Beijing-Dublin International College

SEMESTER 2 FINAL EXAMINATION - 2022/2023

School of Computer Science

COMP3009J Information Retrieval

Dr. Robert Ross
Assoc. Prof. Neil Hurley
Dr. David Lillis *

Time Allowed: 120 minutes

Instructions for Candidates

Answer Question 1 and any two other questions. Question 1 has 30 marks available. All other questions have 35 marks available.

BJUT Student ID: _____ **UCD Student ID: _____**

I have read and clearly understand the Examination Rules of both Beijing University of Technology and University College Dublin. I am aware of the Punishment for Violating the Rules of Beijing University of Technology and/or University College Dublin. I hereby promise to abide by the relevant rules and regulations by not giving or receiving any help during the exam. If caught violating the rules, I accept the punishment thereof.

Honesty Pledge：_____Le Liu_____
(Signature)

Instructions for Invigilators

Candidates are allowed to use non-programmable calculators during this examination.

Question 1:

(a) Describe what is meant by *"information need"* in the context of Information Retrieval. What are the different types of information need?

User only use IR when there is some information that they interesting.
Information need is motivation for using IR
Visceral Need, Conscious Need, Formalised Need, Compromised Need

[6 marks]

(b) Below is part of a positional index relating to the term "same". In creating this index, stopword removal and stemming have not been used. Postings lists begin at 1 for the first term in each document. Which document(s) could contain the phrase "the same day at the same time"? Explain your answer.

```
<same: 41825;
1: 120, 124, 167;
2: 9, 10, 13;
3: 121, 162;
4: 4, 101, 105, 106;
5: 1, 5, 88, 888, 889;
...>
```

1,,4could contain the phrase
Because they each have two occurrenices of "same" that are exactly 4 positions apart. While, In to must have "same same". In 5 same must at begain.

[6 marks]

(c) A *modern Information Retrieval pipeline* may include Boolean searches, simple ranking and reranking based on machine learning. Explain why these are all useful to make an effective Information Retrieval system.

[6 marks]

(d) The *BM25* method of Information Retrieval is based on the belief that a good term weighting scheme is based on three principles. Briefly describe each of these principles.

Inverse Document Frequency(IDF): Terms that appear in fewer documents are more informative.
Term Frequency(TF): The higher frequency, more important
Document Length: Long documents are penalized to avoid unfair advantage from term repetition.

[6 marks]

(e) Compare and contrast the preprocessing steps of *stemming* and *lemmatisation*. In particular, what are the advantages and disadvantages of each?

Stemming is the process reduce the words to a common rot, often by suffix stripping.
Adv: Fast, Simple, Reduces vocabulary size
Disadvan: Overstemming and Homographs
Lemmatisation is a NLP technique for converting word into lemmas
Adv: More accurate, return real word
Dis: slower, need more analysis

[6 marks]

[Total 30 marks]

Question 2:

(a) The *Boolean Model* makes use of the query operators *AND*, *OR* and *NOT*. Explain how these work and how they affect the number of documents returned by an Information Retrieval system. Also show how each of these can be implemented by using operations from Set Theory. 01b

AND is used to narrow a search. More AND, Fewer records
OR is used to broaden a search, Documents contains any term specified will be return
NOT is used to specifically exclude a term from search. More NOT, Fewer records.

[6 marks]

(b) Briefly describe **two** ways in which the process of running Boolean queries can be optimised so that they can be processed more efficiently. 01c

1. Process in order of increasing of frequency: start with smallest, if reach the end of shorter list, we can stop.
2. Skip Pointer: it can reach later parts of list without interating through every element.

[6 marks]

(c) The *probabilistic model* of Information Retrieval makes use of two probabilities relating to query terms. These are $P(k_i|R)$ (the probability that a relevant document will contain the term k_i) and $P(k_i|\bar{R})$ (the probability that a non-relevant document will contain the term k_i). However, these probabilities cannot be calculated directly and must be estimated.

(i) Briefly describe how initial values for these probabilities may be generated.

(ii) Explain how these initial estimates can be improved with user feedback.

[8 marks]

(d) Below is a small document collection, containing three documents. Answer the questions that follow.

1. Lowercasing all word.
2. Remove stopwords
3. Tokenisation

Stopwords: and, be, is, it, to, will

Document 1: It is going to rain and rain and rain today.

Document 2: Today I will be playing sport.

Document 3: I am going to watch the play.

(i) Describe the preprocessing steps you would use when creating an index for these documents.

(ii) Calculate a vector to represent each document, using the TF-IDF weighting system. You should use the stopword list provided, but do not perform stemming.

(iii) Calculate the cosine similarity for each vector using the query "going to play football", and show the final ranked list of documents for this query.

football", and show the final ranked list of documents for this query.

(iv) What effect on the results would you see if you had used stemming for this corpus?

