DUBLIN CITY UNIVERSITY

AUGUST/RESIT 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□ *MCQ Only - Do not publish*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*

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 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. When researching a new story journalists must often examine information from many different sources, including multiple repositories of differing reliability, and in different media types, including text, audio, image and video. Examining all of these sources independently and then collating retrieved information to develop a story, as well as using material from one source to develop a query to search another one, can be a significant burden on journalist as they seek to manage all this information. Your company believe that there is a significant commercial opportunity in the development of an interactive application which would enable journalists to query across multiple information sources using a single interface, and to be able to manage retrieved information so that it can be easily re-found as the journalist works on their story, and to be integrated into a plan or storyboard of the story as the journalist develops it. You are tasked with the development of an initial prototype of a system designed to improve the efficiency of journalist workflow in the development of a story, which will hopefully also improve the quality of developed stories by improving the quality of the available search tools.

2. The convention in commercial web search engines is that the user enters a query expressing their information need in response to which the search engine returns a ranked list of potentially relevant documents. These are shown to the user in batches of 10 ranked documents displayed in the form of a simple snippet. The snippet is generally composed of two or three pieces taken from each potential relevant document containing words from the searcher's query. The purpose of the snippet is to attempt to indicate to the searcher whether the document is relevant. The searcher can then click on a link for any document which they believe might be relevant to open the complete document. Users are very familiar with this model, and it is often used successfully. However, there is much interest in improving this form of presentation, e.g. including summary result cards of popular people, places, events, etc. You are working for a team responsible for the search engine in your company. The standard version of this has the same appearance and operational functionality of a web search engine. However, the system is completely open source. Your team leader decides that focusing on an alternative means of presenting the results of searches made by company employees might improve the effectiveness of the search engine within the company. You are tasked with creating some proposals for prototype alternative interfaces and developing a plan for testing them within the company.

3. Viewing videos is now a commonplace activities on smartphones. However, interfaces for searching for scenes within video archives such as *YouTube* are simple and make navigating through content very inefficient. Develop novel potential interface components that could be integrated into smartphone video viewing applications which could enable efficient location of relevant video content, and their evaluation. It is important to recall when doing this that viewing non-relevant video will be time consuming and often frustrating for users.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

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

2(a) [4 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?

2(b) [6 Marks]

Outline the stages in the workflow in a standard question answering system.

2(c) [4 Marks]

Explain the role of identifying an *Expected Answer Type (EAT)* in a question answering system of this type.

2(d) [3 Marks]

What is the fundamental limitation of using a information extraction (IE) methods in a question answering system?

2(e) [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 2]

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

3(a) [4 Marks]

What is the purpose of an information retrieval system, and how does it seek to fulfil this purpose?

3(b) [3 Marks]

Explain why terms with a medium frequency across a collection of documents usually make the best search terms for information retrieval.

3(c) [6 Marks]

Compare and contrast the features of *Boolean* and *best-match* information retrieval methods. Where appropriate, use examples to illustrate your answers. Your answer should make clear which approach is best suited for different classes of user, for example which would be suitable for use by a professional librarian or by a novice web searcher, and why this is the case.

3(d) [3 Marks]

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

3(e) [4 Marks]

What is *tokenization* in automatic indexing for an information retrieval system?

3(f) [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.

[End Question 3]

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

4(a) [3 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.

4(b) [7 Marks]

By means of a simple example explain the principles of the *PageRank* algorithm. Your example should show how the PageRank value of each page is calculated.

4(c) [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(d) [7 Marks]

In order to evaluate the effectiveness of an information retrieval system for a specific task, 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.

How is the relevance of a document determined for each request?

Why is it not practical to determine the relevance of each document for each query?

Outline the *pooling* method which is used in an attempt to address this problem. What is the key assumption made when using the pooling method.

4(e) [4 Marks]

When pooling is applied in the evaluation of a web search engine will precision or recall be calculated more reliably? Explain your answer.

[End Question 4]

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

5(a) [3 Marks]

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

5(b) [3 Marks]

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

5(c) [5 Marks]

Describe the process of shot boundary detection when preprocessing video data for use in a multimedia information retrieval system.

What problems are typically encountered in shot boundary detection for edited real-world video data such as movies?

5(d) [6 Marks]

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.

5(e) [8 Marks]

i. Explain why searching spoken audio files to locate relevant content within a file by listening to the audio data is generally very inefficient. Use simple scenario examples of a user listening an audio file in an attempt to locate content relevant to their information need to explain your answer.

ii. Sketch a browsing interface for individual spoken documents. Your interface should indicate potential relevant content to enable the user to determine which parts they wish to listen to.

[End Question 5]

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

6(a) [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.

6(b) [6 Marks]

Anchortext within a node of a hypertext forms the launch point from this node to another node, e.g. the linking highlighted anchortext on one web page to another web page or the link to an object such as an image. Explain how anchortexts can be used to enable the retrieval of nodes within a hypertext which have not yet been crawled by a web spider.

6(c) [2 Marks]

Suggest how the presence of anchortext in a sentence could be used as a factor in the calculation of a score for a sentence in generation of snippets.

6(d) [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.

[End Question 6]

[END OF EXAM]