Final Exam for the Course Web Data Processing Systems 2016/2017.

Instructions

You must answer all questions. You get 10 free points. The questions in this exam sum up to the remaining 90 points. The grade will be converted on a decimal scale and rounded to the closest half unit (e.g. 1,1.5,2,2.5, etc.)

Questions marked with *(N)* are intended to evaluate your knowledge of notions discussed during this course. Questions marked with *(A)* are meant to measure your ability to analyze problems and compare different solutions.

Answers must be given in English. No external material can be consulted.

Questions

1. (N) What are the three main issues a crawler needs to face? **[3 pts]**. Can you describe one policy for revisiting web pages? **[3 pts]** Can you report the name of one implementation of a crawler that you would use if you would need to perform this task? **[4 pts]**

   Solution: First answer: freshness, quality and volume. Second answer: Selection policy (or others, see slide 16 of chp 2). Third answer: Apache Nutch (for instance)

2. (N) What is a knowledge base and the RDF model? **[10 pts]** Can you name three popular knowledge bases? **[5 pts]** (A) If you would need to pick one knowledge base for your work, which one would you take? Why? **[3 pts]**

   Solution: First answer: A repository of knowledge written in a machine-readable format. RDF: data model where knowledge is encoded as triples. Second answer: Wikidata, Babelnet, DBPedia. Third answer: It depends on the usecase: I would not consider Freebase because it is discontinued, DBPedia has a strong link to Wikipedia for ideal knowledge. Wikidata has very high accuracy since it is manually verified.

3. (A) What is the main difference between traditional information extraction and open information extraction? **[5 pts]**

   Solution: Traditional Information Extraction: Start from a known set of relations. Open Information Extraction: Extract information without prior information. (see slides)

4. (N) Report at least two examples of context-independent features and context-dependent features **[5 pts]**

   Solution: Context-Independent Features: Exact matching or Hamming Distance between two strings, Entity type as returned by the NER. Context-Dependent: Bag-of-words, links in the webpage, named entities.

5. (N) Report the definition of the F1 measure and briefly discuss why it is so frequently used [**5 pts**]

   Solution: The F1 is the armonic mean of precision and recall (2*prec*recall/(precision+recall). You should know this since you have used it in the assignment. It is widely used because it considers both the precision and the recall so it offers a broad view of the performance of the system.

6. (N) Present two systems to perform Entity Linking. Briefly discuss what is the main idea behind them, their features and disadvantages. [**10 pts**]

   Solution: We discussed AIDA and DBPedia Spotlight. One uses YAGO2 and relies on a graph-based approach, while the other relies on VSM for disambiguation. A disadvantage of both systems is that they cannot link entities which are not yet in the knowledge base or entities that are not common. Moreover, YAGO2 has less coverage than DBPedia so this limits the performance. DBPedia Spotlight relies on VSM therefore it is not as interpretable as the AIDA.

7. (A) What do you believe is harder to perform: Word Sense Disambiguation or Entity Linking? Motivate in detail your answer. [**10 pts**]

   Solution: The two tasks encode different challenges. In Entity Linking, we have the additional problem of detecting entities which might be formed by multiple words. Moreover, the vocabulary used for entity linking is significantly larger than the few thousand senses that are available for WSD. However, WSD is particularly challenging on categories of words like verbs. Also, there is less training data for WSD than for Entity Linking. In general, WSD is considered an AI-complete problem, which means that a system that can solve WSD should encode strong AI. This makes WSD harder. However, I would also accept the argument that Entity Linking is harder due to the above-mentioned reasons, provided the argumentation is clearly stated.

8. (N) Describe the following concepts:

   - open and closed world assumption [**3 pts**]
   - local closed world assumption [**3 pts**]
   - gradient descent [**3 pts**]
   - tensors [**3 pts**]

   Solution: Open world assumption: What is not in the database is considered unknown. Closed world assumption: What is not considered in the database is considered false. Local world assumption: If there is a (s,p) pair in the database, then we assume we know all (s,p) pairs. Gradient Descent: Method to optimize a model given some training data. It calculates the gradient (vector with partial derivatives) and update the parameters to reduce the loss. Tensors: Multi-dimensional arrays. If the dimension is two then it is called a matrix.

9. (A) Suppose you want to execute a full materialization of some Datalog rules. Will the evaluation always terminate or can you come up with a set of rules for which the program will run forever? **[7 pts]**

   Solution: It will always terminate. Since rules are safe and the database contains a finite number of symbols, there is an upper bound to the possible derivations we can produce. So the materialization will be finite.

10. (A) Can you think of a possible way of how we could replicate reasoning with a statistical relational learning approach? **[8 pts]**

    Solution: One possible way to replicate reasoning through inference is to collect a very large amount of data so that the statistical model will make predictions with such high accuracy that could resemble a logical derivation. Another possibility would be to replicate the process of logically deducting the conclusions within the statistical model. For instance, by mining rules and then assigning confidence weights as explained in the lecture.