DA5402 - Assignment 1

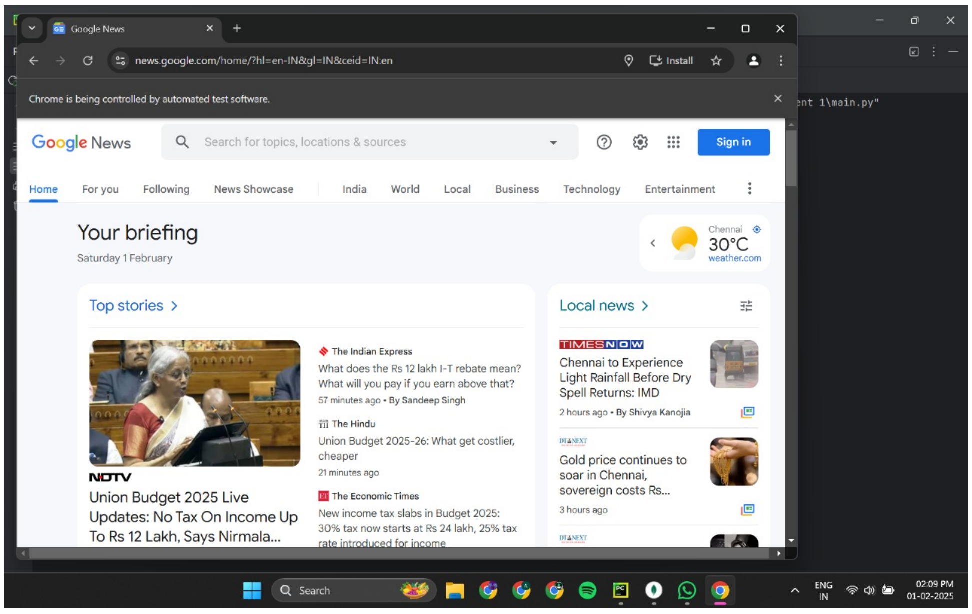
Anirudh Rao be21b004

Module 1

To scrape Google News, the `selenium` library in Python was used. This was chosen as it has the ability to scrape lazy loading websites. It uses a web driver for a web browser to interact with websites and scrape information from them. For this project, the web driver for Google Chrome was used.

The Chrome web driver, powered by `selenium`, opens the homepage of Google News. The URL for this is not hard-coded and is instead configurable via the attached `config.json` file.

To account for lazy loading, a sleep time is incorporated to allow the website to load sufficient content for scraping. This sleep time is also configurable using `config.json`.



NDTV
Union Budget 2025 Live Updates: No Tax On Income Up To Rs 12 Lakh, Says Nirmala...

The Indian Express
What does the Rs 12 lakh I-T rebate mean? What will you pay if you earn above that?
57 minutes ago • By Sandeep Singh

The Hindu
Union Budget 2025-26: What get costlier, cheaper
21 minutes ago

The Economic Times
New income tax slabs in Budget 2025: 30% tax now starts at Rs 24 lakh, 25% tax rate introduced for income

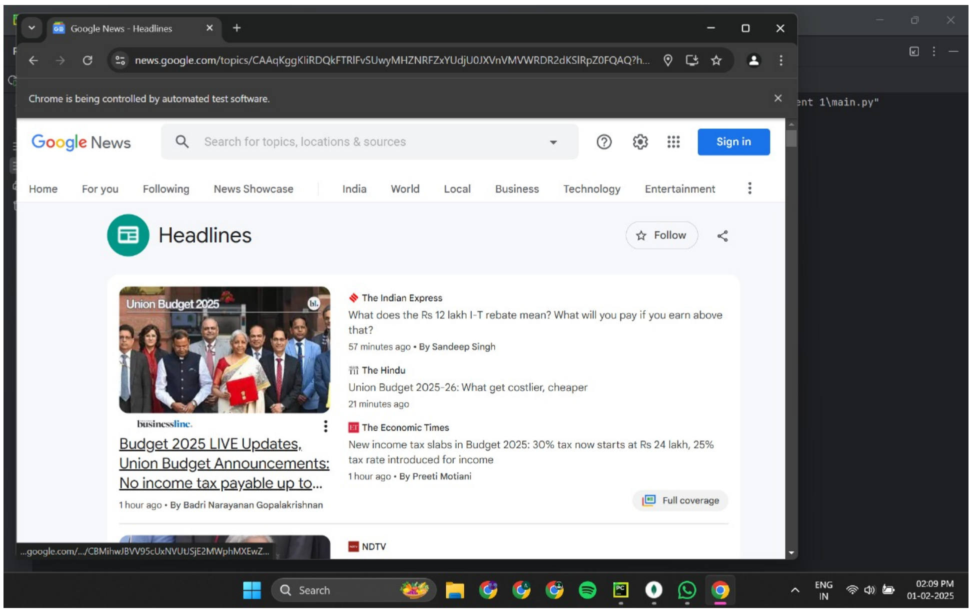
TIMES NOW
Chennai to Experience Light Rainfall Before Dry Spell Returns: IMD
2 hours ago • By Shivya Kanojia

