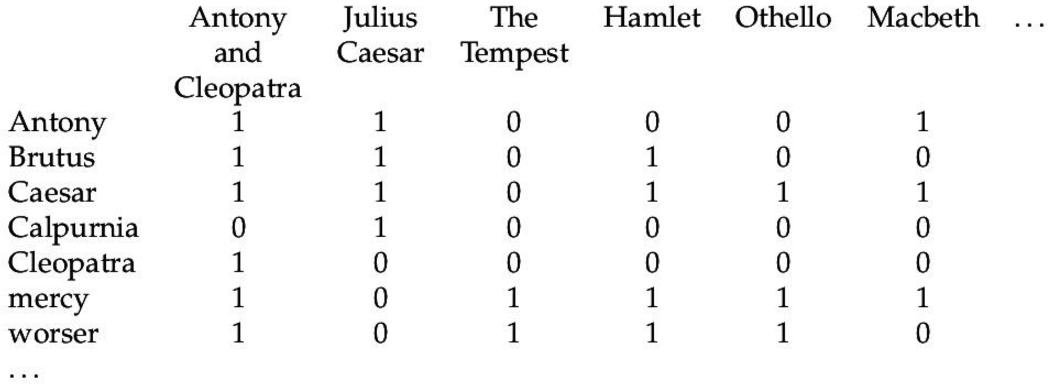
# Problem Solving-1

1. Initialize the following term-incidence matrix. Process the following query: "**Brutus AND Caesar AND NOT Calpurnia"**



| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth | ... |
|---|---|---|---|---|---|---|---|
| Antony | 1 | 1 | 0 | 0 | 0 | 1 | |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 | |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 | |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 | |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 | |
| mercy | 1 | 0 | 1 | 1 | 1 | 1 | |
| worser | 1 | 0 | 1 | 1 | 1 | 0 | |
| ... | | | | | | | |

2. Given four documents:

   Doc1: Breakthrough drug for Schizophernia

   Doc2: New Schizophernia drug

   Doc3: New approach for treatment of Schizophernia

   Doc4: New hopes for Schizophernia patients

   Generate the **term-document incidence matrix**.

3. Construct an Inverted Index for the above specified input document collections.
4. Construct a Sorting based Inverted Index for the above specified input document collections.
5. Process the query **'BRUTUS AND CALPURNIA'** using the **intersect algorithm**.

$$\text{BRUTUS} = 1 \rightarrow 2 \rightarrow 4 \rightarrow 11 \rightarrow 31 \rightarrow 45 \rightarrow 173 \rightarrow 174$$
$$\text{CALPURNIA} = 2 \rightarrow 31 \rightarrow 54 \rightarrow 101$$

6. For the queries below, can we still run through the intersection in time O (x+y), where x and y are the lengths of the postings lists for Brutus and Caesar? If not, what can we achieve?
   a. Brutus AND NOT Caesar
   b. Brutus OR NOT Caesar

7. For a conjunctive query, is processing postings lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it isn't.

8. Consider the collection made of the 3 following documents (one document per line):
   - out of the clear blue sky
   - the blue car next to the entrance
   - sky news: information retrieval is nice

i. propose a stop list and give the index of this collection for this stop-list,
ii. give the positional index of this collection.

9. Are the following statements true or false?
   – Stemming increases the size of the vocabulary.
   – Stemming should be invoked at indexing time but not while processing a query.

10. Assume a biword index. Give an example of a document which will be returned for a query of New York University but is actually a false positive which should not be returned.

11. Consider the following document:
   "**The universe contains many different universities**"
   - How many entries a character trigram index would contain?
   - What is the boolean query on this index for the initial query uni*?
   - How would you process a query such as uni*e* ? Give the detail of the processing.

12. Draw a trie which encodes the following terms: Hawai'i, hare, hiss, hissing, hissed, he, hunger, honey, hello, hallo, Hungary.

13. Compute the Levenshtein matrix for the distance between the strings "apfel" (input) and "poems" (output).

14. Caculate the edit distance between cat – catcat.

15. If $|s_i|$ denotes the length of string $s_i$ , show that the edit distance between $s_1$ and $s_2$ is never more than $\max(|s_1|, |s_2|)$