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