DT NEXT
Gold price continues to soar in Chennai, sovereign costs Rs...
3 hours ago

Figure 1: `selenium` opening the Google News homepage

Module 2

Next, `selenium` clicks on the 'Top stories' link to navigate to that page. This string is not hard-coded and is instead configurable via `config.json`.



news.google.com/topics/CAAqKggKIiRDQkFTRlFvSUwyMHZNRFZxYUdjU0JXVnVMVMVWRDR2dKSlRpZ0FQAQ?h...

Chrome is being controlled by automated test software.

Search for topics, locations & sources

businessline.
Budget 2025 LIVE Updates, Union Budget Announcements: No income tax payable up to...
1 hour ago • By Badri Narayanan Gopalakrishnan

The Indian Express
What does the Rs 12 lakh I-T rebate mean? What will you pay if you earn above that?
57 minutes ago • By Sandeep Singh

The Hindu
Union Budget 2025-26: What get costlier, cheaper
21 minutes ago

The Economic Times
New income tax slabs in Budget 2025: 30% tax now starts at Rs 24 lakh, 25% tax rate introduced for income
1 hour ago • By Preeti Motiani

NDTV

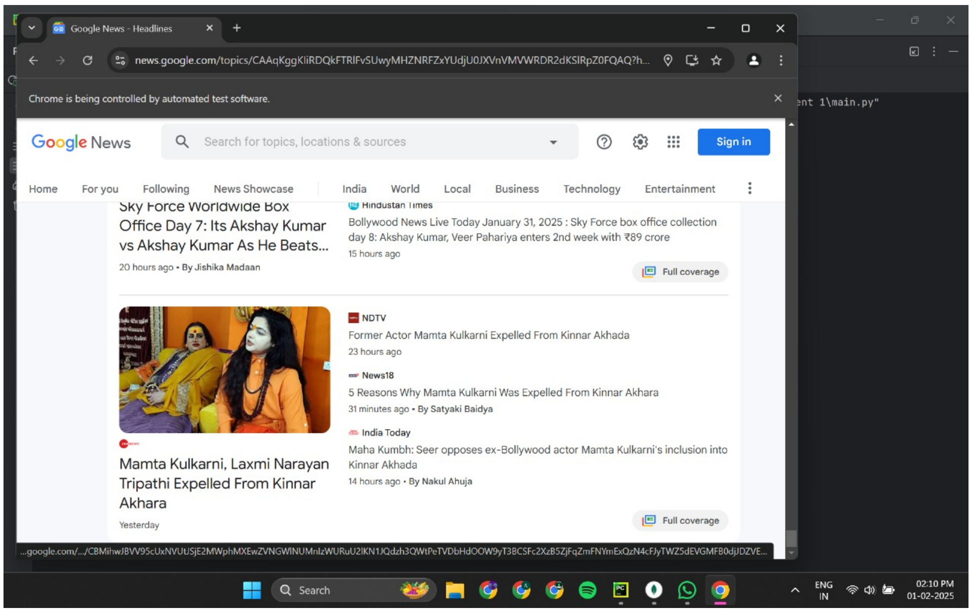
...google.com/.../CBMihwJBVV95cUxNVUUSjE2MWphMXEwZ...

ent 1\main.py"

Figure 2: `selenium` navigating to the 'Top stories' page

Module 3

`selenium` scrolls through the 'Top stories' page until the height of the webpage stays constant. This scrolling is performed using the configured sleep time, keeping in mind the lazy loading of the webpage.



news.google.com/topics/CAAqKggKIiRDQkFTRlFvSUwyMHZNRFZxYUdjU0JXVnVMVMVWRDR2dKSlRpZ0FQAQ?h...

