DCU

DUBLIN CITY UNIVERSITY

SEMESTER ONE - RESIT EXAMINATIONS 2013

MODULE: (Title & Code)	CA437 / CA437A / CA437D / CA437F Multimedia Information Retrieval
COURSE:	B.Sc. in Computer Applications (SE Stream) B.Sc. in Computer Applications (IS Stream) B.Sc. in Computer Applications (Evening) B.Eng. in Digital Media Engineering
YEAR:	4
EXAMINERS: (Including Telephone Nos.)	Dr. Gareth Jones Ext. 5559 Dr. Micheal Manzke Dr. J. Power Prof. Finbarr O'Sullivan Prof. W. Buchanan
TIME ALLOWED:	2 hours
INSTRUCTIONS:	Please answer any 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

QUESTION 1 [TOTAL MARKS: 25]

1(a)

What is the purpose of an information retrieval system?

[3 Marks]

1(b)

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.

[8 Marks]

1(c)

What is conflation and why is it important in information retrieval?

What different types of conflation techniques are available for use in information retrieval systems?

[6 Marks]

1(d)

Outline the design of a data-based question answering system for large document collections such as the World Wide Web. Your outline should make clear how answers could be identified in documents and how data quantity can compensate for quality in systems of this type.

[8 Marks]

[End of Question 1]

QUESTION 2 [TOTAL MARKS: 25]

2(a)

Distinguish between noes, links and anchors in a hypertext.

[4 Marks]

2(b)

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

[3 Marks]

2(c)

What is the PageRank algorithm and why are algorithms such as PageRank important in search engines for the World Wide Web?

[6 Marks]

2(d)

For a web page j, the PageRank $PR(j)$ is calculated as follows,

$$PR(j) = (1-d) + d \sum_{v \in B_j} \frac{PR(v)}{N_v}$$

where
F_j = set of web pages j points to (outedges)
B_j = set of web pages that point to j (inedges)
$N_j = |F_j|$ = number of links from j
d is a scalar constant. d is typically set to 0.15 in operational retrieval systems.

PageRank is the likelihood that a "random surfer" visits a web page.

In terms of the next action of the random surfer after arriving at a web page, what is the meaning of the parameter d?

What does it mean for the next action of the random surfer if d is set $d = 0$ or $d = 1$?

[4 Marks]

2(e)

In the context of web search, discuss each of the following topics:

- query sense identification;
- freshness;
- diversity.

[8 Marks]

[End of Question 2]

QUESTION 3 [TOTAL MARKS: 25]

3(a)

Give the standard definitions of *precision* and *recall* as used in the evaluation of information retrieval systems. **[4 Marks]**

3(b)

Explain, using a simple example, the procedure for calculating the *average precision* of the ranked document output of a best-match information retrieval system. **[6 Marks]**

3(c)

In order to evaluate the retrieval accuracy of a web search engine it has been suggested that *pooling* could be used to identify a set of relevant documents. What is pooling? Explain how pooling could be used to evaluate and compare the retrieval accuracy of a number of available web search engines. **[7 Marks]**

3(d)

What is enterprise search and why is it of increasing importance? **[4 Marks]**

3(e)

What is *faceted search*? Use examples from enterprise search to explain your answer. **[4 Marks]**

[End of Question 3]

QUESTION 4 [TOTAL MARKS: 25]

4(a)

Give a concise definition of a document *summary*. In your answer contrast the possibilities for depth versus coverage in the summary when compared with the document being summarised. **[4 Marks]**

4(b)

Snippet summaries are used in the ranked document output list of search engines to indicate the potential relevance of each document to the user's query. The user can then click on a link to open the full version of a document that they feel to be potentially relevant.

Outline a simple method for generating snippet summaries for use in the ranked document output list of a web search engine.

When might such summaries not fulfill the purpose for which they are designed? **[6 Marks]**

4(c)

GPS sensors embedded in digital cameras and mobile devices containing cameras such as smartphones enable the location at which a photo is taken to be captured and stored in the text metadata of the photo, together with a timestamp of when the photo was taken.

Explain how this location and time of capture information can be used to make searching of personal photo collections easier for users. **[6 Marks]**

What other data might you find useful to capture with a photo, extract from photos, or infer from their metadata to make searching a personal photo collection even easier? **[3 Marks]**

4(d)

Spoken document retrieval refers to searching for relevant spoken documents. Searching the contents of spoken documents for retrieval requires that the contents have been identified using a speech recognition system. Speech recognition is very difficult and even the best current speech recognition systems make mistakes in identifying the words that have been spoken.

(i). 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. **[3 Marks]**

(ii). What effect can these recognition errors have on matching between queries and documents in spoken document retrieval. **[3 Marks]**

[End of Question 4]

QUESTION 5 [TOTAL MARKS: 25]

5(a)

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

[3 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.

[4 Marks]

5(b)

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

How can manual text annotations be used to attempt to overcome the semantic gap in multimedia information retrieval?

[6 Marks]

5(c)

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.

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, as well as, those assigned using automatic content analysis techniques.

[8 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?

[4 Marks]

[End of Question 5]

QUESTION 6

[TOTAL MARKS: 25]

6(a)

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

[4 Marks]

6(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

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, and how these operate when combined in the BM25 combined weight.

[12 Marks]

6(c)

What is relevance feedback in information retrieval?

What are the two standard methods used in relevance feedback?

Why is user based relevance feedback generally more effective than pseudo or blind relevance feedback?

[9 Marks]

[End of Question 6]

[END OF EXAM]