Homework 2

To be submitted 13th june 2021

1. Given

Doc 1: feature engineering used in software engineering

Doc 2: software engineering is fun

i. Draw the posting list for: software and engineering

ii. Draw the term-document incidence matrix

2 Write a query using **Westlaw** syntax which would find any of the words information systems or technology in the same paragraph as a form of the verb study.

3. discuss the effect of stemming in **precision and recall**

4. what is the difference between **web crawler** and **A web scraper, which one is used in information retrieval.**

5. what are the main problem of boolean search

6. compute the **Jaccard coefficient** for each of the two documents below?

– Query: Cairo is the fun

– Document 1: I am having fun at Cairo University

– Document 2: Cairo is the capital of Egypt

7. why do we need log-frequency weight

8. compute the cosine similarity between the following documents given the term raw frequency in each Document

Term	doc1	doc2	doc3
Information	1000	0	100
Systems	100	10	10
FCI	0	10	1
Cairo	10	1	1

9. Compute the $\mathbf{wt}_{,d}$ for the terms/document given in the table in # 8

10. Why The Euclidean distance is a bad idea for measuring similarity between documnts.