Question 1

If for the χ^2 statistics for a binary feature, we obtain $P(\chi^2 |DF = 1) < 0.05$, this means:

a. **That the class labels depends on the feature**
b. That the class label is independent of the feature
c. That the class label correlates with the feature
d. No conclusion can be drawn

Question 2

Given a document collection, if we change the ordering of the words in the documents, which of the following will not change?

a. **Singular values in LSI**
b. The entities extracted using HMM
c. The embedding vectors produced by Word2vec
d. All the previous will change

Question 3

We want to return, from the two posting lists below, the top-2 documents matching a query using Fagin's algorithm with the aggregation function taken as the sum of the tf-idf weights. How many entries (total of both lists) are accessed in the first phase of the algorithm performing round robin starting at List 1 (i.e., before performing the random access)?

List 1			List 2	
document	tf-idf		document	tf-idf
d3	0.8		d1	0.8
d2	0.6		d3	0.6
d1	0.5		d4	0.5
d4	0.4		d2	0.4

a. 3
b. 4
c. **5**
d. 6

Question 4

In the physical representation of an inverted file, the size of the index file is typically in the order of (where n is the number of documents):

a. O(log(n))
b. **O(sqrt(n))**
c. O(n)
d. O(n^2)

Question 5

Which is **true** about the use of entropy in decision tree induction?

a. The entropy of the set of class labels of the samples from the training set at the leaf level is always 0

b. We split on the attribute that has the highest entropy
c. The entropy of the set of class labels of the samples from the training set at the leaf level can be 1
d. We split on the attribute that has the lowest entropy

<u>Question 6</u>

Given the *2-itemsets* {1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 5}, when generating the *3-itemset* we will:

a. Have **4** *3-itemsets* after the join and **4** *3-itemsets* after the prune
b. **Have 4 *3-itemsets* after the join and 2 *3-itemsets* after the prune**
c. Have **3** *3-itemsets* after the join and **3** *3-itemsets* after the prune
d. Have **2** *3-itemsets* after the join and **2** *3-itemsets* after the prune

2019 Final - MCQ

1. Which one of the following about the Expectation-Maximization algorithm is FALSE?

   a) In E step the labels change, in M step the weights of the workers change.
   b) The label with the highest probability is assigned as the new label.
   c) Assigning equal weights to workers initially decreases the convergence time. (CORRECT)
   d) It distinguishes experts from normal workers.

2. Which of the following is TRUE?

   a) Ontologies are used to directly map two schemas in order to overcome semantic heterogeneity.
   b) Graph representation of an RDF statement facilitates exchange and storage.
   c) RDF is a standardized model for encoding ontologies (CORRECT)
   d) XML does not facilitate introducing new terms which are domain specific.

3. Regarding features engineering, which of the following is wrong:

   a. Supervised discretization can merge any two intervals of the same variable. (CORRECT)
   b. Classifiers can be sensitive to the absolute scale of the variables.
   c. Features *filtering* consider single variables, whereas *wrapping* considers features combinations.
   d. Standardisation can produce arbitrarily large values whereas scaling does not.

4. Given the graph 1→2, 1→ 3, 2→3, 3→2, switching from Page Rank to Teleporting PageRank will have an influence on the value(s) of

   a. All the nodes (CORRECT)
   b. Node 1
   c. Node 2 and 3
   d. No nodes. The values will stay unchanged.

5. Which of the following is true:

   a. Modularity is a measure of how communities are connected together
   b. Agglomerative algorithms recursively decompose communities into sub-communities
   c. Divisive algorithms are based on modularity
   d. Girvan-Newman works by removing edges with the highest betweenness measure (CORRECT)

6. Which of the following is true:

   a. The tf-idf weight is the ratio between tf and idf
   b. The idf term decrease the impact of stop-words (CORRECT)
   c. Frequent terms obtain low tf score
   d. The tf term is computed over the whole document collection

7. Which of the following tasks would typically not be solved by clustering?
   a. Community detection in social networks.
   b. Discretization of continuous features.
   c. Spam detection in an email system (CORRECT)
   d. Detection of latent topics in a document collection

8. Which one is false about Label Propagation?

   a. The labels are inferred using the labels that are known apriori.
   b. It can be interpreted as a random walk model.
   c. Propagation of labels through high degree nodes are penalized by low abandoning probability. (CORRECT)
   d. Injection probability should be higher when labels are obtained from experts than by crowdworkers.

