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

i.

software [2] → 1 2

engineering [2] → 1, 2

ii.

	Doc1	Doc2
feature	**1**	**1**
engineering	**1**	**0**
used	**1**	**0**
in	**1**	**0**
software	**1**	**0**
is	**0**	**1**
fun	**0**	**1**

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.

information /1 systems technology /p stud!

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

Stemming decreases the size of the vocabulary and it can increase the retrieved set then it may lower the precision since it may return more documents than needed.
It may cause the recall to increase or stay th same (never reduce it) since it will return the same or mor documents.

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

- **A web crawler sometimes called a “spider,” is a standalone bot that systematically scans the Internet for indexing and searching for content, following internal links on web pages.**

- **A web scraper is a process of extracting specific data. Unlike web crawling, a web scraper searches forspecific information on specific websites or pages.**

web crawler is the one used in information retrieval.

5. what are the main problem of Boolean search
The feast or famine problem

- Boolean queries often result in either too few (≈0) or too many (1000’s) results.
- It takes a lot of skill to come up with a query that produce a manageable number of hits.

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

Jaccard coefficient for Document 1: 2/9 = 0.2222
Jaccard coefficient for Document 2: 3/7 = 0.4286

this is not part of the exercise but will be required in the exam
Document 2 is better match to the query because it has bigger Jaccard coefficient *(not graded)*

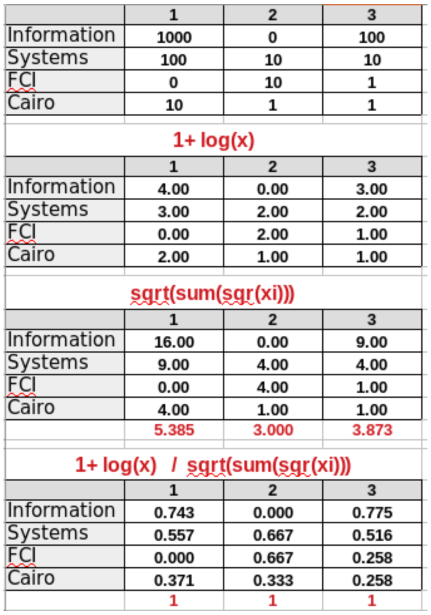
7. why do we need log-frequency weight

The term frequency tf $_{t,d}$ of term **t** in the document **d** is defined as the number of times that t occurs in d.
Raw term frequency is not what we want:
– A document with 10 occurrences of the term is more relevant than a document with 1 occurrence of the term But not 10 times more relevant
i.e. Relevance does not increase proportionally with term frequency
so we use the log-frequency weight is used to dampen the effect of the the increase in term frequency.

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

Cos(1, 2) = 0.50
Cos(1, 3) = 0.96
Cos(2, 3) = 0.60
doc 1 is more similar to doc 3 than any other combination.



	1	2	3
Information	1000	0	100
Systems	100	10	10
FCI	0	10	1
Cairo	10	1	1

1+ log(x)

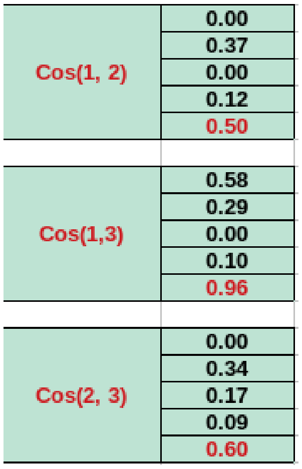
	1	2	3
Information	4.00	0.00	3.00
Systems	3.00	2.00	2.00
FCI	0.00	2.00	1.00
Cairo	2.00	1.00	1.00

sqrt(sum(sqr(xi)))

	1	2	3
Information	16.00	0.00	9.00
Systems	9.00	4.00	4.00
FCI	0.00	4.00	1.00
Cairo	4.00	1.00	1.00
	5.385	3.000	3.873

1+ log(x) / sqrt(sum(sqr(xi)))

	1	2	3
Information	0.743	0.000	0.775
Systems	0.557	0.667	0.516
FCI	0.000	0.667	0.258
Cairo	0.371	0.333	0.258
	1	1	1



Cos(1, 2)	0.00
	0.37
	0.00
	0.12
	0.50
Cos(1,3)	0.58
	0.29
	0.00
	0.10
	0.96
Cos(2, 3)	0.00
	0.34
	0.17
	0.09
	0.60

9. Compute the **$wt_{,d}$** for the terms/document given in the table in # 8

$$w_{t,d} = (1 + \log tf_{t,d}) \times \log_{10}(N / df_t)$$

Term	df	log(N/df)	1+log $_{t,1}$	1+log $_{t,2}$	1+log $_{t,3}$	$W_{t,1}$	$W_{t,2}$	$W_{t,3}$
Information	2	0.176	4	0	3	**0.704**	**0**	**0.528**
Systems	3	0.000	3	2	2	**0**	**0**	**0**
FCI	2	0.176	0	2	1	**0**	**0.352**	**0.176**
Cairo	3	0.000	2	1	1	**0**	**0**	**0**

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

If the size difference between 2 similar documents changes the Euclidean distance becomes quit large. For example givn a document d1 let d2 = d1 + d1 (appended) the eucladean distance bcome quit large despite that d2 is d1 twice so no difference is supposed to appear.