Two hours

UNIVERSITY OF MANCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Documents and Data on the Web

Date: Wednesday 22nd January 2020

Time: 14:00 - 16:00

Please answer BOTH Questions
Each question is worth 30 marks

Use a SEPARATE answerbook for each QUESTION

This is a CLOSED book examination

Electronic calculators may be used in accordance with the University regulations

Question 1

a) Explain what navigational information needs are and provide an example query.

(2 marks)

b) Using an example, discuss how knowing the domain of a document could help in indexing.

(3 marks)

c) Explain what stemming is and justify if each of the following two statements is true or false:

- In a Boolean retrieval system, stemming never lowers precision.
- In a Boolean retrieval system, stemming never lowers recall.

(4 marks)

d) Luhn's hypothesis states that words that are frequent in a given dataset (above an upper cut-off) are considered common and not specific to that dataset. On the other hand, words that are infrequent (e.g. under a lower cut-off) do not contribute significantly to the content of that dataset. Compare and contrast these views with tf*idf and how it is used for document indexing.

(6 marks)

e) Using vector space models (VSMs) in information retrieval relies on bags of words. Explain why is that potentially problematic and how n-gram language models can address that.

(6 marks)

f) One of the approaches in language modelling for information retrieval is to model the query generation process. Explain how documents are ranked in that approach and illustrate it using the following document collection with two documents:

d1: Xyzzy reports a profit but revenue is down
d2: Quorus narrows quarter loss but revenue decreases further

and query with two words: *revenue down.*

Hint: build a mixed model with the maximum-likelihood estimation (MLE) unigram models from the documents and collection, mixed with $\lambda = 1/2$.

(6 marks)

[Question 1 continues on the following page]

[Question 1 continues from the previous page]

g) During an evaluation, an information retrieval system yields a list of 10 documents for a particular test query. This list is ranked, with position 1 being the top-ranked item. The evaluator knows that there are 5 relevant documents in the collection being queried, and noted that the relevant documents returned in the ranked list occurred at positions 2, 3, 4 and 8. Give both the recall and precision values for the system for each of the four returned relevant document positions. Explain your calculations.

(3 marks)

Question 2

a) Given the three relational database tables below, draw the equivalent RDF graph depicting the information available on the three movies shown.

Movie table

movieId	title	director	distributor	releaseDate	userRating
m1	The Laundromat	dir1	dist1	2019-09-27	6.3
m2	El Camino: A Breaking Bad Movie	dir2	dist1	2019-10-11	7.6
m3	Ocean’s Eleven	dir1	dist2	2001-12-07	7.8

Director table

directorId	firstname	lastname	birthDate
dir1	Steven	Soderbergh	1963-01-14
dir2	Vince	Gilligan	1967-02-10

Distributor table

distributorId	name	foundingDate
dist1	Netflix	1997-08-29
dist2	Warner Bros.	1923-04-04

(8 marks)

[Question 2 continues on the following page]

[Question 2 continues from the previous page]

b) Given the following lines written in Turtle, draw an equivalent RDF graph.

```
@prefix prod: <http://example.org/foodproducts/
> . @prefix gr: <http://purl.org/goodrelations/
v1#> . @prefix food: <http://purl.org/
foodontology#> .
prod:SFPNYogurt
   a food:Food ;
   gr:name "Sunnyside Farms Plain Nonfat
   Yogurt" ; food:fatPer100g "0"^^xsd:double ;
   food:proteinsPer100g "21"^^xsd:double ;
   food:energyPer100g "170"^^xsd:double ;
   food:containsIngredient [
         a food:Ingredient ;
         rdfs:label
   "Maltodextrin" .] ;
   food:containsIngredient food:E432 .
```

(5 marks)

c) Below is a vocabulary written in RDF Schema.

```
@prefix fo: <http://purl.org/foodontology#> .

fo:Food rdf:type rdfs:Class ;
      rdfs:label "Food" ;
      rdfs:comment "Something that can be eaten. It may be a basic food
            such as sugar or it may be more complex such as a dish
      resulting from a recipe, for example, a cake" .

fo:Recipe rdf:type rdfs:Class ;
      rdfs:label "Recipe" ;
      rdfs:comment "A set of instructions for preparing a particular
      dish, including a list of the ingredients required" .

fo:Dish rdf:type rdfs:Class ;
      rdfs:label "Dish" ;
      rdfs:comment "Food that is prepared by following a recipe" .

fo:method rdf:type rdf:Property ;
      rdfs:label "method" ;
      rdfs:comment "A procedure for preparing food, for example,
            baking" .
```

[Question 2 continues on the following page]

[Question 2 continues from the previous page]

i. Define a new class **Ingredient** as a subclass of the **Food** class that can be used to represent basic food.

(2 marks)

ii. Define a new property **technique** that can be used to represent a procedure that is applied to ingredients. In your definition, indicate that this property is a type of the **method** property, and that it applies *only* to the **Ingredient** class that you just created above.

(2 marks)

d) Discuss what the 5-star Rating Scheme for Data is by describing what makes data 1-, 2-, 3-, 4- or 5-star, and providing corresponding examples.

(5 marks)

e) Refer to the SPARQL query below, which will be run on a triple store containing information on news articles. What task is it aiming to address and what will it return? Explain what the FILTER and OPTIONAL clauses are for, and why it is important to make use of the `xsd:date` datatype URI in one of the FILTER clauses.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bbc: <http://www.bbc.co.uk/ontologies/creativework/>

SELECT ?headline ?writer ?publicationDate ?primaryTopic
WHERE {
        ?article bbc:createdBy ?writer .
        ?article bbc:publishedWhen ?publicationDate .
        ?article bbc:hasHeadline ?headline .
        OPTIONAL {?article bbc:about ?primaryTopic . }
        FILTER (?publicationDate >= "2019-01-01"^^xsd:date) .
        FILTER (regex(?headline, "Brexit"))
}
```

(5 marks)

[Question 2 continues on the following page]

[Question 2 continues from the previous page]

f) Consider the HTML document below.

```
<html>
  <head>
    <title>How to build your own microservices </title>
  </head>
  <body>
    <h1>Part 1: Setting up Flask</h1>
      Author: <em>Sergio Sola</em> <br>
      Created: <em>2017-03-05</em> <br>
      Version: <em>1.0</em> <br>
      Previous version: <a href="microservices0.8.html">version 0.8</a>
  </body>
</html>
```

You have been asked to turn this HTML document into a Linked Data-compliant one using RDFa, based on the Dublin Core vocabulary (`dcterms:http://purl.org/dc/terms/`). Below is a table showing some of the terms that the Dublin Core vocabulary contains, that you can use.

Term	URI	Definition
created	dcterms:created	Date of creation of the resource
creator	dcterms:creator	An entity primarily responsible for making the resource
subject	dcterms:subject	The topic of the resource
replaces	dcterms:replaces	A related resource that is superseded by the described resource.

Show how the HTML document will look like, with RDFa incorporated.

(3 marks)

END OF EXAMINATION