DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2019/2020

MODULE: CA4009 – Search Technologies

PROGRAMME(S):

CASE	BSc in Computer Applications (Sft.Eng.)
EC	BSc in Enterprise Computing
DS	BSc in Data Science
CPSSD	BSc in ComputationalProblem Solv&SW Dev.
ECSA	Study Abroad (Engineering & Computing)
ECSAO	Study Abroad (Engineering & Computing)

YEAR OF STUDY: 3,4,O,X

EXAMINER(S):

Gareth Jones	(Internal)	(Ext:5559)
Prof. Mathieu d'Aquin	(External)	External
Dr. Hitesh Tewari	(External)	External
Dr. Samia Kamal	(External)	External
Prof. Brendan Tangney	(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 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 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 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. News journalists increasingly rely on rapid access to information relating to stories that they are working on so that they can create new reports very quickly in the environment of 24 hour news services. Journalists need to create stories varying from short rapidly written articles describing details of emerging news stories as they break, to longer more detailed articles describing an ongoing developing situation, e.g. a political or economic story. In order to provide their journalists with the widest variety of information for their research, a large national news agency has decided to develop a new news search tool to provide its journalists with online information from traditional published newspapers, radio, and TV services, as well as social media news services. Using this tool journalist should be able to easily compare related content from different news sources, so that they can choose how to progress their research on a news story and to prepare their own news reports.

2. You work for a company which supplies software to support online student learning. Your company has decided to develop a new product for use with remote delivery of teaching programmes, including online video lectures and notes and readings provided online to students to support the lectures. Your new product should include search functionality to enable students to locate material to help them with their studies, and efficient access to relevant sections of each video lecture, so that students do not have to view complete lectures to find relevant material. The search tool will also be used by lecturers to enable them to search for material related to lectures that they are preparing, so that they do not repeat material in existing lectures, and they can make sure that the content of their lectures is aligned with what students are learning in other courses.

3. A review of information management practices within a software consultancy company 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. This is particularly concerning for the managers of this company since the company undertakes many similar projects, and it is expected that the staff will make extensive use of documented experiences from previous projects when working on new ones; if the staff cannot find the relevant information from the documentation of previous projects, then this cannot happen, and there will be much expense arising from repeating previous work when undertaking new projects. The managers propose to develop a new 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.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

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

2(a) [4 Marks]

The purpose of an information retrieval (IR) system is to satisfy a user's information need. How does a user of an IR system, such as a Web search engine, use entry of a search query and subsequent interaction with the system to attempt to address their information need?

2(b) [4 Marks]

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

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

2(c) Relevance feedback (RF) methods are applied in IR systems following an initial retrieval pass or run. RF seeks to improve the search effectiveness of an IR system in a second retrieval pass by adjusting the parameters of the IR system and/or expanding the user's search query to better express their information need.

i. [5 Marks]

Adjustment of system parameters and query expansion in the application of RF in a retrieval pass requires the identification of relevant documents retrieved from the first retrieval pass.

What **3** ways are available to gather relevance information from the first retrieval pass to enable subsequent use of RF?

Comment on the likely reliability of each of these ways of gathering relevance information.

ii. [5 Marks]

Query expansion for RF using the Okapi probabilistic retrieval model operates using the offer weight $ow(i)$ for a term i, which is defined as follows,

$$ow(i) = r(i) \times rw(i)$$

where

$rw(i) =$ the Robertson/Sparck Jones weight
$r(i) =$ the number of **<u>known</u> relevant** documents term $t(i)$ occurs in,

where the Robertson/Sparck Jones weight is defined as follows,

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

$n(i) =$ the number of documents term $t(i)$ occurs in.
$N =$ the total number of documents in the collection archive.
$r(i) =$ the number of **<u>known</u> relevant** documents term $t(i)$ occurs in.
$R =$ the total number of **<u>known</u> relevant** documents in the collection archive.

$ow(i)$ is calculated for all terms present in the known relevant documents. The terms are then ranked by their $ow(i)$ values, with the highest ranking terms being added to the original query as query expansion terms.

