Data Analytics (CS40003)

Practice Set I
(Topic: Introduction)

I. Concept Questions

1. Which of the following are not necessarily data analytic techniques?

(a) Association rule mining
(b) Clustering
(c) Data compression
(d) Encryption
(e) Regression analysis
(f) Decision tree for classification modeling
(g) ANOVA testing
(h) Data cube modeling
(i) Forecasting
(j) Prediction
(k) Optimization
(l) Sorting and searching

2. Give some examples of data analytics applications with reference to the following data?

(a) Credit card transactions of a bank
(b) Calls in a call center
(c) Geo-satellite data
(d) Facebook messages

3. Discuss whether or not each of the following activities is a data analytic task.

(a) Dividing the customers of a company according to their gender.
(b) Dividing the customers of a company according to their profitability.
(c) Computing the total sales of a company.
(d) Sorting a student database based on student identification number.
(e) Predicting the outcomes of tossing a (fair) pair of dice.
(f) Predicting the future stock price of a company using historical records.
(g) Monitoring the hearts rate of patients for abnormalities.
(h) Monitoring seismic waves for earthquake activities.
(i) Extracting the frequencies of a sound wave.

4. Suppose that you are employed as a data analytic expert for an Internet search engine company. Describe how data analytics can help the company by giving specific examples of techniques, such as clustering, classification, association rule mining and anomaly detection, etc.

5. For each of the following data sets, explain whether or not data privacy is an important issue.

   (a) Census data collected from 1950- 2010.
   (b) IP address and visit times of Web users who visit your websites.
   (c) Images from earth-orbiting satellites.
   (d) Names and addresses of the people from the telephone book.
   (e) Names and addresses of the people from the Web.
   (f) Messages in Social media sites such as Facebook, Twitter, etc.

6. Mention at least six parameters to characterize the complexity of data.

7. What are the issues to process Big data? (Mention at least FIVE issues/challenges).

8. Following are the few real-life applications. In these applications, find the data, which are of primary concern in each.

   (a) Weather forecasting
   (b) Fraud detection(say in banking applications)
   (c) Hand written character recognition
   (d) Categorization of persons from a population
   (e) Anomaly detection

9. Give the 3V values of streaming data when it occurs during streaming of a TV program, for example, broadcasting of a T20 cricket event for a day.

II Objective Questions

1. As on today, which of the following is the largest size of data?
   (a) Wikipedia
   (b) YouTube
   (c) Google gmails
   (d) Facebook

2. The largest size of data is
   (a) Petabyte (PB)
   (b) Exabyte (EB)
   (c) Zyttabyte (ZB)
   (d) Yottabyte (YB)

3. Present size of the digital universe is
- (a) Giabyte (GB)
- (b) Terabyte (TB)
- (c) Petabyte (PB)
- (d) Exabyte (EB)

4. One quintillion byte is equivalent to
- (a) 10^{10} byte
- (b) 10^{18} byte
- (c) 10^{30} byte
- (d) 10^{100} byte

5. Which is/are not the source of data in data analytics
- (a) Scientific instruments
- (b) Social media
- (c) Mobile devices
- (d) Sensor networks

6. Which are not related to the 3V characteristics of Big data?
- (a) Speed
- (b) Complexity
- (c) Size
- (d) Computability

7. Which can better describe a big data?
- (a) Structured data
- (b) Complex statistical analysis
- (c) Business intelligence
- (d) Predictive analytics

8. Elastic is a tool for
- (a) Strong big data
- (b) Processing data with scalable architecture
- (c) A distributed file system
- (d) Cloud security

9. MapReduce is meant for
- (a) Data visualizations
- (b) Massive parallel programming
- (c) Query reporting
- (d) Data storage in Cloud