DA5402 Assignment 1

Name: **Anuj Jagannath Said**
Roll no: **ME21B172**

Indian Institute Of Technology Madras

February 2, 2025

Contents

1 Introduction

This report presents an automated data pipeline designed to generate an image captioning dataset by continuously extracting news data from the Google News website. The pipeline is organized into multiple distinct modules, each responsible for separate functionalities including web scraping, lazy loading, image handling, database connectivity, de-duplication, and overall orchestration.

2 Problem Statement

The data pipeline needs to continuously extract fresh news articles from Google News and build an image captioning dataset comprising `<caption, image>` tuples. The key requirements include:

- Scraping the main Google News homepage using configurable parameters.
- Dynamically identifying and following the link to the "Top Stories" section.
- Extracting article details such as the complete URL, headline, thumbnail image (with lazy-loading support), source, and timestamp.
- Storing these extracted tuples in a PostgreSQL database, using separate tables for news meta-information and image data.
- Implementing de-duplication logic to ensure that duplicate news entries are not inserted into the database.
- Orchestrating the entire sequence in a cascaded style for continuous, automated operation.

3 Module Descriptions and Functionality Overview

3.1 Module1AndModule2Web_scrapping.py

This module is responsible for fetching the Google News homepage and dynamically extracting the link to the "Top Stories" section. Using the **BeautifulSoup** library, it parses the HTML content to locate articles. For each article, it extracts essential details including:

- **Link:** A fully qualified URL pointing to the complete news article (obtained using link https://news.google.com/home).
- **Title:** The headline or caption of the article.
- **Image:** The URL of the thumbnail image, adjusted to handle lazy loading.
- **Source and Timestamp:** Additional metadata such as the publisher and publication time.

The module also uses a **ThreadPoolExecutor** to download images concurrently, significantly reducing the overall processing time.

3.2 Module3.py

Module3 contains functions that support lazy loading and detailed image processing. Its responsibilities include:

- **Configuration Handling:** Reading parameters (e.g., base URL, headers, CSS selectors, query parameters) from a configuration file.
- **Lazy Loading:** The function to factor lazy loading prepares the HTTP request with additional parameters as specified.
- **Metadata Extraction:** Extracting specific components from an article's HTML, such as link, title, image, source, and timestamp.
- **Image Processing:** Handling image download operations with error management.

3.3 Module4.py

This module establishes a connection with a PostgreSQL database and creates the necessary tables if they don't already exist. Key functionalities include:

- **Database Connection:** Using configuration details to connect locally to the PostgreSQL instance.
- **Table Creation:** Setting up two tables:
  - **news_meta_data:** To store news URL, headline, source, and scraping timestamp.
  - **news_image_data:** To store binary image data associated with the corresponding news title.

3.4 Module5.py

Module5 implements the de-duplication logic to ensure data integrity in the database. It checks if a record already exists based on a combination of the news URL and title. This prevents redundant entries during repeated pipeline executions, thereby maintaining a clean dataset.

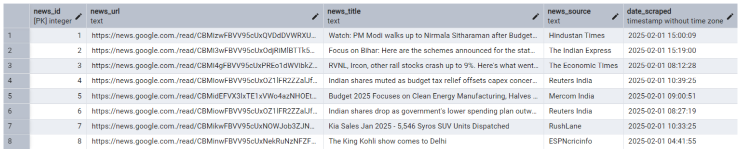
3.5 Module6.py

Serving as the orchestrator of the entire pipeline, Module6 integrates all the previous modules to create an end-to-end process. Its primary functionalities include:

- **Database Setup:** Invoking Module4 to create necessary database tables.
- **Data Extraction:** Calling Module1AndModule2Web_scrapping.py to fetch the latest news data.
- **De-duplication and Insertion:** Using Module5 to verify that only unique news entries are inserted into the database tables.
- **Overall Coordination:** Acting as the main script that orchestrates the sequential execution of tasks, ensuring that the process can be scheduled (for example, via a cron job) for continuous operation.

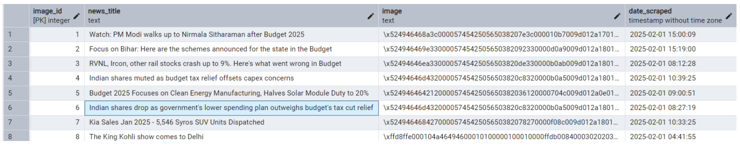
4 Results

Following are the tables obtained by running the orchestration script (Module6.py) script. Table 1 shows the metadata extracted from the Top Stories, which 2 shows image data extracted from news URL and stored in binary format.



