Your Answers:

1

1 / 1 point

Which expression is used to compute the Jaccard similarity of two sets A and B?

- ○ 1 - (A ∩ B) / (A ∪ B)
- ○ 1 - |A ∩ B| / |A ∪ B|
- ○ (A ∩ B) / (A ∪ B)
- ✓ ◉ |A ∩ B| / |A ∪ B|

Feedback

Based on your answer

That is correct. The Jaccard similarity between finite sets is defined as the size of the intersection divided by the size of the union of the sets.

2

3 / 4 points

Briefly explain the shingling technique used to represent a document in the form of set. How to measure similarity of two documents represented as shingle sets? Give a small example to illustrate your answer.

Shingling: Shingling is a technique with the idea to represent a document as k-shingles, to partiton the document into smaller shingles. To generate the k-shingles, a "window" of size k are going through the document to extract tokens of size k as shingles. In a document, tokens can for example be characters or words.

Example:

- D1 = "abacd", then the 2-shingles of document D1 will be = {ab, ba, ac, cd}
- D2 = "abcda", then the 2-shingles of document D1 will be = {ab, bc, cd, da}

✓

Similarity: To measure the similarity between two documents we use a technique called Jaccard Similarity, which is the intersection of the two documents, over the union of the documents *(intersection of D1 and D1 / union of D1 and D2)*

Example: When using the above 2-shingles documents D1 and D2, we see that ab and cd is in both the documents. Hence, the Jaccard Similarity for D1 and D2 is 2/6

Graded

Feedback

Feedback from grader

Incorrect: "is the intersection of the two documents, over the union of the documents (intersection of D1 and D1 / union of D1 and D2)" should be "is **the size (or cardinality) of** intersection of the two documents, over **the size (or cardinality) of** the union of the documents | intersection of D1 and D1 | / | union of D1 and D2| " where |...| is "size of".

3 2 / 2 points

What is a complexity of shingling and computing Jaccard similarity of two documents of $O(n)$ characters each using *k*-shingles? Assume that $n \gg k$ and comparison of two *k*-shingles takes $O(k)$.

- ○ $O(k \cdot \log n)$
- ○ $O(k \cdot n)$
- ● $O(k \cdot n^2)$ ✓
- ○ $O(k^2 \cdot n)$

4 1 / 1 point

Two documents A and B are always identical if the Jaccard similarity of their *n*-shingle sets is equal to 1.

- ○ True
- ● False ✓

5 3.5 / 5 points

Briefly explain the minhashing technique used in finding textually similar documents. Justify the use of minhashing.

Minhashing technique: From the shingles that are generated with the k-shingles, a characteristic matrix is produced. The rows represent the shingles and the columns represent the documents. For each shingle that is in a certain document, there will appear a 1 in the matrix and otherwise 0. In minhashing we aim to create signature vectors that represents permutations, which then build the matrix. To generate the signature matrix, we make some number of permutation of the characteristic matrix, for example 100, which is the same matrix but the order of the rows are now randomly mixed. For each permutation we also have h hash function (h = number of columns). For each of these hash functions we apply them to the shingles of a document and select the smallest value, based on the permutation order. From the hash-values of each permutation, we create a signature matrix, where each row represents the hash values from each permutation. Hence, the columns will correspond to each document, as it does in the characteristic matrix.

Justify the use: Since shingles can be quite big, which has a higher complexity and requires big memory. The minhashing technique reduces the complexity by comparing only the documents based on the matrix instead of comparing all the k-shingles that are produced one by one.

Graded

Feedback

Feedback from grader

Imprecise: Should not mix together "permutations" with "hash functions" as these are two different methods of building minhash signatures from shingle sets (characteristic matrix).

"Since shingles can be quite big..." should be "Since shingle sets can be quite big...", as shingles can be hashed (4B per shingle).

Incomplete: Should explain how to estimate similarity using minhash signatures.

6 2 / 3 points

Minhashing allows estimating similarity of two documents represented by columns in a signature matrix. Select correct statement(s).