Explain why the factors $r(i)$ and $rw(i)$ are useful measures of a term i's potential value as a query expansion term to enhance the user's initial query to improve IR effectiveness.

Following your arguments for the use of $r(i)$ and $rw(i)$ in selecting good expansion terms, why are terms with both high $r(i)$ and $rw(i)$ which get the highest $ow(i)$ values, likely to be the best overall expansion terms?

2(d)

i. [4 Marks]

Give the standard definitions of *Precision* and *Recall* as used in evaluation of information retrieval systems.

ii. [3 Marks]

Will query expansion tend to improve Precision or Recall or both of the output of an information retrieval system? Explain your answer.

[End Question 2]

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

3(a) [8 Marks]

i. Using an example, show how a *knowledge graph* encodes information taken from multiple source texts in a single data structure.

ii. Explain how search engine companies such as *Google* can create information summary "cards" using information taken from knowledge graphs to provide summaries of the attributes and associations of common entities as part of a Search Engine Results Page (SERP).

3(b) [6 Marks]

What is the PageRank algorithm as used in Web search?

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

3(c) [6 Marks]

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

ii. Outline **two** 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.

3(d) [5 Marks]

What is A/B testing as applied to online web applications? Describe a simple example of the use of A/B testing to evaluate a proposed new algorithm for use in an existing web search engine.

[End Question 3]

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

4(a) [10 Marks]

Prior to their entry into the search index of an information retrieval system, documents are generally preprocessed using *stop word* removal and a process of *conflation*.

After construction of the search index, user search requests are processed in the same way to ensure consistency with the preprocessing of the documents.

i. What are *stop words*? Give three examples of English stop words. What are the effects of removing stop words from the search index of an information retrieval system?

ii. Why is *conflation* applied in information retrieval? Outline the processes of conflation by use of *string-similarity* and *stemming*.

4(b) [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 do the k_1 and b factors operate in the equation for the Okapi BM25 model?

4(c)

i. [2 Marks]

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

ii. [5 Marks]

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

[End Question 4]

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

5(a) [6 Marks]

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

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

5(b) [4 Marks]

A summary is a condensed derivative of a source text, where the content is reduced 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.

5(c)

The temporal nature of videos means that browsing complete videos can be time consuming and inefficient. Browsing video snippet summaries consisting of important content taken from the video could it make much quicker for users to locate interesting material in a video.

i. [5 Marks]

Creating a video snippet would require the video to be preprocessed to locate important content within the video for inclusion in the video snippet.

One element of video processing would be to segment the video information stream. 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 professionally edited real-world video data such as movies?

ii. [5 Marks]

Another key component of many videos is the spoken information stream. Outline the key components of an automatic speech recognition system which could be used to transcribe the spoken contents of a video.

What problems are typically encountered by an automatic speech recognition system when seeking to transcribe spoken content?

iii. [5 Marks]

Outline a framework which could be used to create video snippet summaries based on identified video shots, a transcript of the spoken content, and any other audio or visual features which you wish to use.

You may find it useful to base your video segment creation framework on methods for the creation of text snippets of documents in Web search.

[End Question 5]

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

6(a) [4 Marks]

Compare and contrast the objectives of an information retrieval system and a recommender system.

6(b) [6 Marks]

The output of a recommender system generally takes a number of factors into account, including *relevance*, *novelty*, *serendipity*, *diversity*. Give brief definitions of each of these concepts.

A common application area for recommender systems is e-commerce websites, where the objective is for the recommender system to bring to the attention of the user of the website something which they might like to buy. Why is a successful recommender system integrated into an e-commerce website likely to incorporate multiple factors in determining its output?

6(c)

i. [3 Marks]

What is the *cold start* problem in recommender systems?

ii. [7 Marks]

Describe in outline the operation of a recommender system based on *collaborative filtering*. What are the key features of *memory-based* and *model-based* algorithms as used in collaborative filtering?

6(d) [5 Marks]

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

Give **2** examples of enterprise search tasks. What search challenges can you identify in your chosen examples?

[End Question 6]

[END OF EXAM]