Faculty of Computers and Artificial Intelligence

Cairo, Egypt

Course Name: Information Retrieval (IS322)/Data Storage and Retrieval (IS313)

Course ID: 202102.FCI.IS322

Assignment 2 (IR Sheet)

Name	ID	Group Number
Noura Saad Mabrouk Hussien	20180317	IS , S3

1) Given

Doc 1: feature engineering used in software engineering

Doc 2: software engineering is fun

- First Step: Tokenizer

Word	Documents IDs
feature	1
engineering	1
used	1
in	1
software	1
engineering	1
software	2
engineering	2
is	2
fun	2

- Second Step: Sorting

Word	Documents IDs
engineering	1
engineering	1
engineering	2
Feature	1
Fun	2
in	1
is	2
used	1
software	1
software	2

- Third Step: Posting list of 2 Documents.

Word	Documents IDs
engineering	1,2
feature	1
fun	2
in	1
is	2
used	1
software	1,2

i. Draw the posting list for: software and engineering.

Solution:

software → | 1 | 2 |

engineering → | 1 | 2 |

software and engineering → | 1 | 2 |

ii. Draw the term-document incidence matrix.

Solution:

Terms	Doc 1	Doc 2
engineering	1	1
feature	1	0
fun	0	1
in	1	0
is	0	1
software	1	1
used	1	0

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.

Solution:

The query is: information systems technology /P study!

---

3) Discuss the effect of stemming in **precision and recall.**

Solution:

- Precision is: what fraction of the returned results are relevant to the information need?
- Recall is: What fraction of the relevant documents in the collection were returned by the system?
- Effect of Stemming: Stemming enables different variations of the word to be considered in retrieval, which improves the recall. Stemming increases recall and reduces the size of the indexing structure. However, it may hurt precision because many irrelevant documents may be considered relevant. For example: both “cop” and “cope” are reduced to the stem “cop”, However if one is looking for documents about police,a document that contains only “cope” is unlikely to be relevant.
- Understemming lowers recall and overstemming lowers precision. So,since no stemming at all means no over but max understemming errors, you have a low recall there and a high precision.

---

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

Solution:

- 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 for specific information on specific websites or pages.
- We use web crawler in information retrieval.

5) what is the main problem of boolean search?

Solution:

The main problem of Boolean search is results that can be too few (≈0) or to many (1000's) results (feast or femain).
for example:
if I search for this query "stanford user dlink 650" the result will be 20000 hits,
and when I search for this query "stanford user dlink 650 no card found" the result will be 0 hit.
So, it takes a lot of skill to come up with a query that produce a manageable number of hits.
Note that: AND gives too few results and OR gives too many results.

---

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

Solution:

J(q,d1) = |A ∩ B| / |A ∪ B|
= 2/9 = 0.2222

J(q,d2) = |A ∩ B| / |A ∪ B|
= 3/7 = 0.4285

Then d2 is wins.

7) why do we need log-frequency weight?

Solution:

Because Relevance does not increase proportionally with term frequency. when I search for term “Cairo” in document and if the term frequency for Cairo in doc 1 is 10 and in doc 2 is 20. So, we will say that doc2 is more relevant than doc1 for the word “Cairo”. However, if the term frequency of the same term “Cairo” in doc1 is 1000 and doc2 is 2000, at this point, there is no much difference in terms of relevancy anymore because they both contain a very high count for term “Cairo”.

So, we add log to dampen the importance of term that has a high frequency. When we get $\log_{10} 1000$ it will be 3, So it reduced from 1000 to 3.
We also add 1 to the log(tf) because when tf is equal to 1, the log(1) is zero. So, adding one, we distinguish between tf=0 and tf=1.

---

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

Term	doc1	doc 2	doc 3
Information	1000	0	100
Systems	100	10	10
FCI	0	10	1
Cairo	10	1	1

Solution:

- $\mathbf{w_{t,d} = 1 + \log(x)}$

Term	doc1	doc 2	doc 3
Information	4	0	3
Systems	3	2	2
FCI	0	2	1
Cairo	2	1	1

- **sqrt(sum(sqr(xi)))**

Term	doc1	doc 2	doc 3
Information	16	0	9
Systems	9	4	4
FCI	0	4	1
Cairo	4	1	1
	5.39	3	3.88

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

Term	doc1	doc 2	doc 3
Information	0.742	0	0.773
Systems	0.557	0.666	0.516
FCI	0	0.666	0.258
Cairo	0.371	0.333	0.258

Cos(doc 1 ,doc 2) = (0.742*0) + (0.557*0.666) + (0*0.666) + (0.371*0.333) = 0.495

Cos(doc 1 ,doc 3) = (0.742*0.773) + (0.557*0.516) + (0*0.258) + (0.371*0.258) = 0.957

Cos(doc 2 ,doc 3) = (0*0.773) + (0.666*0.516) + (0.666*0.258) + (0.333*0.258) = 0.601

The most two similarity documents is doc1 and doc3, then doc2 and doc3, then
The least two similarity documents is doc1 and doc2.

9) Compute the wt,d for the terms/document given in the table in # 8

Solution:

$$w_{t,d}$$

Term	doc1	doc 2	doc 3
Information	4	0	3
Systems	3	2	2
FCI	0	2	1
Cairo	2	1	1

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

Solution:

Because Euclidean distance is large for vectors of different lengths.
If we draw a 2D with 3 documents and 2 terms we find that The Euclidean distance between q↑ and q↑ is large even though the distribution of terms in the query q↑ and the distribution of terms in the document d2↑ are very similar. So that we can use angle in stead of cosine and distance similarity.