DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2017/2018

MODULE:	CA4009 – Search Technologies	
PROGRAMME(S):		
CASE	BSc in Computer Applications (Sft.Eng.)	
CPSSD	BSc in ComputationalProblem Solv&SW Dev.	
ECSAO	Study Abroad (Engineering & Computing)	
YEAR OF STUDY:	4,O	
EXAMINER(S):	Gareth Jones	(Ext:5559)
	Prof. Brendan Tangney	External
	Dr. Hitesh Tewari	External
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- □ *MCQ Only - Do not publish*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*

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. Your analysis and design should be based on material studied in CA4009 Search Technologies and any other relevant material which you might wish to incorporate.

In answering this question it is suggested that you address 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, but description of a complete solution containing all relevant elements is required to receive full credit.

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

1. You work for a leading supplier of online learning tools to school, colleges and universities. Your company has a successful product which enables educational establishments to record and archive classes and lectures for later reuse. These materials are used by pupils and students for private study to catch up missed classes or to review material which they did not understand in class, but they are also used by teachers and lecturers to review their teaching sessions, looking for things that went well or badly, so that they can work towards improving their teaching. While the current tool is effective for recording teaching sessions, its playback tool is very basic, only allowing playback from the beginning of the file and there is no facility for searching the archive. Based on the success of the

content recording tool, your company has decided to develop a new application to support search of the recorded teaching archive at each teaching establishment. As part of the search process, the new tool is to include an advanced interactive browsing application for individual retrieved recordings.

2. Many students have difficulty in deciding which module options they should take in each year of their degree course. They consider factors such as which modules they will enjoy most, which modules have the best instructors, which modules they are likely to get the highest mark for, and which modules will help them get the “best” job when they graduate. It has been suggested that a recommender system could be employed to recommend modules to students based on their preferences and/or previous performance, and the ratings of other students who have taken the available modules in previous years. You work for the information services group within a university and are given the task of developing a prototype recommender system for modules which could be used by students at the beginning of each year to receive recommendations of modules that they might take for the year. A very desirable feature of the recommender would be to provide reasons underlying the recommendation of each module, as part of your design consider how this feature could be incorporated into your recommender system.

3. FOIL is a company providing services to the oil industry in the area of oil exploration. They employ around 100 people, most are based permanently in their offices in Dublin, and others spend some of their time traveling to meet clients at their offices or onsite around the world. FOIL provide specialist chemical analysis services at their offices in Dublin. They take on long term exploration projects with major oil companies around the world. FOIL have a large archive of documentation from projects which they have worked on since the company was founded more than 10 years ago. When developing new projects they seek to make use of the information captured in this documentation, as well as online information services describing local geological information about the area of the world where their next project is due to take place. Unfortunately, they do not currently have a dedicated search tool to support exploration of their documentation archive and to integrate this information with the available online information services. You work for a company developing enterprise search solutions for specialist companies. FOIL have given your company the contract to develop a new search tool to support them in their development of new projects by searching their available information sources effectively. You are tasked with leading the development of the new search tool for FOIL.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

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

2(a) [3 Marks]

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

2(b) [10 Marks]

i. What is the "semantic gap" in search of visual media and why is it important in image and video search?

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 using each level, and how it relates to the semantic gap and human interpretation of images.

2(c) [12 Marks]

i. Speech and video are *temporal* media. What does this mean?

ii. Locating relevant information within individual speech and video documents is typically much more time consuming than locating relevant information in individual text documents. Explain why this is the case.

iii. Using simple sketches give examples of interactive tools and visualisations that have been developed to enable individual speech and video documents to be searched efficiently for relevant content.

[End Question 2]

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

3(a) [7 Marks]

i. What is the general reason for adding hypertext annotation to a collection of documents or other content items?

ii. Explain the role of **nodes**, **links** and **anchors** in a hypertext.

3(b) [4 Marks]

What does it mean to “jump into a hyperspace”, such as the WWW?

How do web search engines such as Google support users of the WWW to do this?

3(c) [6 Marks]

What is the PageRank algorithm as used in WWW search?

Use a simple example to outline the principles of the PageRank algorithm.

3(d) [8 Marks]

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

ii. 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 3]

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

4(a) [3 Marks]

For what type of user query is a question answering system a suitable means of addressing a user's information need?

4(b) [13 Marks]

i. Using an example, show how a *knowledge graph* encodes information from source texts.

ii. When annotating the features or attributes of an entity, e.g. a named person or place, how can a knowledge graph capture all the important features for this entity if they are not found in an individual source text?

iii. Explain how search engine companies such as *Google* use information summary "cards" created from knowledge graphs to provide information about common entities as part of a Search Engine Results Page (SERP).

4(c) [9 Marks]

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

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

[End Question 4]

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

5(a) [3 Marks]

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

5(b) [8 Marks]

$tf \times idf$ is a popular term weighting mechanism for ranked information retrieval systems.

In the $tf \times idf$ formulation: tf represents within document *term frequency* and idf presented *inverse document frequency*.

Explain the principles of the tf and idf components in term weighting for information retrieval. In your answer, describe the individual key concepts underlying the use of tf and idf, and make clear why they are generally more effective when used in combination.

5(c) [6 Marks]

One of the best known best-match information retrieval models is the *vector space model*.

In the vector space model the user's search query and the documents of the collection are represented as vectors and the similarity between the searcher's query and each document is calculated using the cosine similarity dot product as follows:

$$\mathbf{Q}.\mathbf{D}(\mathbf{j}) = |Q||D(j)| \cos\theta$$

where Q is the searcher's query, and $D(j)$ is a document j from the collection being searched.

Using a simple example explain how the vector space model creates a ranked list of retrieved documents to be returned to the searcher.

5(d) [8 Marks]

Relevance feedback methods are designed to improve the user's search query. To carry out a relevance feedback process, The user first marks documents retrieved by an initial information retrieval run using the user's original query as either relevant or non-relevant to their information need.

The standard relevance feedback method for the vector space model was developed by Rocchio.

The standard Rocchio query modification equation is as follows:

$$Q' = \alpha Q + \beta \frac{1}{R} \sum_{x=0}^{R-1} RD(x) - \gamma \frac{1}{NonR} \sum_{x=0}^{NonR-1} NonRD(x)$$

where:

Q' = modified query vector
Q = original query vector
R = no of $\mathbf{known}$ relevant documents
$NonR$ = no of $\mathbf{known}$ non-relevant documents
$RD(x)$ = relevant document vector x
$NonRD(x)$ = non-relevant document vector x
α, β, γ = are empirically determined constants

With reference to the vector space model explain how the Rocchio method works. In your answer, make clear the contribution of each the three components on the right hand side of this equation, how they are calculated based on the user contributed relevance feedback information, and the use of the three empirically determined constants.

[End Question 5]

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

6(a) [5 Marks]

i. What is the purpose a recommender system?

ii. What are the two main sources of information used by recommender systems?

6(b) [6 Marks]

Explain the following concepts as they apply to the goals of an operational recommender system: *relevance*, *novelty*, *serendipity*, *diversity*.

6(c) [5 Marks]

Describe in outline the operation of a recommender system based on *collaborative filtering*.

6(d) [5 Marks]

i. What is enterprise search?

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

6(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 6]

[END OF EXAM]