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 $O(n\beta)$, 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

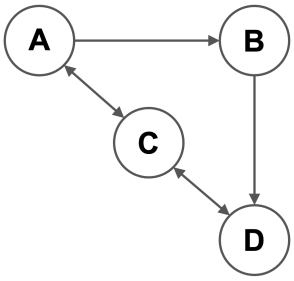
a. CBOW
b. Fasttext
c. Glove (CORRECT)
d. LSI

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.

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

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.

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

Question 6

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

Distributed Information Systems: Spring Semester 2018 - Quiz 1

Student Name: _____
Date: May 18 2018
Student ID: _____
Total number of questions: 8
Each question has a single answer!

_____

1. **Data being classified as unstructured or structured depends on the:**
   - **A. Degree of abstraction**
   - B. Level of human involvement
   - C. Type of physical storage
   - D. Amount of data

2. **Which of the following is an advantage of Vector Space Retrieval model?**
   - A. No theoretical justification is needed why the model works
   - B. Produces provably correct query results
   - **C. Enables ranking of query results according to cosine similarity function**
   - D. Allows to retrieve documents that do not contain any of the query terms

3. **Which of the following is *true*?**
   - A. High precision implies low recall
   - **B. High precision hurts recall**
   - **C. High recall hurts precision**
   - D. High recall implies low precisions

Comment: C) was the intended answer but B is true in some cases. Assume that there are 100 people, 3 of them has cancer (positive) and 97 of them has not. You can force your classification algorithm to find all 3 people with cancer, but in doing so it could also misclassify 1 person as he has cancer as well, so you get %96 precision but %100 recall. Or you can just say everyone is healthy, which awards you with %97 precision but %0 recall (congratz)

4. **Recall can be defined as:**
   - A. P(relevant documents | retrieved documents)
   - **B. P(retrieved documents | relevant documents)**
   - C. P(retrieved documents | number of documents)
   - D. P(relevant documents | number of documents)

5. **Thang, Jeremie and Tugrulcan have built their own search engines. For a query Q, they got precision scores of 0.6, 0.7, 0.8 respectively. Their F1 scores (calculated by same parameters) are same. Whose search engine has a higher recall on Q?**

A. Thang
B. Jeremie
C. Tugrulcan
D. We need more information

Comment: The intended answer was A) Thang. However, the question intended to say that their F1 scores as same as each other, but some students thought their F1 scores are same as their precision scores, which rendered the answer C.

6. **The number of non-zero entries in a column of a term-document matrix indicates:**
   A. how many terms of the vocabulary a document contains
   B. how often a term of the vocabulary occurs in a document
   C. how relevant a term is for a document
   D. none of the other responses is correct

7. **Which one of the following is *wrong*. Schema mapping is used to:**
   A. Overcome semantic heterogeneity
   B. Reconcile different logical representations of the same domain
   C. Optimize the processing of queries
   D. Support schema evolution of databases

8. **In a Ranked Retrieval result, the result at position k is non-relevant and at k+1 is relevant. Which of the following is *always* true (P@k and R@k are the precision and recall of the result set consisting of the k top ranked documents)?**
   A. P@k-1 > P@k+1
   B. P@k-1 = P@k+1
   C. R@k-1 < R@k+
   D. R@k-1 = R@k+1

Distributed Information Systems: Spring Semester 2018 - Quiz 2

Student Name: _____
Date: March 22 2018
Student ID: _____
Total number of questions: XXX
Each question has a single answer!

1. Which of the following is TRUE when comparing Vector Space Model (VSM) and Probabilistic Language Model (PLM)? (Slide 73 Week 2)
   - ☐ a. Both VSM and PLM require parameter tuning
   - ☐ b. Both VSM and PLM use collection frequency in the model
   - ☐ **c. Both VSM and PLM take into account multiple term occurrences**
   - ☐ d. Both VSM and PLM are based on a generative language model

2. Which of the following is WRONG about inverted files? (Slide 24,28 Week 3)
   - ☐ a. The space requirement for the postings file is O(n)
   - ☐ **b. Variable length compression is used to reduce the size of the index file**
   - ☐ c. The index file has space requirement of O(n^beta), where beta is about ½
   - ☐ d. Storing differences among word addresses reduces the size of the postings file

