DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2016/2017

MODULE: CA4009 - Search Technologies

PROGRAMME(S):

CASE - BSc in Computer Applications (Sft.Eng.)
ECSAO - Study Abroad (Engineering and Computing)
CPSSD - BSc in ComputationalProblem SolvandSW Dev.
ECSA - Study Abroad (Engineering and Computing)

YEAR OF STUDY: 4,O,X

EXAMINERS: Gareth Jones (Ph:5559)
Prof. David Bustard
Dr. Ian Pitt

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*
□ *Graph Paper*
□ *Dictionaries*
□ *Statistical Tables*

□ *Thermodynamic Tables*
□ *Actuarial Tables*
□ *MCQ Only - Do not publish*
□ *Attached Answer Sheet*

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 required 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. The university library spends much time answering queries from students and other users. Queues at the library information desk mean that users often have to wait many minutes for answers to their questions. Many of these queries are essentially asking the same question, but users phrase them in different ways depending on their background and understanding of how the library works. The library managers want to improve the efficiency of their service and to free up their staff to help with unusual and complex queries rather than needing to answer the same question multiple times to different users. The proposed solution is to develop an automated online system to answer common questions, and to refer questions online to library staff if the online system is not able to answer the question.

2. A review of information management practices within a small enterprise employing about a dozen people has revealed that employees spend many hours each week looking for information within company archives, emails, professional reference documents, etc to enable them to carry out work on current projects.

The company managers realise that all this time spent looking for information is costing the enterprise a lot of money and slowing down progress with projects and risks overrunning agreed delivery deadlines. The managers propose to develop an enterprise search system to index all information within the company and make it available for online search in the expectation that this will improve the efficiency with which required information can be located.

3. Universities are increasingly creating audio-visual recordings of lectures and making these available online often with the slides used by the lecturer and sometimes links to related learning resources, such as books, research papers, tutorial examples, or practical use cases. In order to make better use of these online materials, the learning development unit at one university has decided to develop an integrated search and access platform for these resources. The expectation is that this platform will increase the use of these resources by students studying at the university and to hopefully improve the learning outcomes of the students. The proposal is that this platform should give access to all of the available resources in a single interface and enable students to smoothly navigate between different resources addressing the same information need. A concern raised by one member of the group is that watching videos of whole lectures within this system to find relevant information will be very inefficient for the students, and they wonder if a creative solution might be found to address this problem.

[End Question 1]

Section A

Answer any 3 of the 5 questions in this section.

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

2(a) [4 Marks]

Give three examples of English stop words, and explain why they are stop words. Why are stop words often removed in information retrieval systems?

2(b) [4 Marks]

$$cfw(i) = \log \frac{N}{n(i)}$$

is the standard equation used to calculate $cfw(i)$ where

$cfw(i)$	=	the collection frequency weight of term i
i	=	the current search term
$n(i)$	=	total number of documents containing term i
N	=	total number of documents in the collection

Would the $n(i)$ value of a stop word typically be greater or less than the average value of $n(i)$ in a collection of documents? Explain your answer.

What would be the impact of stop word removal be on the ranking behaviour of a best-match information retrieval system?

2(c) [2 Marks]

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

2(d) [6 Marks]

Term weighting is an important technique in best-match information retrieval. In addition to the $cfw(i)$ in part (a), what are the other two components typically used in effective term weighting schemes, e.g. the vector-space model and the BM25 model? For each of these two components explain its underlying principles and why it is expected to be beneficial in information retrieval.

2(e) [4 Marks]

Recording proximity of terms within documents in an information retrieval system enables it to take account of 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 potential relevance of a document to a search query containing such a pair of terms?

2(f) [5 Marks]

What are *stemming* algorithms as used in automatic indexing for information retrieval?

Explain what is meant by *under-stemming* and *over-stemming*.

For stemming of English language text, why do we generally want to stem suffixes, but not prefixes?

[End Question 2]

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

3(a) [5 Marks]

What is the difference between a conventional *information retrieval* system and a *question answering* system?

Even if high quality question answering systems were available commercially, why would there still be a need for information retrieval systems?

3(b) [8 Marks]

Describe the design of a data-based question answering system. In your answer explain how quantity of information compensates for the quality of the data in answering questions from data sources of variable quality such as the WWW. Include examples to illustrate your answer.

3(c) [4 Marks]

What is enterprise search? Why is enterprise search of increasing importance?

3(d) [4 Marks]

Discuss the issue of controlling access to content in enterprise search. Use examples to illustrate your answer.

3(e) [4 Marks]

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

[End Question 3]

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

4(a) [5 Marks]

How can search applications on networked mobile devices such as smartphones benefit from the use of multimodal interfaces?

4(b) [5 Marks]

What are the *three* types of errors which can be made by an automatic speech recognition (ASR) system?
Explain how each of these error types can impact on the output of an information retrieval system for spoken content.

4(c) [5 Marks]

Describe using examples how content browsing in spoken content retrieval can be supported by the use of graphics based visualisation.

4(d) [4 Marks]

Give the standard definitions of *precision* and *recall* as used in evaluation of information retrieval systems. Briefly explain what each of these metrics is designed to measure.

4(e) [6 Marks]

What are the three components of an information retrieval test collection? Explain how these should be selected to evaluate the effectiveness of an information retrieval system for a specific task.

[End Question 4]

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

5(a) [5 Marks]

What are the differences between HTML and XML document markup? Use examples to illustrate your answer.

5(b) [4 Marks]

How can XML be used in multimedia information retrieval for items such as images and video?

5(c) [5 Marks]

Retrieving relevant documents at the top of a ranked retrieval list in web search based only on query-document content matching is unreliable. Why is this? To answer this question recall that user queries to web search engines are typically very short and that the world wide web is very large.

How do link-based methods such as *PageRank* improve the reliability of document ranking for web search engines?

5(d) [5 Marks]

By means of a simple example explain the principles of the *PageRank* algorithm.

5(e) [6 Marks]

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.

[End Question 5]

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

6(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.

6(b) [8 Marks]

i. Compare and contrast the objectives of information retrieval systems and recommender systems.

ii. Describe an approach to combining information retrieval and recommender system technologies into a single integrated system.

6(c) [10 Marks]

i. Give a concise definition of a document *summary*. In your answer contrast the possibilities for depth versus coverage in the summary generation process.

ii. Effective snippet summaries are an important part of web search engines. List **four** components that can be used in the selection of sentences for use in snippet summaries in a web search engine. In each case explain your reason for choosing this component.

[End Question 6]

[END OF EXAM]