ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY

DEPARTMENT OF SOFTWARE ENGINEERING

Master's Program (CEP)

News Aggregator

Web Scrapping Project

Fidel Alemayehu Hailemariam GSE0074/17

Submission Date: August, 2025

Submitted To. Dr. Girma

Introduction and Motivation

In the digital era, news is produced at an unprecedented rate across countless online platforms. However, readers often face challenges in accessing trustworthy, diverse, and timely information without being overwhelmed by irrelevant or repetitive content. This project is a Python-based news aggregator that fetches news articles from various sources on the internet using web scraping techniques. It utilizes libraries like Selenium, Beautiful Soup, and Requests to scrape data from different websites and present it in a structured format.

Services/Items to be Provided

- Website Scraping: The project scrapes news articles from predefined websites using techniques like HTTP requests and browser automation.
- Source Selection: Users can select their preferred news sources from a list of supported websites.
- Article Summarization: The application provides a summarized version of each news article to give users a quick overview.
- Category Filtering: Users can filter news articles based on predefined categories such as politics, technology, etc.
- User interface: to interact with the user

Problem Statement

In today's digital environment, news is published across countless online platforms, making it difficult for users to efficiently access relevant, trustworthy, and diverse information. Readers often have to visit multiple websites such as Google News, NDTV, and Addis Insight to stay updated on different topics, leading to information overload, redundancy, and time inefficiency. Furthermore, the absence of a unified platform that categorizes and personalizes news across these sources limits users' ability to quickly compare perspectives or focus on areas of interest such as politics, business, or technology.

This project addresses the challenge by developing a web scraping–based news aggregator that collects and organizes news articles from Google News, NDTV, and Addis Insight into their respective categories, while providing users with a centralized, personalized, and easy-to-navigate platform for streamlined news consumption.

Objectives

The main objective of this project is to design and implement a news aggregator web scraping platform that automatically collects, organizes, and personalizes news articles from Google News, NDTV, and Addis Insight, ensuring that users can access diverse and categorized information in a single interface.

Specific Objectives

1. **Automated News Extraction**

   Develop web scraping pipelines to continuously extract news headlines, summaries, and links from:

   - **NDTV**: Latest, Cities, Education, Trending, Offbeat
   - **Google News**: Technology, Business, World, Ethiopia
   - **Addis Insight**: Latest, Politics, Business, Culture, Opinion

2. **Unified Aggregation**

   Integrate all scraped articles into a centralized platform while minimizing duplication. Ensure that users can view articles source-wise or category-wise.

3. **Personalized User Experience**

   Enable users to select preferred categories (e.g., "Technology" from Google News or "Politics" from Addis Insight) to receive tailored feeds.

Brief Literature Review (personalized/adaptive news systems)

Related work. Personalized news recommendation is widely studied to fight information overload; recent surveys summarize classic approaches (content-based, collaborative filtering, bandits) and deep models for richer personalization. Emerging work explores LLM-driven modeling of both news text and user profile.

Data used. Public benchmarks include MIND (Microsoft News) 1M users and 160k+ English articles with rich text for learning and evaluation and Adressa, a session-based Norwegian dataset with clicks and dwell time (1–10 week releases). These enable reproducible comparisons across models.

Methods.

- **Contextual bandits** (e.g., LinUCB) personalize article selection under fast-changing catalogs and show measurable click-lift on Yahoo! Front Page logs.

- **Deep learning** encodes article semantics and user interests (titles/abstracts/bodies) and dominates recent leaderboards on MIND.
- **Privacy-preserving/Federated learning** appears for training recommenders without centralizing raw user behavior.

Strengths. Bandits adapt quickly to novelty; deep encoders (CNN/RNN/Transformer/LLM) capture nuanced topics and entities; public datasets standardize evaluation; federated setups mitigate privacy risk.

Weaknesses. Many studies rely on English or Western news and may not generalize to Ethiopian outlets; click-focused objectives can amplify popularity/recency bias and narrow exposure; cold-start for new users/long-tail topics remains hard; few works integrate multi-source de-duplication, credibility checks, or scraped-site category alignment end-to-end. User trust and transparency are ongoing concerns.

Gaps (Google News, NDTV, Addis Insight).

