Queen Mary
University of London

***Summer Examination Period 2022 — May/June — Semester B***

***ECS736P INFORMATION RETRIEVAL Duration: 3 hours***

This a 2-hour online exam, which must be started within a 24-hour period. There is an additional 1 hour to resolve any IT issues.

**You MUST submit your answers within 3 hours of your exam start time.**

All instructions and guidelines from the exam page should be followed.

This is an open-book exam and you may refer to lecture material, text books and online resources. The usual referencing and plagiarism rules apply, and you must clearly cite any reference used.

Calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

**Answer ALL questions**

**You MUST adhere to the word limits, where specified in the questions. Failure to do so will lead to those answers not being marked.**

**YOU MUST COMPLETE THE EXAM ON YOUR OWN, WITHOUT CONSULTING OTHERS.**

**Examiners:**
Dr. Q Zhang and Dr. T Roelleke

**Question 1**

(a) There are various IR models, and they build on logic (Boolean expressions), vector algebra and probability theory. Describe how these mathematical foundations are applied in the main IR models. In your explanation, provide some of the main branches of IR models.

**[10 marks]**

(b) Assume a document collection with the following 6 documents:

D1: Jedi use lightsaber.

D2: Jedi Yoda is Jedi master.

D3: Jedi Master Yoda ligthsaber is green.

D4: May the force be with Jedi

D5: Hope Jedi Order's last hope

D6: New hope last Jedi.

And a query ”Jedi Yoda hope force.”

Punctuations are ignored, and small letters and capital letters are treated as the same. Assume no stemming, lemmatization or stopwords in this question.

Draw a term frequency (TF) table, in which columns are query terms ”Yoda”, ”hope”, ”force” and ”Jedi”, and rows are from D1 to D6.

**[10 marks]**

(c) Which similarity metric is usually considered suitable for measuring the similarities between TFIDF vectors? Why?

Using the dataset given in Q1.(b), rank D1-D6 according to their similarity score to the query without document length normalisation. Show the steps.

**[5 marks]**

**Question 2**

(a) Describe the zero-probability problem and how to work around this problem in the context of Language Models (LM).

**[5 marks]**

(b) Without considering the zero-probability problem,

(i) give two most common used types of language models. Does any of these models assume word independence?

(ii) in building a language model, how do you rank the models M1, M2, M3 and M4 containing terms t1, t2, t3, with respect to query Q = t3, t1, t2, t1 based on unigram models?

Table 1: Models and terms.

| **Models** | **t1** | **t2** | **t3** |
|---|---|---|---|
| M1 | 4 | 7 | 1 |
| M2 | 1 | 2 | 3 |
| M3 | 5 | 4 | 6 |
| M4 | 7 | 4 | 4 |

**[10 marks]**

(c) BIRM model: A document collection contains 44 documents, including the following:
$d_1$: “There is a river full of memory.”
$d_2$: “In this river our memory is found.”
$d_3$: “Magic flows in the river songs.”
Consider a query $q$ = “river memory found” ($t_1$:river, $t_2$:memory, $t_3$:found) and the three terms’ document frequencies as specified in the table below. There are a total of 20 relevant documents with respect to this query according to user relevance feedback.

| | Number of relevant documents containing $t_i$* | Number of documents containing $t_i$ |
|---|---|---|
| $t_1$ | the last digit of your student number (before slash“/”) | 15 |
| $t_2$ | the second last digit of your student number (before slash“/”) | 16 |
| $t_2$ | the third last digit of your student number (before slash“/”) | 14 |

*For example, if your student number is 202212345/1, then use 5 for $t_1$, 4 for $t_2$ and 3 for $t_3$.

Please use the binary independence retrieval model (BIRM) to derive a retrieval function based on the information given above, and calculate the relevance scores for the three documents. For the sake of simplicity, use $c_i = \frac{a_i(1-b_i)}{b_i(1-a_i)}$ without the logarithm, and ignore the constant $C$.

**[10 marks]**

**Question 3**

(a) In extended boolean model, given the relevance function for an or-query as

$$R(d, q) = \sqrt{\frac{(w_1^2 + w_2^2 + ... + w_m^2)}{m}}$$

and the relevance function for an and-query as

$$R(d|q) = 1 - \sqrt{\frac{((1 - w_1)^2 + (1 - w_2)^2 + ... + (1 - w_m)^2)}{m}}.$$

Consider three terms $t_1$, $t_2$ and $t_3$ and a query $q = (t_1 \vee t_2) \wedge t_3$, how should the relevance function $R(d, q)$ be written between the query and document $d$? If $t_1$ and $t_3$ are present in $d$ but $t_2$ is not, what is the similarity value?

**[5 marks]**

(b) (i) What assumption employed in the vector space model (VSM) is no longer assumed in the generalised vector space model (GVSM)?

(ii) To implement that, instead of relying on pair-wise orthogonality between term vectors, what other orthogonal vectors are used to form the base of the vector space?

(iii) In the new space, each term vector $\vec{t_i}, i = \{1, 2, ..., 5\}$ is expressed as a linear combination of the new vectors. Write down the representations of term co-occurrence patterns, and the corresponding new vectors to be used in GVSM. (When writing long vectors you can use a short form by replacing some vector elements in the middle by "......". However, make sure you write at least the first 3 and the last elements of the vector).

**[10 marks]**

(c) Following Question 3(b), suppose we have 10 documents:

d1 =(2, 1, 0, 0), d2 =(1, 5, 0, 0), d3 =(1, 3, 1, 1),
d4 =(0, 0, 2, 2), d5 =(0, 4, 1, 2), d6 =(0, 0, 1, 1),
d7 =(0, 0, 2, 1), d8 =(1, 1, 0, 0), d9 =(2, 1, 1, 1), d10 =(0, 1, 2, 2).

(i) What are the vectors that form a base?

(ii) Provide the representation for $\vec{t_2}$.

**[10 marks]**

**Question 4**

(a) 15 documents are retrieved on the basis of query *Q*. Assume that the precision is 0.33 and the recall is 0.25 for this retrieval. What is the total number of relevant documents in the collection for query *Q*?

**[5 marks]**

(b) Discuss the differences between text-based and content-based image retrieval (CBIR) approaches. What are their advantages and disadvantages?

**[10 marks]**

(c) For the following evaluation metrics, which of the attributes (A.to E.) can be captured?

(i) Precision-recall curve:

(ii) Precision at 10:

(iii) F1 measure:

(iv) ROC curve:

(v) NDCG: beyond the part level.

A. Cardinality of irrelevant documents set
B. Relative ranking between relevant documents
C. Recall
D. Precision
E. Relevance of highest ranked documents

**[10 marks]**

**End of questions**