DUBLIN CITY UNIVERSITY

AUGUST/RESIT EXAMINATIONS 2014/2015

MODULE: CA437/F - Multimedia Information Retrieval

PROGRAMME(S):

CASE - BSc in Computer Applications (Sft.Eng.)
DME - B.Eng. in Digital Media Engineering
ECSA - Study Abroad (Engineering and Computing)
CAIS - BSc in Computer Applications (Inf.Sys.)

YEAR OF STUDY: 4,X

EXAMINERS: Dr Gareth Jones (Ext:5559)
Prof. Gerard Parr
Dr. Ian Pitt

TIME ALLOWED: 2 hours

INSTRUCTIONS: Answer 4 questions. All questions carry equal marks.

PLEASE DO NOT TURN OVER THIS PAGE UNTIL INSTRUCTED TO DO SO

The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

Requirements for this paper (Please mark (X) as appropriate)

- □ *Log Tables*
- □ *Graph Paper*
- □ *Dictionaries*
- □ *Statistical Tables*
- □ *Thermodynamic Tables*
- □ *Actuarial Tables*
- □ *MCQ Only - Do not publish*
- □ *Attached Answer Sheet*

QUESTION 1 ***[Total marks: 25]***

1(a) [5 Marks]

What is the purpose of an information retrieval system, and how does it seek to fulfil this purpose?

1(b) [4 Marks]

What is *tokenization* in automatic indexing for an information retrieval system?

1(c)

i. [4 Marks]

Users expect very rapid response from an information retrieval system. Explain the importance of data structures in delivering fast response times in information retrieval systems.

Discuss the trade off between computational cost at indexing time and search time in terms of the complexity of the data structure used for representation of the collection to be searched.

ii. [8 Marks]

Using an example explain the use of inverted files in a text information retrieval system. Your example should illustrate how hashing is used for efficient processing of search terms.

iii. [4 Marks]

Document collections to be searched by an information retrieval system are often dynamic. Documents may be added to the collection or removed from it. When this happens the index of the information retrieval must be updated. Suggest how the inverted file data structure of an information retrieval system might be efficiently updated when documents are added to or removed from the collection.

[End Question 1]

QUESTION 2 ***[Total marks: 25]***

2(a) What does it mean to say that a search term in a best-match information retrieval system has *good selectivity*?

[3 Marks]

2(b) When using the okapi BM25 model for best-match information retrieval, term weights are calculated as follows,

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where

i = the current search term
j = the current document
$cw(i,j)$ = the overall BM25 *combined weight* of search term i in document j
$cfw(i)$ = the *collection frequency weight* of search term i
$tf(i,j)$ = the within document *term frequency* of term i in document j
$ndl(j)$ = the normalised length of document j
k_1 = an experimentally determined constant
b = an experimentally determined constant

i. [9 Marks]

With reference to the okapi model above, explain the key concepts underlying the use of $cfw(i)$, $tf(i,j)$ and $ndl(j)$ in term weighting for best-match information retrieval.

ii. [3 Marks]

Explain the roles of the constants k_1 and b in the okapi BM25 function.

2(c) [4 Marks]

Recording term proximity within documents in an information retrieval system enables it to take into account whether a pair of terms are close together or far apart within a document. Why can term proximity be a useful factor in determining the relevance of a document to a search query containing such a pair of terms?

2(d)

i. [3 Marks]

What is *relevance feedback* as used in information retrieval?

ii. [3 Marks]

How does query expansion in relevance feedback function to potentially improve information retrieval effectiveness?

[End Question 2]

QUESTION 3 ***[Total marks: 25]***

3(a) [4 Marks]

The world wide web is is extremely large, and many documents on the web are very similar in terms of the words that they contain. Explain why this makes effective ranking of relevant documents in response to typical user queries very problematic for web search engines.

3(b) In response to the ranking problem highlighted in part (a), web search engines now adopt a strategy referred to as "learning to rank".

i. [4 Marks]

What is learning to rank?

ii. [12 Marks]

Describe three learning to rank components that can be used to improve the ranked output of a web search engine.

3(c) [5 Marks]

What content selection factors could be used to determine the contents of snippet summaries of retrieved documents in the ranked output list of a web search engine?

[End Question 3]

QUESTION 4 **[Total marks: 25]**

4(a) Careful evaluation of retrieval effectiveness is very important in the development of information retrieval algorithms and systems.

i. [3 Marks]

What are the three components of an information retrieval test collection?

ii. [4 Marks]

State the standard definitions of *precision* and *recall* as applied in evaluation of an information retrieval system.

iii. [6 Marks]

Explain, using a simple example, the procedure for calculating the *average precision* of the ranked document output of a best-match information retrieval system.

4(b) [6 Marks]

What is enterprise search? Give examples of information sources typically indexed and searched in enterprise search systems.

4(c) [6 Marks]

What are the key differences between casual users of search engines among the general public, and those who use search engines in an enterprise setting?

With these differences in mind or otherwise, explain which is typically the more important in enterprise search: maximising precision or recall?

[End Question 4]

QUESTION 5 ***[Total marks: 25]***

5(a) [4 Marks]

Distinguish between nodes, links and anchors in a hypertext.

5(b) [3 Marks]

What does it mean to say that a hypertext containing multiple linked nodes has no beginning and no end?

5(c)

i. [3 Marks]

What is the fundamental difference between HTML and XML markup of document contents?

ii. [4 Marks]

Give a simple example of XML markup of a document. Ensure that your example conforms to the requirements of acceptable XML and illustrates that XML is an open standard.

5(d) Multimedia information retrieval can be based on a combination of manually and automatically generated textual metadata. XML markup can be used to define standards for textual metadata labels of multimedia content. These textual metadata labels can be used for retrieval of relevant multimedia content by using text information retrieval methods.

i. [8 Marks]

Design an XML markup standard that could be used for labeling of images or videos for use in a multimedia information retrieval system based on text information retrieval.

Your design should include examples of metadata labels to be assigned by manual examination of the content and using automatic content analysis techniques.

ii. [3 Marks]

What problems might you expect to observe when using the manually assigned textual metadata for multimedia information retrieval based on text information retrieval methods?

[End Question 5]

QUESTION 6 ***[Total marks: 25]***

6(a)

i. [3 Marks]

What is the *semantic gap* in multimedia information retrieval?

ii. [5 Marks]

Automatic image analysis for multimedia information retrieval is typically broken into three levels: image primitives, iconography and iconology.

Explain these different levels of image processing. In your answer make clear the relative complexity of using each level, and how it relates to the semantic gap and human interpretation of images.

6(b) Multimedia information retrieval for video content often requires automatic analysis of the video content to support retrieval and browsing.

i. [4 Marks]

What is *shot boundary detection*?

Outline basic data analysis methods for automatic shot boundary detection.

ii. [5 Marks]

What problems can occur in accurate shot boundary detection for real-world data such as movies (where the video content has been carefully edited) or press conferences (often attended by large numbers of photographers)

6(c) Spoken document retrieval refers to searching for relevant spoken documents. The contents of spoken documents for retrieval are usually most efficiently identified using a speech recognition system. However, speech recognition is very difficult and even the best current speech recognition systems make mistakes in identifying the words that have been spoken.

i. [3 Marks]

What types of errors can be present in the output of an automatic speech recognition system? Use examples to illustrate how these would appear in the recognised output.

ii. [5 Marks]

What effect can these recognition errors have on matching between queries and documents in spoken document retrieval, and the ranking of relevant spoken documents in the system output.

[End Question 6]

[END OF EXAM]