IR Tutorial Set 2

Question 1 (a)

Consider the following collection of five documents and a query:

- Doc 1: *we wish efficiency in the implementation for a particular application*
- Doc 2: *the classification methods are an application of Li's ideas*
- Doc 3: *the classification has not followed any implementation pattern*
- Doc 4: *we have to take care of the implementation time and implementation efficiency*
- Doc 5: *the efficiency is in terms of implementation methods and application methods*
- Query1: *application of classification methods*
- Query2: *efficiency in implementation of applications*

Now consider that the vocabulary is:

> {*efficiency, implementation, application, classification, methods, ideas, pattern, time*}.

Represent each document and the query using unit normal vectors following the "lnc.ltc" scheme. Rank the 5 documents based on their relevance with the query (most to least) measured via the cosine similarity metric. Consider the following collection of five documents and a query:

Doc 1: we wish efficiency in the implementation for a particular application

Doc 2: the classification methods are an application of Li's ideas

Doc 3: the classification has not followed any implementation pattern

Doc 4: we have to take care of the implementation time and implementation efficiency

Doc 5: the efficiency is in terms of implementation methods and application methods

Query1: application of classification methods

Query2: efficiency in implementation of applications

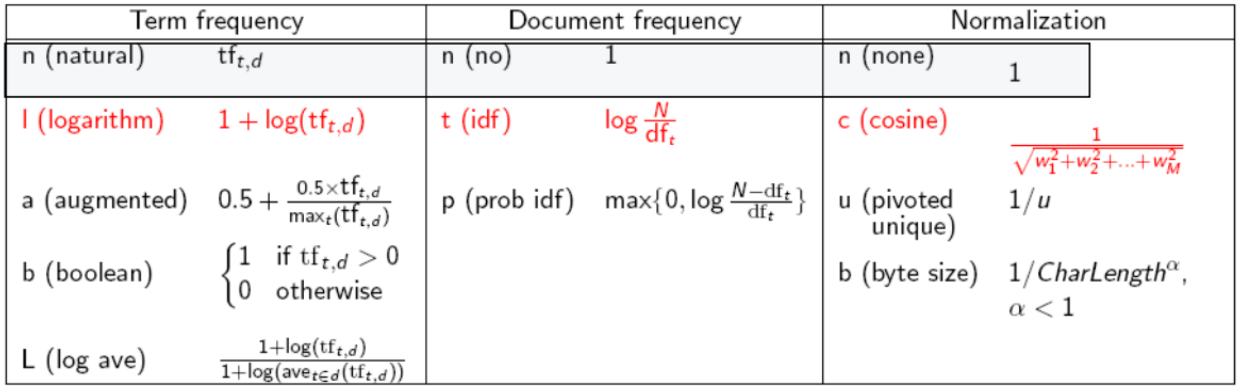
Now consider that the vocabulary is:

{efficiency, implementation, application, classification, methods, ideas, pattern, time}.

1. Represent each document and the query using unit normal vectors following the “lnc.ltc” scheme. Rank the 5 documents based on their relevance with the query (most to least) measured via the cosine similarity metric.

Solution:-

Reference for what do in the lnc/ltc scheme.



Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log\frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0,\log\frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

At first, creating the term frequency matrix

Consider the number of occurrences of a term in a document for the terms in the vocabulary and make the matrix where columns represent the documents (D1 → D5), query (Q1) and the rows represent the words in the vocabulary

	D1	D2	D3	D4	D5	Q1
efficiency	1	0	0	1	1	0
implementation	1	0	1	2	1	0
application	1	1	0	0	1	1
classification	0	1	1	0	0	1
methods	0	1	0	0	2	1
ideas	0	1	0	0	0	0
pattern	0	0	1	0	0	0
time	0	0	0	1	0	0

Second, creating the log frequency weight matrix using the term frequency matrix - Use the formula mentioned below to convert the term frequency matrix to the log frequency weight matrix (consider base 10)

The log frequency weight of term t in d is

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

	D1	D2	D3	D4	D5	Q1
efficiency	1	0	0	1	1	0
implementation	1	0	1	1.301	1	0
application	1	1	0	0	1	1
classification	0	1	1	0	0	1
methods	0	1	0	0	1.301	1
ideas	0	1	0	0	0	0
pattern	0	0	1	0	0	0
time	0	0	0	1	0	0

Calculating the weighted IDF (Considering base 10) -

$$\mathrm{idf}_t = \log_{10} (N/\mathrm{df}_t)$$

	IDF
efficiency	log(5/3) = 0.222
implementation	log(5/4) = 0.097
application	log(5/3) = 0.222
classification	log(5/2) = 0.398

methods	log(5/2) = 0.398
ideas	log(5/1) = 0.699
pattern	log(5/1) = 0.699
time	log(5/1) =0.699

