SIT772 Database and Information Retrieval

Assessment Task 2

This assessment task enables students to demonstrate their proficiency against Unit Learning Outcome 5. ULO5: Demonstrate data retrieval skills in the context of a data processing system.

Assessment 2 (Individual)	**Written report**
Weight (% of total mark)	30
Due date	Monday, 28 May 2018 5 PM AEST
Submission method	Through CloudDeakin via Future Learn
Referencing style	Harvard

Please read the rubric carefully as it outlines what criteria your assessment will be evaluated on.

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 documents (unstructured) and plan to apply an index technique to convert them into an inverted index.

Doc 1：Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing.

Doc 2：Information retrieval is finding material of an unstructured nature that satisfies an information need from within large collections.

Doc 3：Information systems is the study of complementary networks of hardware and software that people and organizations use to collect, filter, process, create, and distribute data.

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

a. Remove all stop words and punctuation, and then apply Porter's stemming algorithm to the documents. The list of stop words for this task is provided as follows:

   Is, The, Of, To, An, A, From, Can, Be, On, Or, That, Within, And, Use

b. Create a merged inverted list including the within-document frequencies for each term.

c. Use the index created in part (b) to create a dictionary and the related posting file.

d. You may like to test the inverted index by using the following keywords: `information, system, index`

e. Please design three Boolean queries, (for example, `web` AND `search`) and list the relevant documents for each query.

f. 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.

i. Target 1: obtain the course information for S779.
ii. Target 2: obtain the price of the new Samsung Tablet.
iii. Target 3: obtain the manual of installing tera term.
iv. Target 4: obtain the oracle SQL tutorial.
v. Target 5: obtain the price of 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. Assume that you have **14** relevant documents in total (retrieved and not-retrieved).

The following questions are based on your search results.

a) List your target, results and designed search queries (You can use any keywords you think are related to the target).

   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 2 in search engine 1. Interpolate them to 11 standard recall levels. Then plot them into the same chart as above.

   Now find the average 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 curves in one single chart.

b) List your target, results and designed search queries

   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.

   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.

   Now find the average 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 curves in one single chart, separate to that of part (a).

c) Plot the averages for Search Engine 1 and Search Engine 2 on a separate chart, and compare the algorithms in terms of precision and recall. Which search engine do you think is superior? Why?