CSCI 6370 IR and Web Search
ASSIGNMENT 1
Due is 06/08/2020 23:59

Ulvi Bajarani

Student ID 20539914

E-mail: ulvi.bajarani01@utrgv.edu

Questions and Answers:

Problem 1. Table 1 lists the index terms and their appearances in a set of documents.:

Doc/Term	retrieval	database	computer	text	information
D1	4	10	2	0	1
D2	3	0	7	4	5
D3	7	2	4	6	8

Table 1: Term Frequencies

We also know that the total number of documents in the set is 1000. Table 2 shows the document frequencies of these terms.

Term	retrieval	database	computer	text	information
Frequency	100	70	220	80	110

Table 2: Document Frequencies

Compute tf-idf for each of the (doc, term) pairs listed in Table 1. List your results in sorted order from the largest value of tf-idf to the smallest value.

Term					
tf/idf					

Table 3: Term Frequency-Inverted Document Frequencies

Answer 1.
Using the formula $IDF_i = \log_2 \frac{N}{n_i}$ for each term k_i, we receive such approximate values of IDF_i:

Term	retrieval	database	computer	text	information
IDF	3.32192809488736	3.83650126771712	2.18442457113743	3.64385618977473	3.18442457113743

After that, the formula of *TF-IDF* , which is

$$\text{TF-IDF} = \begin{cases} (1 + \log_2 f_{i,j}) * \log_2 \frac{N}{n_i} & \text{if } f_{i,j} > 0; \\ 0 & \text{if } f_{i,j} \leq 0. \end{cases}$$

we can find all values of *TF-IDF*. The sorted list is as below (to fit it into the page, the table is divided to the several tables):

Doc,Term	D1,database	D3,text	D3,information	D3,retrieval	D2,text
TF-IDF	16.5810826150176	13.0630877983631	12.7376982845497	12.6477592827908	10.9315685693242
Doc,Term	D2,information	D1,retrieval	D2,retrieval	D2,computer	D3,database
TF-IDF	10.5784294489111	9.96578428466209	8.5870595553759	8.31687964278366	7.67300253543424
Doc,Term	D3,computer	D1,computer	D1,information	D2,database	D1,text
TF-IDF	6.55327371341228	4.36884914227486	3.18442457113743	0	0

Problem 2. Assume we use the tf-idf as the weight in the vector space model, write down the document-term matrix using the results generated from the above problem. Remember a document-term matrix has terms as its columns and docu-

ments as its rows.:

Answer 2.

Instead of sorting the values, the results of calculation might be written here directly as the term-document matrix:

Document	retrieval	database	computer	text	information
D1	9.96578428466209	16.5810826150176	4.36884914227486	0	3.18442457113743
D2	8.5870595553759	0	8.31687964278366	10.9315685693242	10.5784294489111
D3	12.6477592827908	7.67300253543424	6.55327371341228	13.0630877983631	12.7376982845497

Problem 3. Now assume we have a query Q = “computer information”, compute the similarity based on the inner product similarity and the cosine similarity for each of the documents listed in Table 1. Which document is the most relevant in each of the similarity measures? Which one is the least relevant?:

Answer 3.

Let’s assume that Q = “computer information” is the document and try to calculate its term frequency:

Document	retrieval	database	computer	text	information
Computer Information	0	0	1	0	1

Then, let's calculate the IDF of the document:

Term	retrieval	database	computer	text	information
IDF	0	0	2.18442457113743	0	3.18442457113743

After that, the time to calculate TF-IDF step:

	Term				
q="Computer Information"	retrieval	database	computer	text	information
TF-IDF	0	0	2.18442457113743	0	3.18442457113743

In the final step, we can calculate the cosine similarity based on the formula below:

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q_j}}{|\vec{d_j}| \times |\vec{q_j}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

For the easy calculation, Inner product similarity are also provided:

Document	Cosine similarity to q=”Computer Information”
D1	0.253765104212892
D2	0.694053244578525
D3	0.582746893701989

Document	Inner product similarity to q=”Computer Information”
D1	19.6839812632417
D2	51.8538069080455
D3	54.877371518022

As a result, we can define that D2 is the highest ranked (most relevant), while D1 is the lowest ranked (least relevant) document to the query ”Computer Information”. In the case of Inner product similarity, D3 is the highest one, then D2, and finally D1.