[15 marks]

[Total 35 marks]

(ii)

doc1 : going, rain, rain, rain, today

doc2 : today, i, playing, sports

doc3 : i, am, going, watch, play

vocabulary:

am, going, i, play, playing, rain, sport, today, watch

IDF	$\log\frac{3}{1}$	$\log\frac{3}{2}$	$\log\frac{3}{2}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{2}$	$\log\frac{3}{1}$
doc1	0	$\frac{1}{3}$	0	0	0	1	0	$\frac{1}{3}$	0
doc2	0	0	1	0	1	0	1	1	0
doc3	1	1	1	1	0	0	0	0	1

doc1 = [0, 0.1949, 0, 0, 0, 1.584, 0, 0.1949, 0]

doc2 = [0, 0, 0.5849, 0, 1.584, 0, 1.584, 0.584, 0]

doc3 = [1.584, 0.5849, 0.5849, 1.584, 0, 0, 0, 0, 1.584]

(iii)

q = [0, 0.5849, 0, 1.5849, 0, 0, 0, 0]

len(d1) = 1.6038

len(d2) = 2

len(d3) = 2.236

len(q) = 1.6813

cosSim(d1, q) = 0.042

cosSim(d2, q) = 0

cosSim(d3, q) = 0.7555

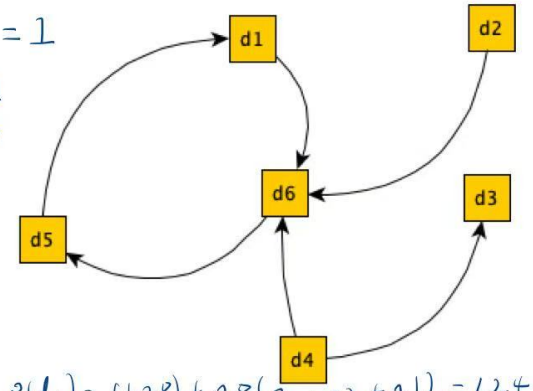
d3, d1, d2

(iv): if stemming, "playing" and "play" to "play".
This would increase the similarity between Doc2. It results Doc2 more relevant in the ranking.

Page 4 of 6

Question 3:

(a) The link structure of some web pages is shown below. There are six web pages shown (d1, d2, d3, d4, d5 and d6), and the arrows show links between the pages (e.g. d4 contains links to d3 and d6).



Show a worked example of how *PageRank* scores are calculated for these documents. Use a damping factor of 0.8 and show at least 3 iterations. [14 marks]

(b) Compare the *MAP*, *bpref* and *NDCG* evaluation metrics. In your answer, outline any advantages or disadvantages of each. For each metric, suggest a situation where it is more appropriate than the others. [9 marks]

(c) Below is a set of results and relevance judgments for a query:

Retrieved = d_{13}, d_{21}, d_{19}, d_{12}, d_6, d_{24}, d_{11}, d_1, d_3, d_{17}, d_9, d_{23}, d_{10}, d_{14}

Relevant = {d_2, d_3, d_7, d_9, d_{12}, d_{15}, d_{17}, d_{23}}

Calculate the following metrics:

(i) Mean Average Precision (MAP)

(ii) Recall

(iii) R-Precision

[12 marks]

[Total 35 marks]

[12 marks]

[Total 35 marks]

Question 4:

(a) The *Rocchio Algorithm* uses a *modified query vector* to achieve relevance feedback. Explain why this is effective in improving the effectiveness of an Information Retrieval system, and how this modified query vector can be calculated.

It moves the query closer to relevant document and away from non-relevant ones in vector space. This helps better capture the user's true information need.
Why works: This method is especially helpful when recall is important, and users can identify relevant results

[10 marks]

(b) The table below shows results from three search engines in response to the same query. Each set of results consists of a ranked list of unique document identifiers (DocID), along with the ranking score. Complete the following tasks, showing your workings for each.

(i) Calculate the ranking score that document D6 would have using *CombSum*.

(ii) Calculate the ranking score that document D10 would have using *CombMNZ.*

(iii) Calculate the ranking score that document D5 would have using *Borda Fuse*.

Engine A

DocID	Score
D10	0.60
D9	0.57
D12	0.48
D11	0.46
D8	0.41
D1	0.37
D6	0.26
D7	0.19

Engine B

DocID	Score
D5	971
D1	936
D11	860
D2	516
D8	414
D6	300
D4	153
D10	99

Engine C

DocID	Score
D12	9.23
D1	9.00
D2	7.88
D6	6.69
D7	5.03
D8	4.22
D5	3.63

[9 marks]

(c) Different levels of *corpus overlap* can influence the design of a data fusion algorithm. Explain why this is the case.

Disjoint databases(Collection Fusion): No document appears in more than one result set. Chorus effect is not applicable. Fusion must treat each result set independently.
Identical databases (Data Fusion): Documents will frequently appear in multiple result sets. Repeated appearances are strong signals of relevance(Chorus Effect). Fusion should give higher scores to such doc
Overlapping databases: Documents may appear in multiple result sets. It hard to judge if absence means irrelevance or unawareness.

[10 marks]

(d) Briefly describe *three* sources of synonyms for use in *query expansion*.

1. Manually-created thesaurus, such as WordNet
2. Automatic-created thesaurus from some external corpus like Wikipedia
3. Word embeddings, where words are represented by vctors, e.g. woerd2vec

[6 marks]

[Total 35 marks]

C = 12

System A: (4+3+2+1)/4 = 2.5

System B give D_5 : 12

System C give D_5 : 6

Borda D_5 = 12+6+2.5 = 20.5

Q4.b) i)

d_6 (Normalized) $\quad \frac{\text{unNormal} - \min}{\max - \min}$

SysA 0.171

SysB 0.231

SysC: 0.546

CombSum (d_6) = 0.171 + 0.231 + 0.546 = 0.948

ii) D_{10}

	System A	System B	System C	Fused
d_{10}	1	0	0	1