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.