- [x] The Jaccard similarity of two shingle sets can be estimated by dividing the number of rows that two corresponding columns agree in the signature matrix, by the number of rows in signature matrix.
- [ ] The Jaccard similarity of two shingle sets can be estimated by dividing the number of rows that two corresponding columns agree in the signature matrix, by the number of distinct values in those columns.
- [x] The probability that two columns have the same value in a given row of the signature matrix equals the Jaccard similarity of the shingle sets corresponding to those columns.
- [ ] The fraction of rows that two columns agree in the signature matrix is an estimate of the true Jaccard similarity of the corresponding shingle sets.

7 1 / 1 point

Consider the following data set of ten market baskets where each basket (identified by a transaction id, TID) is a small set of items a customer (identified by CID) bought in one visit to a shop. Compute the support for the itemset { coke, beer, bread } by treating each TID as a basket.

CID	TID	Items
A	9001	{ milk, beer, bread }
A	9011	{ milk, coke, cereal, bread }
B	9002	{ milk, coke, beer, bread }
B	9012	{ milk, cereal, beer, bread }
C	9003	{ coke, cereal, bread }
C	9013	{ coke, beer, bread }
D	9004	{ cereal, beer }
D	9014	{ milk, coke, cereal }
E	9005	{ milk, beer, bread }
E	9015	{ milk, coke, bread }

NOTE: When entering a numeric answer, please make sure to use point rather than comma for a decimal separator, e.g. 0.99

2

8 2 / 2 points

Consider the following data set of ten market baskets - the same data set as in the previous question - where each basket (identified by a transaction id, TID) is a small set of items a customer (identified by CID) bought in one visit to a shop. Compute the confidence for the association rule { bread } ➔ { coke, beer } by treating each TID as a basket.

CID	TID	Items
A	9001	{ milk, beer, bread }
A	9011	{ milk, coke, cereal, bread }
B	9002	{ milk, coke, beer, bread }
B	9012	{ milk, cereal, beer, bread }
C	9003	{ coke, cereal, bread }
C	9013	{ coke, beer, bread }
D	9004	{ cereal, beer }
D	9014	{ milk, coke, cereal }
E	9005	{ milk, beer, bread }
E	9015	{ milk, coke, bread }

NOTE: When entering a numeric answer, please make sure to use point rather than comma for a decimal separator, e.g. 0.99

✓ 0.25

9 3 / 4 points

Briefly describe the A-priory algorithm to find frequent itemsets. What is an association rule between two itemsets?

A-priory: The A-priory algoritm is used when we want to find frequent itemset. With the algorithm we aim to find the frequent itemsets of size k=1,2,... until we no longer have any frequent itemsets.

The algorithm is based on key ideas that any subset of a frequent itemset, must also be frequent and that if a itemset appears at least **s** *(support threshold)* times, so does the itemsets superset. The algorithm start by extracting all unique singular items that is in the baskets with their corresponding support to a candidate set C1. From C1 we apply a filter that only accepts the frequent itemsets from C1 to be included in the frequent itemset L1. A itemset is frequent if the support is bigger than the support threshold.

When we have C1 and L1, we can start to iterate where we create Ck and Lk.

- Ck: Since any subset of a frequent itemset must be frequent, we use Lk-1 to create Ck by combining the frequent itemsets. *Example:* If we have all frequent singular itemsets in L1, we create all possible itemsets that is of size 2 from L1 and calculate the new itemsets of size 2's support. Hence, we have created C2
- Lk: By filtering Ck and only extract the itemsets of size k that has a support over the support threshold, Lk is created.

The algorithm will stop to generate Ck and Lk when there are no frequent sets to create a new candidate set with.

Association rule: An association rule between two itemset can been seen as an "if-then" rule I --> J *(where I and J are itemsets)*, . If a person buys I , it will most likely also buy J.

Graded

Feedback

Feedback from grader

Incorrect: ..."if a itemset appears at least s (support threshold) times, **so does the itemsets superset.** " should be "so does its subsets" OR "if an itemset is not frequent, then neither are its supersets." OR "The support of an itemset is at least as the support of its superset"

Imprecise: "we use Lk-1 to create Ck by combining the frequent itemsets." -- Ck is typically constructed by combining itemsets from Lk-1 with itemsets from L1.

