جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

INFORMATION RETRIEVAL WITH APPLICATIONS OF NLP
ENCS4130

Assignment #1

Prepared by: Mazen Batrawi - 1190102

Instructor: Dr. Adnan Yahya

Section: 1

BIRZEIT

APR – 2023

My number: 1190102 → even

Dataset:

SET2= {Doc 21 breakthrough drug for schizophrenia
Doc 22 new schizophrenia drug
Doc 23 new approach for treatment of schizophrenia
Doc 24 new hopes for schizophrenia patients}

Part1

1. How many tokens and how many terms you have in your collection?
   We have 18 tokens (18 words), from these 18 we have 10 terms (distinct words).

2. Draw the term-document incidence matrix (1/0 matrix) for your document collection.

	Doc21	Doc22	Doc23	Doc24
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
schizophrenia	1	1	1	1
new	0	1	1	1
approach	0	0	1	0
treatment	0	0	1	0
of	0	0	1	0
hopes	0	0	0	1
patients	0	0	0	1

3. Using the incidence matrix, what are the returned results for the queries for your set.
   For SET2: Q1= chizophrenia AND drug Q2= for AND NOT (drug OR approach).

   Q1: 1111 & 1100 = 1100 = Doc21 + Doc22
   Q2: 1011 & ~(1100 | 0010) = 1011 & ~(1110) = 1011 & 0001 = 0001 = Doc24

4. If we have the operator W1 \B2 W2 to mean W1 must be at most 2 words before W2: can we answer such query from the Incidence Matrix? Why? Why Not?

   It is not possible to determine the presence or absence of a term in a document solely based on the incidence matrix. Additionally, the incidence matrix does not provide information about the specific location of a term within a document.

5. Draw the inverted index that would be built for your document.

breakthrough	1	→	21			
drug	2	→	21	21		
for	3	→	21	23	24	
schizophrenia	4	→	21	22	23	24
new	3	→	22	23	24	
approach	1	→	23			
treatment	1	→	23			
of	1	→	23			
hopes	1	→	24			
patients	1	→	24			

Part2

6. Compute term frequency for each element/document and document frequency for each term then.

	DF	IDF	TF21	TF22	TF23	TF24
breakthrough	1	2	0.25	0	0	0
drug	2	1	0.25	0.33	0	0
for	3	0.42	0.25	0	0.167	0.2
schizophrenia	4	0	0.25	0.33	0.167	0.2
new	3	0.42	0	0.33	0.167	0.2
approach	1	2	0	0	0.167	0
treatment	1	2	0	0	0.167	0
of	1	2	0	0	0.167	0
hopes	1	2	0	0	0	0.2
patients	1	2	0	0	0	0.2

DF = Number of documents containing the term.
IDF = log2(N / DF), N = Number of documents.
Term frequency is the number of occurrences for each term in a document divided by the length of that document.

7. Replace the 1/0 of the incidence matrix by the corresponding tf-idf for that term/document.
   $w_{ij} = tf_{ij} * idfi$

	Doc21	Doc22	Doc23	Doc24
breakthrough	0.5	0	0	0
drug	0.25	0.33	0	0
for	0.105	0	0.07014	0.084
schizophrenia	0	0	0	0
new	0	0.1386	0.07014	0.084
approach	0	0	0.334	0
treatment	0	0	0.334	0
of	0	0	0.334	0
hopes	0	0	0	0.4
patients	0	0	0	0.4

8. For set2 = {chizophrenia drug approach}

$$\text{CosSim}(\boldsymbol{d_j}, \boldsymbol{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\,\right|} = \frac{\sum_{i=1} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}}$$

The query has the terms chizophrenia, drug, and approach. The numbers in the following vectors represent the count of each term in the query and the documents.

Q = [0, 0.33, 0, 0, 0, 0.66, 0, 0, 0, 0]
Doc21 = [0.5, 0.25, 0.105, 0, 0, 0, 0, 0]
Doc22 = [0, 0.33, 0, 0, 0.1386, 0, 0, 0, 0, 0]
Doc23 = [0, 0, 0.07014, 0, 0.07014, 0.33, 0.33, 0.33, 0, 0]
Doc24 = [0, 0, 0.084, 0, 0.084, 0, 0, 0, 0.4, 0.4]

CosSim(Doc21, Q) = (0.25 * 0.33) / $\sqrt{(0.33^2 + 0.66^2) * (0.25^2 + 0.5^2 + 0.105^2)}$ = 0.197.
CosSim(Doc22, Q) = (0.33 * 0.33) / $\sqrt{(0.33^2 + 0.66^2) * (0.33^2 + 0.1386^2)}$ = 0.412.
CosSim(Doc23, Q) = (0.33 * 0.66) / $\sqrt{(0.33^2 + 0.66^2) * (2 * 0.07014^2 + 3 * 0.33^2}$= 0.509.
CosSim(Doc24, Q) = 0 / ... = 0

From the results, we can see that the most relative document is Doc23 (The highest cosine similarity).