COMP631 Introduction to IR
Homework 1

February 7, 2022

Upload one zip file containing the following two files to Canvas-Assignments:
1) a scanned handwritten solution or typed pdf file named "*YourLastName_HW1.pdf*" containing your answer to each question;
2) a zip file named "*YourLastName_HW1P1.zip*" containing your programs for Question 1.
(Note: The output of program for each subquestion in Question 1 should be included in the pdf file "*YourLastName_HW1.pdf*".)

– Due Date: February 17, 2022 11:59pm

1 Programming Question

Given the following documents:

- Today DeepMind released AlphaCode!!!!!! Go AlphaCode!
- The AlphaCode is released today.
- the AlphaCode won last week.
- Find the AlphaCode news last week.
- AlphaCode is a system released by DeepMind.

1) **Preprocessing**: Conduct punctuation removal, stop word removal, case-folding, lemmatization, stemming on the documents. Please provide references or rules you use for each task. Output each document after preprocessing as a term sequence.

2) Construct the term-document incidence matrix based on the result of the question above. Report the resultant incidence matrix.

3) Implement TF-IDF on the term-document incidence matrix. Output the resultant TF-IDF matrix.

4) Measure the similarity between every pair of documents using cosine similarity. Organize the output in a 5×5 matrix.

2 Indexing

1) Consider the documents in Question 1, graphically illustrate the inverted index you will create for the documents.

2) Given the query "release AlphaCode", make use of the index we just build to retrieve the matched documents. Illustrate the query processing procedure in details.

3) Explain the advantages of inverted index over term-document incidence matrices for organizing texts?

3 Tolerant Retrieval

1) Build a permuterm index for "sydney" and "sidney", respectively.

2) Consider the wildcard query "s*dney", explain in details how to search this query in the permuterm indices. What is the rotated wildcard query that we will look up for in the permuterm indices?

3) Compute the edit distance between the terms in pair as below:

- "AlphaCode" and "AlphaGo"
- "november" and "december"
- "condense" and "confidence"

4) Compute the trigram overlap between each pair of terms list above.

4 Evaluation Metrics

For a particular search query, your IR systems returns 125 relevant documents and 50 irrelevant documents. There are a total of 200 relevant documents in the overall collection.

1) What are the *precision* and *recall* of the system on the search?

2) Can you think of a search scenario where *recall* is preferred over *precision*? Explain.

3) An obvious alternative to the *precision/recall* metric is the *accuracy* score, that is, the percentage of its classifications that are correct. Please explain the vulnerability of *accuracy* in evaluating the performance of IR systems alone. Provide one example where a decent accuracy score may be accompanied with terrible IR performance.
(Hint: In most real world IR circumstances, the data is extremely skewed, i.e., normally most of the documents are nonrelevant with the query, no matter what input query is given.)