Project Title: # News Search Engine

Subject: Information Retrieval **DS5603**

Team Members:

1. **Dhruvadeep Malakar**	**[142201026]**
2. **Pusti Vasoya**	**[112201005]**
3. **Shreya Verma**	**[112201007]**
4. **Pramod Sai**	**[112201029]**

Description

In an age characterized by information overload and the rapid dissemination of breaking news, users face significant challenges in finding reliable, timely, and relevant news articles online. With digital news platforms proliferating and social media amplifying headlines instantly, it has become crucial for both casual readers and professional researchers to access news stories that match their interests and provide credible, trustworthy information. Traditional keyword-based search engines, while powerful, often struggle with ambiguity, lack of context, and the dynamic nature of news content.

This project aims to build an advanced news search engine that leverages modern information retrieval (IR) techniques, semantic embeddings, and retrieval-augmented generation (RAG) to enhance the relevance and accuracy of search results. By grounding our work in principles learned throughout the Information Retrieval course—including document cleaning, indexing, ranking, evaluation, and the integration of language models—this project stands at the intersection of academic theory and real-world application.

We want to see how Google and other news aggregator works and how they conclude these things together,

Motivation

1. **Relevance Improvement**: In the current digital landscape, relevance is vital for a positive user experience. Users searching for current events, in-depth reporting, or opinions must be able to efficiently locate high-quality articles matching their queries.

2. **Search Engine Optimization (SEO) Awareness**: News publishers and aggregators employ SEO strategies to maximize visibility. Understanding and optimizing how articles are indexed and ranked not only improves user access but also benefits publishers by increasing engagement and reach.

3. **Timeliness and Trust**: News retrieval systems must balance speed and accuracy, providing up-to-date information while filtering misinformation and duplicate stories. Enhanced retrieval models allow for improved filtering, ranking, and presentation of news content.

Objective

- Collect and preprocess a large-scale, diverse news article corpus (minimum 5000 documents, preferably much more for robust results).
- Build an IR pipeline that incorporates document cleaning, indexing (inverted index), and ranking algorithms (TF-IDF, BM25).
- Augment baseline algorithms with semantic embeddings (BERT, SentenceTransformers) to support context-aware retrieval and improve result quality.
- Integrate retrieval-augmented generation (RAG) techniques to provide more comprehensive and informative answers beyond simple document matching.
- Apply link analysis concepts (PageRank/HITS), especially for major stories cited across sources, to assess article authority and influence.
- Implement standard SEO ranking factors to analyze how technical choices impact result visibility and overall engine effectiveness.
- Evaluate system performance using precision, recall, mean average precision (MAP), and nDCG metrics; compare results across standard baselines.

Dataset Links:

News Category Dataset

News API

Google News Datasets

We will utilize the News Category Dataset available on Kaggle, which contains over 200,000 news articles collected from HuffPost, spanning a variety of categories, timelines, and sources.

Additional datasets can be found via resources such as News API and/or Google News Datasets

Overall Progess

1. Data Collection: Acquire the News Category Dataset from Kaggle, ensuring it includes enough relevant articles spanning various categories and timelines.

2. Data Cleaning: Remove duplicates, irrelevant content, and noise; carry out tokenization, normalization, and stopword removal to prepare the corpus for indexing.

3. Indexing: Build an inverted index for efficient keyword-based lookup.

4. Baseline Ranking: Implement traditional retrieval models like TF-IDF or BM25 for initial search quality assessment.

5. Semantic Embedding: Encode queries and news articles using dense semantic embeddings (e.g., BERT) for context-aware search, improving relevance.

6. Retrieval-Augmented Generation (RAG): Integrate a RAG pipeline to incorporate LLM-based summarization or direct answering over retrieved articles.

7. SEO & Link Analysis: Analyze metadata and SEO factors, and optionally employ link analysis (PageRank/HITS) to rank article authority and relevance.

8. Evaluation: Test system accuracy and relevance using standard IR metrics like Precision, Recall, MAP, and nDCG.

9. Web App/UI: Develop a basic web interface for interactive search and result display, integrating all features for user evaluation

Outcomes

By undertaking the creation of an advanced news search engine, the project team will deepen technical expertise in both classical and state-of-the-art IR techniques. This will not only help users discover timely, trustworthy, and relevant news at scale, but will also shed light on the relationship between technical search engine choices and SEO strategies—ultimately, enhancing our understanding of information dissemination in the digital era.