NUI Galway
OÉ Gaillimh

Autumn Examinations 2019-2020

Course Instance Code(s)	4BCT1, 1CSD1, 1CSD2, 1MAI1, 4BS2
Exam(s)	4th B.Sc. Computer Science and IT
	M.Sc. in Computer Science (Data Analytics)
	M.Sc. in Computer Science (Artificial Intelligence)
	B.Sc. (Hons)
Module Code(s)	CT4100
Module(s)	Information Retrieval
Paper No.	1
External Examiner(s)	Dr.Jacob Howe
Internal Examiner(s)	Professor M. Madden
	*Dr. Colm O'Riordan

Instructions: Answer any 3 questions.
                              All questions will be marked equally.

Duration	2 hours
No. of Pages	
Discipline(s)	School of Computer Science
Course Co-ordinator(s)	Dr D. Chambers

Requirements:

Release in Exam Venue	Yes ☐	No ☐
MCQ Answer sheet	Yes ☐	No ☐
Handout	None	
Statistical/ Log Tables	None	
Cambridge Tables	None	
Graph Paper	None	
Log Graph Paper	None	
Other Materials	None	
Graphic material in colour	Yes ☐	No ☐

PTO

Integrity Statement

In submitting this work I confirm that it is entirely my own. I acknowledge that I may be invited to online interview if there is any concern in relation to the integrity of my exam, and I am aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations*: https://www.nuigalway.ie/media/registry/exams/QA230---Procedures-for-Dealing-with-Breaches-of-Examination-Regulations.pdf*

Q. 1.

a) Outline, with an appropriate example, a suitable indexing strategy to deal with Boolean queries. Discuss an approach for handling queries:

        i)   (term1 OR term2)
        ii)   (term1 AND term2)

Outline an efficient approach to dealing with general Boolean queries.

(10 marks)

b) Outline a suitable data structure to allow searching for the presence of terms and prefixes of terms in a passage of text. Illustrate how the following terms would be placed in such a data structure:

stand, stood, standing, door, star

(8 marks)

c) Normalisation is often used in term weighting schemes. Explain the term *normalisation* in this context and with reference to any well-known weighting scheme discuss any approach to normalisation.

(7 marks)

Q.2.

A company has a large set of scientific articles (each of which contains a title, abstract, authors, key words, year of release, main body of the paper and a bibliography).

a) Suggest a means to measure the similarity between two documents based on:
        i)   Content of the document
        ii)   Authors
        iii)   Bibliographies
        iv)   Content, authors and bibliographies   (8 marks)

b) The company wishes to rank all papers that are relevant to a given query and to then re-order the papers in the answer set according to how authoritative or influential the papers are. Outline an approach that could be used to give a suitable solution for this requirement.   (10 marks)

c) Outline a suitable approach to cluster these documents in the collection into useful sub clusters that may be of use in user search tasks. Briefly list and limitations of the approach.   (7 marks)

Q. 3.

a) The extended Boolean model has been often been used to overcome some of the limitations of the classical Boolean model. Show how you would calculate the relevance to a document *d* to each of the following queries:

      i)   rough AND rowdy
      ii)   rough OR rowdy

                              (10 marks)

b) Given a ranked answer set and evaluation judgements, explain how you would generate a precision-recall graph. (5 marks)

c) Learning mechanisms has been used successfully in information retrieval. Using an approach of our choice, suggest a learning mechanism to identify good weights to apply to terms in documents. Discuss any limitations of your approach (10 marks)

Q. 4.

a) Query expansion is often used to improve the performance of information retrieval systems. Outline an approach to suggesting terms to add to a query in the absence of user feedback. (8 marks)

b) Collaborative filtering systems often struggle to make recommendations to new users to a system (as the new users will not have given any recommendations). Propose and approach to help overcome this problem.

                              (7 marks)

c) The majority of information retrieval systems return a ranked list of results for a given query. Propose an alternative means to present the return set that:

      i)   shows the relationship between the query and the returned documents
      ii)   shows the relationship between the returned documents.

                              (10 marks)