Data Science || CS321 || End Semester Exam 2021

u19cs009@coed.svnit.ac.in Switch account

Draft saved

Your email will be recorded when you submit this form

* Required

Paper

*

The "Twitter datastream" contains tuples of the form:
(messageID, message, userID of posting user, in reply to messageID, time of posting, language of message).

You can assume that messageID and userID are unique, i.e. every message has a unique identifier and every user has a unique identifier. If the message is not posted in reply to any other message, we have in reply to messageID=null.

Examples of tuples in that stream are:
(124324234324, "@Nelly: I had breakfast just now!", 33523232, 122192225674, "28/11/2021", "English").

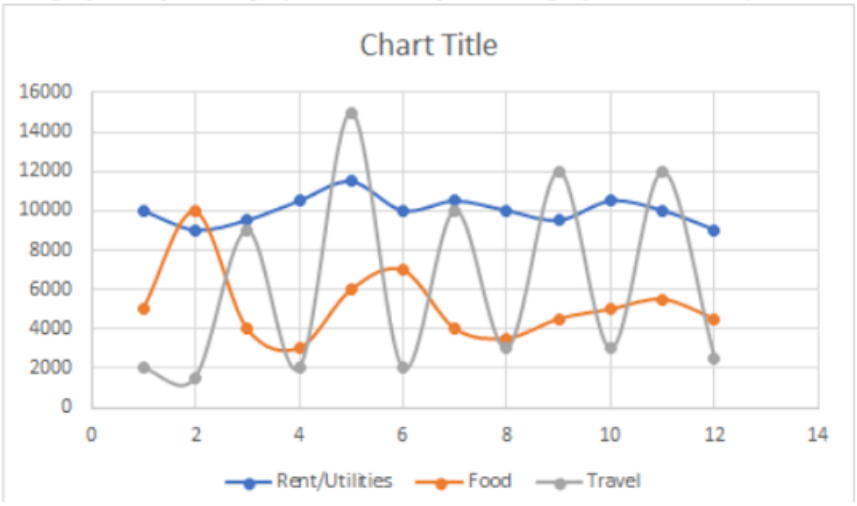
(435345332432, "Sitting in Paris, drinking a coffee", null, 122198435674, "29/11/2021", "English").

We want to answer queries by sampling roughly 1/10th of the data. What is a good sampling strategy to answer the following query: How many replies does a tweet have on average?

- ( ) Sample userIDs and include all messages by a user
- ( ) Sample (userID,in_reply_to_messageID)
- (•) Sample in_reply_to_messageIDs
- ( ) Generate a random number r between 0 and 9 and sample the tuple if r==0

*

This is a scatter plot with smooth lines. Each point on the line shows expense in that month under that category. Analyse the graph and identify the category where the expenditure is most varying.



- ○ Entertainment
- ○ Shopping
- ○ Food
- ◉ Travel

*

The FM-sketch algorithm uses the number of zeros the binary hash value ends in to make an estimation. Which of the following statements is true about the hash tail?

- ○ Only bit patterns with more 0's than 1's are equally suitable to be used as hash tails.
- ○ Only the bit pattern 0000000..00 (list of 0s) is a suitable hash tail.
- ◉ Any specific bit pattern is equally suitable to be used as hash tail.
- ○ Only the bit patterns 0000000..00 (list of 0s) or 111111..11 (list of 1s) are suitable hash tails.

*

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters.
A1=(2,10),A2=(2,5),A3=(8,4),A4=(5,8),A5=(7,5),A6=(6, 4),
A7=(1,2),A8=(4,9)
Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. After iteration 1 what's the center value of all clusters?

- (●) C1 = (2,10), C2 = (6,6), C3 = (1.5, 3.5)
- ( ) C1 = (3,9.5), C2 = (5.5,5.25), C3 = (1.5, 3)
- ( ) C1 = (2.5,10), C2 = (6,6), C3 = (1.5, 3.5)
- ( ) C1 = (3,9.5), C2 = (6,5.25), C3 = (1.5, 3)

*

D1 : I am Sam.
D2 : Sam I am.
(k = 4)-character shingles of D1∪D2 are

- ( ) Not possible to generate character shingles
- (●) {[iams], [amsa], [msam], [sams], [sami], [amia], [miam]}
- ( ) {[i a s s], [a s s i], [s s i a]}
- ( ) {[i am sam sam], [am sam sam i], [sam sam i am]}

*

Point out the wrong statement.

