Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Documents, Services and Data on the Web

Date: Thursday 15th May 2014
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 NOT permitted
1. a) Jansen et al. (2008) classify user needs of Information Retrieval into informational, navigational and transactional needs. Briefly describe these three needs and give two examples of each kind of need. (2 marks)

b) When attempting to classify such user needs expressed via a query, what criteria would you apply for each type of need identified by Jansen et al., and how successful do you think your criteria might be? (2 marks)

c) Consider the following document collection, where each document has a unique identifier (docn) and consists of one sentence:

doc1: Spring tides and gales — storm surges in pm.
doc2: Gale Storm Surges Up 100-Most-Beautiful List.
doc3: Stormin’ Norman Springs “Desert Storm”.
doc4: P.M. Gåle Størm’s surgeon: “Spring-time A.N.D. operation listed”.

i) Explain and justify the decisions you would make to arrive at a set of index terms for these documents, discussing both the impact such decisions might have on retrieval, and the advantages and disadvantages of any alternatives. (4 marks)

ii) Based on your index term decisions, draw up the basic inverted index representation for this collection. (2 marks)

iii) For this collection, using the inverted index you have established, what would be the result of the following query? Demonstrate how you arrive at your answer. (spring OR surge) AND (storm OR gale) (1 mark)

(If the result would be empty with respect to your inverted index, then, for the same mark, briefly explain what changes would be needed to obtain some result, assuming no change by the user to the query.)

d) “Stemming, in general, has not been an unmitigated success in improving Information Retrieval.” (Kantrowitz et al., 2000). To what extent do you agree with this view? Justify your answer. (2 marks)
e) Why are we interested in measuring the inverse document frequency (IDF) for a word, rather than its document frequency? (1 mark)

f) In relation to the Vector Space Model, three documents (D1, D2, D3) are analysed and results reported as follows:

\[
\begin{align*}
\text{cosine}(D1, D2) &= 0.94 \\
\text{cosine}(D1, D3) &= 0.79 \\
\text{cosine}(D2, D3) &= 0.69
\end{align*}
\]

i) What are these results telling us about these three documents? (1 mark)

ii) Why might you be interested in knowing whether or not the L_2 norm had been applied during the analysis? (1 mark)

Figure 1: A graph with 5 nodes

```
A ---- C ---- E
|       |       |
D <--- C <--- B
```

Consider the graph in figure 1. Probability mass is evenly distributed, in the initial state. We apply the PageRank algorithm, but do not use a damping factor and assume no random jumps. What would be the value for node B after 2 iterations? Values may be given to 3 decimal places. Show your working. (3 marks)
h) In a Web environment, a graph will at any one time contain dangling nodes (i.e., nodes with no out links). What is the effect of such nodes if we do not take appropriate action when calculating PageRank? What is the appropriate action that we should take? (1 mark)

i) Consider the following newswire text:

Amsterdam, Netherlands, 9th February, 2013. /WorldEvents Online News/
Floods fly away but it rains cats (and maybe dogs)
By Willem van Nord
Staff Reporter

Dutch scientist Dr Frits van Schenau, University of West Amsterdam, and Graphene Manufacturing plc [LSE:GRMA] today announced successful trials of Bibographene™ on the flooded plains of Leefdal, Brabant, Belgium. Dr van Schenau said that yesterday from 09:15 CET an AviaPlano CD-180E Cropduster aircraft sprayed a fine sheet of Bibographene over 5 km² of inundated fields on several occasions. Once bound to water molecules, the material became lighter than air due to reaction with Hydrogen and thus raised a layer of water into the sky where the wind carried it off. Graphene Manufacturing has announced orders worth over 3bn euros. However, the Belgian Air Traffic Control Agency issued a NOTAM (Notice to Airmen) warning of “Multiple Drifting Near-Invisible Airborne Water Hazards” covering a large area to the NW of 50° 53’ N 4° 29’ E and causing huge disruption to flights. Moreover, across the North Sea, Davinia Gristwistle, MBE, CEO of Norfolk Flax Company said that her company’s linseed crop had been wiped out by a massive Bibographene sheet breaking up overhead. At the same time, an infestation by a strain of Melampsora lini (only found in Belgium) had also hit production. A tiny Singapura kitten fitted with a Belgian chip named Trinetta had a lucky landing in a pint of Neckit Golden Ale being drunk by local farmer Mr Cardew Rednose. The Royal Society for the Protection of Miniature Animals has applied to the High Court for an injunction against “indiscriminate” use of Bibographene.