3. The SMART algorithm for query relevance feedback modifies? (Slide 11 Week 3)
   - ☐ a. The original document weight vectors
   - ☐ **b. The original query weight vectors**
   - ☐ c. The result document weight vectors
   - ☐ d. The keywords of the original user query

4. What is the benefit of LDA over LSI?
   - ☐ a. LSI is sensitive to the ordering of the words in a document, whereas LDA is not
   - ☐ **b. LDA has better theoretical explanation, and its empirical results are in general better than LSI's**
   - ☐ c. LSI is based on a model of how documents are generated, whereas LDA is not
   - ☐ d. LDA represents semantic dimensions (topics, concepts) as weighted combinations of terms, whereas LSI does not

5. How does LSI querying work?
   - ☐ a. The query vector is treated as an additional term; then cosine similarity is computed

☐ b. The query vector is transformed by Matrix S; then cosine similarity is computed

☐ **c. The query vector is treated as an additional document; then cosine similarity is computed**

☐ d. The query vector is multiplied with an orthonormal matrix; then cosine similarity is computed

6. In general, what is true regarding Fagin's algorithm?
   ☐ a. It performs a complete scan over the posting files
   ☐ **b. It provably returns the k documents with the largest aggregate scores**
   ☐ c. Posting files need to be indexed by the TF-IDF weights
   ☐ d. It never reads more than $(k\ n)^{½}$ entries from a posting list

7. Tugrulcan wanted to plan his next summer vacation so he wrote “best beaches” to his favourite search engine. Little did he know, his favourite search engine was using pseudo-relevance feedback and the top-k documents that are considered relevant were about the beaches only in Turkey. What is this phenomenon called?
   ☐ a. Query Bias
   ☐ b. Query Confounding
   ☐ **c. Query Drift**
   ☐ d. Query Malfunction

8. Which of the following statements about index merging (when constructing inverted files) is correct?
   ☐ **a. While merging two partial indices on disk, the inverted lists of a term are concatenated without sorting**
   ☐ b. Index merging is used when the vocabulary does no longer fit into the main memory
   ☐ c. The size of the final merged index file is O (n log2 (n) M)), where M is the size of the available memory
   ☐ d. While merging two partial indices on disk, the vocabularies are concatenated without sorting

Distributed Information Systems: Spring Semester 2018 - Quiz 3

Student Name: _____
Date: April 12 2018
Student ID: _____
Total number of questions: XXX
Each question has a single answer (!)

1. When representing the adjacency list of a Web page in a connectivity server by using a reference list from another Web page, the reference list is searched only in a neighbouring window of the Web page's URL, because:
   - ☐ a. subsequent URLs in an adjacency list have typically small differences
   - ☐ b. typically many URLs in a web page are similar to each other
   - ☐ **c. often many URLs among two pages with similar URL are similar**
   - ☐ d. most extra nodes are found in the neighbouring window.

2. When constructing a word embedding, negative samples are
   - ☐ **a. word - context word combinations that are not occurring in the document collection**
   - ☐ b. context words that are not part of the vocabulary of the document collection
   - ☐ c. all less frequent words that do not occur in the context of a given word
   - ☐ d. only words that never appear as context word

3. Which of the following statements on Latent Semantic Indexing (LSI) and Word Embeddings (WE) is correct
   - ☐ **a. LSI is deterministic (given the dimension), whereas WE is not**
   - ☐ **b. LSI does not take into account the order of words in the document, whereas WE does**
   - ☐ **c. The dimensions of LSI can be interpreted as concepts, whereas those of WE cannot**
   - ☐ d. LSI does take into account the frequency of words in the documents, whereas WE does not.

