Quiz - 2021

1. Information extraction:

- □ Necessarily requires training data.
- □ Is used to identify characteristic entities in a document.
- □ Is always bootstrapped by using ontologies.
- ☑ ~~Can be used to populate ontologies.~~

2. What is **TRUE** regarding Fagin's algorithm?

- □ Posting files need to be indexed by TF-IDF weights
- □ It performs a complete scan over the posting files
- □ It never reads more than (kn)½ entries from a posting list
- ☑ ~~It provably returns the k documents with the largest aggregate scores~~

3. Which of the following statements on Latent Semantic Indexing (LSI) and Word Embeddings (WE) is false?

- □ The dimensions of LSI can be interpreted as concepts, whereas those of WE cannot
- □ LSI does not depend on the order of words in the document, whereas WE does
- □ LSI is deterministic (given the dimension), whereas WE is not
- ☑ ~~LSI does take into account the frequency of words in the documents, whereas WE with negative sampling does not~~

4. When constructing a word embedding, what is **TRUE** regarding negative samples?

- ☑ ~~They are oversampled if less frequent~~
- □ Their frequency is decreased down to its logarithm
- □ They are words that do not appear as context words
- □ They are selected among words that are not stop-words

5. A page that points to all other pages but is not pointed by any other page would have:

- □ Nonzero authority
- □ Zero hub
- ☑ ~~Nonzero PageRank~~
- □ None of the above

6. When computing PageRank iteratively, the computation ends when:

- □ The difference among the eigenvalues of two subsequent iterations falls below a predefined threshold

- ☑ ~~The norm of the difference of rank vectors of two subsequent iterations falls below a predefined threshold~~
- □ The probability of visiting an unseen node falls below a predefined threshold
- □ All nodes of the graph have been visited at least once

7. In Ranked Retrieval, the result at position k is non-relevant and at k+1 is relevant. Which of the following is always true?
Hint: P@k and R@k are the precision and recall of the result set consisting of the k top-ranked documents.

- □ P@k-1>P@k+1
- □ R@k-1=R@k+1
- ☑ ~~R@k-1<R@k+1~~
- □ P@k-1=P@k+1

8. Which of the following is **TRUE** regarding community detection?

- ☑ ~~The high betweenness of an edge indicates that the communities are well connected by that edge~~
- □ The Girvan-Newman algorithm attempts to maximize the overall betweenness measure of a community graph
- □ The high modularity of a community indicates a large difference between the number of edges of the community and the number of edges of a null model
- □ The Louvain algorithm attempts to minimize the overall modularity measure of a community graph

9. What is **WRONG** regarding the Transformer model?

- ☑ ~~Its computation cannot be parallelized compared to LSTMs and other sequential models.~~
- □ It uses a self-attention mechanism to compute representations of the input and output.
- □ Its complexity is quadratic to the input size.
- □ It captures the semantic context of the input.

10. In User-Based Collaborative Filtering, which of the following is **TRUE**?

- □ Pearson Correlation Coefficient and Cosine Similarity have the same value range and return the same similarity ranking for the users.
- ☑ ~~Pearson Correlation Coefficient and Cosine Similarity have different value ranges and can return different similarity rankings for the users~~
- □ Pearson Correlation Coefficient and Cosine Similarity have different value ranges, but return the same similarity ranking for the users
- □ Pearson Correlation Coefficient and Cosine Similarity have the same value range but can return different similarity rankings for the users

11. Which of the following is **TRUE** for Recommender Systems (RS)?

- □ The complexity of the Content-based RS depends on the number of users
- □ Item-based RS need not only the ratings but also the item features
- □ Matrix Factorization is typically robust to the cold-start problem.
- ☑ ~~Matrix Factorization can predict a score for any user-item combination in the dataset.~~

12. Considering the transaction below, which one is **WRONG**?

Transaction ID	Items Bought
1	Tea
2	Tea, Yoghurt
3	Tea, Yoghurt, Kebap
4	Kebap
5	Tea, Kebap

- □ {Yoghurt} -> {Kebab} has 50% confidence
- □ {Yoghurt, Kebap} has 20% support
- □ {Tea} has the highest support
- ☑ ~~{Yoghurt} has the lowest support among all itemsets~~

13. Suppose that in a given FP Tree, an item in a leaf node N exists in every path. Which of the following is **TRUE**?

- □ N co-occurs with its prefixes in every transaction
- □ For every node P that is a parent of N in the FP tree, confidence (P->N) = 1
- ☑ ~~{N}'s minimum possible support is equal to the number of paths~~
- □ The item N exists in every candidate set

14. Which of the following properties is part of the RDF Schema Language?

- □ Description
- □ Type
- □ Predicate
- ☑ ~~Domain~~

15. Which of the following is wrong regarding Ontologies?

- □ We can create more than one ontology that conceptualizes the same real-world entities
- □ Ontologies help in the integration of data expressed in different models
- ☑ ~~Ontologies dictate how semi-structured data are serialized~~
- □ Ontologies support domain-specific vocabularies