

**Faculty of Engineering & Technology – Electrical & Computer Engineering Department**

**Second Semester 2022 – 2023**

**Natural language processing (NLP)-ENCS5342**

**Assignment #1**

**Name: Amal Ziad**

**ID: 1192141**

**Section: 1**

**Instructor: Dr. Adnan Yahya**

**Date: 19th Apr 2023**

SET1= {Doc 11 new home sales top forecasts

Doc 12 home sales rise in july

Doc 13 increase in home sales in july

Doc 14 july new home sales rise}

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?

Tokens: 39 (each word is a token)  
Terms: 17

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | D11 | D12 | D13 | D14 | D21 | D22 | D23 | D24 |
| new | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Home | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| sales | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| top | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| forecasts | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rise | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| July | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| for | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| in | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| increase | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| breakthrough | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| of | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| drug | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| schizophrenia | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| approach | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| treatment | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| hopes | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| patients | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

1. Using the incidence matrix, what are the returned results for the queries for your set:

For SET1: Q1= july AND home 🡪D12, D13, D14

Q2= for AND NOT(increase OR top) 🡪 returns nothing

For SET2: Q1= shizophrenia AND drug 🡪D21, D22

Q2= for AND NOT(drug OR approach) 🡪D24

1. 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?

No we can’t since incidence matrix doesn’t tell us the order of the words.

1. Draw the inverted index that would be built for your document collection as we did in class in figure 1.2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| new | 🡪 | 11 | 14 | 22 | 23 | 24 |  |  |
| Home | 🡪 | 11 | 12 | 13 | 14 |  |  |  |
| sales | 🡪 | 11 | 12 | 13 | 14 |  |  |  |
| top | 🡪 | 11 |  |  |  |  |  |  |
| forecasts | 🡪 | 11 |  |  |  |  |  |  |
| Rise | 🡪 | 12 | 14 |  |  |  |  |  |
| July | 🡪 | 12 | 13 | 14 |  |  |  |  |
| for | 🡪 | 21 | 23 | 24 |  |  |  |  |
| in | 🡪 | 12 | 13 |  |  |  |  |  |
| increase | 🡪 | 13 |  |  |  |  |  |  |
| breakthrough | 🡪 | 21 |  |  |  |  |  |  |
| of | 🡪 | 23 |  |  |  |  |  |  |
| drug | 🡪 | 21 | 22 |  |  |  |  |  |
| schizophrenia | 🡪 | 21 | 22 | 23 | 24 |  |  |  |
| approach | 🡪 | 23 |  |  |  |  |  |  |
| treatment | 🡪 | 23 |  |  |  |  |  |  |
| hopes | 🡪 | 24 |  |  |  |  |  |  |
| patients | 🡪 | 24 |  |  |  |  |  |  |

1. Compute term frequency for each element/document and document frequency for each term then.(tf=)

D11 , max{fij}=1

D12, max{fij}=1

D13 , max{fij}=2, “in”

D14 , max{fij}=1

D21 , max{fij}=1

D22 , max{fij}=1

D23 , max{fij}=1

D24 , max{fij}=1

|  |  |  |
| --- | --- | --- |
| i\j | Term freq= fij/max{fij} | Document freq. |
| new | D11=1, D14=1, D22=1, D23=1, D24=1 | 5 |
| Home | D11=1, D12=1, D13=1/2, D14=1 | 4 |
| sales | D11=1, D12=1, D13=1/2, D14=1 | 4 |
| top | D11=1, | 1 |
| forecasts | D11=1 | 1 |
| Rise | D12=1, D14=1, | 2 |
| July | D12=1, D13=1/2, D14=1 | 3 |
| for | D21=1, D23=1, D24=1 | 3 |
| in | D12=1, D13=2/2, | 2 |
| increase | D13=1/2 | 1 |
| breakthrough | D21=1 | 1 |
| of | D23=1 | 1 |
| drug | D21=1, D22=1 | 1 |
| schizophrenia | D21=1, D22=1, D23=1, D24=1 | 4 |
| approach | D23=1 | 1 |
| treatment | D23=1 | 1 |
| hopes | D24=1 | 1 |
| patients | D24=1 | 1 |

