COMP 38120

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Documents, Services and Data on the Web

Date: Wednesday 23rd May 2018

Time: 14:00 - 16:00

Please answer all Questions.

Use a SEPARATE answer book for each QUESTION

Each question is worth 20 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 collection of N=100000 documents has been indexed. Four of the index terms and their postings list sizes are as follows:

Index term	Posting list size
bread	45000
flour	12000
butter	32000
toast	90000

A user gives the Boolean query:

```
bread AND (NOT flour) AND butter AND (NOT toast)
```

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

(3 marks)

b) A guide to searching states: "Consider using Boolean search operators to broaden or narrow your search".

Discuss to what extent this is a reasonable piece of general advice, taking into account different types of user, and commenting on advantages and disadvantages of the Boolean model.

(4 marks)

c) When building a system to produce inverted indexes, decisions on which text-processing operations to use can affect the size of the indexed vocabulary. For each of the following operations, state whether it would increase or decrease the size of the indexed vocabulary and give a short explanation why. Assume all text is in English.

i) Map all upper case characters to lower case.

ii) Apply the Porter stemmer.

iii) Use a list of stop words.

iv) Indicate the part of a document that a word occurs in (e.g., in the title, the abstract, the body, …).

v) Indicate different subject domains.

(5 marks)

d) 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.

(2 marks)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

e) "Keyword-based search is becoming increasingly ineffective at helping the modern knowledge worker. Many people still think of search as putting words in a box, but this is hugely limiting, as it relies on the user knowing exactly what they are looking for and using the right keywords to do so." (Squirro whitepaper, July 2017)

Techniques such as named entity recognition and fact extraction have existed for some time now. Discuss how and to what extent these techniques, and others supporting semantic search that you may be familiar with, could offer improvement on keyword-based search. In your answer, comment on implications, advantages and disadvantages of deploying such techniques in a Web context.Justify your views and conclusions, giving appropriate examples to back up your arguments.

(6 marks)

[PTO]

2. **Services on the Web**

a) Give two properties of the Hybrid Clouds. (3 marks)

b) In the context of the MapReduce programming model, explain what a *combiner* is and give two of the main differences between a *combiner* and a *reducer*.
(4 marks)

c) Consider a company in need of pulling down its Website for the duration of one hour to allow maintenance of backend servers to be carried out, causing its Website to be completely unavailable. To decide the precise day and hour of the day at which the Website is to become unavailable, it is necessary to identify the day and time when web traffic is at its lowest, to avoid affecting too many users of the site. To be able to make this identification, a Web server Log for each day during the month of March of year 2014 is available. The Log records the activities happening on the site, and because the files are quite large, a MapReduce program is to be written to process the files. Assume that all Log files have the same structure, and an excerpt of one of them is given as follows:

```
64.242.88.10 – – [07/Mar/2014:22:12:28 -0800] "GET
/twiki/bin/attach/TWiki/WebSearch HTTP/1.1" 401 12846

64.242.88.10 – – [07/Mar/2014:22:15:57 -0800] "GET
/mailman/listinfo/hs_rcafaculty HTTP/1.1" 200 6345
```

Write a Map function and a Reduce function for this program using pseudo-code. Note that the Reduce function should calculate the total number of hits for each hour of day and day of the month; the hour of the day and day of the month that have the least number of hits are perfect for the downtime. Recall that a Reduce function typically receives as input a key value pair of the form *<key,[value]>* and, for this example, the combination of date and hour can be used to determine the value of the key. For this program, you are interested only in the timestamp field of the Log files, e.g., *[07/Mar/2014:22:12:28 -0800]*.
(7 marks)

d) To support its growing customer base, three years ago Telecron invested in a complex Customer Relationship Management (CRM) software package for tracking customer opportunities and contracts. However, the package it selected was difficult to use and maintain, and upgrades to enable specific features required extensive system customization and frequent outages.

[Question 2 continues on the following page]

[Question 2 continues from the previous page]

Based on their research, the IT and Sales leadership teams at Telecron believe they should move their CRM instance to a cloud-based platform, going for a new Software as a Service CRM solution. The table below illustrates the result of the teams' research, indicating the TCO-based annual savings as well as the three-year savings that Telecron may be able to gain by moving from its Legacy-model TCO to a SaaS-model TCO.