Comment: The question intended to ask which one is incorrect :)

4. Given the following list of transactions: {apple,milk}, {milk, bread}, {apple, bread, milk}, {bread}
   - ☐ a. milk -> apple has support 1/2 and confidence 1
   - ☐ b. milk -> bread has support 1/2 and confidence 1
   - ☐ c. bread -> milk has support 1/2 and confidence 1
   - ☐ **d. apple -> milk has support 1/2 and confidence 1**

5. Given the 2-itemsets {1,2}, {1,5}, {2,5}, {1,4}, {1,3}, when generating the 3-itemsets we will…
   - ☐ a. Generate 5 3-itemsets after the join and 2 3-itemsets after the prune
   - ☐ **b. Generate 6 3-itemsets after the join and 1 3-itemsets after the prune**
   - ☐ c. Generate 4 3-itemsets after the join and 1 3-itemsets after the prune
   - ☐ d. Generate 4 3-itemsets after the join and 2 3-itemsets after the prune

6. Given the following teleporting matrix (E) for nodes A, B and C:

   [0 ½ 0]
   [0 0 0]
   [0 ½ 1]

   and making no assumptions about the link matrix (R), which of the following is **correct**:

   - ☐ a. A random walker can never reach node A
   - ☐ b. A random walker can never leave node A
   - ☐ c. A random walker can always leave node C
   - ☐ **d. A random walker can always leave node B**

   Reminder: columns are the probabilities to leave the respective node.

7. When computing PageRank iteratively, the computation ends when:
   - ☐ **a. The norm of the difference of rank vectors of two subsequent iterations falls below a predefined threshold**
   - ☐ b. The difference among the eigenvalues of two subsequent iterations falls below a predefined threshold
   - ☐ c. All nodes of the graph have been visited at least once
   - ☐ d. The probability of visiting an unseen node falls below a predefined threshold

8. For his awesome research, Tugrulcan is going to use the Pagerank with teleportation and HITS algorithm, not on a network of webpages but on the retweet network of Twitter! The retweet network is a directed graph, where nodes are users and an edge going out from a user A and to a user B means that "User A retweeted User B". Which one is FALSE about a Twitter bot that retweeted other users frequently but got never retweeted by other users or by itself?
   - ☐ a. It will have a non-zero hub value.
   - ☐ b. It will have an authority value of zero.
   - ☐ **c. It will have a pagerank of zero.**
   - ☐ **d. Its authority value will be equal to the hub value of a user who never retweets other users.**

Comment: The intended answer was C. There is no general rule regarding the self-links as the paper proposing the HITS algorithm did not specify it, so we also accepted d.

Distributed Information Systems: Spring Semester 2018 - Quiz 4

Student Name: _____
Date: April 26 2018
Student ID: _____
Total number of questions: 8
Each question may have a single answer!

1. **In a FP tree, the leaf nodes are the ones with:**
    a. Lowest confidence
    b. Lowest support
    c. Least in the alphabetical order
    d. None of the above

Comment: The intended (and to me, the only correct) answer was b. However it is also possible to say that the leaf nodes are the ones with lowest frequency, that's why we also accepted d.

2. **Suppose that an item in a leaf node N exists in every path. Which one is correct?**
    a. N co-occurs with its prefix in every transaction.
    b. For every node p that is a parent of N in the fp tree, confidence(p->n) = 1
    c. N's minimum possible support is equal to the number of paths.
    d. The item N exists in every candidate set.

3. **Fundamentally, why clustering is considered an unsupervised machine learning technique?**
    a. Number of clusters are not known.
    b. The class labels are not known.
    c. The features are not known.
    d. The clusters can be different with different initial parameters.

4. **Which of the following is true for a density based cluster C:**
    a. Any two points in C must be density reachable. Each point belongs to one, and only one cluster
    b. Any two points in C must be density reachable. Border points may belong to more than one cluster
    c. Any two points in C must be density connected. Border points may belong to more than one cluster
    d. Any two points in C must be density connected. Each point belongs to one, and only one cluster

5. **Suppose that q is density reachable <u>from</u> p. The chain of points that ensure this relationship are {t,u,g,r} Which one is <u>FALSE</u>?**
    a. {t,u,g,r} have to be all core points.
    b. p and q will also be density-connected
    c. p has to be a core point
    d. q has to be a border point

