*Distributed Information Systems: Spring Semester 2015*

**Quiz 3: Vector Space Model + Advanced Retrieval Models**

Student Name: _____ Date: 26 Mar 2015

Student ID: _____ Time: 11:15AM to 11:30AM

*Total number of questions: 8*

**Each question has a single answer!**

---

Let T be the vocabulary and D the set of documents, defined as

$$T = \{race, biology, formula, chemistry, health\}$$

$$\begin{aligned} D = \{&d1 = (biology, chemistry, biology, chemistry, health),\\ &d2 = (formula, chemistry, formula, formula),\\ &d3 = (race, formula),\\ &d4 = (biology, biology, biology),\\ &d5 = (health, chemistry, health, health, health)\} \end{aligned}$$

The idf values are given by the following table:

| term | idf |
|---|---|
| chemistry | 0.51 |
| biology | 0.92 |
| formula | 0.92 |
| health | 0.92 |
| race | 1.61 |

1. What is the ranked result set for query (biology, chemistry) when applying the tf-idf method?

☒ *a*) (d1, d4, d2, d5, d3)
☐ *b*) (d4, d1, d2, d5, d3)
☐ *c*) (d1, d5, d4, d2, d3)
☐ *d*) (d4, d1, d5, d2, d3)

2. Assume that according to the user, documents d1 and d4 are relevant for the result of a certain query. Which one among the following result sets has precision 33% and recall 50%?

☐ *a*) (d3, d2, d5) recall 0
☐ *b*) (d1, d4, d2) recall 1
☐ *c*) (d4, d1, d5) recall 1
☒ *d*) (d3, d4, d5) d4 / d3 + d4 + d5

---

3. Which of the following is **true**?

☐ *a*) The term frequency is normalized with respect to the maximal frequency of all terms occurring within the whole document collection. not whole, but current document
☐ *b*) Stop words, that typically occur in all the documents of a collection, are better removed at the beginning because their term frequencies will normally be 0. are better removed but do not normally have a tf of 0
☒ *c*) The vector model with tf-idf weights assumes independence of index terms.
☐ *d*) The Boolean retrieval model does not match documents unless they contain all keywords appearing in the query.

4. Let V1 and V2 be binary feature vectors with their respective norm greater than 0. If $sim(V1, V2) = 0$, which statement is **wrong**?

   ☒ *a*) Each feature that is 0 in V1 is always 1 in V2. wrong. some elements could be 0 in both v1 v2

   ☐ *b*) The two vectors are orthogonal.

   ☐ *c*) The two vectors have no common entry that is 1 for both.

   ☐ *d*) If V1 is a query, then the document represented by V2 would not be returned in the response.

5. Which statement regarding Vector Space retrieval (VS) and Latent Semantic Indexing (LSI) is **true**?

   ☐ *a*) Like in VS, LSI maps into the same space both documents and queries.

   ☐ *b*) Differently to VS, LSI handles synonymy.

   ☐ *c*) Differently to VS, LSI is a dimensionality reduction method.

   ☒ *d*) All of the three statements are true.

6. Which statement about Single Value Decomposition (SVD) is **true**?

   ☐ *a*) Only a square matrix can be decomposed using SVD.

   ☐ *b*) The eigenvector decomposition and SVD always return the same result.

   ☒ *c*) The singular values matrix is a diagonal matrix.

   ☐ *d*) The singular values matrix is nonnegative.

7. Which of the following is **wrong** in the context of the Rochio method used to account for relevance feedback?

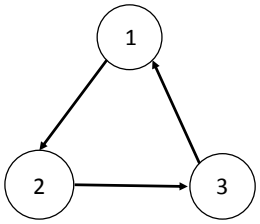
   ☐ *a*) The revised query might contain terms that were not in the original query. q = "a", d = "a,b", q_rev = "a,b"

   ☐ *b*) The revised query might put a weight of 0 on terms that were present in the original query.

   ☐ *c*) The revised query might be the same as the original query.

   ☒ *d*) The terms in the original query will always remain in the revised query, but with different weights.

8. Consider a 3 node graph that forms a directed circle, such as the one below. What is **true** about the page rank of the node 3?




   ☐ *a*) It changes if the jump parameter $q$ changes.

   ☐ *b*) It is undefined, as the algorithm never converges.

   ☐ *c*) It is equal to 1.

   ☒ *d*) None of them is true.