Beijing-Dublin International College

SEMESTER 2 FINAL EXAMINATION - 2023/2024

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.

Instructions for Invigilators

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

All numeric answers should be given in decimal format, and be correct to 3 places of decimals.

Question 1:

(a) A *modern Information Retrieval pipeline* may include Boolean searches, simple ranking (using BM25, for example) and reranking based on machine learning. Explain why each of these are useful to make an effective Information Retrieval system.

[6 marks]

(b) The *Cranfield Paradigm* is typically used as the basis of evaluation strategies for Information Retrieval. Explain in detail what is meant by this.

[6 marks]

(c) The *information need* of a user is said to have 4 stages. Describe these stages.

[6 marks]

(d) When *tokenising* text, the natural language that the documents are written in can influence the strategy being used. Discuss three examples of issues that can arise when tokenising languages other than English.

[6 marks]

(e) Explain what is meant by the phrase *Adversarial Information Retrieval*. In the context of web search, show two examples of where this can occur.

[6 marks]

[Total 30 marks]

Question 2

(a) This question relates to using postings lists to index a document corpus.

(i) Explain why an *inverted index* is a more suitable data structure for representing a document corpus, compared to a *term-document incidence matrix*.

(ii) Describe in detail how a set of *postings lists* can be created to represent a document corpus. Your answer should include details of the data structures that are used during this process.

[10 marks]

(b) This question relates to preprocessing.

(i) What is meant by *stopword removal*, and how can this help to improve the Information Retrieval process?

(ii) In what way is *Zipf's Law* related to stopword removal?

(iii) Explain what is meant by *stemming* and *lemmatisation*? Give advantages and disadvantages of each.

[10 marks]

(c) Below is a small document collection, containing three documents. Answer the questions that follows. All calculations should be presented in decimal format, and be correct to three decimal places.

Stopwords: a, an, is, some, the

Document 1: Her doctor gave her some very, very bad news.

Document 2: A no news day is a good news day.

Document 3: An apple a day keeps the doctor away.

(i) 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 or lemmatisation.

(ii) Calculate the cosine similarity for each vector using the query "What is her news?" and show the final ranked list of documents for this query.

[15 marks]

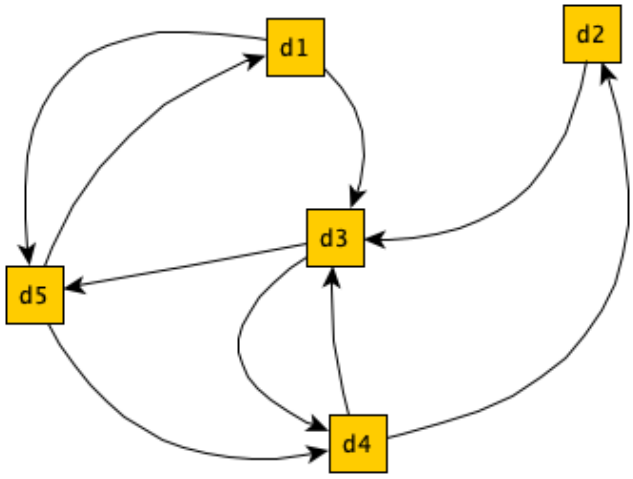
[Total 35 marks]

Question 3

(a) This question relates to PageRank.

(i) In the context of PageRank, what is a *damping factor* and why is it important?

(ii) The link structure of some web pages is shown below. There are five web pages shown (d1, d2, d3, d4 and d5), and the arrows show the links between the pages.



Using this structure as an example, describe in detail how a PageRank score is calculated. Use a damping factor of 0.85 and show at least 3 iterations (the first step of assigning the same initial PageRank to each page does not count as an iteration).

Your answer must include a description of the steps you take, in addition to the calculations.

Answers must be given in decimal format and be correct to three places of decimals.

[15 marks]

(b) This is question is related to the evaluation of Information Retrieval systems.

(i) Compare and contrast the P@10, MAP and NDCG evaluation metrics. In your answer, outline any advantages and disadvantages of each. For each metric, suggest a situation where it is more appropriate than the others.

(ii) Explain how the creators of the *bpref* evaluation metric showed that it is suitable for dealing with *incomplete relevance judgments*?

[12 marks]

(c) Below is a set of results that were returned by a search engine in response to a query.

Retrieved = d7 d1 d10 d14 d8 d2 d18 d17 d4 d20 d21 d3 d11 d12 d5 d24

Judged Relevant = {d1, d3, d6, d7, d8, d13, d17, d19, d21, d24}

Judged Non-relevant = {d9, d10, d11, d12, d18, d25}

Calculate the following evaluation metrics for this query. Your results should be presented in decimal format, and be correct to three places of decimals:

(i) Precision
(ii) Recall
(iii) Mean Average Precision (MAP)
(iv) bpref

[8 marks]

[Total 35 marks]

Question 4

(a) Describe in detail how you would design a *web crawler* to find information that could be included in the index of an Information Retrieval system. In your answer, include details about what standards the crawler should follow, and situations where accessing information is difficult.

[10 marks]

(b) The table below shows results from four 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. Answers should be correct to three places of decimals.

(i) Calculate the score that document d16 would have using *CombMNZ*.
(ii) Calculate the score that document d18 would have using *Borda Fuse*.
(iii) Calculate the score that document d10 would have using *CombSum*.

Engine A	
DocID	**Score**
d19	0.999
d23	0.873
d2	0.784
d25	0.733
d10	0.733
d18	0.608
d9	0.507
d16	0.435
d15	0.055

Engine B	
DocID	**Score**
d22	3.710
d24	3.655
d18	3.421
d12	3.411
d10	3.339
d6	2.976
d11	2.780
d20	2.248
d21	2.149
d16	1.717
d9	1.054

Engine C	
DocID	**Score**
d10	98.762
d6	97.898
d2	89.051
d17	66.673
d8	53.929
d3	22.150
d7	10.985

Engine D	
DocID	**Score**
d15	0.974
d13	0.957
d12	0.911
d24	0.871
d22	0.728
d20	0.690
d23	0.626
d10	0.385
d16	0.163
d18	0.146
d7	0.122
d14	0.081
d8	0.020

[10 marks]

(c) Three *effects* can be exploited when designing a fusion algorithm.
   (i) Describe the three effects.
   (ii) Based on the fusion algorithms you have studied, give *four* examples where one of these effects is exploited, and explain how they exploit this effect.

[9 marks]

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

[6 marks]

[Total 35 marks]