DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2018/2019

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☐	*Dictionaries*	☐	*MCQ Only - Do not publish*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 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. University students are often given assignments which require them to search for material to enable them to complete assignments, be it a student studying history, writing an essay or dissertation, or a student studying computing needing to explore latest developments in algorithms or learning how to work with a new software development platform. An increasing range of information sources are available including traditional (but increasingly online) university library resources, documents on the internet, but also sources such as *YouTube* and MOOCs (Massive Online Learning Courses) which provide full courses on specific subjects, often delivered in video form by international experts. You are

tasked with developing a proposal for a portal to service to be provided by university libraries which provides integrated access to these resources by students. Your proposal is one of several which will be submitted by members of your team of software engineers in the technical support unit of a university. So you are encouraged by your manager to be as creative as possible, but to consider carefully the needs of the student target users. You can assume that the metadata of existing online services such as YouTube and existing MOOCs can be crawled for indexing in a local search engine.

2. News departments of major broadcasts are increasingly integrating their radio, TV and online news desks, with journalists and support staff working across these different media streams, In order to improve efficiency of their operation *Owl* news, a major news international news channel, is developing a new cross-media search engine which will combine all their online, radio and TV news archives, to support preparation of reports on current news stories. Retrieved material will either be used to develop details for inclusion in the current story or may itself be included in the current story. You are employed by a software consultancy who is employed to develop this new search engine, which should be well suited for use by journalists and news researchers in terms of ease of search and exploitation of retrieved relevant content.

3. *Tsurumi* is a specialist Japanese car manufacturer exporting to markets around the world. The central computer control system in Tsurumi cars means that they must be serviced by authorised dealers. These dealers typically also sell one or more major car brands with Tsurumi being a small niche part of their sales. Servicing Tsurumi cars and in particular the central computing system requires considerable expertise. The small volume of sales means that dealers generally do not have to work with these cars on a day to day basis, and so rely heavily on the online documentation provided by Tsurumi. Over the years Tsurumi have dealt with many technical enquiries from their dealers, and now have a considerable archive of responses to specific problems which have been created by their specialist engineers. In order to make use of the knowledge contained in these response documents, Tsurum now wish to index these into a specialist search engine which will also include there standard online documentation, and to make this indexed content available both to their engineers and to staff at their authorised dealers. The hope is that technical staff will be able to provide dealers with ready prepared responses to many enquiries, and that dealers themselves should be able to answer many of their problems without even needing to contact Tsurumi. You are a member of a team hired by Tsurumi to implement their vision of a search engine for technical support for their staff and authorised dealers.

[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 are the differences between HTML and XML document markup? Use examples to illustrate your answer.

2(b) [6 Marks]

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

ii. What is the "semantic gap" in search of visual media? Why does the semantic gap pose a challenge for multimedia search systems?

2(c) [4 Marks]

How can XML be used for content annotation in multimedia information retrieval for items such as images and video? Use examples to illustrate your answer.

2(d) [8 Marks]

Before it can be searched, a collection of video recordings must be analysed to identify its visual and audio features. These extracted features for the video are then entered into a multimedia search system.

Outline a range of analysis techniques that can be applied to the visual and audio content of a video, such as a television news broadcast or sports match, to prepare it for indexing by an interactive video search system.

2(e) [3 Marks]

Suggest how multiple audio and visual features extracted from a video could be combined to identify an event in the video, e.g. a goal in a soccer match, a wedding in a movie, or the key points in a scientific lecture.

[End Question 2]

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

3(a) Relevance feedback (RF) methods are applied in information retrieval (IR) systems following an initial retrieval pass or run. RF is designed to improve the search effectiveness of an IR system 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]

Application of RF requires the identification of relevant documents retrieved in a first retrieval pass.

What **3** sources of relevance information for use in RF are potentially available in an IR system?

Comment on the likely reliability of each of these sources of relevance information.

ii. [4 Marks]

Query expansion for 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 $\underline{\mathbf{known}}$ $\mathbf{relevant}$ documents term $t(i)$ occurs in,

and 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 $\underline{\mathbf{known}}$ $\mathbf{relevant}$ documents term $t(i)$ occurs in.
$R =$ the total number of $\underline{\mathbf{known}}$ $\mathbf{relevant}$ documents in the collection archive.

$ow(i)$ is calculated for all terms present in known relevant documents, which 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?

3(b) [4 Marks]

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

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?

iv. [3 Marks]

Effective application of RF improves the rank of relevant documents in retrieval runs carried out following the application of RF.
Will effective RF improve precision, recall or both? Explain the reasoning underlying your answer.

[End Question 3]

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

4(a) [4 Marks]

Give $\mathbf{three}$ examples of English stop words, and explain why they are stop words. Why are stop words often removed in information retrieval systems?

4(b) [5 Marks]

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?

4(c) [7 Marks]

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

ii. 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.

4(d)

i. [4 Marks]

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

ii. [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 in combination with suitably designed rich user interfaces, to facilitate effective enterprise search?

[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]

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 the Okapi BM25 model?

5(c) [5 Marks]

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

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

5(d)

i. [4 Marks]

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

Even if high quality question answering systems were available commercially, why would there still be a need for information retrieval systems?

ii. [5 Marks]

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

Suggest how a knowledge graph could be used for question answering instead of retrieving documents.

[End Question 5]

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

6(a) [6 Marks]

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

Why is a successful recommender system likely to incorporate all of these factors in determining its output?

6(b) [4 Marks]

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

ii. How does the cold start problem pose a challenge for new items introduced into the catalogue of an e-commerce website?

6(c) [6 Marks]

What are the features of a *knowledge-based recommender system*? For what tasks are knowledge-based recommender systems well suited? Why would a *content-based recommender system* or a *collaborative filtering* approach not be suitable for these tasks?

6(d) [3 Marks]

What is A/B testing as applied to online web applications?

6(e) [6 Marks]

i. Summarization is content reduction through *selection* or *generalisation* on what is important in the source. Using an example, explain what is meant by the concepts of *selection* and *generalisation* in this definition.

ii. How could the suitability of a range of possible summarization methods for items shown in the output of a recommender system be evaluated using A/B testing? Assume that items are selected for inclusion in the output of the recommender system based on a collaborative filtering method.

[End Question 6]

[END OF EXAM]