Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 EXAMINATIONS 2020/2021

MODULE: CA4009 – Search Technologies

PROGRAMME(S):

CASE	BSc in Computer Applications (Sft.Eng.)
EC	BSc in Enterprise Computing
DS	BSc in Data Science
ECSAO	Study Abroad (Engineering & Computing)

YEAR OF STUDY: 3,4,O

EXAMINER(S):

Gareth Jones	(Internal)	(Ext:5559)
Dr. Hitesh Tewari	(External)	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 YOU ARE 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.

There are no additional requirements for this paper.

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 as appropriate the following elements:

- analysis of the search requirements of the end users of the system
- analysis of the domain and the 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 suitable test collection and choice of evaluation metrics

These points are suggestions, you are free to include any issues, 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. Students make extensive use of lecture notes provided online in their studies. These notes are typically based on one or more textbooks and other resources, generally contained in a reading list. If a student wants to learn more about a topic covered in a lecture, they generally need to find the original sources used to create the notes and then to spend time studying them to try to find relevant material relating to the lecture notes. The effort to do this is often off-putting to students who frequently end up relying on the lecture notes and not bothering to consult the more detailed descriptions contained in the source texts. In order to encourage students to study topics more deeply, the IT department where you

work at your local university has decided to create a tool to automatically link the contents of lecture notes to related content. This related content can either be the source for the notes or other sources related to the content of the notes. You are tasked with developing this linking system which will be made available to students to help them in their studies.

2. Morgan Legal Associates provide a range of legal services to their clients, who are individuals and small businesses. Morgan employs around 25 solicitors and legal secretaries and a corresponding number of administrators and secretarial staff. The company seeks to provide a high quality service to its clients, and to do this as efficiently as possible by reusing solutions created for previous clients. To do this, Morgan make extensive use of previous case notes when advising their clients. The cost of searching their records for these notes means that Morgan's staff are often not able to identify all the potentially useful information available in their archives. In order to improve the efficiency and coverage of their archive search, Morgan have decided to commission a new software application to enable indexing and searching of their archives, many of which are currently only available as paper documents. You work for a search consultancy company which provides bespoke search engines to medium sized clients. Morgan have awarded your company the contract to develop their new search system. You have been tasked by your manager with creating a project plan for the development of the system solution for Morgan.

3. Viewing videos is a commonplace activity on smartphones. However, interfaces for searching for scenes within video archives such as *YouTube* are very simple and make navigating through content very inefficient. Your company wants to advance the state-of-the-art in video search and browsing on mobile devices by developing a new interface for engagement with video content. You are tasked with developing novel potential interface components and an integrated prototype video search and browsing application using these components which can be used on a smartphone. It is important to recall when developing your solution 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) **[6 Marks]**

Compare and contrast the purposes of HTML and XML document markup? Use examples to illustrate your answer.

2(b) **[4 Marks]**

What does it mean to "jump into a hyperspace" for a hypertext structure, such as the World Wide Web (WWW)?

Explain how web search engines, such as *Google*, support users of the WWW to do this?

2(c) **[3 Marks]**

Search queries entered into Web Search engines are generally very short, typically consisting of only a few words. Why does this mean that retrieving relevant documents at the top of a ranked retrieval list for a large search collection such as the Web, based only on query-document content matching is unreliable?

2(d) **[8 Marks]**

Term weighting is an important factor in the effective ranking of documents in best-match information retrieval systems. One of the best known and most effective methods of term weighting is the Okapi BM25 term weighting function 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 as described by the equation above, explain the concepts of:

- *collection frequency weighting*,
- *term frequency* weighting,
- *document length normalisation*

How does use of these weighting factors in combination within the Okapi term weighting function enable the computation of effective terms weights for best-match information retrieval?

How do the k_1 and b factors operate in the equation for the Okapi BM25 model?

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. Using an example, explain why term proximity information can be a useful factor in determining the potential relevance of a document to a search query containing such a pair of terms?

[End Question 2]

QUESTION 3 ***[TOTAL MARKS: 25]***

3(a) **[8 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.

Give **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 using an example how the component can contribute to the effective selection of sentences to create a good snippet.

3(b) **[4 Marks]**

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

What do each of these metrics seek to measure about an information retrieval system?

3(c)

i. **[4 Marks]**

What are the three components of an information retrieval *test collection*?

Explain the features that should be taken into consideration for each of these components in order to create a test collection suitable for the evaluation of the effectiveness of an information retrieval system for a specific task.

ii. **[7 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.

iii. **[2 Marks]**

Can either or both of precision and recall be calculated reliably when using an information retrieval test collection created using pooling? Explain your answer.

[End Question 3]

QUESTION 4 ***[TOTAL MARKS: 25]***

4(a) **[5 Marks]**

User engagement with search engines is essentially a cognitive process in which a user seeks to satisfy their information need. When using a search engine, users often encounter difficulties in creating effective search requests due to the "non-specifiability of need"' problem.

What is the non-specifiability of need problem, and why do users encounter this problem?

How does this problem pose a challenge to information retrieval systems?

Why can document relevance for a request only be determined by a human relevance assessor?

4(b) **[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?

4(c)

i. **[6 Marks]**

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

Explain using examples 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?

ii. **[3 Marks]**

Explain using examples why conflation by string comparison methods can be important in information retrieval systems.

4(d)

i. **[2 Marks]**

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

ii. **[5 Marks]**

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 4]

QUESTION 5 ***[TOTAL MARKS: 25]***

5(a) **[6 Marks]**

Knowledge graphs encode information extracted from source texts. A knowledge graph typically describes the relationships between entities and the attributes of the entities.

Create a simple graphical example of a knowledge graph to illustrate these features. Your example should contain at least **4** entities, each with at least **2** attributes. No credit will be given for copying knowledge graph examples from the lecture notes.

5(b)

i. **[3 Marks]**

Sketch the standard workflow of a question answering system based on document retrieval.

ii. **[8 Marks]**

Explain in outline how a question answering system based on document retrieval can seek to use the documents to answer a question entered by a user for: (a) a small collection of documents, (b) a very large collection of documents.

iii. **[3 Marks]**

Why can a question answering system based on the use of a knowledge graph potentially answer questions that a document-based system is unable to?

5(c) **[5 Marks]**

Explain the principles of semantic representation of words using word embedding.

[End Question 5]

QUESTION 6 ***[TOTAL MARKS: 25]***

6(a) **[5 Marks]**

What is the *PageRank* algorithm?

Why can the *PageRank* algorithm be important in the operation of an effective Web search engine?

By means of a simple numerical example explain the principles of the calculation of the value the *PageRank* value of a Web page. No credit will be given for copying the example from the lecture notes.

6(b) **[8 Marks]**

Give and explain in outline **three** complex features typically used in learning-to-rank for web search.

Note: These features should not include the use of standard information retrieval ranking methods or PageRank. No credit will be given for describing information retrieval ranking methods or PageRank in the answer to this part of the question.

6(c)

i. **[3 Marks]**

What is meant by there being a "human-in-the-loop" in image and video search systems?

ii. **[3 Marks]**

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

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

[End Question 6]

[END OF EXAM]