**School of Engineering and Applied Science**

**Takeaway assessment answer booklet**

**As with all graded assessments, you are required to undertake this assessment individually and adhere to the University Regulations on Student Discipline.**

**For your information, and awareness, below are statements describing plagiarism and collusion.**

PLAGIARISM: *“where a student uses, without acknowledgment, the work of other people and presents it as their own which may give an unfair advantage over others. Intentional and unintentional acts of plagiarism (whether reckless or otherwise) will be construed as offences.”*

COLLUSION: *"where two or more people have worked together without permission to produce a piece of work which is then submitted for assessment as the work of only one person, which may give an unfair advantage over others. Action may be taken against a student who has allowed their work to be used as well as against a student who submits work resulting from collusion."*

***By submitting my assignment, I declare that:***

I have personally prepared this assignment and that it has not previously, in whole or in part, been submitted for THIS, OR any other degree or qualification.

The work described here is my own, carried out personally by me unless otherwise stated.

Where applicable all sources of information, including quotations, are acknowledged by means of reference, both in the final reference section, and at the point where they occur in the text.

I understand that plagiarism and collusion are regarded as offences within the University’s Regulations on Student Discipline and may result in formal disciplinary proceedings.

*I understand that by submitting this assessment, I declare myself fit to be able to undertake the assessment and accept the outcome of the assessment as valid*

Please add, your typed answers (where possible) to your takeaway assessment in this answer booklet.

Ensure you clearly indicate which question number your answer relates to, this includes any embedded images. Where applicable, the final reference section should be submitted at the end of the relevant question.

If you have any additional information related to a question, this also needs to be included in this answer booklet at the end of document, again clearly labelled with the corresponding question number.

Once completed, save this answer booklet as a PDF document, and submit within the 24hr timeframe.

More detailed guidance related to this takeaway assessment can be found in the Takeaway assessment folder.

**Question 1**

**A(i)**

Question b is an information retrieval question.

The reason for this is because the structure of the data regarding this question would be unstructured. Question A would be a simple query on the appropriate database, which holds structured data.

**A(ii)**

**What is defined as success (A):** visitor successfully finds what they are looking for.

**Measure:** number of return visits

**What is defined as success (B):** the user successfully buys what they were looking for.

**Measure:** the number of visitors that are converted to buyers.

**A(iii)**

**Example to measure relevance to user query**

* User wants to find information of the causes of the financial crisis.
* Writes query: ‘Causes of the financial crisis’
* Returned result: ‘Article titled: Financial crisis causes a 10% increase in cases of clinical depression.’

The return result, although it is almost a perfect match for the user’s query, it is irrelevant to the user’s information need. Therefore, this result cannot be used to measure user satisfaction of the search query.

**A(iv)**

Ranked retrieval returns the documents in a collection, in order of relevance with respect to the query submitted. This kind of retrieval does not suffer from the feast-or-famine problem meaning it does not return to little or too many results. The user would not be overwhelmed with too many documents being returned, as the documents are ordered, only the top N may be shown. Unlike that of Boolean retrieval, which returns many results in no specific order.

**A(v)**

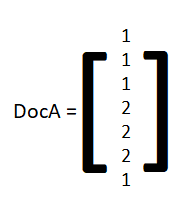
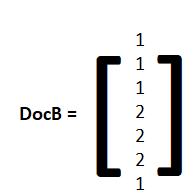
**What is the bag-of-words model for documents?**

**What is the difference from the set-of-words model?**

Every document is represented as a vector of dimension, which is equal to the total number of terms found in the collection dictionary. Every vector element that corresponds to a term in the dictionary, is equal to the total number of times the term appears in the document. In the set of words model, the elements of the vector are either a 1 or 0, 1 being the term appears in the document and 0 being the term does not appear in the document.

**A(vi)**

* **What is the bag-of-words representation of the documents docA:” Alice is quicker than Bob who is quicker than Charlie” docB:” Charlie is quicker than Alice who is quicker than Bob”**
* **Using your answer as illustration, state the main weakness of the bag-of-words representation**

* The Main weakness regarding the bag-of-words representation is that it completely ignores the word order. In the above example, both documents consisted of the same words, only in a different order, therefore, it resulted in both having the same bag of words. This is even though they both have different meanings.

**A(vii)**

**Document clustering**

* Document clustering is the process of grouping documents. The process groups documents into different classes, so that documents within the same class are similar, while documents in different classes are different.
* One application of this would be a better navigation of search results, which improves search recall by returning all the documents within the same cluster, when the query matches a document.
* Another application would be improving the speed of retrieval, which carries out a search on cluster centres rather than the whole collection.