- [ ] Applications typically implement the mapper and reducer interfaces to provide the map and reduce methods.
- [ ] The MapReduce framework operates exclusively on <key,value> pairs.
- [ ] A MapReduce Job usually splits the input data sets into independent chunks which are processed by the map tasks in a completely parallel manner.
- [x] DataNode is aware of the files to which the blocks stored on it belong to.

*

Imagine that a social-networking site has a relation

Friends(User, Friend)

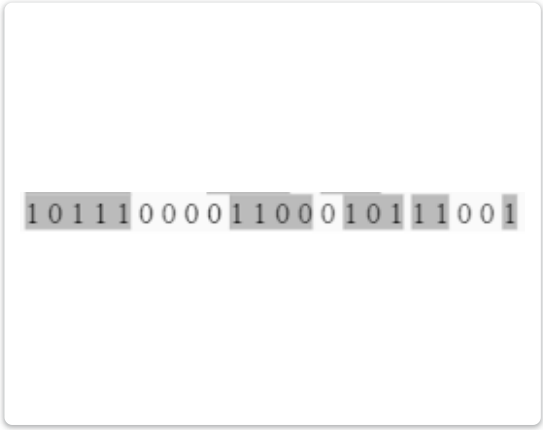
This relation has tuples that are pairs (a, b) such that b is a friend of a. The site might want to develop statistics about the number of friends members have.

How would the site evaluate the same? What is the first step to develop statistics?

- (●) i) using the map and reduce function ii) Their first step would be to compute a count of the number of friends of each user
- ( ) i) using the map and reduce function ii) for each group the count of the number of friends of that user is made
- ( ) i) using the map function ii) Their first step would be to compute a count of the number of friends of each user
- ( ) i) using the map function ii) for each group the count of the number of friends of that user is made

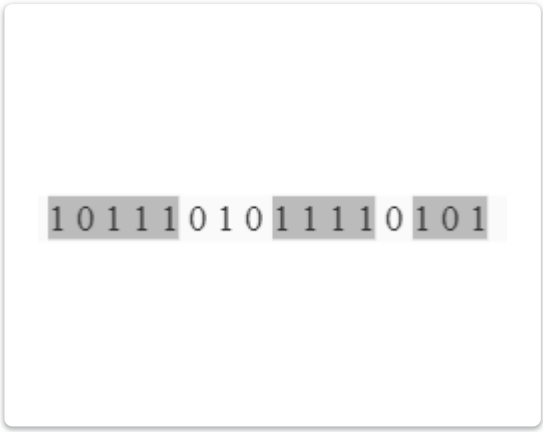
*

Which of the following streaming windows show valid bucket representations according to the DGIM rules?



Option 2

Option 3



Option 1

Option 4

*

Which of the following are classification tasks?

- [ ] Predict the number of copies of a book that will be sold this month
- [ ] Predict the price of a house based on floor area, number of rooms etc.
- [x] Find the gender of a person by analyzing his writing style
- [x] Predict whether there will be abnormally heavy rainfall next year

*

A feature F1 can take a certain value: A, B, C, D, E, & F and represents the grade of students from a college. Here feature type is

- (●) ordinal
- ( ) categorical
- ( ) boolean
- ( ) nominal

*

Few computer rooms in the school are denoted by the letters A to C. Staff records the number of classes held in each room during the first term. What kind of graph would be appropriate to present the frequency distributions of these data?

- ( ) Time line Plot chart
- (●) Bar Plot
- ( ) Line graph
- ( ) Scatter plot

*

How are keys and values presented and passed to the reducers during a standard sort and shuffle phase of MapReduce?

- (●) Keys are presented to the reducer in sorted order; values for a given key are sorted in ascending order.
- ( ) Keys are presented to a reducer in random order; values for a given key are not sorted.
- ( ) Keys are presented to a reducer in random order; values for a given key are sorted in ascending order.
- ( ) Keys are presented to the reducer in sorted order; values for a given key are not sorted.

