Documentation

Github Repository link : https://github.com/vaibhavsemwalwork/NewsAnalysisTool

---

Project Setup:

Installation:

1. **Install Python**: Ensure Python 3.8 or higher is installed on your system.
2. **Install Dependencies**: Run the following command to install the required Python packages.
3. **Set Up ChromeDriver**: Selenium requires ChromeDriver to automate the browser. The webdriver-manager package will automatically handle the installation of ChromeDriver.
4. **API Key Configuration**: Add your Gemini API key to the config.py file:

```
GEMINI_API_KEY = "your_api_key_here"

MODEL_TYPE = "gemini-2.0-flash"  # or any other model type

n = 5  # number of articles to process
```

Step by Step Instructions:

1. Clone the repository
```git clone git@github.com:vaibhavsemwalwork/NewsAnalysisTool.git```.

2. (optional) Create python virtual environment
```python -m venv <environment name>```
and activate the environment
```source <\environment name>/bin/activate```

3. Install the required packages:

```pip install -r requirements.txt```

4. Create a config.py with same content as sample_config.py

5. In config.py set your google gemini api key.

(create google account -> go to google AI studio -> generate api Key -> copy api key -> paste in config.py)

5. In one terminal run flask api for backend
```python api.py```

6. In a new terminal run streamlit ui
```streamlit run app.py```

Running the Application

1. Once UI starts, enter company's name in the Option provided.

2. Article will be Analyzed and displayed in UI, Sentiment distribution will be plotted, Hindi Audio summary can be played.

3. Download the json(analysis) and audio(Hindi).

The script will:

- Scrape articles related to the specified company (e.g., "OpenAI").
- Perform sentiment analysis, summarization, topic extraction, and comparative analysis.
- Generate a final report in JSON format and an audio file in Hindi.

Model Details

Sentiment Analysis

- **Model**: The script uses the transformers library with a pre-trained sentiment analysis model (DistilBERT).
- **Output**: Returns sentiment labels (e.g., positive, negative, neutral) and a sentiment score for each article.

Article Summarization

- **Model**: The script uses the transformers library with a pre-trained summarization model (e.g., DistilBART).
- **Output**: Generates a single-line summary for each article.

Topic Extraction

- **Model**: The script uses Google's Gemini 2.0 Flash model to extract topics from articles.
- **Output**: Returns a list of topics for each article.

Comparative Analysis

- **Model**: The script uses Google's Gemini 2.0 Flash model to perform a comparative analysis of multiple articles.
- **Output**: Returns a comparative summary of the articles.

Text-to-Speech (TTS)

- **Model**: The script uses gTTS (Google Text-to-Speech) to convert text into Hindi speech.
- **Output**: Generates an audio file in Hindi.

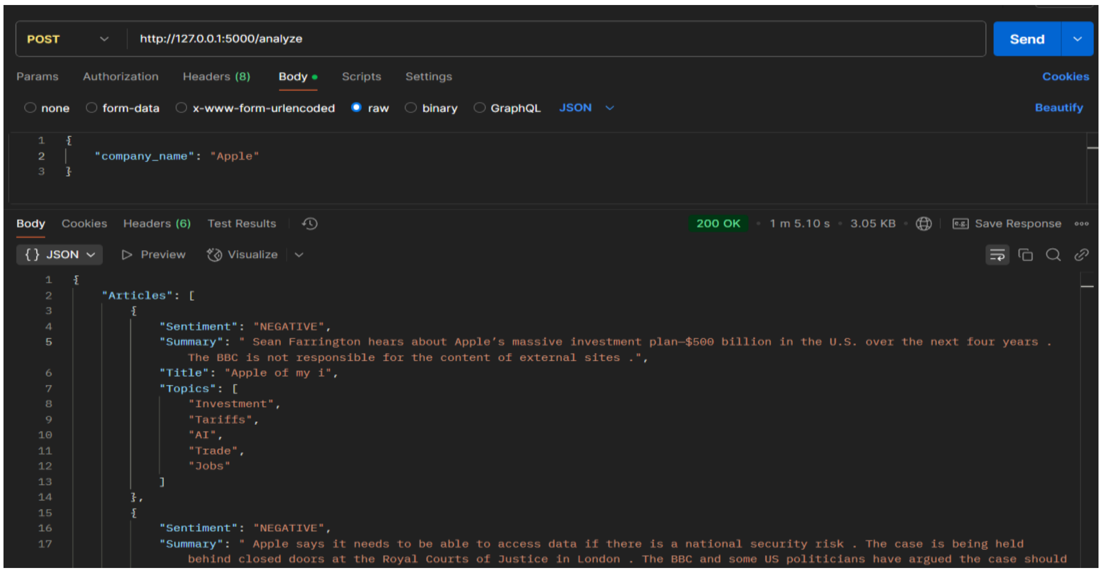
API Development

API Usage

- **Google Gemini API**: Used for topic extraction, comparative analysis, and final sentiment analysis.
  - **Integration**: The API key is stored in config.py and accessed via the google.generativeai library.
  - **Endpoint**: The script interacts with the Gemini API to generate responses based on the input text.
- **gTTS API**: Used for text-to-speech conversion.
  - **Integration**: The gTTS library is used to convert text into speech and save it as an audio file.

