News Document Retrieval (Domain:Environment)

**Project Code**: 62

**Group Members**:

Kattuboina Durga Pavani Sai Harshitha – S20210010116

**Components:**

Indexing

Query processing

Searching

Refining

Relevance feedback

Assessment Component

**Project file drive link:**

<https://github.com/kdpsharshitha/P62-MiniProject-NDRHK>

**Project Description:**

Users will have interest to read some specific type of news documents. They should search for those news documents. we will provide interface using streamlit where user can give a keyword or query according to his interest and search. Our app will take the query and calculate similarity and displays top 10 results and also take user relevance feedback into consideration.

**1.Indexing Component**

It generally refers to the ordering of information, it reduces the documents to the informative terms in them

**2. Query Processsing:**

We take the query input from the user by using Streamlit and performed the following:

- remove all non alphabets regex = [^a-zA-Z],

- remove whitespaces

- convert case to lowercase

- tokenize words

- remove stopwords

- stemming

**3. Searching Component:**

we calculated the tf and idf values for the snippets of the documents and the user query. Then we calculate the similarity between query and each document by using cosine similarity ,based on that we rank the documents and return the top 10 documents.

**4. Refining Component:**

The retrieved list of document information is checked for duplicated and removed if any.

Top 10 list of documents are displayed in the browser.

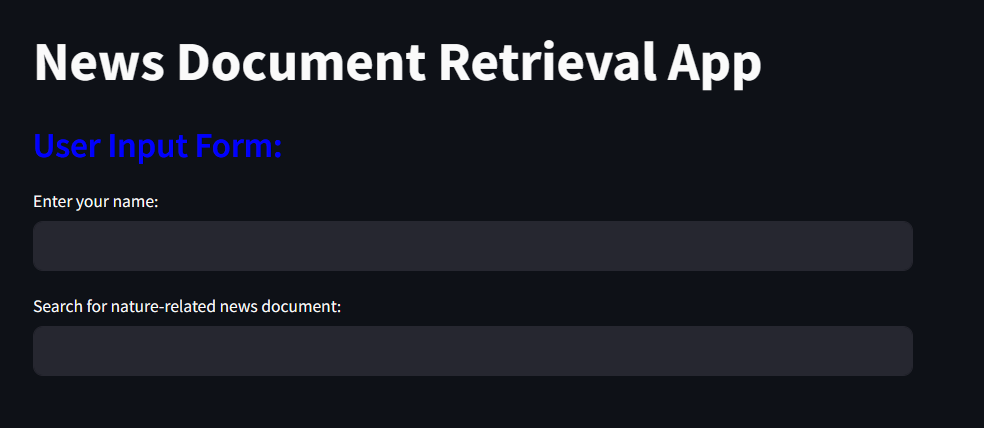
**5. Capturing Relevance feedback -**

Relevance feedback is taken from the user and relevant documents are displayed.

