Two hours

UNIVERSITY OF MANCHESTER  
SCHOOL OF COMPUTER SCIENCE

Documents, Services and Data on the Web

Date: Tuesday 23rd May 2017
Time: 09:45 - 11:45

Please answer any TWO Questions from the THREE Questions provided.

Use a SEPARATE answer book for each QUESTION

Each question is worth 30 marks.

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]
1. **Documents on the Web**

   a) A Web search engine receives the following queries:

   - britishairways.com
   - what is bursitis
   - british airways flights manchester rome
   - navigate manchester newcastle
   - hungry cat jpg
   - buy tickets wimbledon
   - espresso coffee maker
   - bbc news page
   - animals that use sonar but are neither bats nor dolphins

   Discuss to what extent the information needs of users can be determined from queries such as these, using query classification strategies.

   (3 marks)

   b) Consider the following index terms and their postings list sizes:

<table>
<thead>
<tr>
<th>Index term</th>
<th>Posting list size</th>
</tr>
</thead>
<tbody>
<tr>
<td>warm</td>
<td>32000</td>
</tr>
<tr>
<td>lake</td>
<td>230000</td>
</tr>
<tr>
<td>sunny</td>
<td>189000</td>
</tr>
<tr>
<td>halifax</td>
<td>345000</td>
</tr>
<tr>
<td>sea</td>
<td>453000</td>
</tr>
<tr>
<td>dartmouth</td>
<td>21000</td>
</tr>
</tbody>
</table>

   A user gives the Boolean query:

   (warm OR sunny) AND (lake OR sea) AND (halifax OR dartmouth)

   Recommend a query processing order for this query. Justify your recommendation.

   (2 marks)

   c) You are given access to a search engine, and want to find out how many documents have been indexed for it. You are told that indexing has been done on the full (English) texts of the document collection and that the search engine offers Boolean query, where a query returns a list of documents and the number of documents found (hits).

   i) What query or queries would you use to get a reasonable estimate of the number of documents indexed? Explain why.

   (2 marks)

   ii) What query or queries would you use to get the exact number of documents indexed? Explain why.

   (2 marks)

   [Question 1 continues on the following page]
d) “Although the Boolean systems offer very powerful on-line search capabilities to librarians and other trained intermediaries, they tend to provide very poor service to end-users.” (Harman, 1992)

Discuss reasons that would have caused Harman to arrive at this conclusion.

(4 marks)

e) You have been asked to build a system to produce inverted indexes and are at the stage of deciding on steps to apply to the input documents, which could come from any domain and be in any language. You consider the following options:

i) Use an open-source tokenizer, which is reported to achieve 70% F-score on business news text.

ii) Train (or write rules for) your own tokenizer.

iii) Use a list of stop words.

iv) Implement the Porter stemming algorithm.

v) Allow for users typing queries in any mixture of upper/lower case.

vi) Allow for users to specify that some query term should occur near/before/after another term in documents.

Explain what criteria you would use to help you reach a decision on which of these options to choose. Comment on advantages and disadvantages of particular choices and combinations of choices, and on the impact of potential dependencies among choices. Exemplify your answer with appropriate examples. Justify the decisions and the conclusions that you reach.

(5 marks)

f) When preparing an inverted index to be used in a tf-idf perspective, I calculate the inverse document frequency (idf) for a term that occurs in every document of the collection. I then consider whether I should add this term to my inverted list, or add this term to a stop word list. What decision should I take? Justify your answer.

(2 marks)
g) In relation to the Vector Space Model, the result of constructing an inverted index in relation to three terms and three documents is as follows, where real values representing weights have already been normalised (i.e., the corresponding vectors are normalised):

<table>
<thead>
<tr>
<th>Index term</th>
<th>Posting list</th>
</tr>
</thead>
<tbody>
<tr>
<td>cheese</td>
<td>doc1:0.3343, doc2:0.9551, doc3:0.5937</td>
</tr>
<tr>
<td>sandwich</td>
<td>doc1:0.0, doc2:0.30, doc3:0.31</td>
</tr>
<tr>
<td>tomato</td>
<td>doc1:0.94, doc2:0.0, doc3:0.74</td>
</tr>
</tbody>
</table>

i) Determine the following similarities according to the cosine measure. Show your working:

\[
\begin{align*}
\text{sim}(\text{doc1, doc2}) & \\
\text{sim}(\text{doc1, doc3}) & \\
\text{sim}(\text{doc2, doc3}) &
\end{align*}
\]

(2 marks)

ii) What do the results tell us about these three documents and how can this information be used?