Accessing APIs via Postman

- **Google Gemini API**:
  - **Endpoint**: https://generativelanguage.googleapis.com/v1beta/models/{model}:generateContent
  - **Body**: JSON payload with the input text and model parameters.
- **gTTS API**:
  - The gTTS library is used locally, so no external API calls are made.
- **local API to analyze sentiment (POST)**:
  - **Endpoint** : http://localhost:5000/analyze   (Run app.py before accessing locally)
  - **Body** : {"company_name": "Apple"}
  - **Response** : json for sentiment analysis



```
http://127.0.0.1:5000/analyze
```

```
{
    "company_name": "Apple"
}
```

```
{
    "Articles": [
        {
            "Sentiment": "NEGATIVE",
            "Summary": " Sean Farrington hears about Apple’s massive investment plan—$500 billion in the U.S. over the next four years .
                The BBC is not responsible for the content of external sites .",
            "Title": "Apple of my i",
            "Topics": [
                "Investment",
                "Tariffs",
                "AI",
                "Trade",
                "Jobs"
            ]
        },
        {
            "Sentiment": "NEGATIVE",
            "Summary": " Apple says it needs to be able to access data if there is a national security risk . The case is being held
                behind closed doors at the Royal Courts of Justice in London . The BBC and some US politicians have argued the case should
```

{AWS : ip-172-31-30-194.eu-north-1.compute.internal:8080/analyze } {will work only when instance is initiated}

API Usage (Third-Party APIs)

Google Gemini API

- **Purpose**: Used for advanced natural language processing tasks such as topic extraction, comparative analysis, and sentiment analysis.
- **Integration**: The API key is configured in config.py, and the google.generativeai library is used to interact with the API.

gTTS API

- **Purpose**: Used for converting text into speech in Hindi.
- **Integration**: The gTTS library is used to generate audio files from text.

Assumptions & Limitations

Assumptions

1. **Article Availability**: The script assumes that the articles related to the specified company are available on the BBC website.
2. **Language**: The script assumes that the articles are in English and translates the summary into Hindi.
3. **API Limits**: The script assumes that the Google Gemini API key has sufficient quota for the required operations.

Limitations

1. **Scraping**: The script is dependent on the structure of the BBC website. Changes to the website’s HTML structure may break the scraping functionality.
2. **API Dependencies**: The script relies on third-party APIs (Google Gemini and gTTS). Any downtime or changes to these APIs may affect functionality.
3. **Language Support**: The translation and TTS functionality are limited to English and Hindi.
4. **Performance**: The script may take significant time to process a large number of articles due to API calls and model inference.

Codebase Structure

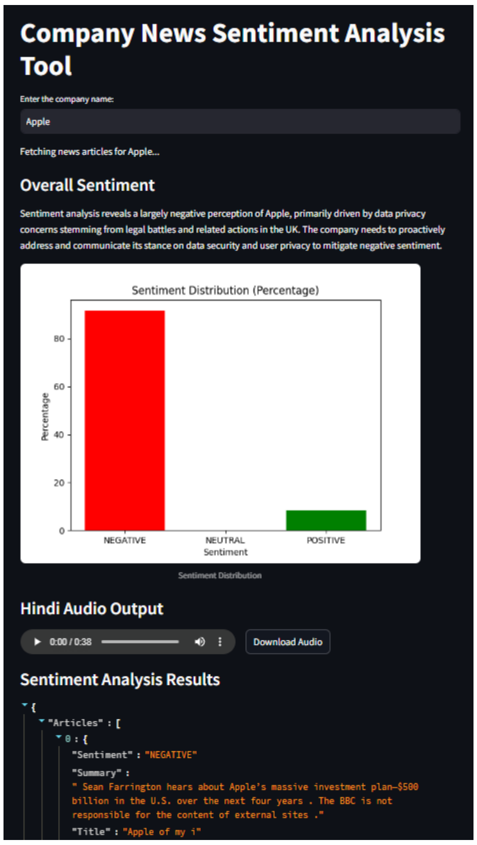
Key Functions

1. **scrape(company_name)**: Uses Selenium to scrape article links from the BBC website.
2. **extract_content(article_urls)**: Uses BeautifulSoup to extract the title and content of articles.
3. **analyze_sentiment(articles)**: Performs sentiment analysis on the articles using a pre-trained model.
4. **summarize_article(content)**: Generates a summary of the article using DistilBART.
5. **topic_of_text(api_key, articles)**: Extracts topics from the articles using the Gemini API.
6. **comparative_analysis(analyzed_articles)**: Performs a comparative analysis of the articles using the Gemini API.
7. **translate_text(text)**: Translates the summary from English to Hindi.
8. **text_to_speech_gtts(text)**: Converts the translated text into Hindi speech using gTTS.
9. **report(company_name, ...)**: Generates a final report in JSON format.

Example Workflow

1. Scrape articles related to "Apple".
2. Extract content from the articles.
3. Perform sentiment analysis and summarization.
4. Extract topics and perform comparative analysis.
5. Translate the summary into Hindi and generate an audio file.
6. Generate a final report in JSON format.

Example Output :



Company News Sentiment Analysis Tool

Enter the company name:

Apple

Fetching news articles for Apple...

Overall Sentiment

Sentiment analysis reveals a largely negative perception of Apple, primarily driven by data privacy concerns stemming from legal battles and related actions in the UK. The company needs to proactively address and communicate its stance on data security and user privacy to mitigate negative sentiment.

Sentiment Distribution

Hindi Audio Output

Sentiment Analysis Results

```
{
  "Articles" : [
    0 : {
      "Sentiment" : "NEGATIVE"
      "Summary" :
      " Sean Farrington hears about Apple's massive investment plan—$500
      billion in the U.S. over the next four years . The BBC is not
      responsible for the content of external sites ."
      "Title" : "Apple of my i"
```