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

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 + {“city”, “river”} = LOCATION*). If you find it useful to introduce other named entity 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)

j) “Research suggests people prefer to state their information need rather than use keywords. […] Information worded as questions is increasing on the Web.” (Hearst, UC Berkeley)

Discuss the challenges and potential solutions for Web-scale document indexing activities that aim to support search beyond use of simple keywords and to help deliver appropriate answers rather than lists of documents to read. Justify your views and conclusions, giving appropriate examples to back up your arguments. (5 marks)
2. a) Discuss the extent to which the following statement is true: “Web Services are a great way to save money”. (3 marks)

b) The idea of sending messages between programs over a network is not new. Past efforts to create distributed computing infrastructure include CORBA, DCOM and Java RMI. However, Web Services have been more widely deployed across the Internet and intranets than the previous attempts. Give two reasons why this is true. (3 marks)

c) A DVD/Blu-Ray seller wishes to provide a Web Service to allow customers to browse, order and pay for films in DVD or Blu-Ray. Before a customer can place an order for the first time, the customer needs to register and provide personal details, which may be retrieved at any point in time using the customer number (i.e., at registration, each customer is assigned a customer number). When a customer places an order, an order number should be received. At the end of each month, customers must pay for orders they have placed during that month. Failure to pay for the orders of the month will incur a penalty over the remaining amount. When a customer wishes to make a payment, he/she must provide a bank account number and the amount he/she wishes to pay. The service must provide the functions described in the following. For each of the functions, specify the inputs, outputs and operations you would expect for this Web Service. You are free to use WSDL in your specification or not (it is not mandatory). Explain your assumptions.

i) Discovery of films (by words in their title) (2 marks)

ii) Ordering of films in DVD or Blu-Ray (by an unique identifier) (2 marks)

iii) Payment on account (by an unique customer number) (4 marks)

d) Explain what a UDDI registry is in terms of its role and the information it holds. (2 marks)

e) Give two examples of Quality of Service factors that reflect customer expectations, explaining each. (2 marks)
f) Explain how Cloud Computing relates to Grid Computing, Utility Computing and Software as a Service. (4 marks)

g) Explain how Cloud Computing contributes with the realization of Ubiquitous Computing. (4 marks)

h) Compare the three cloud computing delivery models, SaaS, PaaS, and IaaS from the point of view of application developers and users, discussing the security of each model and analysing their differences. (4 marks)
3.  
   a) Explain the 5-star deployment scheme for Open Data. (3 marks)
   
   b) What are the four Principles of the Web of Data? (2 marks)
   
   c) Different linked data sources may use different URIs to refer to the same entity. Explain how identity resolution works with linked data. (1 mark)
   
   d) Consider the following Turtle document.

   ```turtle
   @prefix foaf: <http://xmlns.com/foaf/0.1/> .
   @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
   @prefix corp: <http://www.example.org/corp#> .
   
   <http://www.pipian.com/people/pipian/card#me> a foaf:Person;
   foaf:name "Ian Jacobi"@en, "イアン・ジャコービ"@jp;
   foaf:age 24 ;
   
   ```

   List all triples that are represented in this document. (4 marks)
   
   e) RDF/XML, RDFa, Turtle and N-Triples are the most common RDF serialisations. Discuss which is the best one for publishing RDF. (4 marks)
   
   f) You have identified the following open linked data sets:

   - http://data.ordnancesurvey.co.uk, which contains a definitive set of geographical information postcodes in the UK from the Ordnance Survey. For example, the following URI contains information about M13 9PL:
     http://data.ordnancesurvey.co.uk/id/postcodeunit/M139PL
     A postcode URI contains a ‘Distinct’ property, which has the URI
     http://data.ordnancesurvey.co.uk/ontology/postcode/district

   - http://statistics.data.gov.uk, which contains data from the Office of National Statistics (ONS) and is cross-referenced with the Ordnance Survey identifiers.

   - http://education.data.gov.uk, which includes – among other data – data covering all schools in England and Wales. Along with the property named ‘label’ that gives the name of a school, there is a property called ‘districtAdministrative’ that relates the school to the local authority district it belongs to, which refers to the ONS.

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

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

```sparql
PREFIX pc: <http://data.ordnancesurvey.co.uk/ontology/postcode/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX school: <http://education.data.gov.uk/def/school/>

SELECT ?school ?label
WHERE {
  ?ons_district owl:sameAs ?district .
  ?school rdfs:label ?label
}
```

(5 marks)

g) Explain whether you agree or not with the following statement: “Combined with schema languages such as RDF-Schema and OWL, the linked data model allows the use of as much or as little structure as desired, meaning that tightly structured data as well as semi-structured data can be represented”.

(5 marks)

h) You have been asked to advise an e-commerce company on publishing their data as linked data. Discuss the impact that such a scenario can have on various aspects of e-commerce. You may consider the impact on resellers, end consumers, or new business models.

(6 marks)