1. Replace the 1/0 of the incidence matrix by the corresponding tf-idf for that term/document. =*tfij* log2 (*N/ dfi*)

**N=8**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | D11 | D12 | D13 | D14 | D21 | D22 | D23 | D24 |
| new | 1\*log2 (8/5)= 1.0709 |  |  | 1\*log2 (8/5)= 1.0709 |  | 1\*log2 (8/5)= 1.0709 | 1\*log2 (8/5)= 1.0709 | 1\*log2 (8/5)= 1.0709 |
| Home | 1\*log2 (8/4)= 1 | 1\*log2 (8/4)=1 | 0.5\*log2 (8/4)= 0.5 |  |  |  |  |  |
| sales | 1\*log2 (8/4)= 1 | 1\*log2 (8/4)=1 | 0.5\*log2 (8/4)= 0.5 |  |  |  |  |  |
| top | 1\*log2 (8/1)= 3 |  |  |  |  |  |  |  |
| forecasts | 1\*log2 (8/1)= 3 |  |  |  |  |  |  |  |
| Rise |  | 1\*log2 (8/2)= 2 |  | 1\*log2 (8/2)= 2 |  |  |  |  |
| July |  | 1\*log2 (8/3)=1.42 | 0.5\*log2 (8/3)=0.708 | 1\*log2 (8/3)=1.42 |  |  |  |  |
| for |  |  |  |  | 1\*log2 (8/3)=1.42 |  | 1\*log2 (8/3)=1.42 | 1\*log2 (8/3)=1.42 |
| in |  | 1\*log2 (8/2)=2 | 1\*log2 (8/2)=2 |  |  |  |  |  |
| increase |  |  | 0.5\*log2 (8/1)=1.5 |  |  |  |  |  |
| breakthrough |  |  |  |  | 1\*log2 (8/1)=3 |  |  |  |
| of |  |  |  |  |  |  | 1\*log2 (8/1)=3 |  |
| drug |  |  |  |  | 1\*log2 (8/2)=2 | 1\*log2 (8/2)=2 |  |  |
| schizophrenia |  |  |  |  | 1\*log2 (8/4)=1 | 1\*log2 (8/4)=1 | 1\*log2 (8/4)=1 | 1\*log2 (8/4)=1 |
| approach |  |  |  |  |  |  | 1\*log2 (8/1)=3 |  |
| treatment |  |  |  |  |  |  | 1\*log2 (8/1)=3 |  |
| hopes |  |  |  |  |  |  |  | 1\*log2 (8/1)=3 |
| patients |  |  |  |  |  |  |  | 1\*log2 (8/1)=3 |

1. Given the queries

For **SET1**: {july home increase} =**Q1**

For **SET2**: {chizophrenia drug approach} =**Q2**

Find the **most relevant** document to this query in your set using **Cosine Similarity**.

Q1={july home increase}=0\*(new)+1\*(home)+0\*(sales)+0\*(top)+0\*(forcasts)+0\*(rise)+0\*(in)+1\*(july)+1\*(increase)

D11=1\*(new)+1\*(home)+1\*(sales)+1\*(top)+1\*(forcasts)+0\*(rise)+0\*(in)+0\*(july)+0\*(increase)

CosSim(D11,Q1)==0.169

D12=0\*(new)+1\*(home)+1\*(sales)+0\*(top)+0\*(forcasts)+1\*(rise)+1\*(in)+1\*(july)+0\*(increase)

CosSim(D12,Q1)==0.338

D13=0\*(new)+1\*(home)+1\*(sales)+0\*(top)+0\*(forcasts)+0\*(rise)+2\*(in)+1\*(july)+1\*(increase)

CosSim(D13,Q1)==0.6124

D14=0\*(new)+1\*(home)+1\*(sales)+0\*(top)+0\*(forcasts)+1\*(rise)+0\*(in)+1\*(july)+0\*(increase)

CosSim(D14,Q1)==0.338

So that, D12 is the most relevant document for the query Q1

Q2={schizophrenia drug approach}=0\*(breakthrough)+1\*(drug)+0\*(for)+1\*(schizopheria)+0\*(new)+1\*(approach)+0\*(treatment)+0\*(of)+0\*(hopes)+0\*(patients)

D21=1\*(breakthrough)+1\*(drug)+1\*(for)+1\*(schizopheria)+0\*(new)+0\*(approach)+0\*(treatment)+0\*(of)+0\*(hopes)+0\*(patients)

CosSim(D21,Q2)==0.577

D22=0\*(breakthrough)+1\*(drug)+0\*(for)+1\*(schizopheria)+1\*(new)+0\*(approach)+0\*(treatment)+0\*(of)+0\*(hopes)+0\*(patients)

CosSim(D22,Q2)==0.667

D23=0\*(breakthrough)+0\*(drug)+1\*(for)+1\*(schizopheria)+1\*(new)+1\*(approach)+1\*(treatment)+0\*(of)+0\*(hopes)+0\*(patients)

CosSim(D23,Q2)==0.4714

D24=0\*(breakthrough)+0\*(drug)+1\*(for)+1\*(schizopheria)+1\*(new)+0\*(approach)+0\*(treatment)+0\*(of)+1\*(hopes)+1\*(patients)

CosSim(D24,Q2)==0.258

So that, D22 is the most relevant document for the query Q2