

Data Driven Online News Summarizer-ML-NLP

Prepared by: Sohel Rana

Position: Digital Currency Investor & Technical Lead

Nationality: Bangladeshi

1. Introduction

The Data Driven Online News Summarizer-ML-NLP project is designed to automate the collection, storage, and summarization of online news articles using modern machine learning and natural language processing (NLP) techniques. The system leverages FastAPI for web service delivery, SQLite for lightweight data storage, and the TextRank algorithm for extractive summarization.

This project demonstrates how data-driven approaches can be applied to real-world information streams, enabling users to consume concise summaries of complex news articles efficiently.

2. Objectives

- Automate the ingestion of online news via RSS feeds.
- Store articles in a structured database for persistence.
- Apply NLP-based summarization to generate concise, human-readable summaries.
- Provide both JSON APIs and HTML interfaces for accessibility.
- Ensure robustness with error handling and fallback mechanisms.

3. Technology Stack

- Framework: FastAPI (Python)
- Database: SQLite (via SQLAlchemy ORM)
- Data Source: BBC World News RSS feed (feedparser)
- Summarization Engine: TextRank (Sumy library)
- Deployment: Uvicorn ASGI server
- Environment: Visual Studio Code (development IDE)

4. Project Structure

The project follows a simple, modular structure:

Data Driven Online News Summarizer/

```
|
+-- main.py          # FastAPI application code
| +-- Database setup (SQLite + News table)
| +-- Utility functions (clean_text, summarize_text)
| +-- RSS fetcher + local seed fallback
| +-- Startup event (populate DB with articles)
| +-- Endpoints:
|   +-- "/" Homepage (HTML list of articles + Summarize links)
|   +-- "/favicon.ico" (fixes browser favicon request)
|   +-- "/news" (JSON list of articles)
|   +-- "/summarize/{id}" (HTML summary of selected article)
|
+-- requirements.txt  # Python dependencies
| +-- fastapi, uvicorn, sqlalchemy, requests, feedparser, sumy
|
+-- news.db          # SQLite database storing articles
    +-- News table: id, title, content
```

5. Workflow

- Startup Event
 - Fetches the latest news articles from BBC RSS feed.
 - Cleans and stores them in the SQLite database.
 - Falls back to sample local articles if RSS is unavailable.
- Homepage (/)
 - Displays a list of articles with clickable "Summarize" links.
- Summarization (/summarize/{id})
 - Retrieves the article by ID.
 - Applies TextRank summarization.
 - Returns a concise HTML summary.

- API Endpoints
 - /news -> JSON list of all articles.
 - /favicon.ico -> Prevents browser favicon errors.

6. Key Features

- Data Driven: Automatically ingests live news feeds.
- NLP Powered: Uses TextRank for extractive summarization.
- Dual Interface: JSON for developers, HTML for end-users.
- Error Handling: Graceful fallback to local data when feeds fail.
- Lightweight: Runs on SQLite, suitable for local or cloud deployment.

7. Applications

- Media Monitoring: Quick summaries for journalists and analysts.
- Financial Insights: Investors can scan global news faster.
- Educational Use: Students can learn NLP concepts with a practical project.
- Technical Demonstration: Showcases integration of FastAPI, databases, and NLP.

8. Conclusion

The Data Driven Online News Summarizer-ML-NLP project exemplifies how modern NLP techniques can be combined with lightweight web frameworks to deliver practical, real-world applications. By automating news ingestion and summarization, the project reduces information overload and enhances decision-making efficiency.

This project reflects the growing importance of data-driven solutions in both technical and investment domains, aligning with the vision of leveraging AI and ML for smarter information consumption.