NUI Galway
OÉ Gaillimh

Semester II 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 user submits a query, *q*, to an information retrieval system which returns a ranked list to the user. Given the top *k* ranked documents.

i) Explain how, in terms of IR evaluation, one could measure the quality of this answer set. Discuss any limitations of this approach. (6 marks)

ii) Explain how the query *q* could be modified using evidence from the returned set to improve performance. Outlines any potential limitations of this approach. (10 marks)

iii) Given the original query *q* and the modified query *q'*, discuss, in your words, how you might predict which query is likely to perform better. (9 marks)

Q. 2.

a) Outline, with an appropriate example, a suitable indexing strategy to deal with both Boolean and proximity queries. (8 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, with a suitable example, how the data structure operates. (9 marks)

c) Describe, in your own words, with reference to any well-known term weighting scheme, the main constituents of a *good* weighting scheme. (8 marks)

Q. 3.

a) Recommender systems are used to generate recommendations for users on unseen items. Collaborative filtering is one such approach. Explain the main stages of collaborative filtering and illustrate this approach by generating a recommendation for item "Oh Mercy" for user "Jack". (10 marks)

	Dylan	**Desire**	**Saved**	**Tempest**	**Oh Mercy**
Jack	1	5	3	3	
Lily	1	4			4
Rosemary	4	4	3		5
Rita	1	4			

b) In many domains, in addition to collaborative ratings, extra information is often present and may prove useful in the provision of useful recommendations. For example, in the domain of music recommendations, we may have information on each album being rated. So, in addition to each rating present, we have for each album the following set of attributes: year, producers, musicians, genre, sub-genre, songs.

Outline an approach where this extra information can be used to hopefully improve the performance of the system. (8 marks)

c) In your words, describe how you would measure the performance of your system designed in part b). (7 marks)

Q.4.

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 listed
   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)