Group Members:

Pratyush Kumar Singh(20MIA1131)
Naman Grover (20MIA1079)
Anchita Singh (20MIA1125)
Md. Mustaqeem Khan (20MIA1085)

Digital Assignment-II

In the module 4 you got an introduction to boolean queries and how they can be processed using a binary matrix with terms as rows and documents as columns as an index. And the matrix contained 1s and 0s depending on if the term occurred in the document or not.

a. **What are the problems with using such a binary matrix as an index? What approaches can you think of to overcome these problems? Explain your solution with an example.**

Using a binary matrix as an index for search capability in IR systems can lead to several problems. One issue is the high dimensionality of the matrix, which can make it difficult to store and process efficiently. Another problem is that the matrix may be sparse, meaning that many entries are empty, which can also impact efficiency.

To overcome these problems, one approach is to use a technique called "term weighting," where each term in the matrix is assigned a weight that reflects its importance in the document. This can help to reduce the dimensionality of the matrix and improve efficiency.

For example, consider a binary matrix that represents the presence or absence of words in a set of documents. To apply term weighting, we could use a technique like TF-IDF (term frequency-inverse document frequency), which assigns a weight to each term based on how often it appears in the document and how common it is across all documents in the collection. This can help to prioritize terms that are more relevant to the query and reduce the impact of noise or irrelevant terms.

Another approach is to use more advanced indexing techniques, such as inverted indexing or clustering, which can help to group similar documents together and reduce the size of the index. These methods can also help to improve search accuracy by identifying patterns and relationships between documents.

Overall, while binary matrices can be useful for indexing and searching in IR systems, it's important to consider the limitations and explore alternative approaches to optimize performance and accuracy.

b. **One of the variants of Boolean queries is to specify proximity of term occurrences. For example “Hope AND Prosperity occurring within 10 words” from each other. Propose a solution to create posting lists for processing such positional queries.**

To create posting lists for processing positional queries, we can make use of inverted indices. In addition to the term frequency and document frequency, we can store the positions of each occurrence of a term within a document.

To implement proximity-based queries, we can use a sliding window approach. For example, if we want to find occurrences of "Hope" and "Prosperity" within 10 words from each other, we can first retrieve the posting lists for both terms. Then, we can iterate through the posting lists and for each document that contains both terms, we can check if the difference between the positions of the terms is less than or equal to 10.

If the positional index becomes too large to fit in memory, we can use compression techniques like delta encoding or variable-byte encoding to reduce the storage requirements.

We can also use techniques like block-based compression to partition the index into smaller segments for efficient retrieval.

Explain your solution with a below example. Programming languages: The assignment can be implemented in any programming language of your choice.
Text Collections:
Every year Maha Shivratri is celebrated with a lot of pomp and grandeur. It is considered to be a very special time of the year since millions of people celebrate this momentous occasion with a lot of fervour and glee.
Lord Shiva devotees celebrate this occasion with a lot of grandness. It is accompanied by folk dances, songs,prayers, chants, mantras etc. This year, the beautiful occasion of Maha Shivratri will be celebrated on February 18.
People keep a fast on this Maha shivratri, stay awake at night and pray to the lord for blessings, happiness, hope andprosperity. This festival holds a lot of significance and is considered to be one of the most important festivals inIndia.
The festival of Maha Shivratri will be celebrated on February 18 and is a very auspicious festival. This Hindu festival celebrates the power of Lord Shiva. Lord Shiva protects his devotees from negative and evil spirits. He is the epitome of powerful and auspicious energy.
Find index term, Boolean weights and build Boolean retrieval model.

Index terms:
- Maha Shivratri
- celebrated
- pomp
- grandeur
- special
- millions
- fervour
- glee
- Lord Shiva
- devotees
- grandness
- folk dances
- songs
- prayers
- chants
- mantras
- February 18
- fast
- stay awake at night
- pray
- blessings
- happiness
- hope
- prosperity
- significance
- important
- India
- auspicious
- Hindu festival
- power
- negative
- evil spirits
- epitome
- energy

Boolean weights:

- Maha Shivratri: 1
- celebrated: 1
- pomp: 1
- grandeur: 1
- special: 1
- millions: 1
- fervour: 1
- glee: 1
- Lord Shiva: 1
- devotees: 1
- grandness: 1
- folk dances: 1
- songs: 1
- prayers: 1
- chants: 1
- mantras: 1
- February 18: 1
- fast: 1
- stay awake at night: 1
- pray: 1
- blessings: 2
- happiness: 1
- hope: 1
- prosperity: 1
- significance: 1
- important: 1
- India: 1
- auspicious: 2
- Hindu festival: 1
- power: 1
- negative: 1
- evil spirits: 1
- epitome: 1
- energy: 1

Boolean retrieval model:
- Maha Shivratri: {1, 2, 4}
- stay awake at night: {4}
- blessings: {4}
- Maha Shivratri AND (stay awake at night OR blessings): {4}

Explanation:
- The query consists of three terms connected by Boolean operators: "Maha Shivratri", "stay awake at night", and "blessings".
- We first retrieve the posting lists for each term: {1, 2, 4} for "Maha Shivratri", {4} for "stay awake at night", and {4} for "blessings".
- We then apply the Boolean operators to get the final result: {4}, which represents the documents that contain both "Maha Shivratri" and either "stay awake at night" or "blessings".
- Therefore, the document that satisfies the query is the fourth document in the collection, which describes the significance of Maha Shivratri and how people keep a fast, stay awake at night, and pray for blessings, happiness, hope, and prosperity.