Item	Annual Cost/Savings	Three-Year Charge
Legacy-model TCO	£1,970,000.01	£5,910,000.00
SaaS-model TCO	£1,482,499.99	£4,447,500.00
Savings	£487,500.02	£1,462,500.00

Assuming that Telecron has £212,500 as upfront cost towards its SaaS solution (i.e., cost of investment), and using the annual savings shown on the table as gains from investment, as well as a discount rate of 10 percent, ensure that this is indeed a sound use of finances by calculating the Net Present Value (NPV) for this example. Reflect on your analysis.

(6 marks)

[PTO]

3 Data on the Web

a) Give at least three (3) benefits of using the RDF Data Model in the Linked Data context.

(3 marks)

b) Why is it a good idea to use URI aliases, i.e., multiple URIs identifying the same entity?

(3 marks)

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

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:chord="http://purl.org/ontology/chord/">

<rdf:Descriptionrdf:about="http://purl.org/ontology/chord/symbol/F:sus2">
   <rdf:typerdf:resource="http://purl.org/ontology/chord/Chord"/>
   <chord:rootrdf:resource="http://purl.org/ontology/chord/note/F"/>
   <chord:base_chordrdf:resource="http://purl.org/ontology/chord/sus2"/>
</rdf:Description>

<rdf:Descriptionrdf:about="http://purl.org/ontology/chord/sus2"
rdfs:label="sus2">
   <chord:interval>
         <chord:ScaleInterval>
               <chord:degree>1</chord:degree>
         </chord:ScaleInterval>
   </chord:interval>
   <chord:interval>
         <chord:ScaleInterval>
               <chord:degree>2</chord:degree>
         </chord:ScaleInterval>
   </chord:interval>
   <chord:interval>
         <chord:ScaleInterval>
               <chord:degree>5</chord:degree>
         </chord:ScaleInterval>
   </chord:interval>
</rdf:Description>
```

(3 marks)

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

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-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:Classrdf:about = "http://purl.org/rss/1.0/channel"
      rdfs:label = "Channel"
      rdfs:comment = "An RSS information channel.">
      <rdfs:isDefinedByrdf:resource="http://purl.org/rss/1.0/"/>
</rdfs:Class>

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

<rdf:Propertyrdf: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:isDefinedByrdf:resource = "http://purl.org/rss/1.0/"/>
</rdf:Property>

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

and on the followingpage is an example RSS feed:

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

```
<?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/">

<channel>
      <title>BBC Nature Video collections</title>
      <link>http://www.bbc.co.uk/nature/collections</link>
      <description>A vast array of wildlife video clips.</description>

      <items>
            <rdf:Seq>
                  <rdf:li
resource="http://www.bbc.co.uk/nature/collections/p011x1yq"/>
                  <rdf:li
resource="http://www.bbc.co.uk/nature/collections/p01btg6w"/>
      </rdf:Seq>
      </items>
</channel>

<itemrdf:about="http://www.bbc.co.uk/nature/collections/p011x1yq">
      <title>This is Planet Earth</title>
      <link>http://www.bbc.co.uk/nature/collections/p011x1yq</link>
      <description>Narrated by Sir David Attenborough</description>
      <pubDate>Fri, 21 Dec 2012 16:37:42 GMT</pubDate>
</item>


<itemrdf:about=" http://www.bbc.co.uk/nature/collections/p01btg6w">
      <title> Sensational summer wildlife</title>
      <link>http://www.bbc.co.uk/nature/collections/p01btg6w</link>
      <description>Documents amazing opportunities to get close to the
incredible wild neighbours on your doorstep.
      </description>
      <pubDate>Thu, 08 Aug 2013 16:56:23 GMT</pubDate>
</item>

</rdf:RDF>
```

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

(3 marks)

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

(2 marks)

iii. Define property *duration* that can be used to represent an item'sduration in terms of number of seconds. Note that this property only applies to the *video* class that you just created above. Make *duration*a subproperty of http://purl.org/dc/elements/1.1/format

(2 marks)

[Question 3 continues on the following page]

[Question 3 continues from the previous page]

e) Refer to the following SPARQL query. Assuming that thisis meant to be run on a platform that contains product reviews, what task is it aiming to address?

```
PREFIX rev: <http://purl.org/stuff/rev#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?name ?createdOn
WHERE
{
      ?productrdfs:label ?name .
      ?productrev:hasReview ?review .
      ?reviewrev:createdOn ?createdOn .
      FILTER ((?createdOn>= "2016-01-01"^^xsd:date) && (?createdOn<
"2017-01-01"^^xsd:date)) .
      FILTER (regex(?name, "phone"))

}
```

Note that the <xsd:date> data type is used to represent date in the YYYY-MM-DD format. Why is it important to cast values to <xsd:date> in the above query?

(4 marks)

END OF EXAMINATION