Chrome is being controlled by automated test software.

Search for topics, locations & sources

Sky Force Worldwide Box Office Day 7: Its Akshay Kumar vs Akshay Kumar As He Beats...
20 hours ago • By Jishika Madaan

Hindustan Times
Bollywood News Live Today January 31, 2025 : Sky Force box office collection day 8: Akshay Kumar, Veer Pahariya enters 2nd week with ₹89 crore
15 hours ago

Mamta Kulkarni, Laxmi Narayan Tripathi Expelled From Kinnar Akhara
Yesterday

NDTV
Former Actor Mamta Kulkarni Expelled From Kinnar Akhada
23 hours ago

News18
5 Reasons Why Mamta Kulkarni Was Expelled From Kinnar Akhara
31 minutes ago • By Satyaki Baidya

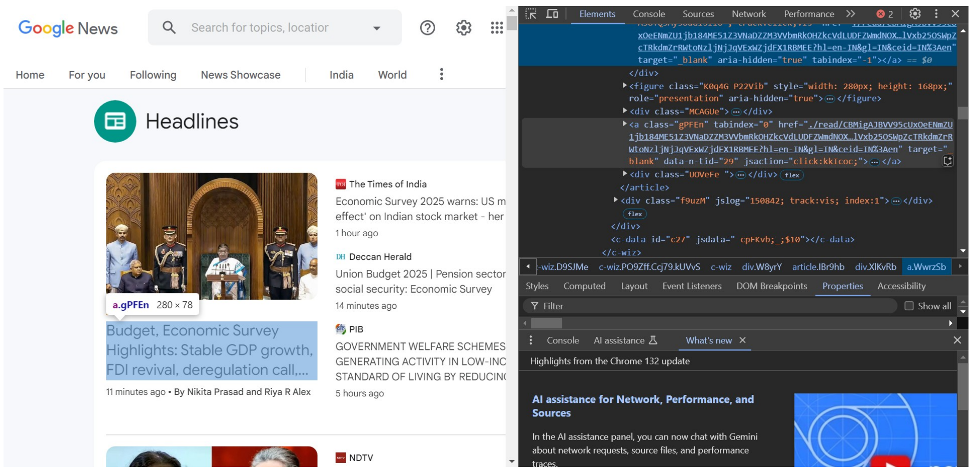
India Today
Maha Kumbh: Seer opposes ex-Bollywood actor Mamta Kulkarni's inclusion into Kinnar Akhada
14 hours ago • By Nakul Ahuja

...google.com/.../CBMihwJBVV95cUxNVUUSjE2MWphMXEwZVNGWlNUMnIzWURuU2lKN1JQdzh3QWtPeTVDbHdOOW9yT3BCSFc2XzB5ZjFqZmFNYmExQzN4cFlyTWZ5dEVGMFB0djlDZVE...

ent 1\main.py"

Figure 3: `selenium` navigating to the bottom of the 'Top stories' page

Finally, `selenium` uses CSS selectors to scrape the headline, thumbnail, newspaper, article URL, and article date. To account for Google News changing its layout, these CSS selectors are not hard-coded and are instead configurable via `config.json`.



```
xOeENmZU1jb184ME51Z3VNaDZZM3VVbmRkOHZkcVdLUDFZWmdNOX…lVxb25OSWpZ
cTRkdmZrRWtoNzljNjJqVExWZjdFX1RBMEE?hl=en-IN&gl=IN&ceid=IN%3Aen"
target="_blank" aria-hidden="true" tabindex="-1"></a> == $0
</div>
▸<figure class="K0q4G P22Vib" style="width: 280px; height: 168px;"
role="presentation" aria-hidden="true">⋯</figure>
▸<div class="MCAGUe">⋯</div>
▸<a class="gPFEn" tabindex="0" href="./read/CBMigAJBVV95cUxOeENmZU
1jb184ME51Z3VNaDZZM3VVbmRkOHZkcVdLUDFZWmdNOX…lVxb25OSWpZcTRkdmZrR
WtoNzljNjJqVExWZjdFX1RBMEE?hl=en-IN&gl=IN&ceid=IN%3Aen" target="_
blank" data-n-tid="29" jsaction="click:kkIcoc;">⋯</a>
▸<div class="UOVeFe ">⋯</div> flex
</article>
▸<div class="f9uzM" jslog="150842; track:vis; index:1">⋯</div>
flex
</div>
<c-data id="c27" jsdata=" cpFKvb;_;$10"></c-data>
</c-wiz>
```

