**Question #1 of 3**

## ***NOTE: You can choose to answer ONE OUT OF THREE Kinds of Questions in the Data Science Test***

**Vehicle Prediction Test**

Given a corpus of metal-content readings and controlled testing environments, build a classification engine that can then be used to predict the type of vehicle travelling on top of that meta. Assume a single metal strip and assume all vehicles are travelling at around 15KM/hr. Assume that input files are provided one set each for a (a) bus (b) car (c) motorbike, with each category having distinct average metal readings.

For example, the corpus has the following readings for various time instances, where subsequent time instances are separated by 1 second. First column is time, second is Ampere rating, third column says what vehicle it is.

Input data sample for one metal reading:

1, 0.0, -

2,-0.05, -

3,0.1, -

4, 0.5, Motorbike

5, 0.4, Motorbike

6, 0.0, -

7, 1.5, Bus

8, 1.55, Bus

9, 1.29, Bus

10, 1.62, Bus

11, -0.09, -

12, -0.01, -

13, 5.5, Car

14, 5.66, Car

15, 4.58, Car

16, 0.3, -

17, 0.2, -

Question 3.1: Build a classification engine for being able to correctly classify given a set of readings the category they belong to. Note that, as per data above, different vehicles induce readings of differing lengths on the metal.

Question 3.2: Demonstrate how that classification engine can be used for prediction given unclassified data.

Question 3.3: How would you change the model building approach if the problem setting is such that there are multiple metal detectors each covering a width of about 1 feet? Note that while a motorcycle is of width < 1 feet, a bus or a car have a width of at least 3 feet and so would span to cover multiple metal detectors.

Question 3.4: How would your predictive model (built in Question 2) perform when there are 1000s of vehicles moving at 60km/hr every second on a national highway?

**Question #2 of 3**

## ***NOTE: You can choose to answer ONE OUT OF THREE Kinds of Questions in the Data Science Test***

**Keystroke Prediction**

Question 2.1: Keystroke prediction

Given a corpus of sentences, build a prediction engine that provides the top 3 choices of next character given the current typed sequence.

For example, if the corpus has 2 sentences "quick brown fox is lazy" and "queue is long", then (a) when the user types "qu" next, the engine should provide as prediction "i" followed by "e".

(b) when the user types "i", the engine should provide as prediction "s" only.

Question 2.2:

(i) How does your code perform when the input corpus is really large in size, say a billion sentences?

(ii) What are the parameters that determine how quickly you can predict the next character?

(iii) What would you change in your code if you were asked to predict the next character in real-time as the user types his word?

Question 2.3:

Extend your code above to do whole word prediction.

For example, if the corpus is the same as above, when the user next types "is", then the engine should predict "lazy" and "long" as the options.

**Question #3 of 3**

## ***NOTE: You can choose to answer ONE OUT OF THREE Kinds of Questions in the Data Science Test***

**What comes next?**

Given a sequence of website activity by anonymous users for various sessions, the task is to predict what page will be visited next given a sequence of pages.

For example, input data has 1000s of rows of the form:

Time, Sessionid, PageName

10:00, 1, Home

10:01, 1, Buy

10:02, 1, Search

10:04, 1, Checkout

10:00, 2, Home

10:01, 2, Search

10:02, 2, Search

10:01, 3, Search

10:02, 3, Search

10:03, 3, Search

10:03, 3, Buy

Given a prediction input sequence of “Home->Buy->?”, your task is to predict output as “Search” etc. Assume there are 1000s of rows, not just these 10 shown here as a sample.

Question 1.1: Outline an algorithm for the same.

Question 1.2: Implement the algorithm in a programming language of your choice. It can be SQL, Java, R, Python, or any Hadoop-based language like Spark, MR.

Question 1.3: Suggest ways of effectively visualising this output.

Question 1.4: What would you change in your approach if this was website activity from a popular website like Flipkart?