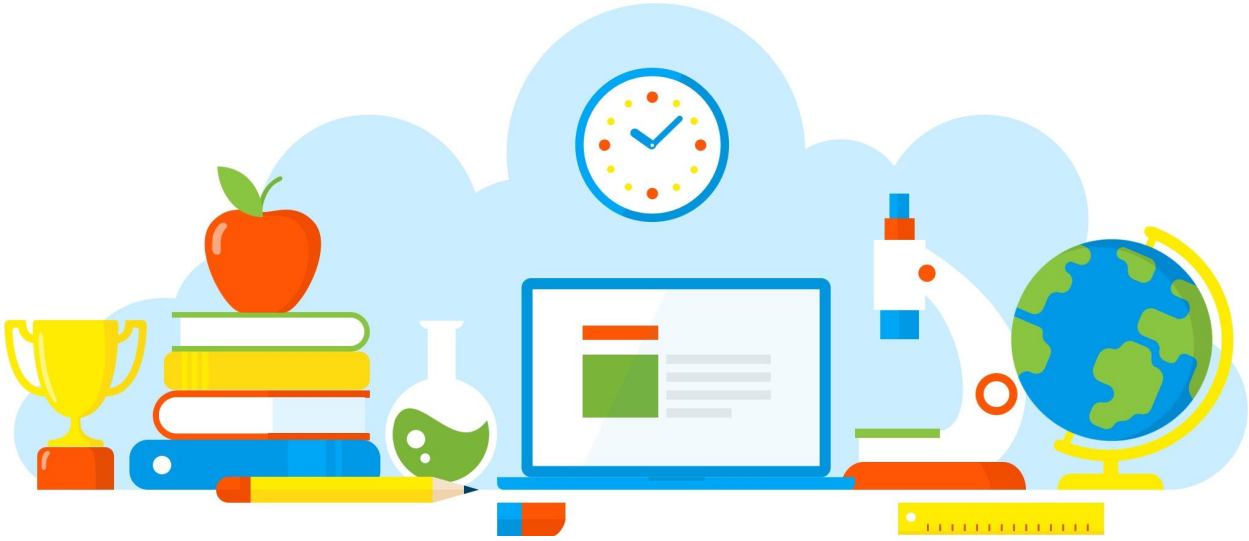

News Document Retrieval

Group - P09

Information Retrieval
12/17/2021

Project Description

News Document Retrieval is an IR system where the main focus of this search engine is to collect news articles from a collection of articles over 30,000.

The data is crawled from famous news article publishers like Times of India and The Hindu.

Tasks

- Indexing the dataset using hadoop
- Ranking the retrieved documents
- Evaluating the retrieved documents according to the relevance feedback

Sub Tasks

- Crawling data from the web in the news domain.
- Querying the collection / Searching component.
- Relevance feedback.

Team Details

Group: P09

Members:

SURYA TEJA TANGIRALA	S20190010174
MAHABOOB SHAIK	S20190010159
PRATHYUSH SIRIMALLE	S20190010165
SRI PRANAV YINTI	S20190010201

Contributions

Surya Teja:

- **Task:** Indexing the whole dataset using hadoop map-reduce.
- **Sub-Task:** Crawling the news articles from the web by creating a web crawler in python.

Mahaboob:

- **Task:** Ranking the queried documents by using necessary techniques.
- **Sub-Task:** Querying the dataset with the help of inverted index.

Prathyush:

- **Task:** Evaluating the ranked documents based on PR curves and F1 scores.
- **Sub-Task:** Taking the relevant feedback of the ranked search results from the user.

Pranav:

- **Task:** Searched relevant documents for the retrieved documents for some queries in the evaluation part.