Web Scraping to create Image Captioning Dataset

R.Rohith
EP21B030

1 CAUTION

Ensure that you have Google Chrome and MongoDB installed. Additionally, verify that all packages listed in the `requirements.txt` file are installed. Simply run `main.py` once you are set.

2 Config

The configuration JSON file includes the following three parameters:

- *home_page_url*: The URL of the Google News home page.
- *keywords*: A list of potential keywords to search for the top stories page, containing strings that could represent **Top stories**.
- *image_dir*: The directory, relative to the project directory, where images are saved locally for future reference.

3 Module 1

Module 1's goal is to scrape the HTML content from the home page of Google News. The `fetch_page.py` Python file accomplishes this task. Given the *home_page_url*, it returns the HTML content of the home page using the `requests` library.

4 Module 2

Module 2 aims to retrieve the URL of the top stories page from the HTML content of the home page. This is handled by the `fetch_topstories_url.py` file, which, when given the HTML content, searches through the text of all $< a >$ tags. It returns the `href` attribute (the URL of the top stories page, relative to the home page) when the text matches a keyword in the *keywords* list.

5 Module 3

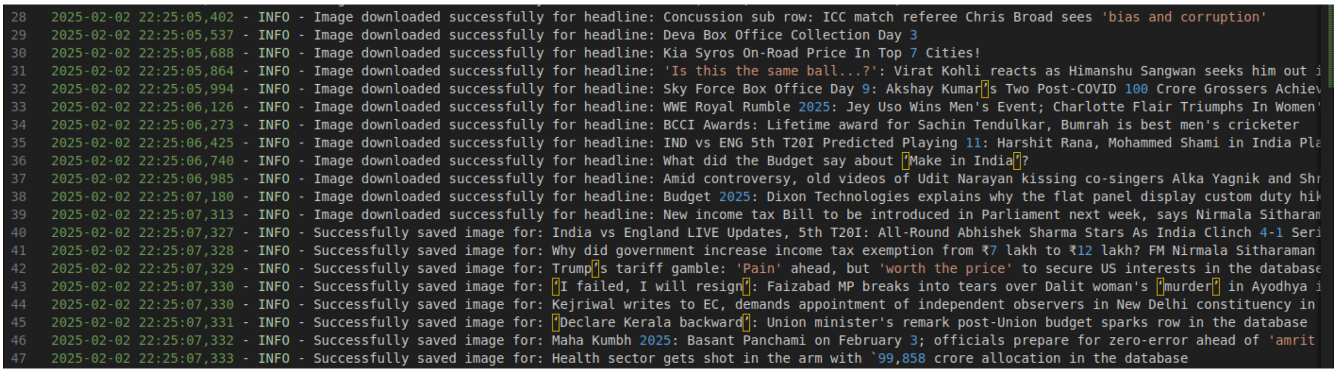
Module 3 focuses on scraping images and captions from the top stories page. It uses the `selenium` package to handle lazy loading by scraping the HTML content of the top stories page through the Chrome browser. The script then iterates over all `article` tags in the page that contain figures to gather the images, captions, and publication date-time.

6 Module 4 & Module 5

Module 4 receives data from Module 3 (including the caption, image URL, date-time, and image data) and passes it through Module 5. Module 5 checks whether an entry with the same caption and date-time already exists in the database. The data is then stored in the **image_captioning_database**. Images are saved in the **images collection**, and the captions with metadata are stored in the **captions collection**. The images are associated with the unique ID of their metadata in the captions collection, ensuring a one-to-one mapping between the entries in both collections. Both are implemented in the `save_db.py` file.