*

The self-organizing maps can also be considered as the instance of _____ type of learning.

- (●) Unsupervised learning
- ( ) Supervised learning

*

Consider the (k = 2)-shingles for each D1, D2, D3, and D4:
D1 : [I am], [am Sam]
D2 : [Sam I], [I am]
D3 : [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham]
D4 : [I do], [do not], [not like], [like them], [them Sam], [Sam I], [I am]
Find the Jaccard similarity between documents.
JS(D2, D3) =? , JS(D1,D4) =? and JS(D3,D4) = ?

- (●) JS(D2, D3) = 0, JS(D1, D4) = 0.128, JS(D3, D4) = 0.286
- ( ) JS(D2, D3) = 0, JS(D1, D4) = 0.125, JS(D3, D4) = 0.273
- ( ) JS(D2, D3) = 0.285, JS(D1, D4) = 0.285, JS(D3, D4) = 1
- ( ) JS(D2, D3) = 0.01, JS(D1, D4) = 0.333, JS(D3, D4) = 0.286

*

Which of the following statements about standard Bloom filters is correct?

- [ ] A Bloom filter always returns the correct result.
- [ ] It is possible to delete an element from a Bloom filter.
- [x] A Bloom filter always returns TRUE when testing for a previously added element.
- [x] An empty Bloom filter (no elements added to it) will always return FALSE when testing for an element.
- [ ] It is possible to alter the hash functions of a full Bloom filter to create more space.

*

Which of the following statements about sampling are correct?

- [ ] Sampling increases the amount of elements in a data stream.
- [ ] Sampling reduces the diversity of the data stream.
- [ ] Sampling increases the amount of data fed to a data mining algorithm.
- [x] Sampling reduces the amount of data fed to a subsequent data mining algorithm.
- [ ] Sampling algorithms often need multiple passes over the data.
- [x] Sampling aims to keep statistical properties of the data intact.

*

In the sentence, “In Surat I took my hat off. But I can’t put it back on.” Total number of word tokens are:

- ( ) 13
- ( ) 15
- (•) 14
- ( ) 16

*

For what purpose, the analysis tools pre-compute the summaries of the huge amount of data?
Suppose that to get some information about something, you write a keyword in Google search. Google's analytical tools will then pre-compute large amounts of data

- ○ For authentication
- ○ In order to maintain consistency
- ◉ To obtain the queries response
- ○ For data access

*

What is the shortcoming of a content based recommender system?

- ◉ As it is based on similarity among items and users, it is not easy to find the neighbouring users.
- ○ Users will only get recommendations related to their preferences in their profile and the recommender engine may never recommend any item with other characteristics.
- ○ It needs to find a similar group of users, so suffers from drops in performance, simply due to growth in the similarity computation.

*

Which of the following is/are not true about DBSCAN clustering algorithm:

- ☐ It does not require prior knowledge of the no. of desired clusters
- ☑ It has strong assumptions for the distribution of data points in dataspace
- ☑ It has substantially high time complexity of order O(n^3)
- ☐ For data points to be in a cluster, they must be in a distance threshold to a core point

*

Which describes how a client reads a file from HDFS?

- The client queries the NameNode for the block location(s). The NameNode returns the block location(s) to the client. The client reads the data directory off the DataNode(s).
- The client contacts the NameNode for the block location(s). The NameNode then queries the DataNodes for block locations. The DataNodes respond to the NameNode, and the NameNode redirects the client to the DataNode that holds the requested data block(s). The client then reads the data directly off the DataNode.
- The client queries all DataNodes in parallel. The DataNode that contains the requested data responds directly to the client. The client reads the data directly off the DataNode.
- The client contacts the NameNode for the block location(s). The NameNode contacts the DataNode that holds the requested data block. Data is transferred from the DataNode to the NameNode, and then from the NameNode to the client.

Approximately, what is your average monthly expenditure? *

Consider you are planning to reduce your monthly expenses. You have the month-wise data of all the expenses for the past year.

