The University of Hong Kong
Faculty of Engineering

Department of Computer Science

COMP7306, Web technologies

Date:15 May 2013 Time: 6:30pm – 8:30pm

You have two hours to answer the questions of this test. No documents, computers, or communicating devices are allowed during the examination. Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script.

The exam consists in two independent exercises, and is graded out of 20 points (8 point for the first exercise, 12 for the second one).

1 The Web: A Variety of Technologies (8 points)

1. (2 points) Consider the following two news headlines:
   a) *Google has blocked millions of .co.cc websites from its search results after it deemed the subdomain providers to be proliferating malware.*
   b) *Baidu Signs Deal for English Results in China with Microsoft's Bing.*

   Explain and comment these two news items in a few sentences. In both cases, what is technically carried out (as far as you can tell from the title)? What is the motivation for these decisions?
2. (4 points) Give brief answers (a couple of short sentences at most) to the following questions:
   a) Why are absolute value units (cm, in, etc.) *not* appropriate in CSS stylesheets for display in a browser?
   b) Under which name is JavaScript standardized?
   c) What is the role of software such as Microsoft IIS?
   d) Give two advantages of CSS stylesheets linked by the `<link>` tag with respect to CSS stylesheets included within a `<style>` tag.
   e) What is the dominating browser on the market?
   f) Which agreement have Yahoo! and Microsoft reached with respect to their respective search engines?
   g) Give two differences between the relational data model and the semi-structured data model.
   h) Explain the MVC software engineering paradigm.
3. (2 points) Imagine you are hiring someone to create, from scratch, the Web site of your company. You have one short last question to ask her before taking a decision. Put forward such a question, and describe what would be a "good" answer (that would convince you to hire her) and what would be a "bad" answer (that would make you reject her application).

2 Attractions of European cities (12 points)

Consider the following corpus of 7 documents:

d_1 The Eiffel tower is the most famous monument in Paris.
d_2 If you go to Paris, do not miss the Louvre museum!
d_3 The Anatomy museum, in Pavia, is very interesting.
d_4 The Ashmolean museum (in Oxford) is the oldest museum in Europe.
d_5 Paris has numerous museums, such as the Orsay museum.
d_6 The jewels of the Tower of London are famous.
d_7 The Luxembourg gardens are in Paris, not in Luxembourg.

1. (2 points) Apply to this collection the following preprocessing steps: tokenization, morphological stemming, stop word removal. Give, for each document, the sequence of resulting terms. It is not required to give the output of each preprocessing step, just the final output.

2. (2 points) Construct an inverted index from the preprocessing obtained in the previous question. In order to reduce the size of the index, only terms appearing in at least two documents are recorded. The weighting function used should be tf-idf but fractions and logarithms can be kept as is, without computing their numerical values. Be sure to sort documents by decreasing order of weight. As a reminder, tf-idf can be computed as follows:

$$\text{tf-idf}(t,d) = \frac{n_{t,d}}{\sum_{t'} n_{t',d}} \cdot \log \frac{|D|}{|\{d' \in D \mid n_{t,d'} > 0\}|}$$

$n_{t,d}$ number of occurrences of t in d
D set of all documents

3. (3 points) Assume the corpus is distributed on a cluster of machines, and such an inverted index is to be constructed using the MapReduce programming paradigm. Propose (in pseudo-code) Map and Reduce functions computing an inverted index in a distributed manner. These Map and Reduce functions should cover everything done in the two steps above.

4. (1 point) One would like to build an information extractor to recognize city names appearing in the corpus. Propose one practical solution, and discuss its advantages and inconvenients.

5. (2 points) One would like to extract from the corpus the following facts:
   - locatedIn(Eiffel tower, Paris)
   - locatedIn(Louvre museum, Paris)
   - locatedIn(Anatomy museum, Pavia)
   - locatedIn(Ashmolean museum, Oxford)
   - locatedIn(Ashmolean museum, Europe)
   - locatedIn(Orsay museum, Paris)
   - locatedIn(Tower of London, London)
   - locatedIn(Luxembourg gardens, Paris)

   Put forward an information extraction approach that reaches a recall level of at least 50% and a precision of at least 80% on the example corpus. Explain as precisely as possible the proposed approach. Compute the exact recall and precision obtained on the example corpus.

6. (2 points) Assume all facts from the previous questions have been extracted. Create an RDF graph with (1) all entities organized in a taxonomy of at least 3 classes; and (2) an example of relation using the `:locatedIn` predicate. Use the standard predicates `rdf:type` and `rdfs:subClassOf`.

END OF PAPER