SPPU-TE-COMP-CONTENT – KSKA Git

Total No. of Questions : 8] **SEAT No. :**

PA-1449 **[Total No. of Pages : 3**

[5926]-65

T.E. (Computer Engg.)

DATA SCIENCE AND BIG DATA ANALYTICS

(2019 Pattern) (Semester-II) (310251)

Time : 2½ Hours] ***[Max. Marks : 70***

Instructions to the candidates:

1) ***Answer Q1 or Q2, Q3. or Q4, Q5 or Q6, and Q7 or Q8.***
2) ***Neat diagram must be drawn wherever necessary.***
3) ***Figures to the right indicate full makrs,***
4) ***Use of logarithmic tables slide rule, mollier charts, electronic pocket calculator and steam tables is allowed.***
5) ***Assume suitable data if necessary.***

Q1) a) Draw the diagram of data analytics life cycle in big data and briefly explain its phases. **[8]**

b) Explain in detail how the model building phase is built by team in data analytics life cycle? **[9]**

OR

Q2) a) List and explain the steps in data preparation phase of data analytics life cycle. **[8]**

b) Write short note on the following: **[9]**

i) ETL

ii) Common tools for the model building.

iii) Model selection for data analytics.

Q3) a) What are the types of analytics in big data? Explain in brief. **[9]**

b) Calculate the support and confidence value for all the possible item sets. **[9]**

Transaction ID	Items bought
1	Onion, Potato, Cold drink
2	Onion, Burger, Cold drink
3	Eggs, Onion, Cold drink
4	Potato, Milk, Eggs.
5	Potato, Burger, cold drink, Milk eggs.

OR

P.T.O.

CTCGP013091 49.248.216.238 14/01/2023 09:40:46 static-238

Q4) a) Explain the use of logistic function in logistic regression in detail. **[9]**

b) Write short note on the following:

i) Removing duplicates from data set.

ii) Handling missing data

iii) Data transformation. **[9]**

Q5) a) Suppose that the given data the taste is to cluster points (With (x.y) representing location) into three cluster, where the points are.

A1(2,10), A2(2,5), A3(8,4), B1 (5,8)

B2(7,5) B3(6,4), C1(1,2), C2(4,9)

The distance function is Euclidean distance suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. use the k-means algorithm to show only the three cluster centers after the first round of execution with steps. **[9]**

b) Explain the following text analysis steps with suitable example. **[8]**

i) Part of speech (POS) tagging

ii) Lemmatization

iii) Stemming

OR

Q6) a) Given the confusion matrix, calculate accuracy. precision, Recall, Error rate with description on heart attact risk. **[8]**

		Predicted classes	
	Classes	Heart-Attack Risk-yes	Heart Attack Risk-No
Actual Classes	Heart Attack Risk-yes	80	220
	Heart Attack Risk-No	150	9,500

b) Explain the TF/IDF (term frequency-inverse document frequency) terms in text analysis with suitable example. **[9]**

CEGP013091 49.248.216.238 14/01/2023 09:40:46 static-238

Q7) a) List the data visualization tools and discuss any four applications of data visualization along with the use of the suitable plot. **[9]**

b) List the challenges of data visualization explain the types of visualization with example. **[9]**

OR

Q8) a) Explain in detail the Hadoop Ecosystem with suitable diagram **[9]**

b) Write a short note on the following **[9]**

i) Map reduce.

ii) Pig

iii) Hive

● ● ●

CEGP013091 49.248.216.238 14/01/2023 09:40:46 static-238