COMP38120

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Documents, Services and Data on the Web

Date: Monday 6th June 2016

Time: 14:00 - 16:00

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) "We still lack a full understanding of information needs […] when it comes to web searching." (Lewandowski, 2015)

Discuss to what extent we can hope to identify a user's information need when he uses a Web search engine. In your answer, comment on the effectiveness of strategies that are typically employed to identify information needs. Do you share Lewandowski's view? Give reasons why or why not.

(3 marks)

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

Index term	**Posting list size**
biscuit	213313
chocolate	87010
lemonade	107914
orange	271659
sugar	46654
toast	316813

A user gives the query:

(sugar OR toast) AND (lemonade OR orange) AND (chocolate OR biscuit)

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

(2 marks)

c) You formulate a Boolean query for an information retrieval system based on the Boolean model, using 3 terms linked by AND. You receive some results. You then formulate another query, by appending to the original query a further AND plus a fourth term. You receive no results. You note that all terms and operators are correctly spelled.

i) Explain why you should not be surprised to receive no results.

(1 mark)

ii) Give four disadvantages of the Boolean model.

(2 marks)

iii) The following is a sample query to the Westlaw search engine, which uses the Extended Boolean model:

disab! /p access! /s (work-site OR work-place OR (employment /3 place))

Discuss to what extent the use of such extended operators in the Westlaw query language helps to overcome the disadvantages you identified in 1) c) ii).

(2 marks)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

d) 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 available off-the-shelf language recogniser, which is reported to achieve 95% F-score on general language text.

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

iii) Use the Porter stemmer algorithm.

iv) Carry out case folding.

v) Produce a positional inverted index.

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)

e)

i) What is the inverse document frequency for a term that occurs in every document of a collection?

(1 mark)

ii) Can the tf-idf weight of a term in a document exceed 1.00? Justify your answer.

(1 mark)

f)

i) I form a new document by taking a copy of a document and appending it to itself. This new document evidently has the same semantic content as the original. What steps do I have to take to ensure that a cosine similarity calculation does not cause me to think that these two documents are very different?

(1 mark)

ii) A user gives a query where one of his query terms does not occur in the retrieval system's indexed terms. This implies that a vector for this query would not be in the vector space of the collection. How should we adapt the representation of vector space to handle this case? Give your reasoning.

(1 mark)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

iii) In relation to the Vector Space Model, the results of processing a query in relation to three documents is reported as follows:

cosine(Doc1, Query) = 0.64
cosine(Doc2, Query) = 0.78
cosine(Doc3, Query) = 0.93

What are these results telling us about these three documents?

(1 mark)

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

(1 mark)

g) Consider the graph in figure 1. We apply the PageRank algorithm, but do not use a damping factor and assume no random jumps. After 2 iterations, what would be the value i) for node **E** and ii) for node **A**? Values may be given to 3 decimal places. Show your working.

(3 marks)

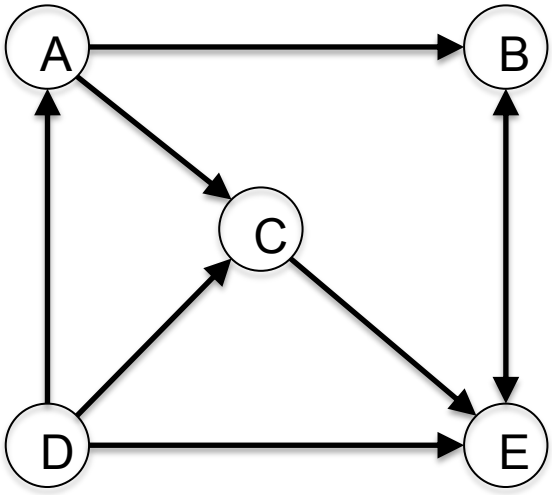

Figure 1: A graph with 5 nodes

h) If a node with IN links but no OUT links is introduced in figure 1, state what effect this could have on the PageRank calculation and how that effect could be overcome.

(1 mark)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

i) Answer only **ONE** of the following alternatives:

EITHER

"The accuracy delivered by entity driven search brings increased satisfaction among users. They will see documents that are about a specific semantic concept, with concrete properties, and not about a keyword that can be ambiguously interpreted." (Benedetti & Perez, Zaizi Ltd, 2014)

Any search engine relies on some kind of index. Entity-driven search relies on an entity-based index. Discuss issues and challenges in building entity-based indexes, and describe how such indexes may be used in types of search applications that you are aware of.

(5 marks)

OR

Consider the following newswire text:

Paris, France, 16th February 2016. /Innov-First News/
How Clever Do You Smell
By Marie-Claire France
Staff Reporter

MurkSys S.A. (ENX:MKSY) today announced its novel product IQ-Snif®. This latest addition to MurkSys's product range, which cost €7.5m to develop, uses "e-nose" technology to sniff your breath – to detect how clever you are. Dr. Vuzette Scitéau, Chief Technology Officer, demonstrated IQ-Snif to a select audience. The 75-year-old President of the French Science Institute, Prof. Abbou de Souffle, struggled to maintain the necessary breath Peak Flow Rate (PFR) of 600 L/min as he was suffering from Farmer's Lung disease, but was declared "very clever". Madame Sibelle Servelle was declared "below average clever". But Servelle, winner of the 2015 World Prize for Physics for research using CERN's Large Hadron Collider (Meyrin, Geneva), said, "I am not surprised. The average PFR for females is around 400 L/min, so this product gives false readings for females.". MurkSys Spokeswoman Portia Parole claimed that, by end of Q3 2016, cleverness of "all ages and genders" would be detectable by IQ-Snif. Later, a demonstrator, Livide Affollé, denied a job during field trials of IQ-Snif by Top Jobz Recruitment Agency, was arrested by the River Seine, after throwing 15 kg of strong-smelling Puefort cheese at MurkSys employees shouting "this cheese is more clever than you!".