1. Multilingual/regional: multilingual support is not available and news based on regions is not available also
2. Responsible personalization: diversify recommendations (serendipity) and add credibility signals; optionally explore bandits with diversity constraints.
3. Searching method is not available: I only used filtering mechanism
4. Variety of news resources: for this project I only used 3 different news resources but it is better to have variety of news resources and multiple categories
5. Date and time filtering: it would be nicer if this project has date and time filtering.

Opportunities for the Project

1. Expansion of Sources: Add more Ethiopian and international news outlets.
2. Personalization: Allow users to select topics and get customized feeds.
3. Mobile/Cloud Deployment: Host as a web or mobile app for wider accessibility.

Methodology

The development of the news aggregator application followed a systematic process, beginning with data collection and progressing through data preprocessing, exploratory analysis, feature extraction, and presentation. The methodology is outlined as follows:

1. Data Collection

- Libraries used**:** requests, BeautifulSoup, selenium

- Different news sources were targeted, including NDTV, Google News, and Addis Insight.
- For static websites such as NDTV and Addis Insight, the requests library was used to fetch HTML pages, and BeautifulSoup was employed to parse the HTML and extract relevant content (headlines, links, and article bodies).
- For dynamic websites like Google News, which load content asynchronously, selenium with a headless Chrome browser (basically Google Chrome running without its graphical user interface (GUI)) was used to simulate scrolling and fully load the page before scraping.

2. Data Preprocessing

- Libraries used**:** BeautifulSoup, time
- Extracted data often contained HTML tags, whitespace, and irrelevant elements. These were cleaned using HTML parsing with BeautifulSoup.
- Headlines and links were standardized into a consistent format.
- The first few paragraphs of each news article were extracted as summaries, ensuring users could preview the article before reading.
- A scrolling function was implemented with Selenium and time.sleep() to dynamically load and capture complete sets of news headlines.

3. Parallel Data Retrieval

- **Libraries used:** joblib (Parallel, delayed)
- To optimize scraping speed, multiple pages were fetched in parallel using joblib's parallel processing.
- This significantly reduced the time required to scrape paginated categories such as NDTV's "Latest" and "Cities" sections.

4. Exploratory Data Handling

- Extracted news data was organized into tuples containing:
    - ✓ Headline (title of the news article)
    - ✓ Link (URL to the full article)
    - ✓ Summary text (a short excerpt from the article)
- Duplicate entries were removed to maintain clean results.
- Articles were grouped by categories (Politics, Business, Culture, etc.) for easy user navigation.

5. Feature Engineering (Presentation Layer)

- Libraries used**:** streamlit
- The application’s front-end was developed using Streamlit.
- Users can:
    - ✓ Select the news source (NDTV, Google News, Addis Insight).
    - ✓ Choose a category (For Google News Business, World, Technology, Ethiopia for NDTV Latest, Cities, Education, Trending, Offbeat for Addis Insight Latest, Politics, Business, Culture, Opinion).
    - ✓ View headlines, summaries, and clickable links.
- News was displayed in a structured, user-friendly format with styled headlines, summaries in colored text, and "Read More" buttons for full articles

6. Evaluation & Testing

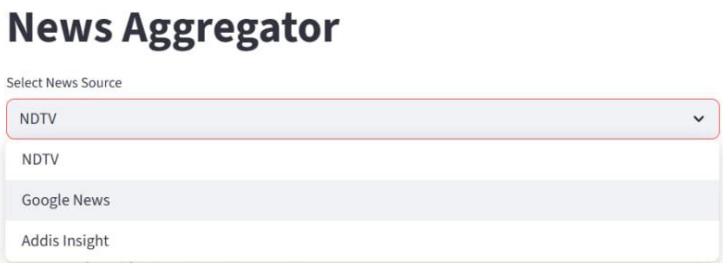
- The scraper was tested against multiple categories and pages to ensure reliability across sources.
- Verification was done by comparing the scraped headlines and links with the actual website content.
- Google News scraping was evaluated separately using Selenium to confirm that all dynamically loaded headlines were captured.