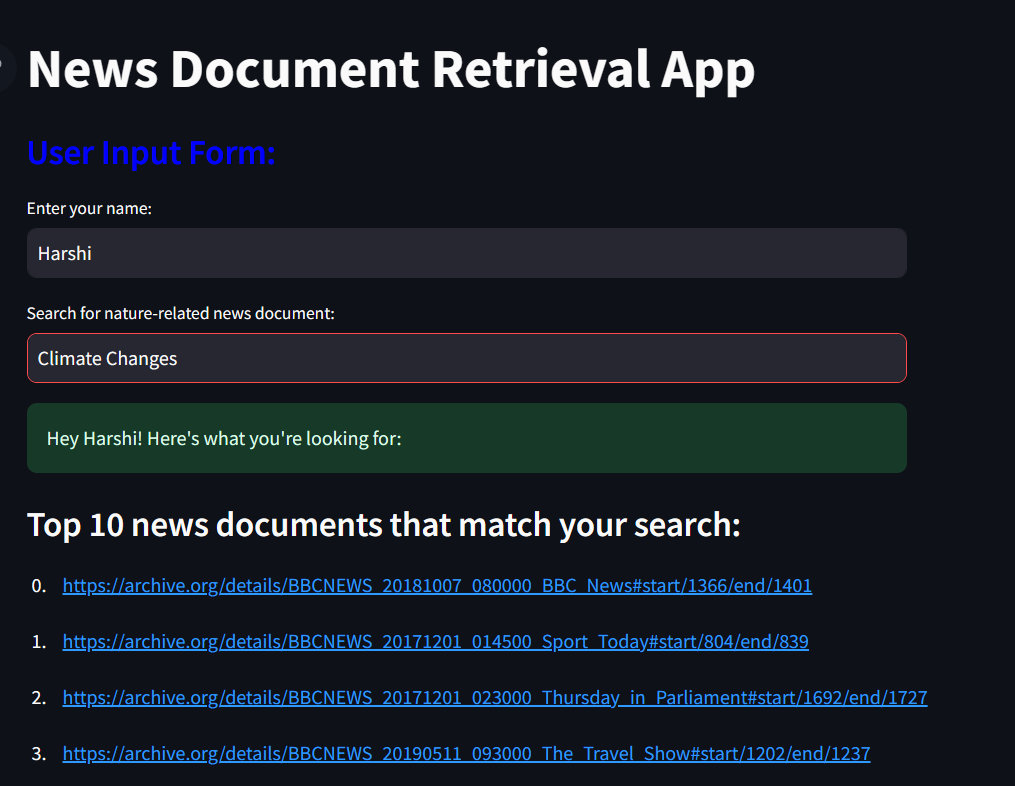
**6. Assessment Components** (Precision,Recall,P-R curve) -

We calculate the Precision, Recall and plot P-R Curve using matplotlib.

**Initial User Interface:**

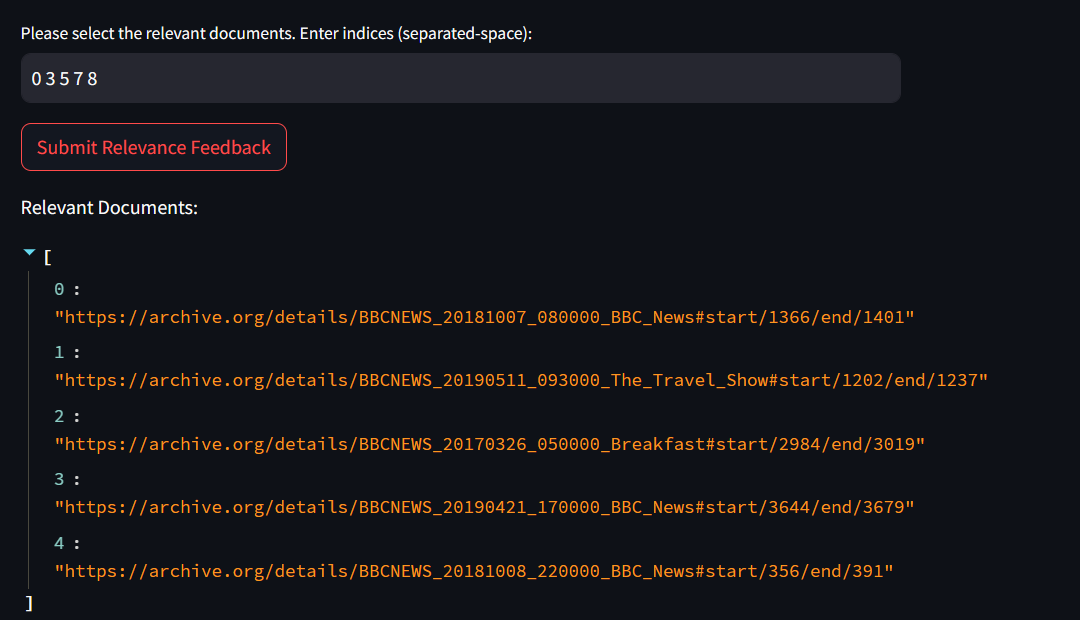


**Search Results:**





**Relevance Feedback and displaying relevant documents:**



**Precision-Recall Curve:**

