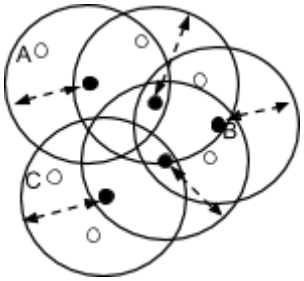
Distributed Information Systems: Spring Semester 2017 - Quiz 3

Student Name: _____
Date: April 13 2017
Student ID: _____
Total number of questions: 8
Each question has a single answer!

_____

1. Which of the following statements is **false** for k-means algorithm:
   - ☐ a. The number of clusters needs to be known in advance
   - ☐ **b. If a point is randomly selected to be a centroid of a cluster in the initial round, then the point remains part of that cluster**
   - ☐ c. Initial partitioning of the data can be random, following iterations iteratively assign points to the closest centroids
   - ☐ d. Converges fast, but often it may terminate at a local optimum

2. According to the following figure. Black points are core points, white points are border points, circles represent the neighborhood covered by core points, the dotted lines represent the radius of the neighborhood.



   - ☐ a. A is density reachable from B. B is density reachable from A
   - ☐ b. B is density reachable from C. A is density reachable from B
   - ☐ c. B is density reachable from A. B is density reachable from C
   - ☐ **d. A is density reachable from B. C is density reachable from B**

3. Which of the following is **true** for a density based cluster C:
   - ☐ a. Any two points in C must be density reachable. The set of clusters is unique
   - ☐ b. The set of clusters is unique. Each point belongs to one, and only 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

4. Which of the following methods used in advanced information retrieval is sensitive to the ordering of the words in a document:
   - ☐ a. Latent Dirichlet Allocation
   - ☐ b. Latent Semantic Indexing
   - ☐ **c. Word Embeddings**
   - ☐ d. SMART relevance feedback algorithm

   because it uses context information

5. When obtaining negative samples, if p_w is the probability of word *w* in collection, why are they (in practice) sampled with p_w^{alpha} with alpha < 1?
   - ☐ a. To favor more frequent words and increase the quality of the model
   - ☐ **b. To sample infrequent words more often**
   - ☐ c. To get in average the same number of samples for frequent and infrequent words
   - ☐ d. To increase the number of words that are sampled

6. What is the effect of the parameter s, the number of singular values retained in LSI?
   - ☐ **a. A larger value of s means that a more precise approximation of the term-document matrix is used in the construction of the concept space**
   - ☐ b. A larger s means that the resulting concept space has lower dimension.
   - ☐ c. A larger s means that the number of results for a query will be larger
   - ☐ d. A larger s means that the vocabulary considered in the construction of the concept space is larger.

7. How does LSI querying work?
   - ☐ a. The query vector is treated as an additional term. Then cosine similarity is computed
   - ☐ b. Matrix S changes depending on the query vector, apply this transformation to S. Then cosine similarity is computed
   - ☐ **c. The query vector is treated as an additional document. Then cosine similarity is computed**
   - ☐ d. Depending on the situation, query vector can be treated as an additional document or as an additional term. Then cosine similarity is computed

8. What is the benefit of LDA over LSI?
   - ☐ a. LSI's empirical results are in general better than LDA's
   - ☐ **b. LDA has better theoretical foundations, 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.