30.4.19

Roll No. | | | | | | | | | | |

B.E. / B.Tech (FT) END SEMESTER EXAMINATIONS – APRIL / MAY 2019

COMPUTER SCIENCE AND ENGINEERING
Sixth Semester

CS 7005 Big Data Analytics

(Regulation 2015)

Time: 3 Hours | Answer ALL Questions | Max. Marks 100

PART-A (10 x 2 = 20 Marks)

1. What is big data analytics? What is the need for big data analytics in the banking sector?
2. List any two advantages of YARN.
3. In *k*-means clustering algorithm, how can the number of clusters be chosen?
4. What is entropy? Why is entropy used in the construction of decision trees?
5. Suppose there are 50 items, numbered 1 to 50, and also 50 baskets, also numbered 1 to 50. Item *i* is in basket *b* if and only if *i* divides *b* with no remainder. What is the confidence for the association rule $\{5, 8\} \rightarrow 2$.
6. Suppose you have the following set of documents:
   Doc1: cat cat dog
   Doc2: cat dog cat cat
   Doc3: cat cat dog dog dog
   Doc4: mouse
   Doc5: cat mouse dog

   Calculate the TF*IDF score for the term 'cat' in Doc1.
7. List a few examples of stream sources.
8. Compute the surprise number for the stream 3, 1, 4, 1, 3, 2, 4, 2, 1, 3.
9. What is meant by sharding?
10. Why were schema-less models developed?

Part – B (5 x 13 = 65 marks)

11. a) (i) Explain the different components of the Hadoop Distributed File System (HDFS). Explain the functions of HDFS in regard to monitoring, maintaining integrity and replication. (10)
    (ii) The default block size of HDFS is 64 MB/128 MB while the default block size of UNIX/Linux is 4KB/8KB. Why? What implication, do you think, this will have in the design of the NameNode? (3)

(OR)

b) (i) Explain how MapReduce works. (10)

(ii) Suppose there is an online music website where users listen to various tracks. Suppose data is collected and stored in log files in the format Userid | Trackid | Shared-or-not. Design a mapper and a reducer that will get the number of times a track is shared with others. (3)

12. a) Consider the following dataset consisting of the speed and distance features of five drivers:

Sl. No.	DriverID	Speed	Distance
1.	35314	24	71
2.	44523	30	50
3.	12345	25	42
4.	17564	35	89
5.	27389	18	55

Use k-means algorithm to classify the drivers into two groups.

(OR)

b) Consider the following data set:

Name of animal	Give birth	Can fly	Live in water	Have legs	Class
Human	Yes	No	No	Yes	Mammal
Python	No	No	No	No	Non-mammal
Salmon	No	No	Yes	No	Non-mammal
Whale	Yes	No	Yes	No	Mammal
Frog	No	No	Sometimes	Yes	Non-mammal
Komodo	No	No	No	Yes	Non-mammal
Bat	Yes	Yes	No	Yes	Mammal
Pigeon	No	Yes	No	Yes	Non-mammal
Cat	Yes	No	No	Yes	Mammal
Shark	Yes	No	Yes	No	Non-mammal
Turtle	No	No	Sometimes	Yes	Non-mammal
Penguin	No	No	Sometimes	Yes	Non-mammal
Porcupine	Yes	No	No	Yes	Mammal
Eel	No	No	Yes	No	Non-mammal
Salamander	No	No	Sometimes	Yes	Non-mammal
Gila-monster	No	No	No	Yes	Non-mammal
Platypus	No	No	No	Yes	Mammal
Penguin	No	No	Sometimes	Yes	Non-mammal

Use Naïve-Bayes' classifier to classify an animal that gives birth, cannot fly, lives in water and doesn't have legs as mammal or non-mammal.

13. a) State the Apriori principle. Apply Apriori algorithm to find the frequent itemsets with support threshold = 2 for the following data:

Transaction ID	Items
T1	{M, O, N, K, E, Y}
T2	{D, O, N, K, E, Y}
T3	{C, O, O, K, I, E}
T4	{C, A, K, E}
T5	{D, U, C, K}
T6	{R, A, C, K}

(OR)

b)

The following table gives the number of times a user has listened to a particular artist. Using this information, construct the utility matrix with the users as the rows and the artists as the columns and calculate the ratings for the utility matrix. If a user has not listened to an artist, then a rating of 0 is given. If a user has listened to an artist for 1 – 5,000 times, a rating of 1 is given; if a user has listened to an artist for 5,001 – 10,000 times a rating of 2 is given, and so on.

Recommend artists to a user who has listened to The Beatles for 3000 times and Pink Floyd for 22458 times.

Use cosine distance to find the similarity.

User	Artist	Plays
User 1	The Beatles	39000
User 2	Muse	45000
User 2	Coldplay	800
User 2	Radiohead	900
User 3	Muse	48000
User 3	Coldplay	6000
User 3	Radiohead	480
User 4	PinkFloyd	32000
User 4	The Beatles	14900
User 5	The Beatles	31520
User 5	PinkFloyd	5890
User 6	Muse	42888
User 7	The Beatles	34897
User 7	Radiohead	2450
User 7	Pink Floyd	2399
User 8	Coldplay	31776
User 8	Radiohead	18543
User 8	Muse	687
User 9	Coldplay	12999
User 9	The Beatles	4
User 10	Muse	44976
User 10	Coldplay	150

14. a) What is graph analytics? What characteristics of business problems make them suitable for using graph analytics solutions? Discuss any four types of graph analytics algorithmic approaches.

(OR)

b) Explain the problem of counting distinct elements in a stream. Explain the Flajolet-Martin algorithm to estimate the number of distinct elements, with appropriate examples.

15. a) Discuss in detail any four NoSQL datastores.

(OR)

b) What is the need for HBase? Explain the HBase architecture.

PART – C (1 x 15 = 15 marks)

16. (i) While using a bloom filter, suppose you have n bits of memory available and the set of key values S has m members. Instead of using k hash functions, divide the n bits into k arrays and hash once to each array. Find the probability of a false positive as a function of n, m and k. (5)

(ii) Construct a decision tree for the dataset given in Question 12.b. Use the decision tree to classify an animal that does not give birth, cannot fly, does not have legs and sometimes lives in water. (10)