COLLEGE OF ENGINEERING, PUNE

(An Autonomous Institute of Government of Maharashtra.)

END Semester Examination

Programme: B.Tech	Semester: VI
Course Code: CT- 17025	Course Name: Data Science
Branch: Computer Engineering	Academic Year: 2017-18
Duration: 3hrs	Max Marks: 60
Student PRN No.	□□□□□□□□□

Instructions:

1. Figures to the right indicate the full marks.
2. Mobile phones and programmable calculators are strictly prohibited.
3. Writing anything on question paper is not allowed.
4. Exchange/Sharing of stationery, calculator etc. not allowed.
5. Write your PRN Number on Question Paper.

			Marks	CO	PO
Q 1	a)	What are dataspaces ? Discuss it with respect to the following points : data, processing, storage, agility, security, users	05	1	a, d, g
	b)	What is data pre-processing? Why is data pre-processing important? Explain at least 3 tasks of data pre-processing ?	05	1	a, d, g
Q 2	a)	The download time of a resource web page is normally distributed with a mean of 6.5 seconds and a standard deviation of 2.3 seconds. a) What proportion of page downloads take less than 5 seconds? b) What is the probability that the download time will be between 4 and 10 seconds? c) How many seconds will it take for 35% of the downloads to be completed?	06	2	d , g
	b)	The arrival rate of cars at a gas station is $\lambda = 40$ customers per hour. (That is, the inter arrival times are exponentially distributed with rate 40 per hour.) i) What is the probability of having no arrivals in a 5- minute interval? ii) What are the mean and variance of the number, N, of arrivals in 5 minutes? iii) What is the probability for having 3 arrivals in a 5- minute interval?	04	2	d, g
Q 3	a)	Discuss the Page-Rank algorithm for ranking pages, used by Google?	04	3	d,

Q 3 b) Suppose we are building a classifier that says whether a text document is about sports or not. Our training set has 5 sentences: 06 5 d, g

DOC-ID	Text	Category
D1	A great game	Sports
D2	The election was over	Not sports
D3	Very clean match	Sports
D4	A clean but forgettable game	Sports
D5	It was a close election	Not sports

Classify using Naive Bayes algorithm to which category does the test document belongs to? [Hint- Remove stopwords {A, The, was, but, it }. Apply laplace Smoothing, i.e. $\cdot\cdot\frac{\cdots+1}{\cdots+V}$ where V is the distinct vocabulary of the collection]

DOC-ID	Text
test	*A very close game*

Q 4 a) What is Association Mining? Explain the Apriori principle? Define the following : 05 5 d,

i) Frequent Itemset

ii) Support

iii) Confidence