BIRZEIT UNIVERSITY
Electrical and Computer Engineering Department

ENCS5342: “Information Retrieval, Web Search and NLP” Assignment #1:

Instructor: Dr. Adnan H. Yahya, **Index Construction** Due: April 20, 2023.

Problem: Given **SET1** for students with ODD numbers and **SET2** for students with EVEN numbers

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? 18 , 10

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

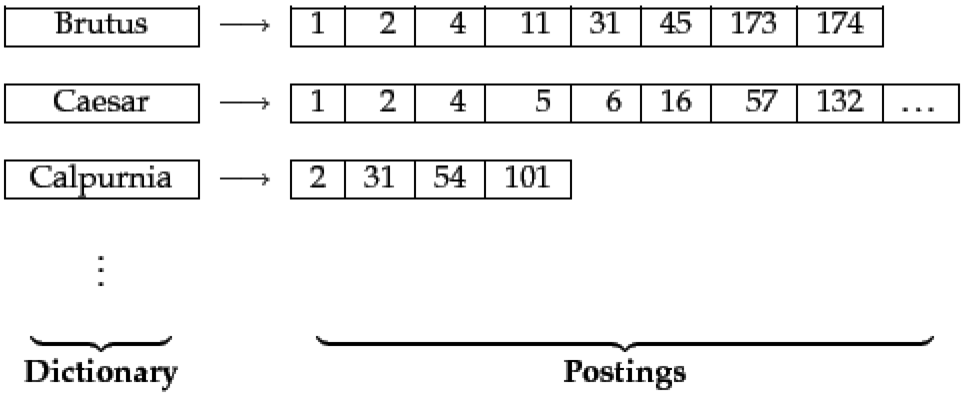
3- Using the incidence matrix, what are the returned results for the queries for your set:

For **SET1**: **Q1**= july AND home Q2= for AND NOT(increase OR top)

For **SET2**: **Q1=** chizophrenia AND drug **Q2=** for AND NOT(drug OR approach)

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?

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



► **Figure 1.2** The two parts of an inverted index. The dictionary is commonly kept in memory, with pointers to each postings list, which is stored on disk.

End of part 1, Please solve by 12/4/2023 (no submission needed). Just Practice!

Part2: Please Submit all parts by 20/4/2023

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

7- Replace the 1/0 of the incidence matrix by the corresponding tf-idf for that term/document.

8- Given the queries

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

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

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

Good Luck