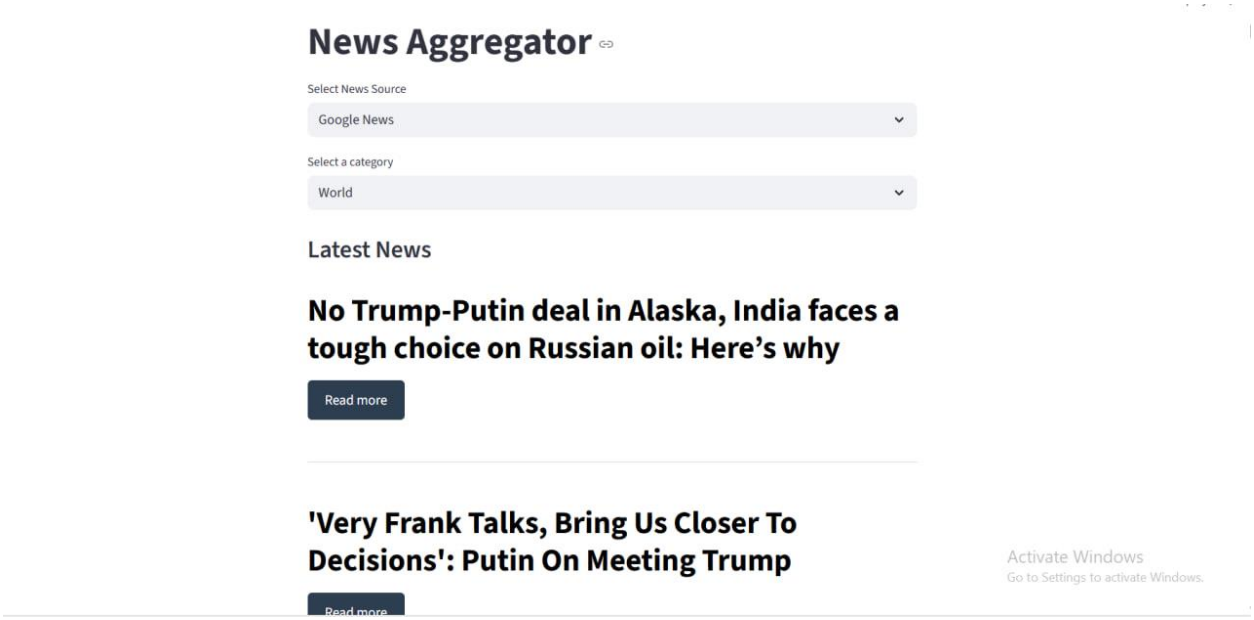
Data Summary

The web scraping process continuously collects all available articles from the specified sources (Google News, NDTV, and Addis Insight) and assigns them to their respective categories. Unlike sampled datasets, the collection is comprehensive and not limited to predefined ratios

Images from the project

.





News Aggregator

Select News Source

Google News

Select a category

Ethiopia

Latest News

Improving Fisheries Governance in the IGAD Region through Strengthening MCS

Read more

East and Horn of Africa exploring ways to

Discussion of Findings

1. **Structured Information from Unstructured Data**
    - Converted messy HTML pages into clean, structured news records with headlines, links, and short summaries.
2. **Source and Category Patterns**
    - Identified that different sources emphasize different areas (e.g., Addis Insight focuses on Politics & Culture, Google News emphasizes World & Business).
3. **Efficiency in Data Collection**
    - Learned how to handle both static content (NDTV, Addis Insight) and dynamic JavaScript content (Google News) using the right tools (BeautifulSoup, selenium).
4. **Real-Time Aggregation**
    - Built a system that automatically pulls live updates, ensuring users always access the most recent news.
5. **Usability and Evaluation**
    - Through Streamlit, provided users with an organized, interactive news hub, making raw data easy to explore and read.

Conclusion

This project built a functional News Aggregator using Python libraries like BeautifulSoup, Selenium, Requests, and Streamlit. It successfully scraped and organized news from NDTV, Google News, and Addis Insight, presenting them in an interactive interface. The system makes real-time news more accessible and demonstrates how web scraping can address information overload.

References

Google News

Get Latest News, India News, Breaking News, Today's News - NDTV.com

Home - Addis Insight

Microsoft News Recommendation Dataset - Azure Open Datasets | Microsoft Learn

Recommender systems using LinUCB: A contextual multi-armed bandit approach | by Yogesh Narang | TDS Archive | Medium

Adressa - adressa.no