10 2 / 2 points

Which of the following statements about sampling from a data stream are correct? Select correct statement(s).

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

11 1 / 1 point

How does the probability of an element to be included in the fixed-size sample (reservoir) change with the increasing number of elements seen so far in the data stream?

- ( ) increases
- (•) decreases
- ( ) does not change

12 0 / 3 points

Which of the following statements about Bloom filter are correct? Select correct statement(s).

- [x] A Bloom filter guarantees no false negatives.
- [ ] A Bloom filter guarantees neither false positives nor false negatives.
- [ ] A Bloom filter always returns TRUE when testing for a stream element with a key previously added to the set.
- [ ] A Bloom filter guarantees no false positives.
- [ ] A Bloom filter always returns FALSE when testing for a stream element with a key that is not in the set.
- [ ] A Bloom filter may return TRUE when testing for a stream element with a key that is not in the set.
- [ ] It is possible to delete a key from a Bloom filter.

13 1 / 1 point

What state needs to be stored in order to answer the standing query about a data stream "***What is the average value ever seen in the stream?***"

- ( ) The elements of the last 1h to compute the rolling average.
- (x) One value for the current number of elements in the stream observed so far, and one value for the current sum of the elements.
- ( ) One value for the current average updated whenever a new stream element arrives.
- ( ) Last 1000 elements of the stream to compute the rolling average.

Feedback

Based on your answer

Correct! The average can be computed by dividing the sum by the number of stream elements.

14 1.5 / 3 points

Select correct answer(s)

- [x] Katz centrality takes global graph topology into account.
- [ ] Degree centrality takes global graph topology into account.
- [ ] Degree centrality is computationally less expensive than closeness centrality
- [ ] Betweenness centrality takes only local graph topology into account

15 0 / 4 points

Select correct answer(s)

- [ ] A graph should always have at least one edge
- [x] Bipartite graphs cannot have triangles
- [x] Graph Adjacency matrices are always symmetric

Selected Answer - Incorrect

- [ ] Every node in giant component has a path to every other node in the same component
- [ ] Each graph has to have at least one bridge edge
- [x] Directed Graphs can have sum of all in-degrees larger than sum of all out-degrees

Selected Answer - Incorrect

- [ ] Number of edges in a complete bipartite graph is N(N-1)/2, where N is number of nodes
- [ ] Clustering coefficient of each node in a bipartite graph is always "1"
- [x] Clustering coefficient of a node with degree "k" is always equal to "1" if there are k number of connections between neighbors of that node.

Selected Answer - Incorrect

16 2 / 2 points

Erdos-Renyi Random graphs with (log(N))/N> p > 1/N exhibit:

- [ ] small average clustering coefficient and power-law degree distribution
- [ ] short diameter and large average clustering coef.
- [x] one large connected component and small average clustering coefficient
- [ ] large diameter and binomial degree distribution

17 0 / 2 points

Assume a graph constructed by Watts-Strogatz Small-World model, with an average degree of O(logN) and rewiring probability such that on average one edge per node gets "rewired".
Such graph will:

- [ ] will exhibit large hubs (very high-degree nodes)
- [ ] exhibit low diameter and large average clustering coefficient
- [x] is likely to be disconnected.

Selected Answer - Incorrect

18 4 / 4 points

Select correct answer(s)

- [ ] PageRank is the same as Eigenvector centrality with teleportation.
- [ ] SimRank (random walk with restarts) teleports a random walker to a node with the lowest pagerank score.
- [x] Hubs and Authorities (HITS) algorithm assigns two scores to each node.
- [x] Link Farm effectiveness depends on teleportation probability in PageRank.
- [ ] Hubs and Authorities (HITS) algorithm requires teleportation in order to converge.

! This question has been regraded.