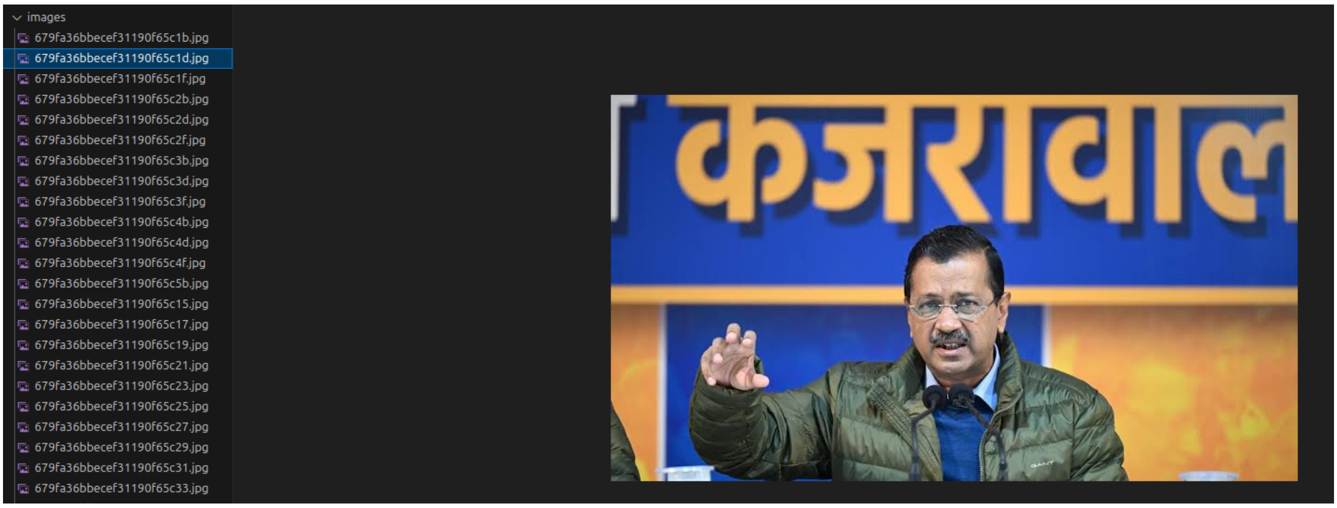
7 Module 6

Module 6 orchestrates all the previous modules and handles logging of actions along with their timestamps. The logs are saved in the directory named `logs`.



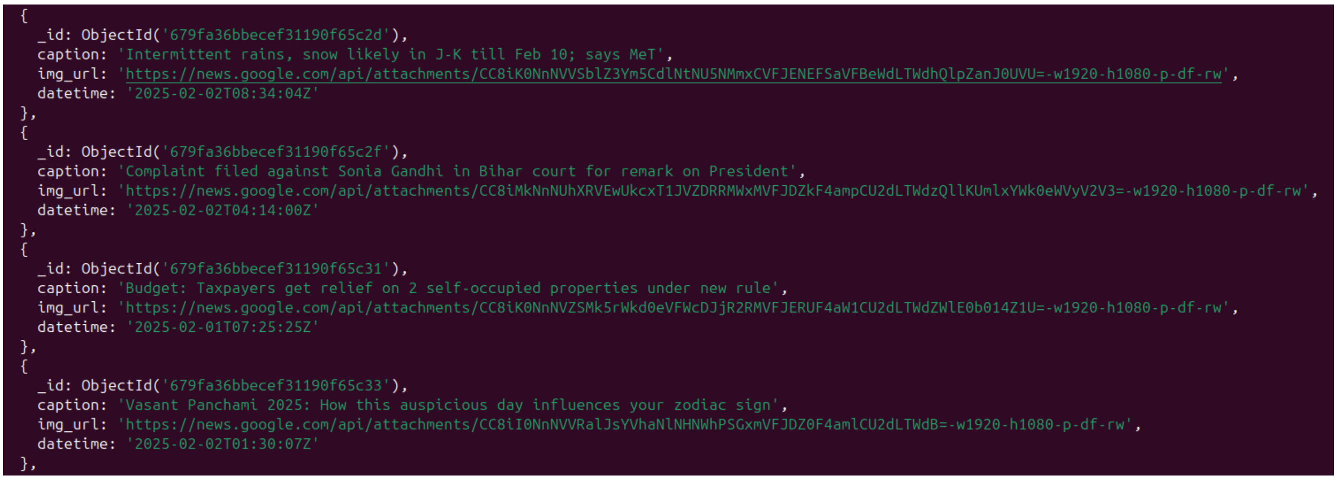
```
2025-02-02 22:25:05,402 - INFO - Image downloaded successfully for headline: Concussion sub row: ICC match referee Chris Broad sees 'bias and corruption'
2025-02-02 22:25:05,537 - INFO - Image downloaded successfully for headline: Deva Box Office Collection Day 3
2025-02-02 22:25:05,688 - INFO - Image downloaded successfully for headline: Kia Syros On-Road Price In Top 7 Cities!
2025-02-02 22:25:05,864 - INFO - Image downloaded successfully for headline: 'Is this the same ball...?': Virat Kohli reacts as Himanshu Sangwan seeks him out i
2025-02-02 22:25:05,994 - INFO - Image downloaded successfully for headline: Sky Force Box Office Day 9: Akshay Kumar’s Two Post-COVID 100 Crore Grossers Achiev
2025-02-02 22:25:06,126 - INFO - Image downloaded successfully for headline: WWE Royal Rumble 2025: Jey Uso Wins Men's Event; Charlotte Flair Triumphs In Women'
2025-02-02 22:25:06,273 - INFO - Image downloaded successfully for headline: BCCI Awards: Lifetime award for Sachin Tendulkar, Bumrah is best men's cricketer
2025-02-02 22:25:06,425 - INFO - Image downloaded successfully for headline: IND vs ENG 5th T20I Predicted Playing 11: Harshit Rana, Mohammed Shami in India Pla
2025-02-02 22:25:06,740 - INFO - Image downloaded successfully for headline: What did the Budget say about ‘Make in India’?
2025-02-02 22:25:06,985 - INFO - Image downloaded successfully for headline: Amid controversy, old videos of Udit Narayan kissing co-singers Alka Yagnik and Shr
2025-02-02 22:25:07,180 - INFO - Image downloaded successfully for headline: Budget 2025: Dixon Technologies explains why the flat panel display custom duty hik
2025-02-02 22:25:07,313 - INFO - Image downloaded successfully for headline: New income tax Bill to be introduced in Parliament next week, says Nirmala Sitharam
2025-02-02 22:25:07,327 - INFO - Successfully saved image for: India vs England LIVE Updates, 5th T20I: All-Round Abhishek Sharma Stars As India Clinch 4-1 Seri
2025-02-02 22:25:07,328 - INFO - Successfully saved image for: Why did government increase income tax exemption from ₹7 lakh to ₹12 lakh? FM Nirmala Sitharaman
2025-02-02 22:25:07,329 - INFO - Successfully saved image for: Trump’s tariff gamble: 'Pain' ahead, but 'worth the price' to secure US interests in the database
2025-02-02 22:25:07,330 - INFO - Successfully saved image for: ‘I failed, I will resign’: Faizabad MP breaks into tears over Dalit woman's ‘murder’ in Ayodhya i
2025-02-02 22:25:07,330 - INFO - Successfully saved image for: Kejriwal writes to EC, demands appointment of independent observers in New Delhi constituency in
2025-02-02 22:25:07,331 - INFO - Successfully saved image for: ‘Declare Kerala backward’: Union minister's remark post-Union budget sparks row in the database
2025-02-02 22:25:07,332 - INFO - Successfully saved image for: Maha Kumbh 2025: Basant Panchami on February 3; officials prepare for zero-error ahead of 'amrit
2025-02-02 22:25:07,333 - INFO - Successfully saved image for: Health sector gets shot in the arm with `99,858 crore allocation in the database
```