TF-IDF - This shows the calculation the weight using the tfidf. We only need to calculate this for the query vector because documents are following the lnc rule.

$$TF\text{-}IDF\ (t, d) = TF\ (t, d) * IDF(t)$$

	D1	D2	D3	D4	D5	Q1
efficiency	1	0	0	1	1	0
implementation	1	0	1	1.301	1	0
application	1	1	0	0	1	0.222
classification	0	1	1	0	0	0.398
methods	0	1	0	0	1.301	0.398
ideas	0	1	0	0	0	0
pattern	0	0	1	0	0	0
time	0	0	0	1	0	0
	1.732	2	1.732	1.922	2.166	0.605

TF-IDF normalized - Now we normalise using the cosine part in the first table.

$$norm(v) = sqrt(v_1^2 + v_2^2 + \ldots + v_{|v|}^2)$$

	D1	D2	D3	D4	D5	Q1
efficiency	0.577	0	0	0.520	0.462	0
implementation	0.577	0	0.577	0.677	0.462	0
application	0.577	0.5	0	0	0.462	0.367
classification	0	0.5	0.577	0	0	0.658
methods	0	0.5	0	0	0.601	0.658
ideas	0	0.5	0	0	0	0
pattern	0	0	0.577	0	0	0
time	0	0	0	0.520	0	0

Dot product - Finally calculate the cosine similarity.1

	D1	D2	D3	D4	D5
efficiency	0	0	0	0	0
implementation	0	0	0	0	0
application	0.212	0.184	0	0	0.170
classification	0	0.329	0.380	0	0
methods	0	0.329	0	0	0.395
ideas	0	0	0	0	0
pattern	0	0	0	0	0
time	0	0	0	0	0
sum	0.212	0.842	0.380	0	0.565

Final ranks -- D2 > D5 > D3 > D1 > D4

Question 2 (Evaluation Metrics)

Consider a corpus of 10 documents, and two retrieval systems S1 and S2.

For each row of the table below, the documents are ranked from left to right based on relevance (most to least) with some query. For the ground truth, the relevance grade is given in brackets (4-very relevant, 0 - irrelevant).

For binary relevance, consider non-zero relevance grades as relevant, zero grade as irrelevant.

Find the following metrics for both systems: (i) Avg. Precision (ii) NDCG

GT	1(4)	3(4)	2(3)	4(2)	8(1)	9(1)	5(0)	6(0)	7(0)	10(0)
S1	1	2	3	4	5	6	7	8	9	10
S2	3	2	4	1	6	10	9	7	5	8

Solution to Question 2

GT	1(4)	3(4)	2(3)	4(2)	8(1)	9(1)	5(0)	6(0)	7(0)	10(0)
S1	1	2	3	4	5	6	7	8	9	10
S2	3	2	4	1	6	10	9	7	5	8

NDCG Calculation

For different lists (GT, S1, S2), keep the docs in the same order as in the lists. For each rank, compare with the gold standard relevance score.

$$DG@i = R^{GT}(i) / \log_2 i$$

Sum up to get the DCG of the system

GT	1(4)	3(4)	2(3)	4(2)	8(1)	9(1)	5(0)	6(0)	7(0)	10(0)	Sum
R^{GT}	4	4	3	2	1	1	0	0	0	0	

DCG^{GT}	4	4	1.893	1	0.431	0.387	0	0	0	0	11.711
S1	1	2	3	4	5	6	7	8	9	10	
R^{GT}	4	3	4	2	0	0	0	1	1	0	
DCG^{S1}	4	3	2.524	1	0	0	0	0.333	0.315	0	11.172
S2	3	2	4	1	6	10	9	7	5	8	
R^{GT}	4	3	2	4	0	0	1	0	0	1	
DCG^{S2}	4	3	1.262	2	0	0	0.356	0	0	0.301	10.919

Calculate NDCG by dividing DCG of system by DCG of GT

$NDCG^{S1}$ = 0.954

$NDCG^{S2}$ = 0.932

Average Precision Calculation

Calculate precision at each rank as

P@i = No. of relevant docs upto rank *i* / *i*

Calculate AP by summing up *P@i* values (only for positions of relevant docs)

GT	1(4)	3(4)	2(3)	4(2)	8(1)	9(1)	5(0)	6(0)	7(0)	10(0)
S1	1	2	3	4	5	6	7	8	9	10
Rel	1	1	1	1	0	0	0	1	1	0
P@	1	1	1	1	-	-	-	0.625	0.667	-
S2	3	2	4	1	6	10	9	7	5	8
Rel	1	1	1	1	0	0	1	0	0	1
P@	1	1	1	1	-	-	0.714	-	-	0.6

AP^{S1} = 0.882

AP^{S2} = 0.886