19 Previous score 0 / 4 points Regrade score 4 / 4 points

At a large organisation a researcher tries to estimate the proportion of staff infected with diseases Covid19 and common flu by performing calls on staff through contact tracing in a random walk manner. I.e., The experiment starts by contacting a known sick person and requesting to give the names of all the colleagues that he/she interacted with within the last week. The researcher then contacts one of these persons randomly, inquires about their health and repeats the procedure until information from 100 people is collected. (i.e., the researcher performs sampling through random walks on the staff-interaction-graph).
After the experiment the researcher identified:

3 people with covid19 who interacted with 12 other people each (had a degree 12 each)
4 people with covid19 who interacted with 16 other people each
5 people with covid19 who interacted with 10 other people each
2 people with common flu who interacted with 8 other people each
3 people with common flu who interacted with 6 other people each
1 person with common flu who interacted with 4 other people each
15 healthy people who interacted with 5 other people each
20 healthy people who interacted with 4 other people each
14 healthy people who interacted with 2 other people each
33 healthy people who interacted with 1 other person each

What is your estimate on the fraction of the staff with covid19 in the organisation?

NOTE: when entering a numeric answer to Canvas, please make sure you use "correct" decimal separator: i.e, by using "decimal comma" (e.g., "0,99") instead of "decimal point" (e.g., "0.99") Canvas might interpret your fractional part as integer(e.g., change your answer to "99").

✓ 0.02

20 4 / 4 points

Assume you got access to an undirected non-bipartite graph G that you analyze by performing spectral analysis, i.e., extracting eigenvalues and associated eigenvectors of the adjacency matrix A representing graph G, as well as extracting eigenvalues and associated eigenvectors of the Laplacian matrix L representing graph G and extracting eigenvalues and associated eigenvectors of normalized Laplacian matrix NL ($NL = D^{-½}AD^{-½}$, where D is a degree matrix, and A is the adjacency matrix of G).

You are given a task to cluster the graph into k clusters ***optimizing on node average internal degree*** in each cluster.

Which of the following you will need to complete your task:

- Eigenvectors associated with k largest eigenvalues of L
- PageRank vector
- ✓ Eigenvectors associated with k largest eigenvalues of A
- Any k eigenvectors of A
- Eigenvectors associated with k smallest eigenvalues of NL
- All Eigenvalues of L

21 4 / 4 points

1. You continue to investigate Graph G from the question above (question on "Spectral graph Analysis 1"). You learn that three smallest eigenvalues of L are equal to zero. Select correct answer(s)

- Graph G is likely constructed by an Erdos-Renyi Random Graph model
- ✓ Random walk will not have a unique stationary distribution on G
- Graph G is likely constructed following Preferential attachment model
- Graph G is very good expander

22 0 / 3 points

There is a relationship between the spectral gap (e.g., the absolute difference between the first two eigenvalues of the Laplacian of the graph) and the convergence of random walk:

- The larger the gap the slower the convergence
- ✗ There is no relationship.
- The smaller the gap the slower the convergence

23 0 / 3 points

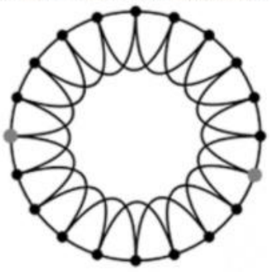
$\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of the Laplacian matrix of a graph depicted in the picture below ($\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$). Select correct answer(s):

- [ ] $\lambda_1 - \lambda_n \approx 0$
- [ ] $\lambda_1 - \lambda_2 \approx 0$
- [ ] $\lambda_1 = \lambda_2$
- [x] Spectral gap of the graph is very large

Selected Answer - Incorrect

24 0 / 3 points

Given Graphs A (d-regular) and B (random) as shown below select correct answer(s).



Graph A

Graph B

- [x] Spectral gap of A is much larger than spectral gap of B

Selected Answer - Incorrect

- [ ] Spectral gap of A is exactly the same as spectral gap of B
- [ ] Spectral gap of A is almost the same as spectral gap of B
- [ ] Spectral gap of A is much smaller than spectral gap of B

25 2 / 2 points

Spectral gap depends only on the degree distribution of the graph

- ○ True
- ◉ False

26 4 / 4 points

Dimensionality Reduction, Q1

You have performed SVD decomposition of a sparse matrix M of size n x m ($n \neq m$) into a product of three matrices USV^T such that the middle matrix S has size d x d, where d is << #rows and d is << #columns of M. Select correct answer(s)

- [ ] M is exactly equal to USV^T if the rank of M is larger than d.
- [ ] U is a diagonal matrix.
- [x] The columns of V are eigenvectors of matrix A^TA
- [ ] Matrix U is always a square matrix.

