

❑ Explanation

It is predicted that the first candidate obtained grade 1 in Mathematics, the second candidate achieved grade 3 in Mathematics, and so on.

❑ Scoring

For each of the N records in the input file, we will compute:
p = abs(Predicted Grade Point in Mathematics - Actual Grade Point in Mathematics)

Where ‘abs’ indicates the Absolute Value or Magnitude. If p=0 or 1 your answer for that particular student record will be considered correct. I.e, we allow a tolerance of one grade point away from the correct answer, to take into consideration the marginal errors which might occur during the testing or grading process.
Score = 100 * ((C-W)/N)
Where C = Number of Correct predictions, not more than one grade point away from the actual grade point assigned.
W = Numbers of wrong (incorrect) predictions and
N = Total number of records in the input.

2. Explore meaningful information from Twitter dataset

❑ Introduction

This dataset is collected from the tweets of an early stage startup named *Bounce.* Bounce is a company which provides mobility solutions in metropolitan cities like Bangalore. We have collected the tweets on Twitter that contain the word ‘*bounce*’ in their Tweet. The dataset contains 1200+ tweets from 01-01-19 to 21-10-2019. This dataset contains much information you need like; *date-time of tweet, username, mentions, tweets, hashtags.*

❑ The Challenge

Your task is to find out meaningful information from the dataset. This dataset can help you find out like:

- Which hashtags are used mostly quarterly?
- What words are most often used in the tweets monthly or quarterly?
- Find Tweets (if any) which share personal information like (email id, mobile number)?
- Plot a time-series of number of tweets monthly?
- Which period shows the highest growth in tweets.
- Find the most popular bi-grams or tri-grams used in the tweets?
- ...You can figure out more such insights about dataset.

Optional (Not compulsory, but can help you in extra scoring)

- Find out the sentiment of the tweets using Transfer Learning?
- Clustering the tweets based on similar topics.