**Anonymous Information Retrieval for News**

Dwaipayan Mondal || Parnita Bokade || Manish Kumar || Piyush Kumar Saini || Pragati Agrawal

Updated Problem Formulation:

This proposal introduces an innovative news aggregation platform designed to prioritize user privacy and deliver personalized, up-to-date news content. Unlike traditional aggregators, our system employs dynamic web crawling, advanced natural language processing (NLP), and machine learning, ensuring a tailored news experience without storing identifiable user data. This addresses the increasing demand for secure and reliable information sources that respect user anonymity.

Looking ahead, we plan to transition from manual news article collection to implementing an efficient web crawling mechanism using the Map Reduce Programming model. This advancement, deployed under the Hadoop framework, aims to enhance speed, scalability, and overall performance.

Furthermore, during the web scraping phase, our system focuses on meticulous data cleaning. This includes the examination and refinement of URLs, headlines, and dates, ensuring real-time resolution of any issues related to date accuracy or inconsistencies. By committing to data quality from the outset, our platform provides users with a reliable and accurate news experience.

Literature Review:

**1.)** In 2019[1], Gnana Venkata Naga Sai Kalyan Karumudi; Rohit Sathyajit; Sandhya Harikumar in their paper “Information retrieval and processing system for news articles in English” present a tool designed for news analysts or users to efficiently extract crucial information from diverse news articles, eliminating the need to navigate each news site individually. Their primary goal was information retrieval from news articles based on date and keywords. The tool also offers direct links to detailed news articles, empowering users to access the original content effortlessly.

A few shortcomings of this paper were:

* Lack of Multilingual support
* For instance, during the web scraping phase without the cleaning of URLs, headlines, and date there will be issues with date at the later stage.

**2.)** In 2017[2], Dhruv Khattar; Vaibhav Kumar; Vasudeva Varma in their paper “Leveraging Moderate User Data for News Recommendation” propose a model which is similar to Item-based CF which leverages the information present in the sequence in which the articles are read by the users. Along with it, they also incorporate the aspect of freshness and similarity between different articles to capture the overall interests of users. The similarity between different articles was captured using a semantic measure. They use MDP to model this altogether.

A few shortcomings of this paper were:

* Large dataset is required for this system to work.

**3.)** In 2019[3], Christos Bouras; Vassilis Poulopoulos; Panagiotis Silintziris in their paper “Personalized News Search in WWW: Adapting on User's Behaviour” employ a streamlined fusion of query expansion and results processing to generate personalized output. Diverging from previous implementations, their query expansion occurs after the initial retrieval of an unranked set of results, actively contributing to the result processing phase. In this outlined procedure, they enhanced the user's query using keywords selected by the engine based on the user's past search sessions and the categorization system of PeRSSonal. Through assigning varied importance weights to certain keywords, the engine gains insights into the user's preferences. Consequently, a relevance factor is computed for each resulting article, derived from keyword weights, facilitating the re-ranking of articles based on user preferences.

A few shortcomings of this paper were:

* The system is slow as employing this technique on a large dataset takes significant time.
* Lacks personalised search system

**4.)** In 2016[4], K. Santhiya, V. Bhuvaneswari in their paper “Map reduce programming model: Construction of inverted index for automated document clustering” endeavour to implement a map-reduce architecture for the automated partitioning of diverse crime articles. The key contributions of this study can be delineated as follows:

Firstly, the construction of an Inverted Index was pursued using the Vector Space Model. This involves a systematic approach to indexing documents based on their content, laying the groundwork for effective information retrieval.

Subsequently, the research extends its contributions to the realm of Inverted Index Construction by employing the MapReduce paradigm. Leveraging the scalability and parallel processing capabilities of MapReduce, this approach aims to enhance the efficiency and speed of constructing the Inverted Index, particularly suited for large-scale datasets.

A few shortcomings of this paper were:

* Here the news articles are collected manually, thus lacking relevant news articles many a times.

**5.)** In 2021[5], Husna Sarirah Husin; Shahrinaz Ismail in their paper “Agent-mediated Recommender System with Detect-Determine-Direct Model for News Website” look into the possibility of deploying software agents in the proposed recommendation system to cater to the users' needs of personalized contents. A conceptual framework that adopts the Detect-Determine-Direct Model was presented in this paper as the outcome of the research purpose.

A few shortcomings of this paper were:

* Dynamic online news environment is not supported in the above paper.
* real-time data retrieval is lacking

How our system plans to tackle the above issues:

1. Extend the application to support scraping from multiple news websites with different structures,

thus, tackling the lack of Multilingual support.

2. We plan to implement efficient web crawling using the Map Reduce Programming model and then

deploy under Hadoop framework.

3. We would like to further enhance the whole system with a more accurate search personalization

algorithm in order to make the whole procedure faster and in order to omit any results that are of

very low user interest.

