**NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY**

**School of Electrical Engineering and Computer Sciences (SEECS)**

**Logo

Description automatically generated**

**COURSE**

**CS 250: Data Structures & Algorithms**

**PROJECT PROPOSAL**

**Members**

**Abdul Hadi Asad  (464594)**

**Musab Abdullah (463963)**

**Rida Waheed (462119)**

**Dataset Details**

For this project I will be using Three Datasets:

1. **Harvard Weekly News Dataset:**

**Description:** This Dataset contains two sets of weekly Global News from 2018 and 2017. It includes all types of articles like Arts and Humanities; Earth and Environmental Sciences; Engineering; Social Sciences.

**Data Size:** The dataset contains approximately 1 million articles, but I will use a subset of around 200,000 articles to ensure manageability within system constraints.

**Link:** [https://dataverse.harvard.edu/dataset](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ILAT5B)

1. **Reddit Data Science Posts ( Kaggle ):**

**Description:** A dataset containing links to [Reddit](https://www.reddit.com/) posts specifically related to data science topics. The dataset spans posts from 2010 to 2023, but I will limit it to posts from 2016 to 2023 (~200,000+ posts) for relevance and to align with modern data science trends.

**Data Attributes:** Includes post titles, URLs, associated subreddits, and timestamps.

**Link:** [https://www.kaggle.com/datasets/reddit-data-science-posts](https://www.kaggle.com/datasets/maksymshkliarevskyi/reddit-data-science-posts)

1. **Global News Dataset ( Kaggle ):**

**Description:** This Dataset contain a collection of global news articles categorized by topics such as politics, sports, and technology. It contains over **200,000** articles.

**Data Attributes:** Each record includes titles, descriptions, publication dates, categories, and URLs.

**Link:** [https://www.kaggle.com/datasets/global-news-dataset](https://www.kaggle.com/datasets/everydaycodings/global-news-dataset)

**Languages and Frameworks**

**Languages:**

1. **Python**:

* **Use Case**: Data preprocessing, dataset parsing, and generating indices like forward and inverted indices.
* **Reason**: Rich library support (e.g., **pandas, nltk, BeautifulSoup**) and ease of handling structured and unstructured data.

1. **Java**:

* **Use Case**: Core implementation of search functionalities, including efficient ranking algorithms, optimization techniques (like **barrels**), and API backend development.
* **Reason**: Known for robustness, scalability, and multithreading capabilities, which are essential for handling high query volumes.

1. **C++**:

* **Use Case**: Potentially used for performance-critical tasks, such as real-time **ranking** of search results or computationally intensive operations.
* **Reason**: Offers speed and memory efficiency for processing large datasets.

**Framework:**

1. **API Development**:

**Purpose**:

* + - Retrieve metadata (titles, descriptions, URLs) for incomplete records in the datasets.
    - Enable secure and structured communication between modules (e.g., between the indexer and the search algorithm).

**Tools**:

* + - **FastAPI** (Python): For preprocessing APIs.
    - **Spring Boot** (Java): For search functionality and backend APIs.

1. **User Interface**:

**Purpose**:

* + - Provide an intuitive search interface for users.
    - Display search results dynamically, including article titles, descriptions, and URLs.

**Technologies**:

* + - **Frontend**: JavaScript, HTML, and CSS.
    - **Integration**: APIs will be integrated into the UI using libraries like **Axios** or **Fetch**.