**NEWS DOCUMENT RETRIEVAL FOR TECH NEWS**

**(TECHNO-TALES)**

The goal of this project is to create a search engine tailored for retrieving technology news articles. The dataset comprises around 1,09,000 documents. The final result is an operational search engine capable of efficiently retrieving relevant technology news articles based on user queries.

**TASKS:**

1. **Data Collection:**
2. **Links Collection from tech-crunch:**

All the articles released on a particular date are arranged under the URL which looks link <https://www.techcrunch.com/dd-mm-yyyy/>.

We sent request to this link by iterating the date from June 6th 2005 to November 25th 2023.

1. **Fetching Data from links:**

After collecting all the links we iteratively sent requests to every link and parsed the response using beautiful-Soup library.

1. **Data Pre-Processing:**
2. Text-Cleaning
3. Tokenization
4. Lemmatization
5. Stop Word Removal
6. **Indexing**
7. **Ranking**

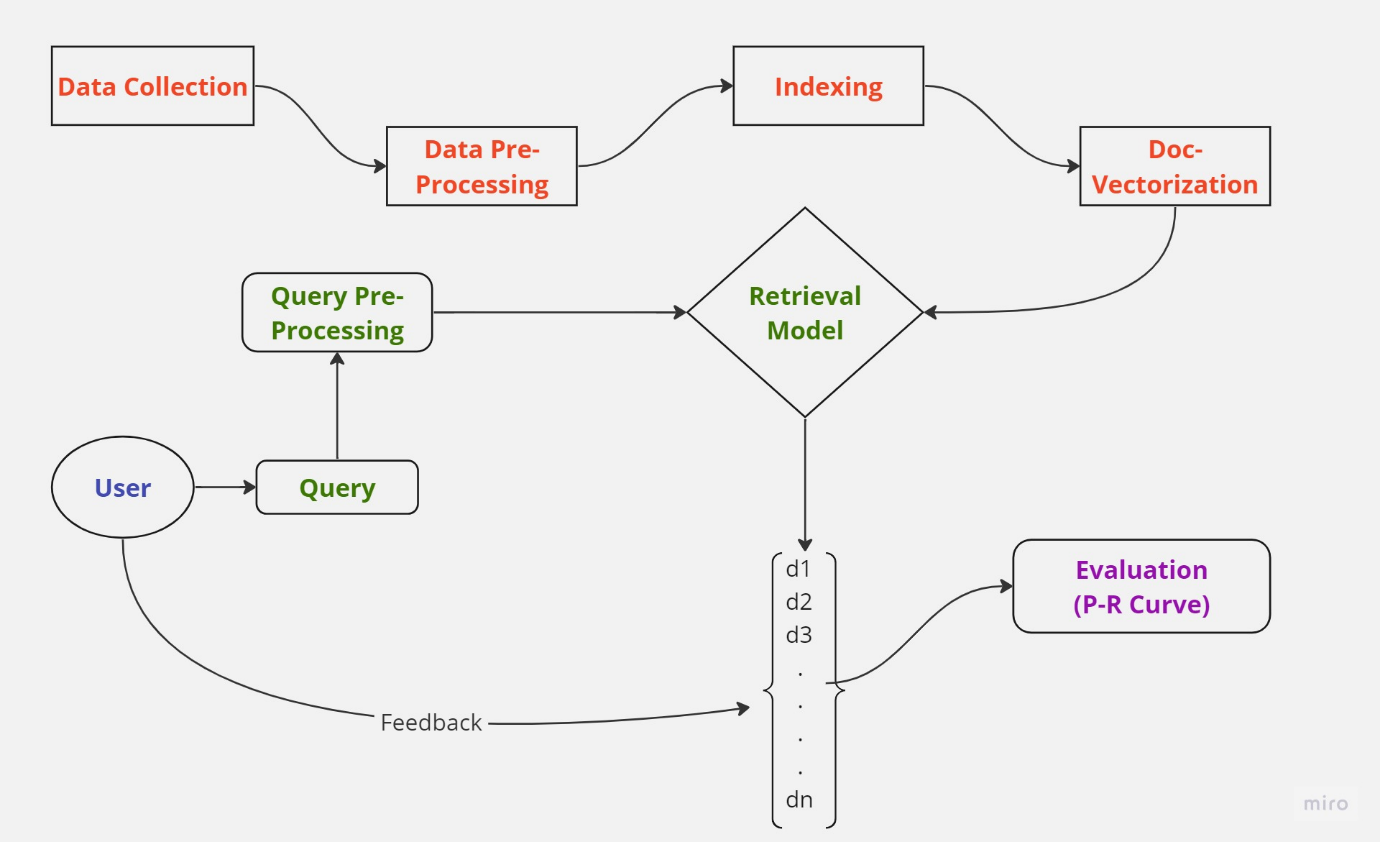
Using Vector Space Model (With doc vectors and query vectors)

1. **Search Engine Implementation:**
2. User Input
3. Query Processing
4. Cosine Similarity (Using Doc- Vectors and Query Vectors)
5. **User Interface**
6. **Feedback Mechanism:**

User Feedback collection

1. **Evaluation:**

P-R curve for query after the retrieved documents are marked as relevant/non-relevant

****

**Team Details:**

1. Rupesh Chowdary Peddineni – S20210010173
2. Varun Peta – S20210010176