DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2015/2016

MODULE: CA4009 - Search Technologies

PROGRAMME(S):

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

YEAR OF STUDY: 4,O

EXAMINERS:

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

TIME ALLOWED: 3 hours

INSTRUCTIONS: Candidates should answer Question 1 in Section A and any 3 questions from the 5 questions in Section B.

All questions are worth a maximum of 25 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☐ Dictionaries	☐ MCQ Only - Do not publish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

Section A

Question 1 is COMPULSORY.

QUESTION 1   *[Total marks: 25]*

[25 Marks]

Question Overview This question requires you to analyse a scenario for which a new search application is required, and then to propose the design of a new search application for this situation based on material studied in CA4009 Search Technologies and any other relevant technologies which you might wish to incorporate.

In answering this question it is suggested that you include the following elements:

- analysis of the search requirements of the end users of the system

- analysis of the domain and search expertise of the end users

- consideration of the types of queries that might be entered by the users

- available search technologies that could be used in a new search application to address this problem

- selection of a set of the components for your new search application and how these would be combined or used within the new system

- how the new system could be evaluated, including the features of a test collection and choice of evaluation metrics

These points are suggestions, you are free to include any topics or materials that you wish to in your answer.

Scenarios Answer this question by selecting one of the following scenarios requiring a new search application.

1. Feedback from users of the university library indicates that they are not satisfied with the quality of the search services for locating relevant materials in the library. Traditional searching using titles, authors, categories, etc, works well, but users now want better search based on the content of books, journals, and other online resources. In response to the user feedback, the university library is now seeking to install a new search system to meet the requirements of the users.

2. A new online news portal is seeking to provide its journalists with efficient access to information for use in writing news stories. These stories will vary from short rapidly written articles describing details of emerging news stories as they break to longer more detailed articles describing a political or economic situation. In order to provide their journalists with the widest variety of information for their research, the news agency wants to provide them with information from traditional published newspapers, radio, TV and social media.

3. A video-on-demand supplier is seeking to increase sales by incorporating recommendation into their search service. At present customers can search for movies using metadata descriptions of the movies including information such as titles, main actors, genre and descriptions of the plot. The company wishes to extend this to include recommendations based on purchases made by "similar" customers, where similar customers could include those who have entered similar queries or who have a purchase history similar to that of the current customer.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

QUESTION 2 *[Total marks: 25]*

2(a) [4 Marks]

What is the purpose of an information retrieval system? How does a standard information retrieval system attempt to achieve this purpose?

2(b) [8 Marks]

The BM25 term weighting function for best-match information retrieval is given by the following equation:

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

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 motivation for the collection frequency weighting, term frequency, and document length normalisation components in best-match information retrieval, and how these operate within the equation for the BM25 model.

2(c) [6 Marks]

    i. What is relevance feedback in information retrieval?

    ii. What are the two components used in information retrieval systems to implement relevance feedback?

    iii. How does each of these two components seek to improve the effectiveness of information retrieval system?

2(d)   [7 Marks]

     i.   Why is the use of suitable data structures vital for the implementation of effective search systems.

     ii. Using an example, explain the use of inverted files in text search systems. Your answer should illustrate how hashing is used for efficient processing of search terms.

[End Question 2]

QUESTION 3   *[Total marks: 25]*

3(a)   [5 Marks]

Explain why accurate automatic speech recognition (ASR) is difficult to achieve.

3(b)   [4 Marks]

How do errors in ASR impact on the behaviour of a speech search system and the experience of users of these systems?

3(c)   [6 Marks]

    i. What is meant by "human-in-the-loop" in image and video search?

    ii. What is the "semantic gap" in search of visual media?

3(d)   [10 Marks]

Before it can be searched, a collection of video recordings must be analysed and structured for indexing by a multimedia retrieval system. Describe the analysis techniques that can be applied to the visual content of a video, such as a television news broadcast, to prepare it for indexing by an interactive video search system.

[End Question 3]

QUESTION 4 *[Total marks: 25]*

4(a) [6 Marks]

     i. What is enterprise search?

     ii. Compare and contrast enterprise search with Web search in terms of user requirements and system specifications.

4(b) [4 Marks]

Explain the use of "early binding security" 'and "late binding security" for access controls in enterprise search.

4(c) [4 Marks]

How can the facets often associated with enterprise content be used with suitably designed rich user interfaces to facilitate effective enterprise search?

4(d) [6 Marks]

     i. Give the standard definitions of precision and recall as used in information retrieval.

     ii. By considering the information needs and working context of the users of enterprise search and Web search engines, explain which of precision and recall is generally more important to users of each of these types of search systems.

4(e) [5 Marks]

Pooling is a popular method used to identify a set of relevant documents when constructing an information retrieval test collection. Describe the pooling procedure as it is used to identify relevant documents for an information retrieval test collection. In your answer, identify the assumptions made in the pooling procedure.

[End Question 4]

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

5(a)   [7 Marks]

     i. What does a recommender system seek to do, and what information does it make use of in order to do this?

     ii. What are the two main classes of recommender system? Explain in outline how each of these operate.

5(b)   [2 Marks]

In what situation would it be appropriate to use a question answering system to address a user's information need?

5(c)   [8 Marks]

     i. Describe the typical workflow of a question answering system.

     ii. Explain in outline the knowledge-based and data-based approaches to question answering.

5(d)   [8 Marks]

IBM Watson provides a framework for developing question answering systems. Using the IBM Watson system designed to play Jeopardy or otherwise, give examples of practical constraints that need to be taken into account when deploying a Watson system in a real-world setting, and outline the sources from which a Watson system can acquire knowledge and how this knowledge might be applied to answer questions.

[End Question 5]

QUESTION 6 [*Total marks: 25*]

6(a) [4 Marks]

What are the fundamental challenges of Web search?

6(b) [6 Marks]

Using an example explain the principles and operation of the PageRank algorithm as applied in Web search.

6(c) [10 Marks]

   i. Explain the concept of "learning-to-rank" as used in Web search.

   ii. Give and explain in outline **three** features typically used in learning-to-rank for Web search.
Note: These features should be in addition to the use of standard information retrieval ranking methods and PageRank. No credit will be given for describing information retrieval ranking methods or PageRank in the answer to this part of the question.

6(d) [5 Marks]

   i. Summarization is content reduction 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.

   ii. Snippet summaries of retrieved items are an important component of the ranked list output of a web search engine. What is the principle purpose of these snippets? Give one method for measuring how effective they are at achieving this.

[End Question 6]

[END OF EXAM]