6. **What is a correct pruning strategy for decision tree induction?**
    a. Apply Maximum Description Length principle
    b. Stop partitioning a node when either positive or negative samples dominate the samples of the other class
    c. Choose the model that maximizes L(M) + L(M|D)
    d. Remove attributes with lowest information gain

7. **Given the distribution of positive and negative samples for attributes A1 and A2, which is the best attribute for splitting?**

A1	P	N
a	7	0
b	1	4
A2	P	N
x	5	1
y	3	3

    a. A1
    b. A2
    c. They are the same
    d. There is not enough information to answer the question

You can use this website to compute the binary entropy:
http://www.wolframalpha.com/widgets/view.jsp?id=4e095a8fa96257fbf9da2529e930ccd3

8. **When using bootstrapping in Random Forests, the number of different data items used to construct a single tree is:**
    a. smaller than the size of the training data set, with high probability
    b. of order square root of the size of the training set, with high probability
    c. the same as the size of the training data set
    d. subject to the outcome of the sampling process, and can be both smaller or larger than the training set

Distributed Information Systems: Spring Semester 2018 - Quiz 5

Student Name: _____
Date: May 17 2018
Student ID: _____
Total number of questions: 8
Each question hopefully has a single answer!

1. Which of the following is **correct** regarding *Louvain* algorithm?
   - ☐ a. It creates a hierarchy of communities with a common root
   - ☐ b. *Clique* is the only topology of nodes where the algorithm detects the same communities, independently of the starting point
   - ☐ **c. If *n* cliques of the same order are connected cyclically with *n-1* edges, then the algorithm will always detect the same communities, independently of the starting point**
   - ☐ d. Modularity is always maximal for the communities found at the top level of the community hierarchy

2. In *User-Based Collaborative Filtering*, which of the following is **correct,** assuming that all the ratings are positive?
   - ☐ a. *Pearson Correlation Coefficient* and *Cosine Similarity* have different value range, but return the same similarity ranking for the users
   - ☐ **b. If the ratings of two users have both variance equal to 0, then their *Cosine Similarity* is maximized**
   - ☐ c. *Pearson Correlation Coefficient* and *Cosine Similarity* have the same value range, but can return different similarity ranking for the users
   - ☐ d. If the variance of the ratings of one of the users is 0, then their *Cosine Similarity* is not computable

3. Which of the following is **correct** regarding *Crowdsourcing*?
   - ☐ a. *Random Spammers* give always the same answer for every question
   - ☐ b. It is applicable only for binary classification problems
   - ☐ c. *Honey Pot* discovers all the types of spammers but not the sloppy workers
   - ☐ d. **The output of *Majority Decision* can be equal to the one of *Expectation-Maximization***

4. Which of the following is **correct** regarding prediction models?
   - ☐ a. Training error being less than test error means overfitting
   - ☐ b. Training error being less than test error means underfitting
   - ☐ **c. Complex models tend to overfit, unless we feed them with more data**
   - ☐ d. Simple models have lower bias than complex models

5. In the χ^2 statistics for a binary feature, we obtain $P(\chi^2 \mid DF = 1) > 0.05$. This means in this case, it is assumed:
   - ☐ 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. None of the above

6. Which is an appropriate method for fighting skewed distributions of class labels in classification?
   - ☐ a. Include an over-proportional number of samples from the larger class
   - ☐ b. Use leave-one-out cross validation
   - ☐ **c. Construct the validation set such that the class label distribution approximately matches the global distribution of the class labels**
   - ☐ d. Generate artificial data points for the most frequent classes

		Class	
		Fraud	¬Fraud
Classified	Fraud	20	20
	¬Fraud	10	60

7. Considering the results of this fraud classifier, which of the following is **correct**?
   - ☐ **a. The classifier has a precision of 50% and a recall of 66.6%**
   - ☐ b. The classifier has a precision of 75% and a recall of 50%
   - ☐ c. The classifier has a precision of 50% and a recall of 75%
   - ☐ d. The classifier has a precision of 66.6% and a recall of 75%

The following question is cancelled:

8. Which of the following is **correct** regarding community detection?
   - ☐ a. High betweenness of an edge indicates that the communities are well connected by that edge
   - ☐ b. The *Louvain* algorithm attempts to minimize the overall modularity measure of a community graph
   - ☐ c. 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
   - ☐ d. The *Girvan-Newman* algorithm attempts to maximize the overall betweenness measure of a community graph

