Beijing-Dublin International College

SEMESTER 2 FINAL EXAMINATION - 2016/2017

School of Computer Science

COMP3009J Information Retrieval

Prof. Pádraig Cunningham
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： ____ **(Signature)**

Instructions for Invigilators

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

Question 1:

(a) In the context of Information Retrieval, what is meant by "*relevance*"?

[6 marks]

(b) What is the role of a *web crawler* in a web search engine?

[6 marks]

(c) When *tokenising* text, the natural language being processed affects the strategy used. Give three examples of issues that can arise when tokenising languages other than English.

[6 marks]

(d) What is *stopword removal*? In your answer, explain how this benefits the Information Retrieval process and also what disadvantages it may have.

[6 marks]

(e) Discuss why you might use *postings lists* to represent a document corpus instead of using a *term-document incidence matrix*.

[6 marks]

[Total 30 marks]

Question 2:

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

Stopwords: a, the, to

Document 1: I went to the shop to buy a football.

Document 2: I saw Chicago Fire play a football match.

Document 3: Jack used a match to light the fire.

(i) Create postings lists to represent these documents. In your answer, you should use stopword removal but do not use stemming.

(ii) Using the query "fire AND match" as an example, explain how two postings lists can be efficiently merged to perform an AND operation.

(iii) What is the time complexity of merging two postings lists? Explain your answer.

[10 marks]

(b) 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]

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

Stopwords: from, in, the

Document 1: Player strike threatens top flight fixtures.

Document 2: Top strike from Mane settles the derby.

Document 3: Derby weekend in top flight.

Document 4: Top clubs in strike threat.

(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 “top flight results”, and show the final ranked list of documents for this query.

(iv) Describe what changes you would make so that users can search for two-word phrases (e.g. “top flight”).

[17 marks]

[Total 35 marks]

Question 3:

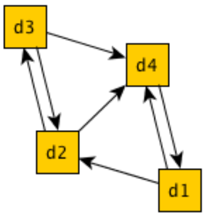

Figure 1

(a) Figure 1 shows the link structure of some web pages. There are four web pages shown (d1, d2, d3 and d4), and the arrows show links between the pages (e.g. d1 contains links to d2 and d4).

Show a worked example of how *PageRank* scores are calculated for these documents. Use a damping factor of 0.85 and show at least 3 iterations.

[12 marks]

(b) What is meant by *precision* and *recall*? Explain why they are generally not used individually to measure the performance of IR systems.

[5 marks]

(c) Modern document collections suffer from the problem of *incomplete relevance judgments*.

(i) What is meant by the phrase “incomplete relevance judgments”?

(ii) What assumption must be made so that traditional evaluation metrics can be applied to collections that have incomplete relevance judgments?

(iii) Explain how the *bpref* evaluation metric deals with this situation.

[6 marks]

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

Retrieved = d_5, d_1, d_4, d_2, d_3, d_6

Below is a set of relevance judgments for the same query. Assume a 0-3 scale of graded relevance judgments where 3 is a highly relevant document and 0 is a non-relevant document. All documents not included have been judged to have a relevance of 0.

Relevant = {[d_1,2], [d_3,1], [d_4,3], [d_5,3], [d_6,2]}

Calculate the following metrics:

(i) Mean Average Precision (MAP)

(ii) Discounted Normalised Cumulated Gain (NDCG)

[12 marks]

[Total 35 marks]

Question 4:

(a) There are three *effects* that may be exploited by a Fusion algorithm.

(i) Briefly describe each of these effects.

(ii) Describe how the *SlideFuse* algorithm exploits these effects.

[9 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 score that was used for ranking. Complete the following tasks, showing your workings for each.

(i) Normalise the scores for each ranked list.

(ii) Fuse the results using the *CombSUM* algorithm.

(iii) Fuse the results using the *linear combination model*, using the following weights:

Engine A: 3, Engine B: 2, Engine C: 1

(iv) Explain why it is necessary to perform normalisation for score-based fusion.

Engine A

DocID	Score
D10	0.97
D12	0.96
D9	0.88
D8	0.87
D1	0.69
D2	0.25
D4	0.09
D5	0.05

Engine B

DocID	Score
D2	980
D11	803
D8	801
D3	622
D10	423
D12	376
D5	85
D4	75

Engine C

DocID	Score
D4	9.12
D11	8.15
D2	6.70
D8	6.21
D6	5.53
D7	4.07
D5	3.25
D12	0.51

[17 marks]

(c) Explain what is meant by the phrase *adversarial information retrieval*. In the context of web search, show two examples of where this is occurs.

[9 marks]

[Total 35 marks]