Beijing-Dublin International College

SEMESTER 2 FINAL EXAMINATION - 2017/2018

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) Information retrieval is defined as dealing with the "representation, storage, organization of, and access to information items". Explain why *representation* is an important part of Information Retrieval.

[6 marks]

(b) Below is part of a positional index relating to the term "fear". In creating this index, stopword removal and stemming have not been used. Which document(s) could contain the phrase "the only thing we have to fear is fear itself"? Explain your answer.

```
<fear: 215230;
1: 2, 4, 130;
2: 20;
3: 134, 199;
4: 7, 100, 102, 156, 279;
5: 8, 88, 888, 890, 891;
...>
```

[6 marks]

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

[6 marks]

(d) Explain what is meant by *stemming*? In your answer, include examples of situations where performing stemming is useful, along with examples of where it may have a negative effect on retrieval.

[6 marks]

(e) The evaluation of Information Retrieval systems generally follows the *Cranfield Paradigm*. Explain in detail what is meant by this.

[6 marks]

[Total 30 marks]

Question 2:

(a) One method to represent an inverted index is to use *postings lists*.

(i) Describe the structure of a postings list.

(ii) What data structures would you choose to represent this type of index in a Python program? Explain your choices.

(iii) Write pseudocode to show how two postings lists can be efficiently merged to perform an OR operation.

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

[10 marks]

(b) *BM25* retrieval is based on the belief that good term weighting comes from three principles. Describe these principles and identify what part of the BM25 formula (below) relates to each one.

$$sim_{BM25}(d_j, q) \sim \sum_{k_i \in d_j \wedge k_i \in q} \frac{f_{i,j} \times (1+k)}{f_{i,j} + k\left((1-b) + \frac{b \times len(d_j)}{avg\_doclen}\right)} \times log\left(\frac{N\text{-}n_i + 0.5}{n_i + 0.5}\right)$$

[8 marks]

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

Stopwords: he, his, in, was, is

Document 1: He washed his coat in New York.

Document 2: My dog's coat was washed yesterday.

Document 3: My new coat is very, very warm.

(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 scheme. You should use the stopword list provided, but do not perform stemming.

(iii) Calculate the cosine similarity for each vector using the query "his new coat", and show the final ranked list of documents for this query.

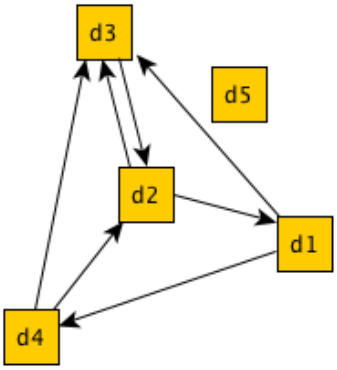
(iv) Describe what changes you would make so that users can search for two-word phrases (e.g. "new coat").

[17 marks]

[Total 35 marks]

Question 3:

(a) 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 links between the pages (e.g. d1 contains links to d3 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) When evaluating Information Retrieval systems, when is a document considered to be *relevant*?

[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 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) Precision at 5 (P@5)

(iii) R-Precision at R=0.3

[12 marks]

[Total 35 marks]

Question 4:

(a) *Adversarial Information Retrieval* occurs when search engine crawlers are presented with document contents that are different to what users see. Explain why this is a problem for search engines, and outline *three* ways in which false content can be presented to crawlers.

[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 *Borda Fuse* algorithm.

(iv) Briefly describe the three *effects* that can be exploited by a Fusion algorithm. Which of these are exploited by CombSUM and Borda Fuse?

Engine A

DocID	Score
D10	92.85
D17	92.01
D18	89.58
D4	86.59
D8	84.84
D16	84.06
D14	78.56
D6	77.49

Engine B

DocID	Score
D18	0.93
D14	0.88
D17	0.79
D5	0.78
D3	0.75
D16	0.44
D20	0.35
D4	0.3

Engine C

DocID	Score
D14	782
D20	711
D18	507
D9	377
D11	331
D12	110

[17 marks]

(c) Modern Information Retrieval systems have user interfaces that make it easier for users to satisfy their information need. Briefly discuss *three* ways in which a user interface can help users in this way.

[9 marks]

[Total 35 marks]