2021 Final - MCQ

Question 1

We want to return, from the two posting lists below, the top-2 documents matching a query using Fagin's algorithm with the aggregation function taken as the sum of the tf-idf weights. How many entries (total of both lists) are accessed in the first phase of the algorithm performing round robin starting at List 1 (i.e., before performing the random access)?

List 1

Document	**tf-idf**
d3	**0.8**
d2	**0.6**
d1	**0.5**
d4	**0.4**

List 2

Document	**tf-idf**
d1	**0.8**
d3	**0.6**
d4	**0.5**
d2	**0.4**

a. 4
b. 8
c. 6 (CORRECT)
d. 2

Question 2

Which of the following is **true** regarding inverted files?

a. Storing differences among word addresses reduces the size of the postings file (CORRECT)
b. Varying length compression is used to reduce the size of the index file
c. The space requirement for the postings file is O(nβ), where β is generally between 0.4 and 0.6

d. Inverted files prioritize efficiency on insertion over efficiency on search

Question 3

The number of non-zero entries in a column of a term-document matrix indicates:

a. none of the other responses is correct
b. how relevant a term is for a document
c. how often a term of the vocabulary occurs in a document
d. how many terms of the vocabulary a document contains (CORRECT)

Question 4

Which of the following is **true** for community detection in social graphs?

a. If n cliques of the same order are connected cyclically with n edges, then the Louvain algorithm will always detect the same communities, independently of the order of processing of the nodes. (CORRECT)
b. The Louvain algorithm always creates a hierarchy of communities with a common root.
c. The Louvain algorithm is efficient for small networks, while the Girvan-Newman algorithm is efficient for large networks.
d. The result of the Girvan-Newman algorithm can depend on the order of processing of nodes whereas for the Louvain algorithm this is not the case.

Question 5

When using matrix factorization for information extraction the entries of the matrix are obtained

a. from both text and a knowledge base (CORRECT)
b. from a knowledge base represented as text
c. from text
d. from a knowledge base

Question 6

For the number of times the apriori algorithm and the FPgrowth algorithm for association rule mining are scanning the transaction database the following is true

a. fpgrowth and apriori can have the same number of scans (CORRECT)
b. apriori cannot have fewer scans than fpgrowth
c. fpgrowth has always strictly fewer scans than apriori
d. all three above statements are false

Question 7

Why is non-discounted cumulative gain used as evaluation metrics for recommender systems

a. because it is more accurate than retrieval metrics, like precision and recall
b. because often only the top recommendations are considered by the user (CORRECT)
c. because it allows to consider the financial value of recommended items
d. because it considers the predicted ratings of all items that have not been rated by the user

Question 8

Which of the following models for generating vector representations for text require to precompute the frequency of co-occurrence of words from the vocabulary in the document collection

a. CBOW
b. Fasttext
c. Glove (CORRECT)
d. LSI

Quiz - 2021

1. Information extraction:

- □ Necessarily requires training data.
- □ Is used to identify characteristic entities in a document.
- □ Is always bootstrapped by using ontologies.
- ☑ ~~Can be used to populate ontologies.~~

2. What is **TRUE** regarding Fagin's algorithm?

- □ Posting files need to be indexed by TF-IDF weights
- □ It performs a complete scan over the posting files
- □ It never reads more than (kn)½ entries from a posting list
- ☑ ~~It provably returns the k documents with the largest aggregate scores~~

3. Which of the following statements on Latent Semantic Indexing (LSI) and Word Embeddings (WE) is false?

- □ The dimensions of LSI can be interpreted as concepts, whereas those of WE cannot
- □ LSI does not depend on the order of words in the document, whereas WE does
- □ LSI is deterministic (given the dimension), whereas WE is not
- ☑ ~~LSI does take into account the frequency of words in the documents, whereas WE with negative sampling does not~~