27 4 / 4 points

Dimensionality Reduction, Q2

You have performed CUR decomposition of the sparse matrix M of size n x m ($n \neq m$) into a product of three matrices CUR such that the middle matrix U has size d x d, where d is << #rows and d is << #columns of M. Select correct answer(s)

- [ ] The columns of C are orthonormal
- [x] Matrix R is sparse
- [ ] Matrix R is always a square matrix
- [ ] Matrix U is sparse

28 2 / 2 points

Dimensionality Reduction, Q3

You have a matrix M representing ratings given by users to the movies.

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7	Movie 8
Alice								
Bob								
Carol								
David								
Erin								
Frank								

The matrix was populated by user ratings (each cell was assigned a value from 0 to 5) and you were given a task to identify main concepts that describe matrix M by reducing the dimensionality of M. The SVD decomposition of $M=USV^T$ provided you with these matrices:

U=					
-0.71	0.02	-0.02	-0.62	0.16	0.3
-0.7	0.06	-0.06	0.61	-0.15	-0.32
-0.03	0.03	0.69	-0.35	-0.25	-0.59
-0.05	0.04	0.72	0.36	0.23	0.54
-0.05	-0.71	0.03	0	-0.65	0.28
-0.03	-0.71	0.03	0.04	0.64	-0.29

29 2 / 2 points

Consider the setting from "Dimensionality Reduction Q3" question. With how many main concepts (dimensions) can matrix M be identified?

3

30 3 / 3 points

Consider the setting from "Dimensionality Reduction Q3" question. Select most similar user to Alice from the list below:

- Erin
- Bob
- Frank
- David
- Carol

31 3 / 3 points

Consider the setting from "Dimensionality Reduction Q3" question. Select the most similar movie to Movie1 from the list below:

- Movie 2
- Movie 3
- Movie 4
- Movie 5
- Movie 6
- Movie 7
- Movie 8

32 2 / 2 points

Consider the setting from "Dimensionality Reduction Q3" question. Did any users rank the movies in exactly the same way (gave the exactly same scores)?

- ○ True
- ◉ False ✓

33 4 / 6 points

Select correct answer(s):

- ☐ Item-Item CF recommender systems perform better than User-user CF recommender systems
- ☐ In order for Content-based Recommender System to recommend items for a user U, it needs data from other users
- ☑ Content-based Recommender Systems are better in recommending new and unpopular items than Collaborative filtering systems ✓
- ☑ Content-based Recommender Systems cannot provide good recommendations for new users. ✓
- ☐ Collaborative filtering Recommender Systems require feature extraction.

35 0 / 3 points

Consider the setting from "Graph Representation Learning (GRL), Q1" question.

You feed a sequence corpus (collection) to Skipgram, where each sequence contains a fixed number of uniformly sampled random nodes as opposed to a random walk sampled following the graph structure. Choose the correct answer.

- ○ If two nodes u and v have the same role, then $z_u \approx z_v$
- ◉ If two nodes u and v belong to the same community, then $z_u \approx z_v$ ✗
- ○ If two nodes u and v have the same role and belong to the same community, then $z_u \approx z_v$
- ○ None of the above

34 0 / 3 points

Graph Representation Learning (GRL), Q1

Recap on the notations on GRL:

- A graph denoted by $G=(V,E)$ or adjacency matrix $\mathbf{A}=[0,1]^{N\times N}$
- Number of nodes N
- A graph representation learning algorithm, $f:V\rightarrow\mathbf{R}^d$
- Similarity in the input space: $SIM_G:V\times V:\mathbf{R}$
- Similarity in the embedding space $\mathrm{SIM}_E:\mathbf{R}^d\times\mathbf{R}^d\rightarrow\mathbf{R}^d$
- The mapping of node u, $f(u)=z_u$

What makes *f* a good quality mapping function? (choose the correct answer)

- ✕ (selected) None of the above
- A certain property of the graph should be preserved by *f*
- It should always preserve the community assignment of nodes.
- $d=N$