Unit Information: SIT772 Database and Information Retrieval
Trimester: 2021 T3
Assessment 2: Information Retrieval Problem Solving Task

This document supplies the detailed information on assessment tasks for this unit.

Key information

- **Due: Sunday, 06 Feb 2022**, 23:59 (AEST)
- **Weighting:** 30%
- **Submit:** Through CloudDeakin

Learning Outcomes

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
ULO 5: Demonstrate data retrieval skills in the context of a data processing system.	GLO 1: Discipline-specific knowledge and capabilities

Purpose

This task evaluates the student's technical skills in the management of unstructured data, with potential usage in real applications. This assessment supports student understandings of the techniques related to unstructured data management and data processing

Instructions and Submission Guide

This is an **individual** assessment task. Students are required to submit ONE written report.

- Read these instructions and the following questions.
- ONE written report with the name as using student ID_givenname_A2.pdf, e.g., 123456_Kevin_A2.pdf)
- **The report must be submitted via CloudDeakin assessment portal. The wrong submission venue or the wrong submitted file may lead to the penalty.**

Question 1: (6- 4+2)

Try and find a Query of the form [Query-term-1, Query-term-2] (without quotes) that, when run on Google, produces at least one result that contains only one of TWO terms. That is, try to find an example where Google does not interpret a the-term query as a conjunction. (If you have difficulty with finding an appropriate query, try one that produces very few hits, say, fewer than 20.)

(i) Take screenshot of the first page of Google results (or more if you want to) and mark each result with 2 (both terms occur on the page), 1 (one term occurs on the page) or 0 (neither term occurs on the page)

(ii) Based on this evidence, does Google interpret all queries as a Boolean conjunction? Explain.

Question 2: (16: 8+4+4)

Recall and Precision are two important evaluation metrics that we use to analyze a set of unranked results. Precision and Recall metrics consider the differences between set of documents retrieved for given query and the set of documents that are relevant to the user's need.

A) Compute Recall, Precision and precision@5 for the following retrieval against Queries Q1, Q2 and Q3

	Relevant document	**Retrieved Document**
Q1	1,14,17,23, 24, 33,54, 55, 59, 74,101,103	2,5,7,23, 33,50, 55, 59, 77,98, 99, 101, 103, 110,120
Q2	14,19, 25, 27,30,39, 42, 63, 769, 790,1563	14, 21, 25, 26,27, 38, 42, 63, 569, 769, 790, 1565, 1589
Q3	8, 11,32,54,67,69,78, 79, 91,99,111,122	11, 13, 17, 19, 21, 32,77,79, 99,102,111,122
Q4	4,26, 38, 63, 569, 769, 790, 1565, 1589	14, 21, 25, 26,27, 38,63, 88, 769, 790,

B) Recall and Precision are often discussed together as their focus is on complementary information. If precision is important, the we don't not want to see any non-relevant documents. That is, whatever is retrieved, should be relevant. If recall is important, we want to see all the relevant documents, even if it requires sifting through some non-relevant ones. Provide and Justify two information-seeking tasks where precision may be considerably more important than recall. Similarly, Provide and Justify two information-seeking tasks where recall may be more important than precision. [Don't forget to justify your choices: Justification will be graded, not the particular choices].

C) The trade-off between Recall and Precision may be user-specific i.e. some users may be interested in precision than recall and vice versa. How the search engine try to guess without asking, whether user cares more about precision than recall, or vice versa? Think of different ways, users interact with a search engine and be creative!

Question 3: (6: 3x3)

(a) Consider, we have three collections C1, C2, and C3 that have 500, 15,000 and 300,000 documents respectively. We have added C1 or C2 into C3. Which collection is likely to have more new terms added to its vocabulary (C1+C3 or C2+ C3) and why?

(b) Calculate the tf-idf for below documents.

a. D1: Sweets Potatoes are Sweet
b. D2: Sweet Oranges are sour and Sweet
c. D3: I have sweet Apple, Sweet Orange, Sweet Potatoes

Question 4: (10-5x2)

Doc-id	house	for	sale	in	Geelong	Melbourne
1	39	11	32	22	22	4
2	19	19	3	15	16	21
3	19	20	1	3	21	9
4	12	20	14	1	13	13

➔ (houses OR for OR sale OR in OR Geelong OR Melbourne)
➔ (houses AND for AND sale AND in AND Geelong OR Melbourne)

Suppose these are issued to a search engine that uses the ranked Boolean retrieval model. Assume, for simplicity, only four documents in the collection (with document ids 1-4).
Answer the following questions. The above table gives the number of times each query-term occurs in each document.

(i) Compute the document scores and the ranking associated with the query (houses OR for OR sale OR in OR Geelong OR Melbourne).

(ii) How is the ranking produced probably sub-optimal and why does this happen?

(iii) Compute the document scores and the ranking associated with the query (houses AND for AND sale AND in AND Geelong OR Melbourne).

(iv) How is the ranking produced probably sub-optimal and why does this happen?

(v) How would you extend the Boolean retrieval model to handle AND NOT constraints (e.g., houses AND NOT Geelong)? Your proposed solution should give a higher score to documents that contain fewer occurrences of the term to the right of the AND NOT (e.g., Geelong). Please be as mathematical as possible. In other words, saying: "I would reduce the score for documents that contain the word to the right of AND NOT." is too vague.

(vi) Using the index, what would be the Boolean retrieval model scores given to documents 1-4 by your proposed scoring method for the query "houses AND NOT Geelong"?

Question 5: (12-4x3)

Doc1: A book is considered a good book that makes the reader feels better.
Doc2: I love reading good books to feel better.
Doc3: One can feel better after reading Tom's recent book.

Query-1: I love books that are good
Query -2: reading good books make you feel better
Stop Word Dictionary=[is, can, after, a, to, I, the, about, that]

i. Explain the similarity scores of both Query -1 and Query -2 using TF-IDF.
ii. How would the result change if TF-IDF is used instead of TF as Query?
iii. What do prefer using TF or TF-IDF as Query (Support your claim using F-score).

Assessment feedback
General feedback to the class will be provided via CloudDeakin-Discussion Forum. The formal assessment feedback will be released with the marks in CloudDeakin altogether.

Extension requests
Requests for extensions should be made to Unit/Campus Chairs 3 days early before the assessment due date.

Special consideration
You may be eligible for special consideration if circumstances beyond your control prevent you from undertaking or completing an assessment task at the scheduled time.

See the following link for advice on the application process: http://www.deakin.edu.au/students/studying/assessment-and-results/special-consideration

Assessment feedback
Detailed written feedback and results will be provided within two weeks of submission.

Referencing
You must correctly use Harvard referencing in this assessment. See the Deakin referencing guide.

Academic integrity, plagiarism and collusion
Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.

Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.

Collusion occurs when a student obtains the agreement of another person for a fraudulent

purpose, with the intent of obtaining an advantage in submitting an assignment or other work.

Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: https://www.deakin.edu.au/students/study-support/referencing/academic-integrity