4. When constructing a word embedding, what is **TRUE** regarding negative samples?

- ☑ ~~They are oversampled if less frequent~~
- □ Their frequency is decreased down to its logarithm
- □ They are words that do not appear as context words
- □ They are selected among words that are not stop-words

5. A page that points to all other pages but is not pointed by any other page would have:

- □ Nonzero authority
- □ Zero hub
- ☑ ~~Nonzero PageRank~~
- □ None of the above

6. When computing PageRank iteratively, the computation ends when:

- □ The difference among the eigenvalues of two subsequent iterations falls below a predefined threshold

- ☑ ~~The norm of the difference of rank vectors of two subsequent iterations falls below a predefined threshold~~
- □ The probability of visiting an unseen node falls below a predefined threshold
- □ All nodes of the graph have been visited at least once

7. In Ranked Retrieval, the result at position k is non-relevant and at k+1 is relevant. Which of the following is always true?
Hint: P@k and R@k are the precision and recall of the result set consisting of the k top-ranked documents.

- □ P@k-1>P@k+1
- □ R@k-1=R@k+1
- ☑ ~~R@k-1<R@k+1~~
- □ P@k-1=P@k+1

8. Which of the following is **TRUE** regarding community detection?

- □ The high betweenness of an edge indicates that the communities are well connected by that edge
- □ The Girvan-Newman algorithm attempts to maximize the overall betweenness measure of a community graph
- ☑ ~~The high modularity of a community indicates a large difference between the number of edges of the community and the number of edges of a null model~~
- □ The Louvain algorithm attempts to minimize the overall modularity measure of a community graph

9. What is **WRONG** regarding the Transformer model?

- ☑ ~~Its computation cannot be parallelized compared to LSTMs and other sequential models.~~
- □ It uses a self-attention mechanism to compute representations of the input and output.
- □ Its complexity is quadratic to the input size.
- □ It captures the semantic context of the input.

10. In User-Based Collaborative Filtering, which of the following is **TRUE**?

- □ Pearson Correlation Coefficient and Cosine Similarity have the same value range and return the same similarity ranking for the users.
- ☑ ~~Pearson Correlation Coefficient and Cosine Similarity have different value ranges and can return different similarity rankings for the users~~
- □ Pearson Correlation Coefficient and Cosine Similarity have different value ranges, but return the same similarity ranking for the users
- □ Pearson Correlation Coefficient and Cosine Similarity have the same value range but can return different similarity rankings for the users

11. Which of the following is **TRUE** for Recommender Systems (RS)?

- □ The complexity of the Content-based RS depends on the number of users
- □ Item-based RS need not only the ratings but also the item features
- □ Matrix Factorization is typically robust to the cold-start problem.
- ☑ ~~Matrix Factorization can predict a score for any user-item combination in the dataset.~~

12. Considering the transaction below, which one is **WRONG**?

Transaction ID	Items Bought
1	Tea
2	Tea, Yoghurt
3	Tea, Yoghurt, Kebap
4	Kebap
5	Tea, Kebap

- □ {Yoghurt} -> {Kebab} has 50% confidence
- □ {Yoghurt, Kebap} has 20% support
- □ {Tea} has the highest support
- ☑ ~~{Yoghurt} has the lowest support among all itemsets~~

13. Suppose that in a given FP Tree, an item in a leaf node N exists in every path. Which of the following is **TRUE**?

- □ N co-occurs with its prefixes in every transaction
- □ For every node P that is a parent of N in the FP tree, confidence (P->N) = 1
- ☑ ~~{N}'s minimum possible support is equal to the number of paths~~
- □ The item N exists in every candidate set

14. Which of the following properties is part of the RDF Schema Language?

- □ Description
- □ Type
- □ Predicate
- ☑ ~~Domain~~

15. Which of the following is wrong regarding Ontologies?

- □ We can create more than one ontology that conceptualizes the same real-world entities
- □ Ontologies help in the integration of data expressed in different models
- ☑ ~~Ontologies dictate how semi-structured data are serialized~~
- □ Ontologies support domain-specific vocabularies

