Task Sheet 3
NLP, RDF, Pattern Matching

1. What is base- and stem-form reduction and how does it influence recall and precision of an IR system?
2. What methods can be used for stemming? What are their respective advantages / disadvantages?
3. What is under- and overstemming?
4. What is the meaning of the term *terminological control*?
5. Provide an overview of the most important problems regarding natural language that IR systems have to deal with!
6. What is an *n-gram*? Describe the advantages and disadvantages!
7. What is *RDF* and what is it used for? Also explain the terms *subject*, *predicate* and *object*. Why does this division into three exist?
8. Describe the difference between *word search* and *classical, exact pattern matching*!
9. An exemplary IR system is holding 500,000 documents (total size: 1.5 GB). On average, a document contains 100 different words. The ID of any document is 8 Byte big. Estimate the memory usage of the used inverted list!
10. What is *Zipf's Law*?