Month	Rent/Utilities	Food	Travel	Shopping	Entertainment	Total
January	10000	5000	2000	8000	2000	27000
February	9000	10000	1500	15000	4000	39500
March	9500	4000	9000	10000	2000	34500
April	10500	3000	2000	11000	2500	29000
May	11500	6000	15000	13000	3500	49000
June	10000	7000	2000	12000	2500	33500
July	10500	4000	10000	11000	2000	37500
August	10000	3500	3000	10000	1500	28000
September	9500	4500	12000	9000	2500	37500
October	10500	5000	3000	11000	2000	31500
November	10000	5500	12000	10000	1500	39000
December	9000	4500	2500	12000	2500	30500
Average	**10000**	**5167**	**6167**	**11000**	**2375**	**34708**

- ○ 10000
- ○ 11000
- ○ 49000
- ◉ 35000

*

Pick the stemming actions

1. studied -- > study
2. plays-->play
3. troubled --> troubl
4. university → univers

- (●) 2,3 and 4
- ( ) 1 and 2
- ( ) 1, 2 and 3
- ( ) 2 and 3
- ( ) 1,2,3 and 4

*

To develop the Collaborative Filtering Based Recommendation System, the User vs. movie utility matrix is given below, utility matrix representing users' ratings of movies on a 1–5 scale, with 5 the highest rating. Blanks represent the situation where the user has not rated the movie. The movie names are HP1, HP2, and HP3 for Harry Potter I, II, and III, TW for Twilight, and SW1, SW2, and SW3 for Star Wars episodes 1, and 2. The users are represented by capital letters A through D. For this given matrix, do the following.

a. Calculate the similarity between User A and User B i.e. sim(A,B) using Jaccard Similarity method.
b. Calculate the similarity between User A and User C i.e. sim(A,C) using Jaccard Similarity method.
c. Which movies should be recommended to the User A (Consider the movie rating threshold as 3.0 and user similarity threshold as 0.33)

	HP1	HP2	HP3	TW	SW1	SW2
User A	4			5	1	
User B	5	5	4			
User C				2	4	5
User D		3				3

- ○ a) 4/5 , b) 1/2 , c) HP2, HP3 and SW2
- ○ a) 1/5 , b) 1/3 , c) SW2
- ○ a) 4/5 , b) 1/3 , c) HP2, HP3 and SW2
- ◉ a) 1/5 , b) 1/2 , c) HP2 and HP3

*

_____ technique re-scales a feature or observation value with distribution value between 0 and 1.

- ○ Hamming Distance
- ○ Euclidean Distance
- ○ Standardization
- ◉ Min-Max Normalization

*

Probability of a false positive in Bloom filters depends on

- ☑ The number of hash functions
- ☐ The density of 0's in the array
- ☑ The density of 1's in the array

*

Language Biases are introduced due to historical data used during training of word embeddings, which one amongst the below is not an example of bias

- ◉ New Delhi is to India, Beijing is to China
- ○ Man is to Computer, Woman is to Homemaker
- ○ None

*

When is the earliest point at which the reduce method of a given Reducer can be called?

- ( ) As soon as at least one mapper has finished processing its input split.
- ( ) As soon as a mapper has emitted at least one record.
- (•) Not until all mappers have finished processing all records.
- ( ) It depends on the InputFormat used for the job.

*

D1 : I am Sam.
D2 : Sam I am.
D3 : I do not like green eggs and ham.
D4 : I do not like them, Sam I am.
The (k = 2)-shingles of D1∪D2∪D3∪D4 are:

- ( ) {[I am], [Sam Sam], [do not], [like green], [eggs and], [ham I], [like them], [Sam I]}
- (•) {[I am], [am Sam], [Sam I], [am I], [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham], [ham I], [like them], [them Sam]}
- ( ) {[I am], [Sam do], [not like], [green eggs], [and ham], [them]}
- ( ) {[I am], [am Sam], [Sam Sam], [Sam I], [am I], [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham], [like them], [them Sam]}

*

In a corpus of N documents, one randomly chosen document contains a total of T terms and the term "hello" appears K times.
What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "hello" appears in approximately one-third of the total documents?