(1 mark)

iii) How else can we use cosine similarity scores with the Vector Space Model?

(1 mark)

h) “At least 20–30% of queries to Bing are simple named entities, and 71% of queries contain named entities.” (Microsoft internal report)

“Keyword-based search has become the de facto standard for information discovery on the Web. However, as the Web evolves […] the challenges to the feasibility of keyword-based search increase. But these have to be overcome, if the latent socio-economic impact of […] data sources is to be unlocked by end-users.” (Nguyen et al., 2016)

Discuss challenges to keyword-based search, and assess to what extent techniques of named entity recognition and fact extraction, and other techniques you are familiar with, can help improve the search experience. Also, comment on implications, advantages and disadvantages of employing such techniques when constructing indexes to support search. Justify your views and conclusions, giving appropriate examples to back up your arguments.

(6 marks)
2. **Services on the Web**

a) Explain how each of the following computing paradigms manifest themselves in Cloud Computing, emphasizing the characteristics of each paradigm and illustrating your answer with an example.

   i) **SaaS**

   ii) **Utility Computing**

   iii) **Grid Computing**

b) Provide two examples of innovations that have changed the way in which businesses and/or individuals collaborate and that have been driven by the emergence of Cloud Computing. Justify your answer.

c) Explain the property of Cloud Computing of being *virtual* and how Cloud service providers can financially benefit themselves by exploring this property.

d) “One of the main advantages of Cloud Computing, the shared infrastructure, could also have a negative impact.” (D. C. Marinescu, 2013)

   Discuss three reasons that would have caused Marinescu to arrive at this conclusion.

e) An IT company decides to provide free access to a public cloud dedicated to higher education. Which one of the three cloud computing delivery models, *SaaS*, *PaaS*, or *IaaS*, should it embrace, and why? What applications would be most beneficial for the students? Will this solution have an impact on distance learning? Why or why not?

f) ATN is a company that provides network equipment to telecommunications industries across the globe. Over the years, ATN has grown considerably and their product portfolio has expanded to accommodate several acquisitions, including companies that specialize in infrastructure components for Internet and cellular providers. ATN is now a leading supplier of a diverse range of telecommunications infrastructure.

[Question 2 continues on the following page]
In recent years, market pressure has been increasing, ATN has begun looking for ways to increase its competitiveness and efficiency by taking advantage of new technologies, especially those that can assist in cost reduction. As a result, ATN decided to explore the potential of adopting Cloud Computing. However, subsequent to their initial inquiries they became overwhelmed by the plenitude of cloud providers and cloud-based products.

Considering a £1,000,000 investment in a cloud computing solution with an additional £100,000 investment on consultancy for deciding which applications to migrate, etc., with cash inflows of £500,000 each year for three years, build a ROI evaluation of the potential move from ATN’s existing infrastructure and legacy systems to a cloud-based solution, considering a period of 3 years, and explain why you think this move is a good or a bad idea.

(4 marks)

g) In relation to the MapReduce programming model, is it possible to start the execution of reducers while mappers are still running? Why or why not?

(4 marks)

h) Describe a Reduce function (using pseudo-code and text) to join two tables with one-to-one relationships, whose schemas are defined in the following. Explain the assumptions you make for the corresponding Map function.

*Schema for Table A:* manager_id, manager_name, dept_id
*Schema for Table B:* departm_id, departm_name, departm_postcode
*Join operation:* A.dept_id = B.departm_id

(4 marks)
3 Data on the Web

a) Explain why re-using existing vocabularies in your linked data set would be beneficial. Provide an example.  

(2 marks)

b) Different linked data sources may use different URIs to refer to the same entity. Explain how the identity resolution works with linked data. 

