# Case Study Assignment-III for IR:

# Implement to retrieve documents by Probabilistic, Non-Overlapped List and Proximal Nodes Models

# Assignment Steps:

**A step-by-step process for retrieving documents using the Probabilistic, Non-Overlapped List and the Proximal Nodes Models:**

## Probabilistic Retrieval Model:

- Let's specifically focus on the Binary Independence Model (BIM) for probabilistic information retrieval.

1. **Binary Independence Model (BIM):**

- **Select the BIM Model:**
  - Choose the Binary Independence Model (BIM) as the probabilistic retrieval model for your information retrieval task.
- **Preprocessing:**
  - Preprocess the document collection and the user's query. This may involve tasks like tokenization, stemming, and removing stop words to prepare the text for analysis.

2. **Term Weighting:**
   - Assign weights to terms in the document collection and query. In BIM, each term is typically assigned a binary weight (1 if the term is present, 0 if it's absent).
3. **Query Representation:**
   - Represent the user's query as a binary vector, where each dimension corresponds to a unique term in the collection. The vector elements are set to 1 if the query term is present and 0 if it's absent.
4. **Document Scoring:**
   - Calculate a score for each document based on the query's binary vector and the binary term vectors of documents. BIM often uses the Jaccard coefficient or the Dice coefficient to measure the similarity between the query and documents.
5. **Ranking:**
   - Rank the documents based on their scores in descending order. Documents with higher similarity scores are considered more relevant to the query.

.

6. **Retrieve Top-K Documents:**
   - Select the top-K documents from the ranked list to present to the user. The value of K can be determined based on system settings or user preferences.
7. **Presentation of Results:**
   - Present the selected top-K documents to the user as search results, along with any additional information such as snippets or document titles.
8. **User Interaction (Optional):**
   - Allow users to interact with the results, such as providing feedback on document relevance. User feedback can be used to refine future searches.
9. **Evaluation (Optional):**
   - Evaluate the performance of the BIM model using relevant metrics like Precision, Recall, or F1-score to assess its effectiveness in retrieving relevant documents.

# Non-Overlapped List Model:

**1. Identify Terms of Interest:**

- Determine the specific terms or keywords you are interested in. Let's say, for example, you're interested in finding documents related to "machine learning" and "data visualization."

**2. Retrieve Documents per Term:**

- Retrieve the list of documents associated with each term separately from your document database or corpus. For "machine learning," retrieve the list of documents (e.g., $D_{machine\_learning}$).

  Similarly, retrieve the list for "data visualization" (e.g., $D_{data\_visualization}$).

**3. Combine Lists for Non-Overlapping Results:**

- Combine the lists of documents for each term using set union ∪ to get non-overlapping results:

  $D_{NonOverlap} = D_{machine\_learning} \cup D_{data\_visualization}$

  $D_{NonOverlap}$ now contains documents that contain either "machine learning" or "data visualization."

**4. Present Results:**

- Present the documents in $D_{NonOverlap}$ to the user as the non-overlapping set of documents related to the specified terms.

# Proximal Nodes Model:

**1. Define Proximal Nodes:**

- Identify proximal nodes or entities that are likely to be related to the desired information. For instance, if you're interested in "space exploration," proximal nodes could be "NASA," "astronauts," and "space missions."

**2. Explore Network Relationships:**

- Navigate the network or graph of interconnected nodes (representing documents, entities, or data points) and identify documents connected to the proximal nodes

**3. Retrieve Connected Documents:**

- Retrieve documents that are directly connected to the identified proximal nodes in the network. These connected documents are considered relevant in the Proximal Nodes Model.

**4. Present Results:**

- Present the retrieved documents connected to the proximal nodes as relevant to the user's query based on the network relationships.

Any
Questions
?