# DA5402 Assignment 1

Name: **Anuj Jagannath Said**

Roll no: **ME21B172**


**Indian Institute Of Technology**
**Madras**

February 1, 2025

# Contents

# 1 Introduction

The report contains my approach to creating an automated data pipeline design for generating an image captioning dataset. The pipeline leverages the Google News website as a source for obtaining a continuous feed of `<caption, image>` tuples.

# 2 Problem Statement

The task is to build a data pipeline that continuously extracts news data from Google News and creates an image captioning dataset. The pipeline comprises several modules:

- **Module 1:** Scrape the homepage of Google News using configurable parameters.
- **Module 2:** Dynamically extract the URL for the "Top Stories" section without hard-coding the link text (done via python library **BeautifulSoup**).
- **Module 3:** Extract the thumbnail images and headlines from the Top Stories page. The implementation handles lazy loading of content.
- **Module 4:** Store the extracted tuples in a database (I have **PostgreSQL**). Two separate tables are maintained: one for image data and another for meta-information such as headlines, URLs, and timestamps.
- **Module 5:** Implement de-duplication logic to avoid inserting redundant data into the database.
- **Module 6:** Create an orchestration script that executes the pipeline modules in a cascaded style and logs execution details.

# 3 Web Scraping Code Overview (web_scrapping.py)

The web scraping module is responsible for obtaining the top news stories from Google News. Key features include:

## 3.1 Configuration Handling

- The script reads a configuration file (`config_file.txt`) that provides essential parameters such as the base URL, target URL, HTTP headers, CSS selectors for scraping various components (article, link, title, image, source, and timestamp), and query parameters.
- This design ensures that changes in page structure or URLs can be easily managed without modifying the source code.

## 3.2 Extracting Top Stories

- The function `extract_top_stories()` is the core routine.

- It first loads the homepage and uses **BeautifulSoup** to locate the element (using a CSS selector from the configuration) that points to the Top Stories section.
- The code concatenates the extracted relative URL to form the full URL for the Top Stories page so as to extract the URL for the image of the corresponding news.

### 3.3 Data Extraction and Lazy Loading

- From the Top Stories page, the script iterates over each article using the provided CSS selectors.
- For each article, the script extracts:
  - **Link:** The complete URL pointing to the full news article.
  - **Title:** The headline text.
  - **Image:** The URL of the thumbnail, derived from an attribute (e.g., `srcset`) with adjustments to handle lazy-loaded images.
  - **Source and Timestamp:** Additional metadata such as the publishing source and the article's publication time.
- To optimize image retrieval, the script uses a `ThreadPoolExecutor` to download images concurrently (decreasing the execution time).

## 4 Database Connection and Data Insertion (connecting_db.py)

The database module connects to a PostgreSQL instance and handles data storage. Its notable aspects include:

### 4.1 Database Connection and Table Creation

- The function `hosting_db_locally()` establishes a connection to the local PostgreSQL database.
- Two tables are created if they do not exist:
  - **news_meta_data:** Stores meta-information (news URL, title, source, and scraping timestamp). Table 1 is the a screenshot of the database **news_meta_data:**

| | 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/CBMiowFBVV95cUxOZ1IFR2ZZaUf... | 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/CBMidEFVX3IxTE1xVWo4azNHOEt... | Budget 2025 Focuses on Clean Energy Manufacturing, Halves ... | Mercom India | 2025-02-01 09:00:51 |
| 6 | 6 | https://news.google.com./read/CBMiowFBVV95cUxOZ1IFR2ZZaUf... | 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 meta data

– **news_image_data:** Stores the binary image data linked to the corresponding news title. Table 2 is the a screenshot of the database **news_image_data:**

| | 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 | \x524946468a3c000057454250565038207e3c000010b7009d012a1701... | 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 | \x52494646ea3300005745425056503820de330000b0ab009d012a1801... | 2025-02-01 08:12:28 |
| 4 | 4 | Indian shares muted as budget tax relief offsets capex concerns | \x5249464d4320000574542505650382 0c8320000b0a5009d012a1801... | 2025-02-01 10:39:25 |
| 5 | 5 | Budget 2025 Focuses on Clean Energy Manufacturing, Halves Solar Module Duty to 20% | \x52494646421200005745425056503820361200007 04c009d012a0e01... | 2025-02-01 09:00:51 |
| 6 | 6 | Indian shares drop as government's lower spending plan outweighs budget's tax cut relief | \x5249464d4320000574542505650382 0c8320000b0a5009d012a1801... | 2025-02-01 08:27:19 |
| 7 | 7 | Kia Sales Jan 2025 - 5,546 Syros SUV Units Dispatched | \x52494646842700005745425056503820782700000f08c009d012a1801... | 2025-02-01 10:33:25 |
| 8 | 8 | The King Kohli show comes to Delhi | \xffd8ffe000104a4649460001010000010001000 0ffdb0084000302020 3... | 2025-02-01 04:41:55 |

Figure 2: Table of news image data

### 4.2 Data Insertion with De-duplication

- Before inserting a news, the function `is_news_in_database()` checks whether the news item already exists in the database by checking whether **news URL** and **headline** matches or not (as these two fully specifies that news article are not repeated in our dataset).
- New records are inserted into both tables only if they pass the de-duplication check.

## 5 Cascaded Execution and Orchestration

The entire pipeline is designed to operate in a cascaded fashion:

- The `connecting_db.py` script sequentially calls the modules:
  – First, it ensures that all necessary database tables exist (necessary check prior inserting entries into the table).
  – Next invokes the function `extract_top_stories()` from the web scraping module to fetch the latest data.
  – Finally, it processes the data by checking for duplicates and inserting new records.
- This end-to-end process can be scheduled as a cron job to maintain a continuous feed of fresh data.

## 6 Conclusion

The web scraping module reliably extracts news information from Google News using dynamic selectors and handles lazy loading efficiently. The database module ensures that the data is stored with appropriate de-duplication mechanisms, and the orchestrated execution guarantees continuous, automated operation.