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!

_____

- Collection Frequency, cf
  - Define: The total number of occurences of the term in the entire corpus
- Document Frequency, df
  - Define: The total number of documents which contain the term in the corpus

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