

Roll No

CS-7005 (2) (CBGS)

B.E. VII Semester

Examination, November 2018

Choice Based Grading System (CBGS)

Data Science and Big Data

Time : Three Hours

Maximum Marks : 70

Note: i) Attempt any five questions.

ii) All questions carry equal marks.

iii) All parts of each question are to be attempted at one place.

1. a) What is Data Wrangling process? Explain with help of a suitable example. <https://www.rgpvonline.com>
b) What do you understand by exploratory analysis? Explain with help of a suitable example.
c) Explain Data Transformation and Cleaning process with help of a suitable example.
2. a) What is Feature Extraction? Discuss advantages of the Feature Extraction process in data analytics.
b) What do you understand by Data Visualization? Discuss some Python's data visualization tools such as Box plots, Pie Charts and Bar Charts in brief.
<https://www.rgpvonline.com>
3. a) What is supervised learning? Explain the working of Logistic regression algorithm with help of a suitable example.
b) Discuss overfitting and underfitting with help a suitable example.

29

4. a) What do you mean by Clustering? Explain working of any clustering algorithm.
b) What do you understand by dimensionality reduction? Explain working of the PCA.
c) Discuss key idea of the Support Vector Machines (SVMs).
5. a) Explain the Vector model based information retrieval scheme. <https://www.rgpvonline.com>
b) What do you mean by Tokenization? Discuss Term Frequency (TF) and Inverse Document Frequency (IDF) with its definition and formula.
c) Assume you have 10 million documents and the word cat appears in one thousand of these documents. Consider a document containing 100 words wherein the word cat appears 3 times. Compute Term Frequency (TF) and Inverse Document Frequency (IDF) for given documents.
6. Write short notes on any two of the following:
a) Multiple hypothesis testing
b) Crawling and Indexes
c) Distributed Architecture of Search Engine
7. a) What is Big Data? Discuss 5V characteristics of Big Data with suitable examples.
b) What is Link analysis? Explain any Page Rank algorithm with a suitable example. <https://www.rgpvonline.com>
8. a) Explain the process of Data storage in Hadoop Distributed File System (HDFS) with help of a suitable example.
b) What is Map Reduce programming model? Explain its Execution process and working with help a suitable flow diagram and example.

30