**Question 2**

**a)**

**Precision:** = **0.75**

**Recall:** = **0.33**

**b)**

A search engine can obtain the maximum possible recall by returning more documents. This can be accomplished by adding more keywords. Recall is a non-decreasing function of the number of docs retrieved. A system that returns all docs has 100% recall.

**c)**

Precision is a trade-off from recall. A search engine can almost certainly obtain the maximum possible precision by restricting the positive predictions. Easy to get high precision for very low recall.

**d)**

**A close up of text on a white background

Description automatically generated**

**e)**

**A close up of a piece of paper

Description automatically generated**

**f)** Based on the curves shown in the above precision curves, we can assume that search engine B is better than search engine A. this is because it has a perfect precision, and the recall is at its maximum (1.0,1.0). engine b is located more towards the top right, meaning it has better results.

**Question 3**

**a)**

- When using the Boolean model, the query written is given a set of terms joined with Boolean operators. For the document to be evaluated against the query, each term in the query is substituted with either a true or a false. This substitution is made depending on whether the term is in the document. After this, the truth value of the resulting Boolean expression is computed. True values deem the document to be relevant

- Examples of the usage of Boolean retrieval would be email searching, library catalogues etc.

**b) Term-document incidence matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Doc1** | **Doc2** | **Doc3** | **Doc4** |
| **Breakthrough** | **1** | **0** | **0** | **0** |
| **Drug** | **1** | **1** | **0** | **0** |
| **For** | **1** | **0** | **1** | **1** |
| **Schizophrenia** | **1** | **1** | **1** | **1** |
| **New** | **1** | **1** | **1** | **0** |
| **Approach** | **0** | **0** | **1** | **0** |
| **Treatment** | **0** | **0** | **1** | **0** |
| **Of** | **0** | **0** | **1** | **0** |
| **Hopes** | **0** | **0** | **0** | **1** |
| **Patients** | **0** | **0** | **0** | **1** |

**c) inverted index representation**

Doc1

Breakthrough

Doc2

Doc1

Drug

Doc4

Doc3

Doc1

For

Doc4

Doc3

Doc2

Doc1

Schizophrenia

Doc3

Doc2

New

Doc1

Doc3

Approach

Treatment

Doc3

Doc3

Of

Patients

Doc4

Hopes

Doc4

**d) –** inverted indices is much more efficient than the full term-document incidence as it strictly tells us which documents the term can be found in(stores only the non-zero positions), making it more efficient is terms of memory and speed. Whereas, the full term document tells us were the document is AND where it is not (storing the non-zero and zero values simultaneously). An example using my illustration would be that the term ‘breakthrough’ in the inverted index representation tells me that it is in doc1, whereas, in the full term-document matrix, it tells me it is in doc1 and also tells me it is not in docs 2, 3 or 4.

**e)**

**i)** schizophrenia AND drug = doc1, doc2

**ii)** for AND NOT (drug OR approach) = doc4

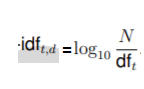
**Question 4**

**a)**

**tf** = term frequency

**df** = document frequency

**N** = total number of documents in the collection.

**b)**

**c)**

**The weight given to a term that appears in all the documents in the collection is zero (0)**

**d) Purpose of idf scores**

**-** When using idf, it is assumed the terms that appear in fewer documents carry more relevant information content, therefore, these terms would be given a higher score when the vector space representation of a document is form. Whereas, idf puts a penalty on terms that would appear in many documents in the collection, giving them a zero weight. Examples of words like this would be ‘the’, ‘a’, ‘if’ etc.

**e)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dictionary** | **tf** | **df** | **idf** | **Tf-idf** | **tf** | **df** | **idf** | **Tf-idf** |
| **and** | 1 | 1,000,000 | 0 | 0 | 1 | 1,000,000 | 0 | 0 |
| **cameras** | 1 | 10,000 | 2 | 2 | 1 | 10,000 | 2 | 2 |
| **digitals** | 1 | 100 | 4 | 4 | 1 | 100 | 4 | 4 |
| **video** | 1 | 10,000 | 2 | 2 | 2 | 10,000 | 2 | 4 |
| **players** | 1 | 10 | 5 | 5 | 0 | 10 | 5 | 0 |

**f)**

A picture containing text, whiteboard

Description automatically generatedThe unnormalized vectors of tf-idf weights for the 2 documents.

**A picture containing text, whiteboard

Description automatically generated**

The norms of the vectors are as follows. You need to take each value in the tf-idf weights vector, square them and add them all together. Then square root the result to get the norms…

**A picture containing text, whiteboard

Description automatically generated**

The cosine similarity is as follows. The first value of each(squared) + second value of each(squared) and so on, divided by the norms attained from the step above multiplied.

**Answer =**