	news_id [PK] integer	news_url text	news_title text	news_source text	date_scraped timestamp without time zone
1	1	https://news.google.com./read/CBMizwFBVV95cUxQVDdDVWRXU...	Watch: PM Modi walks up to Nirmala Sitharaman after Budget...	Hindustan Times	2025-02-01 15:00:09
2	2	https://news.google.com./read/CBMi3wFBVV95cUxOdjRiMlBTTk5...	Focus on Bihar: Here are the schemes announced for the stat...	The Indian Express	2025-02-01 15:19:00
3	3	https://news.google.com./read/CBMi4gFBVV95cUxPREo1dWVibkZ...	RVNL, Ircon, other rail stocks crash up to 9%. Here's what went...	The Economic Times	2025-02-01 08:12:28
4	4	https://news.google.com./read/CBMiowFBVV95cUxOZ1lFR2ZZalJf...	Indian shares muted as budget tax relief offsets capex concer...	Reuters India	2025-02-01 10:39:25
5	5	https://news.google.com./read/CBMidEFVX3lxTE1xVWo4azNHOEt...	Budget 2025 Focuses on Clean Energy Manufacturing, Halves ...	Mercom India	2025-02-01 09:00:51
6	6	https://news.google.com./read/CBMiowFBVV95cUxOZ1lFR2ZZalJf...	Indian shares drop as government's lower spending plan outw...	Reuters India	2025-02-01 08:27:19
7	7	https://news.google.com./read/CBMikwFBVV95cUxNOWJob3ZJN...	Kia Sales Jan 2025 - 5,546 Syros SUV Units Dispatched	RushLane	2025-02-01 10:33:25
8	8	https://news.google.com./read/CBMinwFBVV95cUxNekRuNzNFZF...	The King Kohli show comes to Delhi	ESPNcricinfo	2025-02-01 04:41:55

Figure 1: Table of news metadata



	image_id [PK] integer	news_title text	image text	date_scraped timestamp without time zone
1	1	Watch: PM Modi walks up to Nirmala Sitharaman after Budget 2025	\x524946468a3c00005745425056503820 7e3c000010b7009d012a1701...	2025-02-01 15:00:09
2	2	Focus on Bihar: Here are the schemes announced for the state in the Budget	\x524946469e3300005745425056503820923300000d0a9009d012a1801...	2025-02-01 15:19:00
3	3	RVNL, Ircon, other rail stocks crash up to 9%. Here's what went wrong in Budget	\x52494646ea33000057454250565038 20de330000b0ab009d012a1801...	2025-02-01 08:12:28
4	4	Indian shares muted as budget tax relief offsets capex concerns	\x52494646d43200005745425056503820c8320000b0a5009d012a1801...	2025-02-01 10:39:25
5	5	Budget 2025 Focuses on Clean Energy Manufacturing, Halves Solar Module Duty to 20%	\x5249464642120000574542505650382036120000704c009d012a0e01...	2025-02-01 09:00:51
6	6	Indian shares drop as government's lower spending plan outweighs budget's tax cut relief	\x52494646d43200005745425056503820c8320000b0a5009d012a1801...	2025-02-01 08:27:19
7	7	Kia Sales Jan 2025 - 5,546 Syros SUV Units Dispatched	\x5249464684270000574542505650382078270000f08c009d012a1801...	2025-02-01 10:33:25
8	8	The King Kohli show comes to Delhi	\xffd8ffe000104a46494600010100000100010000ffdb00840003020203...	2025-02-01 04:41:55

Figure 2: Table of news image data

5 Regarding Config_file.txt

5.1 Configuration to setup before execution

The configuration file (config_file.txt) centralizes all parameters the data pipeline requires. This allows modifications without changing the code. Below is a brief explanation of each variable and its use.

- **url:** The base URL for the Google News homepage. It is the starting point for scraping.
- **target_url:** The prefix used to convert relative URLs (extracted during scraping) into absolute URLs for articles and images.
- **headers:** HTTP headers (such as User-Agent) sent with each request to mimic browser behavior and avoid blocking.
- **params:** Additional query parameters for HTTP requests. Critical for retrieving lazy-loaded content.
- **selectors:** A collection of CSS selectors that define how to locate different HTML elements (e.g., article, link, title, image, source, and timestamp).
- **database:** The name of the PostgreSQL database where scraped data is stored.
- **user:** The username used to authenticate with the PostgreSQL database.
- **password:** The password corresponds to the PostgreSQL user.

- **host:** The address (typically localhost) of the PostgreSQL server.
- **port:** The port number on which the PostgreSQL server is listening.

5.2 Usage in the Pipeline

- **Web Scraping:** The `url`, `target_url`, `headers`, `params`, and `selectors` drive the scraping process to retrieve news articles and images.
- **Database Connectivity:** The `database`, `user`, `password`, `host`, and `port` are used to establish a secure connection with the PostgreSQL server for storing data.

6 Conclusion

The designed pipeline successfully automates the extraction of news data from Google News while ensuring robust handling of dynamic content through lazy loading. By segregating responsibilities among modular components and enforcing de-duplication during database insertion, this solution maintains an up-to-date and non-redundant image captioning dataset. The cascading orchestration guarantees seamless integration and continuous automated operation, fulfilling the requirements of the assignment.