- (●) K * Log(3) / T
- ( ) T * Log(3) / K
- ( ) Log(3) / KT
- ( ) KT * Log(3)

*

In NLP, The algorithm decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

- ( ) Latent Dirichlet Allocation (LDA)
- ( ) Word2Vec
- (●) Inverse Document Frequency (IDF)
- ( ) Term Frequency (TF)

*

The prices at several different stores for a pair of shoes are shown below.
$68, $69, $71, $72, $73
Two days later, another store had the same shoes on sale for $55. How does the new price affect the data?

- ○ Only the range was affected.
- ○ The median increased, and the range increased.
- ○ The mean increased, and the range increased.
- ◉ The mean decreased and the range increased.

*

Which one of the following can be considered as the final output of the hierarchical type of clustering?

- ○ None of the above
- ○ Assignment of each point to clusters
- ○ Finalize estimation of cluster centroids
- ◉ A tree which displays how the close thing are to each other

*

Which of the following is NOT a type of metadata in NameNode?

- ○ File access control information
- ◉ No. of file records
- ○ List of files
- ○ Block locations of files

*

Consider the following corpus of 4 documents:

Documents	Terms
D_1	NLP is an interesting subject
D_2	Many students are interested in learning NLP
D_3	ANN plays an important role in NLP applications
D_4	Do you play tennis?

The TF*IDF for the word NLP for D_1, D_2, D_3, D_4 is

- None of the above
- Option 1: $[\frac{1}{5}\ \frac{1}{7}\ \frac{1}{8}\ 0]\ log_{10}(\frac{3}{4})$
- Option 2: $[\frac{1}{6}\ \frac{1}{7}\ \frac{1}{8}\ 0]\ log_{10}(\frac{3}{4})$
- Option 3: $[\frac{1}{5}\ \frac{1}{7}\ \frac{1}{8}\ 0]\ log_{10}(\frac{4}{3})$

*

The application master monitors all Map Reduce applications in the cluster

- (●) False
- ( ) True

*

I am the marketing consultant of a leading e-commerce website. I have been given a task of making a system that recommends products to users based on their activity on Facebook. I realize that user interests could be highly variable. Hence I decide to a. First, cluster the users into communities of like-minded people and b. Second, train separate models for each community to predict which product category (e.g. electronic gadgets, cosmetics, etc) would be the most relevant to that community. The first task is a/an _____ learning problem while the second is a/an _____ problem. Choose from the options:

- (●) Unsupervised and supervised
- ( ) Unsupervised and unsupervised
- ( ) Supervised and unsupervised
- ( ) Supervised and supervised

*

Which of the following are NOT big data problem(s)?

- ( ) Parsing 5 MB XML file every 5 minutes
- ( ) Processing T20 ICC World Cup tweet sentiments
- (•) both (a) and (c)
- ( ) Processing online bank transactions

*

When are the members of two sets more common relatively?

- ( ) Jaccard Index is Closer to -1
- ( ) Jaccard Index is Closer to 0
- (•) Jaccard Index is Closer to 1

*

The "Twitter datastream" contains tuples of the form:
(messageID, message, userID of posting user, in reply to messageID, time of posting, language of message).

You can assume that messageID and userID are unique, i.e. every message has a unique identifier and every user has a unique identifier. If the message is not posted in reply to any other message, we have in reply to messageID=null.

Examples of tuples in that stream are:
(124324234324, "@Nelly: I had breakfast just now!", 33523232, 122192225674, "28/11/2021", "English").

(435345332432, "Sitting in Paris, drinking a coffee", null, 122198435674, "29/11/2021", "English").

We want to answer queries by sampling roughly 1/10th of the data.
What is a good sampling strategy to answer the following query: What fraction of users post only in English?

- (●) Sample userIDs and include all messages by a user
- ( ) Sample by the combined key (userID,language_of_message)
- ( ) Generate a random number r between 0 and 9 and sample the tuple if r==0
- ( ) Sample language_of_message and include all messages of a language

*

Which ONE of the following are regression tasks?

- ( ) Predict whether a document is related to science
- ( ) Predict whether the price of petroleum will increase tomorrow
- (●) Predict the age of a person
- ( ) Predict the country from where the person comes from

*

For YARN, the ------- Manager UI provides host and port information.

