SIT772: Database and Information Retrieval
Assessment 2: Information Retrieval Techniques Problem Solving Task

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

Key information

- **Due:** Sunday, 9 February 2020, 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

- Read these instructions.
- Answer as many questions as possible.
- Place your name, ID and answers in your document.
- Please submit your Word file with your answers and graphs (embedded) where appropriate as a SINGLE document in the Submission Portal.
- Do not submit PDF files.

Question 1 (15 marks)

Suppose you have joined a search engine development team to design a search algorithm based on both the Vector model and the Boolean model.

You have collected the following (3) documents (unstructured) and plan to apply an index technique to convert them into an inverted index.

Doc 1 : data science is a field to use scientific method, process, algorithm, system to extract knowledge.

Doc 2 : data mining is the process to discover pattern in large data to involve method at the database system.

Doc 3 : information system is the study of network of hardware and software that people use to process data.

To answer the below questions, you have to provide the detailed procedures step by step.

Question 1.1: In the process of creating the inverted index, please complete the following steps:

Remove all stop words and punctuation. The list of stop words for this task is provided as follows:

Is, An, That, Use, And, To, From, In, Both, Of, At, The

Question 1.2: Create a merged inverted list including the within-document frequencies for each term.

Question 1.3: Use the index created as above to create a dictionary and the related posting file.

Question 1.4: Please design three Boolean queries, (e.g., **web AND search**) and list the relevant documents for each query. Each query must contain at least two keywords while no one keyword appears in one document only.

Question 1.5: Please use the Vector model to query on the inverted index, and compare the result with the Boolean model. (Hint: you can use cosine similarity and set a similarity threshold).

Question 2 (IR Evaluation) (15 marks)

In this question, you are required to evaluate the performance of different search engines. First, please find two search engines you are familiar with, such as Google, Bing, Yahoo!, etc.

Second, please choose one target from the following list, and design two queries to search in both search engines. So both query 1 and query 2 have to be tested in both search engines.

Target 1: obtain the new features of the new iPad.
Target 2: obtain the user manual for installing Tera Term.
Target 3: obtain a tutorial how to install Oracle SQL.
Target 4: obtain the features of the new Xbox one.

Third, select the first 20 results in both search engines, if they return the target, then mark them as relevant documents, otherwise, they are irrelevant. We can assume there are **12** relevant documents in total (retrieved and not-retrieved). If you think there are more relevant documents to be searched, you can use higher expected relevance as threshold.

The following questions are based on your search results.

Question 2.1: List your target, results and designed search queries (You can use any keywords you think are related to the target). For each result, you can click the link and go to the page, and take the screenshot if you think this result is relevant. At your report, you are required to provide the screenshots and detailed explanation why they are relevant to your queries.

Question 2.2: Get the precision and recall values for 20 documents for query 1 in search engine 1. Interpolate them to 11 standard recall levels. Then plot them into a chart. Get the precision and recall values for 20 documents for query 1 in search engine 2. Interpolate them to 11 standard recall levels. Then plot them into a chart.

Question 2.3: Get the precision and recall values for 20 documents for query 2 in search engine 1. Interpolate them to 11 standard recall levels. Then plot them into the same chart as above. Get the precision and recall values for 20 documents for query 2 in search engine 2. Interpolate them to 11 standard recall levels. Then plot them into the same chart as above.

Question 2.4: Now find the average interpolated precision of query 1 and query 2 for search engine 1 and plot it into the same chart. So you will have total of 3 interpolated curves in one single chart. Now find the average interpolated precision of query 1 and query 2 for search engine 2 and plot it into the same chart. So, you will have total of 3 interpolated curves in one single chart.

Question 2.5: Plot the average interpolated values for Search Engine 1 and Search Engine 2 on one single chart, and compare the algorithms in terms of precision and recall. Which search engine do you think is superior? Why?

Assessment feedback

General feedback to the class will be provided via FutureLearn and Discussion Forum in CloudDeakin. Students will have the opportunity to seek additional feedback during the fortnightly seminar sessions.

Assessment 2: Information Retrieval Techniques Problem Solving Task

Submission details

The assessment file should be a Microsoft Word file with the filename formatted as: FirstName_Surname_StudentID_CourseCode_A2 (eg. Susan_Wolf_123456_SIT772_A2.docx).

The document is to be submitted electronically via Assessment 2 in CloudDeakin by the due date. The submission link is on the FutureLearn SIT772 Assessments tab on the Program page.

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 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

Question 1	Excellent	Good	Fair	Unsatisfactory
Remove all stop words and punctuation	Output suitable word list on all of the three documents. **1 mark**	Output suitable word list on two of the three documents. **0.5 mark**	Output suitable word list on one of the documents **0 .25 mark**	Fail to output stemming results on any document. **0 mark**
Create a merged inverted list including the within-document frequencies list index	Successfully created the merged inverted list index. **3 marks**	2/3 of terms are correctly indexed. **2 marks**	Half of the terms are correctly indexed. **1 mark**	Fail to create an inverted list index. **0 mark**
Create a dictionary and the related posting file	Successfully created the dictionary. All terms are pointing to the posting file **3 marks**	2/3 of the dictionary file is correct. All related terms are pointing to the posting file **2 marks**	Half of the dictionary file is correct. All related terms are pointing to the posting file **1 mark**	Fail to create a dictionary and the related posting file. **0 mark** **0 marks**
Boolean model	The correct documents for all of the three queries are correct. **4 mark**	The correct documents for two of the three queries are correct. **2 mark**	The correct documents for one of the three queries are correct. **1 mark**	Fail to get the correct documents for any of the queries. **0 marks**
Vector Model	The correct documents for all of the three queries are correct. Provide compare result. **4 marks**	The correct documents for two of the three queries are correct. Provide the compare result. **2 marks**	The correct documents for one of the three queries are correct. Provide the compare result**.** **1 mark**	Fail to get the correct documents for any of the queries. Fail to compare vector and Boolean models**.** **0 marks**
Question 2	**Excellent**	**Good**	**Fair**	**Unsatisfactory**
List the target, results and related queries	List target, both queries and results for both queries. **3 marks**	List target and results, but only one query**.** **2 marks**	List target, but not queries/results**.** **1 mark**	Fail to list the target and queries**.** **0 marks**
Query 1: Plot recall and precision curve to interpolated levels for search engines 1 and 2.	Both Precision and recall values and curves are correct. **3 marks**	Precision and recall values for both queries are correct, but related curves are incorrect. **2 marks**	Precision and recall values for one query are correct, but another one is Incorrect **1 mark**	Curves are missing or wrong. **0 marks**

IT Help @ Deakin

Question 2 (cont.)	**Excellent**	**Good**	**Fair**	**Unsatisfactory**
Query 2: Plot recall and precision curve to interpolated levels for search engines 1 and 2.	Both values and curves are correct. **3 marks**	Precision and recall values for both queries are correct, but related curves are incorrect. **2 marks**	Precision and recall values for one query are correct, but another one is incorrect. **1 mark**	Curves are missing or wrong. **0 marks**
Average curves and compare queries for search engines 1 and 2.	Provide correct interpolations for both curves. **3 marks**	Provide approximate correct interpolation but with some key points missing. **2 marks**	Provide one correct interpolation but other's a wrong **1 mark**	Fail to provide the interpolation **0 marks**
Average curves and compare two engines 1 and 2	Obtain correct comparison result with full explanation and the average curves. **3 marks**	Obtain correct comparison result, but does not provide enough explanation/curves. **2 marks**	Compare part of result, but the result in doubtable **1 mark**	Fail to compare them **0 marks**