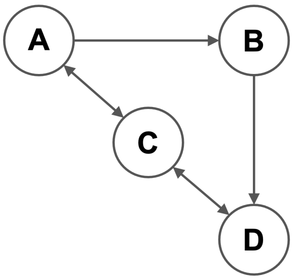
1. Assume documents d_1 and d_2 are two documents where d_1 is the string S and d_2 is the string S+" "+S. Given a query Q, which document will be ranked higher by a cosine-distance-based TF-IDF retrieval model?

a. d_1
b. d_2
c. They will have the same rank (CORRECT)
d. It depends on the query Q

Explanation: the two document will have exactly the same direction, therefore given any vector representation for Q, they will always have the same rank (as it's a cosine-distance-based retrieval)

2. Run the PageRank algorithm on the following graph. What would be the node with the highest PageRank after the 2nd iteration?

Note: In Iteration 0 we assign uniform weights to all nodes. You need to run for the next two iterations: Iteration 1 and Iteration 2.



a. A
b. B
c. C (CORRECT)
d. D

Explanation:

	Iteration 0	Iteration 1	Iteration 2	Ranking
A	¼	1/12	1.5/12	4
B	¼	2.5/12	2/12	3
C	¼	**4.5/12**	**4.5/12**	**1**
D	¼	4/12	4/12	2

3. You have a chain of pages where each page links to the next. Additionally, every page in the chain links back to the first page. How will the PageRank probability of the first page behave, using basic PageRank without random jumps, as the chain growths?

a. It will converge to 0
b. It will converge to ½ (CORRECT)
c. It will converge to 1
d. It will converge to infinity

4. What is TRUE regarding Item-based Collaborative Filtering?

a. It does leverage item description
b. It can recommend niche or new items (CORRECT)
c. It recommends items by finding similar users
d. None of the above

5. Using Matrix Factorization we have ended up with two matrices representing the user preferences (for 4 users: A, B, C, D) and item preferences (For 5 items: 1, 2, 3, 4, 5) as shown below.

$$\mathbf{U} = \begin{bmatrix} 3 & 0 \\ 2 & 2 \\ 4 & 4 \\ 0 & 4 \end{bmatrix} \qquad \mathbf{I} = \begin{bmatrix} 1 & 3 \\ 0 & 4 \\ 4 & 3 \\ 1 & 2 \\ 4 & 0 \end{bmatrix}$$

Which top 2 users, the item 3 should be recommended to?

a. A, B
b. B, C (CORRECT)
c. C, D
d. A, D

Explanation:
The UxI matrix will be the following:

	I1	I2	**I3**	I4	I5
UA	3	0	12	3	12
UB	8	8	**14**	6	8
UC	16	16	**28**	12	16
UD	12	16	12	8	0

And the normalized version of it:

	I1	I2	**I3**	I4	I5
UA	0	0	1.5	0	1.5
UB	1	1	**2**	1	1
UC	2	2	**3.5**	3	2
UD	1.5	2	1.5	1	0

6. You have the following sentence: “The **dollar** index dropped today around 0.5% in New York Stock Exchange” and you want to do Entity Linking for the word “**dollar**”.
You retrieve in the KG different nodes that can be related to the mention of “dollar” and end up in an entity graph with the following properties:

KG Node	**Out-degree**	**In-degree**
United States dollar	2	4
Canadian dollar	7	0
Australian dollar	0	6

Using Personalized Pagerank ranking, starting in the entity graph with a node related to the mention “New York Stock Exchange”, which one of the following is always TRUE:

a. P(“Canadian dollar”) <= P(“Australian dollar”) (CORRECT)
b. P(“United States dollar”) < P(“Australian dollar”)
c. P(“United States dollar”) <= P(“Canadian dollar”)
d. None of the above

Explanation: Since the in-degree of the Canadian dollar is zero, in PPR, we never reach this node, for the prob will be zero.