4. We plan to introduce a user grouping mechanism to create a personalized search system in order

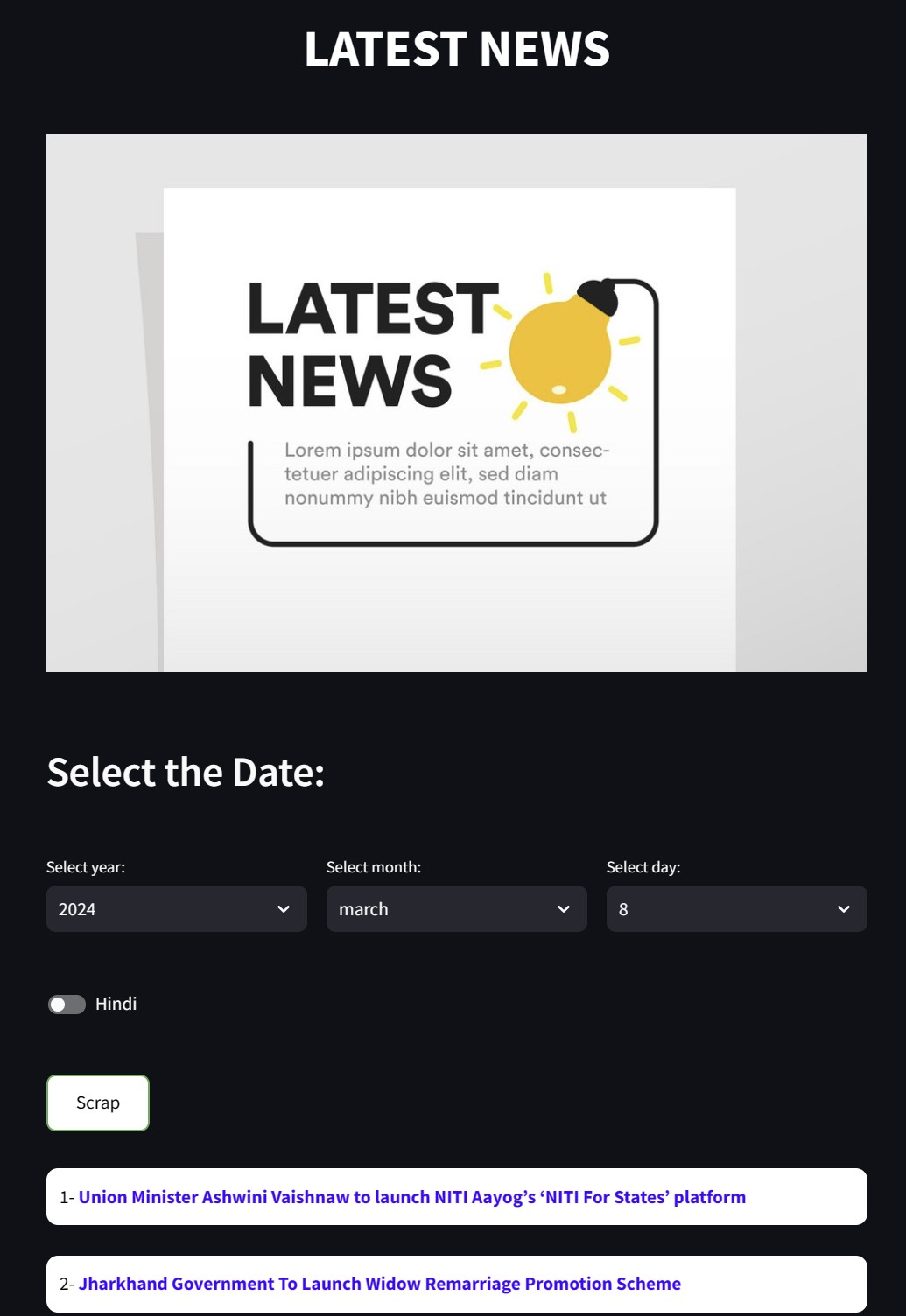
to propose some articles on same or similar queries that were selected from users with similar

profile.

5. We plan to create a caching subsystem in order to have the results of the user's queries stored and

thus, the results could be presented faster which would increase the speed of our overall system.

Screenshots:



This Python script utilizes the Streamlit framework to create a web application for fetching and displaying the latest news articles from a specified website. The application's interface includes options for selecting the desired date, a toggle button for switching between English and Hindi languages, and a "Scrap" button to initiate the news scraping process.

Upon clicking the "Scrap" button, the script sends a request to the "https://sarkaripariksha.com" website, specifically targeting a page containing GK (General Knowledge) and current affairs news. The date for scraping is determined based on the user's selection of year, month, and day. Additionally, the user can choose to view the news in either English or Hindi using the toggle button.

The script employs the BeautifulSoup library for web scraping, extracting relevant information from the HTML structure of the targeted webpage. It iterates through the retrieved news items, extracting titles and hyperlinks, and then displays them in a stylized manner using Streamlit components. The displayed news articles are organized in custom-styled boxes, providing a visually appealing presentation. The script dynamically adjusts its appearance, incorporating a blue background and white text for better visibility, along with other style modifications.

References

- [[1]](https://ieeexplore.ieee.org/document/8986223) Information retrieval and processing system for news articles in English

- [[2]](https://ieeexplore.ieee.org/document/8215736) Leveraging Moderate User Data for News Recommendation

- [[3]](https://ieeexplore.ieee.org/document/5072508) Personalized News Search in WWW: Adapting on User's Behaviour

- [[4]](https://ieeexplore.ieee.org/document/7887971) Map reduce programming model: Construction of inverted index for automated document

Clustering

- [[5]](https://ieeexplore.ieee.org/document/9567856) Agent-mediated Recommender System with Detect-Determine-Direct Model for News Website