Distributed Information Systems: Spring Semester 2017 - Quiz 5

Student Name: _____
Date: May 18 2017
Student ID: _____
Total number of questions: 8
Each question has a single answer!

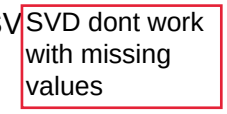
_____

1. When you have a small corpus, what is the **best** practice for Document Classification using Word Embeddings (WE):
   - ☐ a. Train WE using the *test set* in order to cover every word in the vocabulary of the test documents
   - ☐ b. Train WE using the *train set* in order to cover every word in the vocabulary of the train documents
   - ☐ c. Train WE using samples from both the *train* and the *test set*
   - ☐ **d. Use pre-trained WE which may not cover the full vocabulary of the *train* and the *test set***

2. In User-Based Collaborative Filtering, which of the following is **true**:
   - ☐ a. *Pearson Correlation Coefficient* and *Cosine Similarity* have different value range, but return the same similarity ranking for the users
   - ☐ **b. If the variance of the ratings of two dissimilar users is 0, then their *Cosine Similarity* is maximized**
   - ☐ c. *Pearson Correlation Coefficient* and *Cosine Similarity* have the same value range, but can return different similarity ranking for the users
   - ☐ d. If the variance of the ratings of one of the users is 0, then their *Cosine Similarity* is not computable

3. Which of the following is **true** for the Recommender Systems (RS):
   - ☐ a. In *Matrix Factorization* we can decompose the user-item matrix using SV
   - ☐ **b. The complexity of the *Content-based RS* does not depend on the number of users**
   - ☐ c. *Item-based RS* need not only the ratings but also the item features
   - ☐ d. User's age can be a useful feature in *User-based RS*

4. Which of the following is **true** for the *Fastext classifier*:
   - ☐ a. It uses word n-grams in order to create feature vectors for unseen words
   - ☐ b. It can create feature vectors only for the words of its vocabulary
   - ☐ **c. Word n-grams are used to capture the meaning of whole phrases**
   - ☐ d. Character n-grams are useful for languages with one-character words or contractions (e.g., French)



5.

```
<?xml version="1.0" encoding="UTF-8"?>
<people>
    <person>
      <name>Bryan Mills</name>
      <birthdate>June 15, 1957</birthdate>
    </person>
    <person>
<name>Kimberly Mills</name>
    </person>
</people>
```

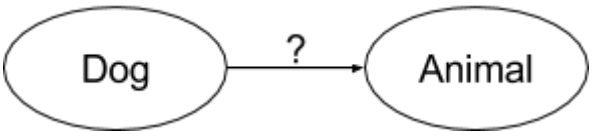
This XML document is:

- ☐ **a. well-formed**
- ☐ b. not well-formed
- ☐ c.well-formed but not valid
- ☐ d. unstructured

6. Which of the following is **true** for Ontologies:
   - ☐ a. They help in the integration of data expressed in different encodings
   - ☐ **b. They help in the integration of data expressed in different models**
   - ☐ c.They do not support domain-specific vocabularies
   - ☐ d. They dictate how semi-structured data are serialized

7. "John said that *Liam Neeson* stars in *Taken*". With how many statements we can express this sentence using *RDF Reification*:
   - ☐ a. We cannot
   - ☐ b. 1
   - ☐ c.3
   - ☐ **d. 5**

8.



What is the **appropriate** property to connect these two classes:

- ☐ a. partOf
- ☐ b. superClassOf
- ☐ **c. subClassOf**
- ☐ d. domainOf