Comment: a) Is not correct, yes an edge with high betweenness connects two communities and these two communities will be distinct and won't be connected by many edges. But these communities are not "well-connected" and also answer does not specificy which communities are that edge connects. (Though I must admit this sentence is not well-formed.) b) and d) are not correct.

c) was the intended answer. Suppose that we take one of the community detected by Louvain algorithm. Then randomize the edges of the whole graph, which we call null model, and take the nodes which belong to the community we found. The community extracted from the original graph would have more edges between the nodes that belong to that community, then the community extracted from the null model. However in c, we accidentally meant the whole graph by the null model, which has the same number of edges as the original graph, which would have more edges than a community.

Distributed Information Systems: Spring Semester 2018 - Quiz 6

Student Name: _____
Date: May 18 2018
Student ID: _____
Total number of questions: 8
Each question really has a single answer!

1. Which of the following is **wrong** regarding *Ontologies*?
   - ☐ a. We can create more than one ontology that conceptualize the same real-world entities
   - ☐ b. Ontologies help in the integration of data expressed in different models
   - ☐ c. Ontologies support domain-specific vocabularies
   - ☐ d. **Ontologies dictate how semi-structured data are serialized**

2. Which of the following is **correct** regarding the use of *Hidden Markov Models (HMMs)* for entity recognition in text documents?
   - ☐ a. HMMs cannot predict the label of a word that appears only in the test set
   - ☐ b. If the smoothing parameter *λ* is equal to *1*, the emission probabilities for all the words in the test set will be equal
   - ☐ **c. When computing the emission probabilities, a word can be replaced by a morphological feature (e.g., the number of uppercase first characters)**
   - ☐ d. The label of one word is predicted based on all the previous labels

3. Which of the following statements is **correct** concerning the use of *Pearson's Correlation* for user-based collaborative filtering?
   - ☐ **a. It measures whether different users have similar preferences for the same items**
   - ☐ b. It measures how much a user's ratings deviate from the average ratings
   - ☐ c. It measures how well the recommendations match the user's preferences
   - ☐ d. It measures whether a user has similar preferences for different items

4. Which of the following statements is **wrong** regarding RDF?
   - ☐ a. An RDF statement would be expressed in SQL as a tuple in a table
   - ☐ **b. Blank nodes in RDF graphs correspond to the special value NULL in SQL**
   - ☐ c. The object value of a type statement corresponds to a table name in SQL
   - ☐ d. RDF graphs can be encoded as SQL databases

5. Which of the following statements is **correct** in the context of information extraction?
   - ☐ **a. A confidence measure that prunes too permissive patterns discovered with bootstrapping can help reducing semantic drift**
   - ☐ b. The bootstrapping technique requires a dataset where statements are labelled
   - ☐ c. Distant supervision typically uses low-complexity features only, due to the lack of training data
   - ☐ d. For supervised learning, sentences in which NER has detected no entities are used as negative samples

6. Which of the following is **correct** regarding schemas and ontologies?

- ☐ a. An ontology is created from constructing mappings between schemas
- ☐ **b. Ontologies can be used for reasoning about different schemas**
- ☐ c. Ontologies always require a schema
- ☐ d. Semi-structured data cannot have a schema

7. *Dude said "I like bowling"*. With how many statements can we express this sentence using *RDF Reification*?
   - ☐ a. We cannot
   - ☐ b. 1
   - ☐ c. 3
   - ☐ **d. 5**

8. Modularity of a social network always:
   - ☐ a. Increases with the number of communities
   - ☐ b. **Increases when an edge is added between two members of the same community**
   - ☐ c. Decreases when new nodes are added to the social network that form their own communities
   - ☐ d. Decreases if an edge is removed

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

- [x] ~~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
- [x] ~~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
- [x] ~~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?

- [x] ~~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.
- [x] ~~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
- [x] ~~Ontologies dictate how semi-structured data are serialized~~
- [ ] Ontologies support domain-specific vocabularies