DCU

DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2014/2015

MODULE: CA437/CA437F - Multimedia Information Retrieval

PROGRAMME(S):

CASE - BSc in Computer Applications (Sft.Eng.)
CAIS - BSc in Computer Applications (Inf.Sys.)
DME - B.Eng in Digital Media Engineering
ECSA0 - Study Abroad (Engineering & Computing)

YEAR OF STUDY: 4

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

TIME ALLOWED: 2 hours

INSTRUCTIONS: This paper contains 6 questions.
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*	☐	*Thermodynamic Tables*
☐	*Graph Paper*	☐	*Actuarial Tables*
☐	*Dictionaries*	☐	*MCQ Only - Do not publish*
☐	*Statistical Tables*	☐	*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)

i. [5 Marks]

What is conflation and why is it important in information retrieval? Illustrate your answer using examples.
Your examples only need to illustrate the principle that you describe, and do not have to be linguistically accurate.

ii. [5 Marks]

Why is it important to split compounds in languages such as German, and to segment sentences in agglutinating languages such as Chinese and Japanese? In your answer make clear the effects that you would expect to see if these languages were not tokenised in this way.

1(c) [10 Marks]

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 BM25 model, explain the use of collection frequency weighting, term frequency, and document length normalisation for best-match information retrieval.

What role do k_1 and b play in the operation of the okapi BM25 model?

[End Question 1]

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

2(a) [5 Marks]

A summary is a condensed derivative of a source text, where the content is reduced through *selection* or *generalisation* on what is important in the source.

Explain what is meant by the concepts of *selection* and *generalisation* in this definition.

2(b) [8 Marks]

Document snippets are often returned in ranked retrieval lists by search engines. These snippets are intended to indicate the potential relevance of each retrieved document to the searcher's query. What factors might be taken into account to form effective snippet summaries for assessing relevance of retrieved documents in a search engine?

2(c) [6 Marks]

Shot boundary detection and keyframe extraction are used to structure and represent the content of video archives.

i. Describe a simple process of shot boundary detection. What problems are typically encountered in shot boundary detection? Your answer should include at least one solution to each of these problems.

ii. Suggest at least four approaches that can be taken to the extraction of a keyframe to represent the contents of a shot.

2(d) [6 Marks]

Suggest how the factors for text snippet creation from 2(b) and the identified shot boundaries and keyframes for video could be combined using transcripts of the spoken content in the video to form video snippets for use in a video retrieval system.

[End Question 2]

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

3(a)

i. [3 Marks]

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

ii. [4 Marks]

Explain how the mechanisms of *term reweighting* and *query expansion*, as used in relevance feedback, are able to improve the rank of relevant documents in the output of an information retrieval system.

iii. [3 Marks]

What are the <u>three</u> ways in which relevance of a document can be input into a relevance feedback algorithm?

3(b)

i. [4 Marks]

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

ii. [7 Marks]

In order to evaluate the effectiveness of an information retrieval system for a specific task, such as web search, medical record search or patent search, a test collection representative of the task must be constructed. An information retrieval test collection consists of:

- a set of documents representative of the task,
- a set of search requests of the form that a user of the system would be expected to use,
- a set of relevance data which identifies the documents relevant to each request.

Describe the *pooling* method which can be used to identify relevant documents. In your answer be sure to identify any assumptions made when using pooling to construct an information retrieval test collection.

3(c) [4 Marks]

i. What would you expect the effect of applying relevance feedback to be on the positions of relevant documents in the ranked list retrieved by an information retrieval system, compared to their positions before the application of relevance feedback?

ii. How would these changes affect the precision and recall of the ranked list?

[End Question 3]

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

4(a) [5 Marks]

Image and video retrieval systems often adopt a "'human-in-the-loop" operational paradigm where multiple interactive retrieval passes are typically required to locate relevant items. Explain why this is the case.

4(b) [8 Marks]

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

ii. 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 each level, issues of how general or domain specific they are, how they relate to attempting to close the semantic gap, and human interpretation of images.

4(c) Spoken Document Retrieval (SDR) systems typically make use of Automatic Speech Recognition (ASR) technologies to identify the words spoken in the content to be retrieved.

i. [7 Marks]

Automatic Speech Recognition is a very hard task, and ASR systems make mistakes in recognition of the words spoken. Give four reasons why Automatic Speech Recognition is difficult. What types of errors can ASR systems make?

ii. [5 Marks]

What effects can errors in ASR have on the behaviour of an SDR system compared to the behaviour of the same system working with a perfect accurate transcript of the content.

[End Question 4]

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

5(a) [5 Marks]

For what types of user queries would you recommend using a question answering system? When would you favour using a standard best-match information retrieval system instead?

5(b) [4 Marks]

Why is it hard for a web search engine to rank relevant documents reliably at the top of a retrieved list based only on the matching of a user search query to the available documents?

5(c)

To address the difficulties of web search based only on query document matching highlighted in part 5(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.

[End Question 5]

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

6(a) [8 Marks]

i. What is enterprise search?

ii. Explain using examples why enterprise search is of growing economic and practical importance?

6(b) [8 Marks]

What difficulties are often encountered in providing reliable ranking of relevant content in the output of enterprise search systems? In your answer you should consider issues relating to both the characteristics of documents typically indexed in an enterprise search system, and the ability of the searcher to use the system effectively. With respect to the searcher, your answer could make reference to the anomalous state of knowledge (ASK) model in information retrieval.

6(c) [6 Marks]

How are document clustering and faceted search used to support user search activities in enterprise search?

6(d) [3 Marks]

Explain the role of access controls in enterprise search.

[End Question 6]

[END OF EXAM]