(2 marks)

c) Draw an RDF graph equivalent to this RDF/XML document:

```xml
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:exterms="http://www.example.org/terms/">

    <rdf:Description rdf:about="http://www.example.org/index.html">
        <exterms:creation-date>August 16, 1999</exterms:creation-date>
        <dc:language>en</dc:language>
        <dc:creator rdf:resource="http://www.example.org/staffid/85740"/>
    </rdf:Description>

    <rdf:Description rdf:about="http://www.example.org/staffid/85740">
        <dc:title>James</dc:title>
    </rdf:Description>

</rdf:RDF>
```

(2 marks)

d) RSS (Rich Site Summary, also known as Really Simple Syndication or RDF Site Summary) is used to deliver regularly changing web content in a standard XML format that conforms to the W3C's RDF Specification. It allows the syndication of lists of hyperlinks, along with other information that helps viewers decide whether they want to follow the link. RSS is extensible via the associated XML-namsepace (http://purl.org/rss/1.0/) and has an RDF Schema (RDFS) that defines, among other things, classes channel and item, and properties items and description. Here are their definitions:

```xml
<rdfs:Class rdf:about="http://purl.org/rss/1.0/channel"
    rdfs:label = "Channel"
    rdfs:comment = "An RSS information channel.">
    <rdfs:isDefinedBy rdf:resource="http://purl.org/rss/1.0/"/>
</rdfs:Class>

<rdfs:Class rdf:about="http://purl.org/rss/1.0/item"
    rdfs:label = "Item"
    rdfs:comment = "An RSS item.">
    <rdfs:isDefinedBy rdf:resource = "http://purl.org/rss/1.0/"/>
</rdfs:Class>
```

[Question 3 continues on the following page]
and here is an example RSS feed

```xml
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns = "http://purl.org/rss/1.0/
    xmlns:dc = "http://purl.org/dc/elements/1.1/"

    <channel rdf:about = "http://example.com/news.rss">
        <title>Example Channel</title>
        <link>http://example.com/</link>
        <description>My example channel</description>

        <items>
            <rdf:Seq>
                <rdf:li resource="http://example.com/2002/09/02/"/>
            </rdf:Seq>
        </items>
    </channel>

    <item rdf:about="http://example.com/2002/09/01/">
        <title>News for September the First</title>
        <link>http://example.com/2002/09/01/</link>
        <description>other things happened today</description>
        <dc:date>2002-09-01</dc:date>
    </item>

    <item rdf:about="http://example.com/2002/09/02/">
        <title>News for September the Second</title>
        <link>http://example.com/2002/09/02/</link>
        <dc:date>2002-09-02</dc:date>
    </item>

</rdf:RDF>
```
[Question 3 continues from the previous page]

i. Explain what RDF(S) is used for and what classes and properties are. Use the above examples to illustrate your points.

(3 marks)

ii. Why is RDF-Schema insufficient as a knowledge representation? Use the above examples to illustrate your points.

(3 marks)

iii. Define a new subclass (advert) of the item class that can be used to represent items that are advertisements.

(2 marks)

iv. Define property Title that can be used to represent an item’s title. Make Title a subclass of http://purl.org/dc/elements/1.1.title

(2 marks)

e) Explain what is the task that the following SPARQL query aims to address:

```sparql
PREFIX dbpedia: <http://dbpedia.org/ontology/>

SELECT ?f ?d
WHERE
{
  ?f rdf:type dbpedia:Film .
  ?f dbpedia:releaseDate ?d .
  FILTER (?d >= "2000-01-01"^^xsd:date) && (?d < "2000-02-01"^^xsd:date))
}
```

Note that the <xsd:date> data type is used to represent date in the YYYY-MM-DD format. What is the role of casting (^^xsd:date) in the above query?

(4 marks)

f) A local authority wants to provide a single reference point for a growing collection of things they collect about the community they serve (e.g. education data, business environment, health and social services, sport facilities, environment, democratic services, etc.). They would like to publish the data they use and produce to build new experiences for the citizens, and to link to more complete, authoritative or canonical open data sources wherever possible (e.g. data from the Department of Health). You have been asked to advise on benefits and issues with the use of Semantic Web technologies and the Linked Data Platform.

[Question 3 continues on the following page]
[Question 3 continues from the previous page]

i. Explain how ontologies can be used to represent the types of data they would need to maintain.

   (3 marks)

ii. Explain why SKOS (Simple Knowledge Organisation System) is a simpler model compared to ontologies and OWL, and whether you would or would not recommend using it in this case study.

   (3 marks)

iii. Discuss the impact that Semantic Web technologies can have on citizens (consumers) and producers (i.e. the local authority).

   (4 marks)

END OF EXAMINATION