D5: Databases and Information Systems
Advanced Topics in IR, SS 2016
Dr. Vinay Setty and Dr. Jannik Strötgen
Assignment 4, due: **16 Jun 2016**

max planck institut informatik

Please submit your solution as a PDF and source code together in a single tar/zip file, name your pdf file in the following format: "Lastname-Immatrikulationnumber-AssigmentX.pdf", also include Immatrikulationnumber in the title of your document and email it to `atir16@mpi-inf.mpg.de` *before the due date mentioned above!*

Semantic Concepts (10 Points)

Problem 1. There are many terms referring to quite similar things that are related to semantics. However, they should not be used interchangeably.

- Please give precise definitions of the concepts of (i) a glossary, (ii) a lexicon, (iii) a taxonomy, and (iv) an ontology? Be careful: there might be different definitions of these concepts on the Web so take into account which sources you want to trust and cite them!
- Which of the concepts is the best way to represent the relations between the four concepts?

schema.org (20 Points)

Problem 2. In the lecture, we talked about URI and RDF. What we did not discuss in the lecture is schema.org.
a) Explain both, URI and RDF, in two sentences.
b) What is schema.org?
c) What are the main differences between RDF and schema.org? Also explain advantages and disadvantages of RDF and schema.org.

Semantic Search (Programming Assignment) (30 Points)

Problem 3.
In the previous assignments, you already indexed our sample corpus using elasticsearch. Now, we want to add "simple" semantic search functionality. For this, we will do Named Entity Recognition. To keep things simple, we will not perform Named Entity Normalization. The idea is as follows:

- go to http://nlp.stanford.edu/software/CRF-NER.shtml and get the NER tool;
- process the data with the Stanford NER tool (use the three class version);
- chose between one of the following options and briefly explain why you used the one option and not the other; (do not use any preprocessing such as stopword removal or lemmatization)
  - **either:** create an additional index for your corpus, which contains only the named entities that were extracted by the NER tool instead of all terms.
  - **or:** use elasticsearch's update API to update your existing index (`https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-update.html`); that is, let elasticseach add the named entities in addition to the text.

- you have already implemented query functionality, the plan is to now offer a second query functionality, namely that you can search for named entities; think about different strategies to realize this, and realize one of them for your own search engine.

Please also describe what you have to take care of when indexing the named entities (e.g., can you search for full names only? ...). What would you have to change if you performed named entity recognition and normalization instead of just recognition?
If you run into any issues: create a (very small) subset of the corpus. Try again, to process the documents with the Stanford NER tool. Provide a useful example query that returns some results on your subcorpus.
If you write some code to process the documents with the NER tool programmatically, please submit the code. You may also write a brief description of the things that you did. If you ran into any trouble, please describe what you tried and what problems you experienced. You still can get "serious" points although you might successfully develop the semantic search functionality!