PES UNIVERSITY

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE18CS322

NOVEMBER 2020: IN SEMESTERASSESSMENT B Tech V SEMESTER TEST – 2

UE18CS322(4 credit subject) - BIG DATA

Time: 80 min	Answer All Questions Please answer all questions in the order asked	Max Marks: 40

1.	a)	Classify the following streaming queries into ad-hoc and standing (i) given a stream of stock sales, compute the total sales for the day (ii) identify the total number of shares sold for company XYZ inc before a change in the CEO	2
	b)	What are stateful operations in Streaming Spark? What types of stateful operations are supported. Give an example to illustrate.	4
	c)	Consider we are processing messages using Kafka in a college forum and have two topics – one for *academic* and one for *non-academic* (like timings, fees etc.). We want to use 3 servers for this with each topic having 6 partitions and no replicas per partition. The users are in a single consumer group with 3 instances. Draw an Kafka deployment diagram illustrating the server, partitions, partition assignments and consumer instances for the above case.	4
2.	a)	Given the following sequence of inputs received, compute the estimate of number of unique elements seen using the Flajolet Martin algorithm – 23, 129, 156, 192, 128, 32, 48, 54	2
	b)	Consider an ad server that is receiving requests from different web-pages (URLs) accessed by different users from different countries. The ad-server wants to identify the user browse pattern by country using 20% sample. (i) what are the keys that are used for hashing (ii) illustrate how hashing can be used to generate the sample and how does hashing help in speeding up generation of the sample?	4
	c)	You are accessing a bloom filter with two hash functions for a given input key and the values from the bloom filter turn out to be 0 and 1 for the two hash functions. Can you say for sure if the key has been seen before by the bloom filter (justify)? Had the bloom filter returned 1 and 1 instead – what would your conclusion be and why?	4
3.		Consider a set of images taken from drones of houses in a city and a machine learning application that identifies the amount of green cover in each house of the city into *high, medium, and low*. While the initial dataset is labeled, it needs to be used in production using a decision tree in MLLib to do the classification.	
	a)	If the features extracted from this include the r,g,b values of the house image, what will the dataframe look like after the feature extraction is done.	2
	b)	Show the transformers, estimators and evaluators that are used for setting up the training pipeline?	4
	c)	Suppose you would like to change from using a decision tree model to a svm model, what changes are required to be done to both the training and testing pipelines?	4
4.	a)	For the following use cases given below, with justifications, choose either supervised or unsupervised learning for solving the problem (a) classify types of vehicles as motorbikes, sedans and SUVs. (b) classify different sounds made on the roads of a city	2
	b)	Given the following sequence of numbers – 232, 102, 41, 983, 101, 430, 245, 301, 500, 822 spread on two servers (odd numbers on server 1 and even on server 2). Use Map Reduce k-means to cluster these into 3 clusters. Assume initial centroids are 100, 300, 900. Do one iteration of map-reduce and show what the keys will be in the map/reduce stage and the computation at both stages	4
	c)	For the k-means problem given above, if we introduce a combiner, how will the key and values change at the mapper stage. What changes will we have to introduce a combiner?	4