Figure 1: A snapshot of the log file



```
images
  679fa36bbecef31190f65c1b.jpg
  679fa36bbecef31190f65c1d.jpg
  679fa36bbecef31190f65c1f.jpg
  679fa36bbecef31190f65c2b.jpg
  679fa36bbecef31190f65c2d.jpg
  679fa36bbecef31190f65c2f.jpg
  679fa36bbecef31190f65c3b.jpg
  679fa36bbecef31190f65c3d.jpg
  679fa36bbecef31190f65c3f.jpg
  679fa36bbecef31190f65c4b.jpg
  679fa36bbecef31190f65c4d.jpg
  679fa36bbecef31190f65c4f.jpg
  679fa36bbecef31190f65c5b.jpg
  679fa36bbecef31190f65c15.jpg
  679fa36bbecef31190f65c17.jpg
  679fa36bbecef31190f65c19.jpg
  679fa36bbecef31190f65c21.jpg
  679fa36bbecef31190f65c23.jpg
  679fa36bbecef31190f65c25.jpg
  679fa36bbecef31190f65c27.jpg
  679fa36bbecef31190f65c29.jpg
  679fa36bbecef31190f65c31.jpg
  679fa36bbecef31190f65c33.jpg
```

Figure 2: A sample of scraped image along with the local image directory



```
{
  _id: ObjectId('679fa36bbecef31190f65c2d'),
  caption: 'Intermittent rains, snow likely in J-K till Feb 10; says MeT',
  img_url: 'https://news.google.com/api/attachments/CC8iK0NnNVVSblZ3Ym5CdlNtNU5NMmxCVFJENEFSaVFBeWdLTWdhQlpZanJ0UVU=-w1920-h1080-p-df-rw',
  datetime: '2025-02-02T08:34:04Z'
},
{
  _id: ObjectId('679fa36bbecef31190f65c2f'),
  caption: 'Complaint filed against Sonia Gandhi in Bihar court for remark on President',
  img_url: 'https://news.google.com/api/attachments/CC8iMkNnNUhXRVEwUkcxT1JVZDRRMWxMVFJDZkF4ampCU2dLTWdzQllKUmlxYWk0eWVyV2V3=-w1920-h1080-p-df-rw',
  datetime: '2025-02-02T04:14:00Z'
},
{
  _id: ObjectId('679fa36bbecef31190f65c31'),
  caption: 'Budget: Taxpayers get relief on 2 self-occupied properties under new rule',
  img_url: 'https://news.google.com/api/attachments/CC8iK0NnNVZSMk5rWkd0eVFWcDJjR2RMVFJERUF4aW1CU2dLTWdZWlE0b014Z1U=-w1920-h1080-p-df-rw',
  datetime: '2025-02-01T07:25:25Z'
},
{
  _id: ObjectId('679fa36bbecef31190f65c33'),
  caption: 'Vasant Panchami 2025: How this auspicious day influences your zodiac sign',
  img_url: 'https://news.google.com/api/attachments/CC8iI0NnNVVRalJsYVhaNlNHNWhPSGxmVFJDZ0F4amlCU2dLTWdB=-w1920-h1080-p-df-rw',
  datetime: '2025-02-02T01:30:07Z'
},
```

Figure 3: Few entries from the **captions collection** in MongoDB Database