DUBLIN CITY UNIVERSITY

AUGUST/RESIT 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*	☐	*Thermodynamic Tables*
☐	*Graph Paper*	☐	*Actuarial Tables*
☐	*Dictionaries*	☐	*MCQ Only - Do not publish*
☐	*Statistical Tables*	☐	*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 techniques or methods 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 of the search 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 test collection and choice of evaluation metrics

These points are suggestions, you are free to include any topics or materials that you wish in your answer, but the description of a complete solution containing all relevant elements is required to receive full marks.

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

1. Laurence Prentice Legal (LPL) provides a range of legal services to individuals and small to medium size organisations. LPL employs around 30 solicitors and legal secretaries and corresponding number of administrators and secretarial staff. The company has been in operation for more than 30 years, but has grown rapidly in recent years, and seeks to provide cost effective, high quality, services to its clients. In order to maximise the efficiency of its operations, LPL makes extensive use of previous case notes in order to set up new contracts, determine how best to proceed based on the progress of previous cases, and to advise its clients. The cost of searching these records means that they are often not able to access the potentially useful information available in their archives. In order to

improve the efficiency and coverage of their archive search, they have decided that they should install a new service to enable indexing and searching of their archives, many of which are only available as paper documents. You work for a search consultancy company providing bespoke search engines for medium seized enterprise clients. LPL approaches your company to propose a design of a search system to enable effective search of their archives by solicitors, legal secretaries and their office support staff. You are tasked with specifying a potential system solution for LPL.

2. You work for a news broadcaster providing content for traditional radio and TV news services, but also their online services in the form of an interactive web site and specialist content created for internet distribution either in video or audio only formats. Researchers working for the broadcaster use previous materials created by the broadcaster to help them develop new articles for broadcast, e.g. news stories and specialist documentaries. This can include directly finding information in previous broadcasts or seeking information in their archived production notes, Excerpts from previous broadcasts can also be reused within new content. The broadcaster also maintains an archive of materials recorded but never used in previous broadcasts, e.g. parts of interviews that were recorded but not included in edited broadcasters. The broadcaster has decided that it wants to make all of these materials available online to its news researchers via a new multimedia search platform. You work in the technical development division of the broadcaster and are tasked with managing a team to develop this new multimedia search platform.

3. Lecture notes provided online to students are typically based on one or more textbooks and other resources. If a student wants to learn more about a topic covered in a lecture, they typically need to find out what these books or resources are and then to spend time studying them to try to find relevant material and to relate it 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 the source text. In order to encourage students to bread more deeply in the source material, the IT department where you work at your local university has decided to create tool to automatically link the contents of lecture notes to related content, that is either the source of the notes or is related to it and describes the topics covered in individual slides in more detail. Your are tasked with developing a system to be made available to students which will automatically link the contents of sets of lecture notes provided to them for the modules that they are studying, to available online digital resources, and to be able to efficiently browse retrieved content to find relevant material to hep them in their studies.

[End Question 1]

Section B

Answer any 3 of the 5 questions in this section.

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

2(a) [4 Marks]

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

2(b) [8 Marks]

The Okapi BM25 term weighting function for best-match information retrieval is 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 Okapi BM25 model?

2(c) [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 taking term proximity into account be a useful factor in determining the potential relevance of a document to a search query containing such a pair of terms?

2(d) Preprocessing the contents of documents prior to indexing them in an information retrieval system can give significant benefits for search effectiveness.

i. [4 Marks]

A frequently used preprocessing step is *stop word* removal. Give $\mathbf{three}$ examples of English stop words, and explain why they are stop words.

Give two reasons why are stop words often removed in information retrieval systems.

ii. [5 Marks]

Another commonly used preprocessing step is *stemming*. 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) Hypertext is a vital technology in enabling exploitation of digital content.

i. [4 Marks]

What are *nodes*, *links* and *anchors* as applied in a hypertext?

ii. [3 Marks]

What does it mean to say that a hypertext containing multiple linked nodes has no beginning and no end?

3(b) [8 Marks]

What is the PageRank algorithm as used in WWW search? Use a simple example to outline the principles of the PageRank algorithm.

3(c)

i. [2 Marks]

What is enterprise search?

ii. [3 Marks]

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

3(d) [5 Marks]

Metadata can be used to annotate enterprise content with facets relating to the content items. Give three examples of typical facets in enterprise content.

How can facets be used to support search of partially remembered content in enterprise search?

[End Question 3]

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

4(a) [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.

4(b) [8 Marks]

Document snippets are often returned in ranked lists of documents returned by search engines. These snippets are intended to indicate the potential relevance of each retrieved document to the searcher's query.

Described four which factors might be taken into account to form effective snippet summaries for assessing relevance of retrieved documents in a search engine. Explain how each of these factors contributes to the creation of effective snippets.

4(c) [4 Marks]

i. Give the standard definitions of precision and recall as used in information retrieval.

ii. By considering the information needs and working context of the users of enterprise search and Web search engines, explain which of precision and recall is generally more important to users of each of these types of search systems.

4(d)

i. [6 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.

ii. [3 Marks]

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

[End Question 4]

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

5(a) [5 Marks]

i. What is the purpose a recommender system?

ii. Recommender systems typically use profiles of the preferences of individual users and the "ratings" of other users. Explain how each of these information sources is useful to a recommender system.

5(b) [6 Marks]

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

Why is using a combination of these features in determining the output of a recommender system likely to be effective?

5(c) [6 Marks]

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

5(d) [8 Marks]

Web search is a challenging information retrieval task. Effective search engines combine multiple signals in a process referred to a "'learning-to-rank".

i. Explain the principles 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 5]

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

6(a) [3 Marks]

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

6(b) [8 Marks]

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

i. Give a simple example of illustrate these features of a knowledge graph.

ii. When annotating the features or attributes of an entity, a knowledge graph can capture important features for this entity even if they are not found in an individual source text. Explain how this is possible.

6(c) Spoken document retrieval refers to searching for relevant spoken documents. Searching the contents of spoken documents generally requires that the contents have been identified using an automatic speech recognition system. Automatic speech recognition is very difficult and even the best current speech recognition systems make mistakes in identifying the words that have been spoken.

(i). [3 Marks]

What types of errors can be present in the output of an automatic speech recognition system? Use examples to illustrate how these would appear in the recognised output.

(ii). [3 Marks]

What effect can these recognition errors have on matching between queries and documents in spoken document retrieval?

6(d) [8 Marks]

Multimedia information retrieval based on the content of images and video involves automatically processing them to identify visual features which enable relevant content to be retrieved by matching them to example visual images used as queries.

i. Using examples of image analysis techniques, explain why the difference between human and computer-based analysis and interpretation of multimedia content is referred to as the *semantic gap*.

ii. Browsing temporal media such as video can be extremely time consuming, since the user often needs to view large amounts of video material in order find relevant content. This situation can be improved by the use of *shot boundary detection* and *keyframe* extraction methods to structure and represent the content of video archives. Describe a simple process of shot boundary detection. What problems are typically encountered in shot boundary detection? Your answer should include at least one solution to each of these problems.

[End Question 6]

[END OF EXAM]