- (●) Resource
- ( ) Replication
- ( ) NameNode
- ( ) DataNode

*

TF-IDF helps you to establish?

- (●) most important word in the document
- ( ) most frequently occurring word in the document

*

The steps involved in data analytics are given in random order. Order the steps correctly.
1-Preprocessing the data, 2- Preparing questionnaire, 3- Analysis of the data, 4- Collection of data

- ( ) 4 1 3 2
- (●) 2 4 1 3
- ( ) 4 1 2 3
- ( ) 4 2 1 3

*

The following given statement can be considered as the examples of_____
Suppose one wants to predict the number of newborns according to the size of storks' population by performing supervised learning

- ○ Clustering
- ◉ Regression
- ○ Classification

*

Which of the below questions are exploration oriented?

- ○ Will Person X buy Product Y from you?
- ○ None of the above
- ◉ Why do customers love your products?
- ○ Is your product liked by the customer?

*

Which of the following statements is TRUE?

- ☐ Outliers is a data point that is significantly close to other data points.
- ☑ The nature of our business problem determines how outliers are used.
- ☐ Outliers should be identified and always removed from a dataset.
- ☐ Outliers can never be present in the testing dataset.

*

Hive can be used for real time queries.

- ○ True if data set is small
- ○ TRUE
- ◉ FALSE
- ○ True for some distributions

*

"You may also like these.....","People who liked this also liked....", this type of suggestions are from the

- ○ Filtering system
- ○ Collaborative system
- ○ Amazon system
- ◉ Recommendation system

*

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters.
A1=(2,10),A2=(2,5),A3=(8,4),A4=(5,8),A5=(7,5),A6=(6,4),
A7=(1,2),A8=(4,9)
Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7.
After iteration 1, new clusters

- 1: {A1,A8}, 2: {A3,A4,A5,A6}, 3: {A2,A7}
- 1: {A1}, 2: {A3,A4,A5,A6,A8}, 3: {A2,A7}
- 1: {A1,A8}, 2: {A3,A5,A6}, 3: {A2,A4,A7}
- 1: {A1,A4,A8}, 2: {A3,A5,A6}, 3: {A2,A7}

*

Which of the following statements about the standard DGIM algorithm are false?

- [ ] DGIM operates on a time-based window.
- [x] The buckets contain the count of 1's and each 1's specific position in the stream.
- [ ] In DGIM, the size of a bucket is always a power of two.
- [ ] DGIM reduces memory consumption through a clever way of storing counts.
- [x] The maximum number of buckets has to be chosen beforehand.

*

Which of these are categorical features?

- ○ Height of a person
- ○ Price of petroleum
- ◉ Mother tongue of a person
- ○ Amount of rainfall in a day

*

You are given reviews of a few netflix series marked as positive, negative and neutral. Classifying reviews of a new netflix series is an example of

- ○ semi supervised learning
- ○ reinforcement learning
- ○ unsupervised learning
- ◉ supervised learning

*

Which of the following are examples of unsupervised learning?

- ☐ Segment online customers into two classes based on their age group – below 25 or above 25
- ☑ Make clusters of books on similar topics in a library
- ☑ Group news articles based on text similarity
- ☐ Filter out spam emails

*

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters.
A1=(2,10),A2=(2,5),A3=(8,4),A4=(5,8),A5=(7,5),A6=(6,4),
A7=(1,2),A8=(4,9)
Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. How many iterations are needed to finalize cluster members?

- ○ 2
- ○ 1
- ○ 4
- ◉ 3

*

Suppose your problem statement is to establish the relationship between the GDP and life expectancy of a country. What would be the more suitable data collection method for this problem?

- ○ Secondary as a large amount of data is required and difficult to gather by itself
- ○ Primary as data is reliable and fresh
- ○ None of the above
- ◉ Secondary data, as it can be readily found on open source websites

*

In linguistic morphology _____ is the process for reducing inflected words to their root form.

- (●) Stemming
- ( ) Both Rooting & Stemming
- ( ) Rooting
- ( ) Text-Proofing

Back Submit Clear form

Never submit passwords through Google Forms.