Instances of the following named entities are to be identified:
PERSON, LOCATION, DATE, ORGANISATION, COMPANY, ARTEFACT, RATE, WEIGHT_AMOUNT, MONEY_AMOUNT, MEDICAL_CONDITION.
(ARTEFACT = something that is man-made)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

What patterns (including contextual clues) would you use to help you write rules to identify the maximum number of instances of the above named entities in the text? For each pattern you specify, state which instances it would match. You may specify patterns informally (e.g., *one or more capitalised tokens followed by the string "city" or the string "river" = LOCATION*). If you find it useful to introduce other types to aid your analysis, do so. Note any problematic aspects of the text that may cause your patterns to recognise too much, too little, or nothing, in certain cases.

(5 marks)

2. **Services**

a) How are Cloud resources configured and controlled for use by the service consumer? Illustrate your answer with one example for each Cloud service model. (4 marks)

b) Explain why Cloud resources may be accessed by heterogeneous client platforms and resources appear to be unlimited. (4 marks)

c) Explain the relationship between Cloud Computing and each of the following:

i) Grid Computing
ii) Utility Computing
iii) Software as a Service (3 marks)

d) Enumerate and explain three metrics used to help in the decision of whether to use a cloud or a local computational resource, in addition to Total Cost of Ownership (TCO), emphasising how effective each is. (3 marks)

e) Consider the following purchase costs related to technology and personnel for a growing IT department, which hosts a database, as well as research libraries, and sales and finance applications.

Disk Storage: $1,500,000 (with an operational lifetime of 1 year).

Disk maintenance: 25% of cost of total Disk Storage cost, each year.

Firewalls: $10,000 per year.

Network switches: $10,000 per year.

Server Hardware: $200,000 (with an operational lifetime of 4 years).

Server Maintenance: 10% of cost of the total Server Hardware cost, each year.

Software Licences: $200,000.

Labour: $900,000 per year.

Calculate the Total Cost of Ownership (TCO) considering a period of three years and discuss the limitations of TCO as the sole metric for supporting the decision as to whether or not to move to a Cloud solution. (7 marks)

f) Explain the fundamental problem of processing big data and how MapReduce addresses this problem. (4 marks)

g) How is the partitioning of the output from mappers performed and why? (5 marks)

3. **Data on the Web**

a) What are the four Principles of the Web of Data? Explain your answer.

(4 marks)

b) Resource Description Framework (RDF) provides a common data model and a mechanism for annotating i.e. describing data and resources in the linked data world. Does RDF provide a way to represent meaning (e.g. predicates)? Explain your answer.

(2 marks)

c) RDF/XML, RDFa, Turtle and N-Triples are the most common RDF serialisations. Discuss which is the best one for publishing RDF.

(4 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, or metadata, that helps viewers decide whether they want to follow the link. RSS is extensible via the associated XML-namespace (http://purl.org/rss/1.0/) and has an RDF Schema (RDF(S)) that defines, among other things, classes *channel* and *item*, and properties *items* and *description.* Here are their definitions:

```
<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>

<rdf:Property rdf:about="http://purl.org/rss/1.0/items"
        rdfs:label="Items"
        rdfs:comment="Points to a list of rss:item elements that are members
                        of the subject channel.">
        <rdfs:isDefinedBy rdf:resource="http://purl.org/rss/1.0/"/>
</rdf:Property>

<rdf:Property rdf:about="http://purl.org/rss/1.0/description"
        rdfs:label="Description"
        rdfs:comment="A short text description of the subject.">
        <rdfs:subPropertyOf
                rdf:resource="http://purl.org/dc/elements/1.1/description"/>
        <rdfs:isDefinedBy rdf:resource="http://purl.org/rss/1.0/"/>
</rdf:Property>
```

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

and here is an example RSS feed

```
<?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/2002/09/01/"/>
           <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>
```

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

(2 marks)

iii. Explain what is the task that the following SPARQL query aims to address:

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

```
PREFIX rss:   <http://purl.org/rss/1.0/>
PREFIX xsd:   <http://www.w3.org/2001/XMLSchema#>
PREFIX  dc:   <http://purl.org/dc/elements/1.1/>
SELECT ?item_title ?pub_date
WHERE {
    ?item rss:title ?item_title .
    ?item dc:date ?pub_date .
    FILTER xsd:dateTime(?pub_date) >= "2005-04-01T00:00:00Z"^^xsd:dateTime &&
        xsd:dateTime(?pub_date) < "2005-05-01T00:00:00Z"^^xsd:dateTime
    }
```

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

(5 marks)

e) A large public media company wants to provide a single reference for a growing collection of *things* that matter to their audience (e.g. sport results, news, programmes, wildlife, politicians, food, etc.). They would like to publish the data they use and produce to build new experiences for the audience, and to link to more complete, authoritative or canonical open data sources wherever possible. You have been asked to advise on benefits and issues with the use of Semantic Web technologies and the Linked Data Platform.

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 consumers, producers (e.g. journalists, editors) and production.

(4 marks)

END OF EXAMINATION