Final Exam for the Course Web Data Processing Systems 2018/2019

Instructions

The questions in this exam sum up to 100 points. The grade will be converted on a decimal scale. Answers must be given in English. No external material can be consulted.

Questions

1. You are asked to extract all sentences that mention a list of products from a 50TB+ collection of documents and to group the sentences by product. Which frameworks/technologies would you use for this task? [**5 pts**] Sketch the pseudocode of a potential implementation of your proposed solution. [**5 pts**]

2. Describe the difference between manifest and latent knowledge. Report an example of knowledge of both types. [**8 pts**]

3. Consider a graph-based representation of a RDF knowledge base. What do nodes and edges represent in such a graph? [**6 pts**] Report a small example of RDF triples and draw the corresponding graph. [**3 pts**]

4. Describe the main operations in a standard NLP pipeline. Which are the most important tasks that we typically perform before we can further process the data? [**4 pts**] What is an example of a framework that we can use for this purpose? [**3 pts**]

5. Describe three main operations that are necessary for performing Entity Linking, and one possible implementation of each of them. [**10 pts**]

6. Let $m_1, m_2, \ldots, m_n$ be a number of entity mentions discovered in a Web page and be $E_{m_1}, E_{m_2}, \ldots, E_{m_n}$ sets of potential entity matches. What is the coherence and how can we measure it? [**10 pts**]

7. What do the embeddings represent in the *word2vec* model and how can we train them? [**7 pts**] What is the difference between the Skip-Gram and the CBOW models? [**5 pts**]

8. What is the main difference between RESCAL and TranSE? [**6 pts**] Which one would you pick and why? [**3 pts**] Describe two limitations of both approaches. [**5 pts**]

9. Describe the process of materializing a knowledge graph with a set of rules and report the formal definition. [**8 pts**] How can we parallelize the execution of the rules? [**6 pts**]

10. Describe one method to determine whether a user is exposing some privacy-sensitive information on the Web. Suppose such method has discovered some problematic content (e.g., a post in a forum). Can you think of a procedure to change the content to reduce the exposure? If so, then describe it. [**6 pts**]