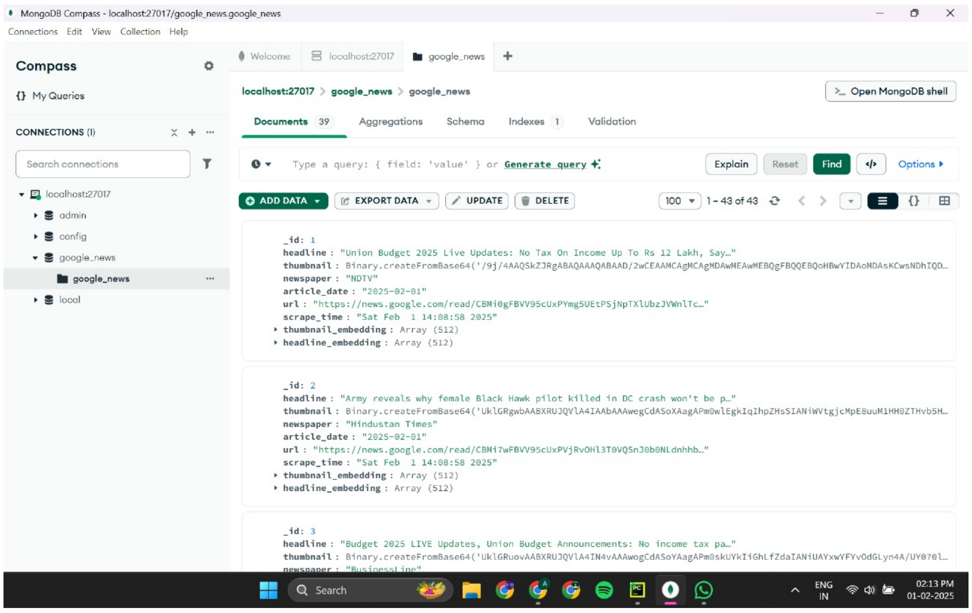
Figure 4: Example of a CSS selector using the class (`gPFEn`) of the headline

Module 4

MongoDB was chosen as the database to store the scraped data. The library `pymongo` is used to interact with MongoDB programmatically. The database is stored locally and not on the cloud.

Connecting to the MongoDB database requires some configurations, which are not hard-coded and are specified in the `config.json` file. In addition to this, MongoDB Compass must be installed locally on the system being used.

The scraped data and metadata, along with the scrape time, are added by the script to the database as a 'document', which is equivalent to a Python dictionary. The thumbnail is stored in a binary format that is friendly to MongoDB. The original image can be obtained from the binary data using the `PIL` library.



```
_id: 1
headline : "Union Budget 2025 Live Updates: No Tax On Income Up To Rs 12 Lakh, Say…"
thumbnail : Binary.createFromBase64('/9j/4AAQSkZJRgABAQAAAQABAAD/2wCEAAMCAgMCAgMDAwMEAwMEBQgFBQQEBQoHBwYIDAoMDAsKCwsNDhIQD…
newspaper : "NDTV"
article_date : "2025-02-01"
url : "https://news.google.com/read/CBMi0gFBVV95cUxPYmg5UEtPSjNpTXlUbzJVWnlTc…"
scrape_time : "Sat Feb  1 14:08:58 2025"
▸ thumbnail_embedding : Array (512)
▸ headline_embedding : Array (512)

_id: 2
headline : "Army reveals why female Black Hawk pilot killed in DC crash won't be p…"
thumbnail : Binary.createFromBase64('UklGRgwbAABXRUJQVlA4IAAbAAAwegCdASoXAagAPm0wlEgkIqIhpZHsSIANiWVtgjcMpE8uuM1HH0ZTHvb5H…
newspaper : "Hindustan Times"
article_date : "2025-02-01"
url : "https://news.google.com/read/CBMi7wFBVV95cUxPVjRvOHl3T0VQSnJ0b0NLdnhhb…"
scrape_time : "Sat Feb  1 14:08:58 2025"
▸ thumbnail_embedding : Array (512)
▸ headline_embedding : Array (512)

_id: 3
headline : "Budget 2025 LIVE Updates, Union Budget Announcements: No income tax pa…"
thumbnail : Binary.createFromBase64('UklGRuovAABXRUJQVlA4IN4vAAAwogCdASoYAagAPm0skUYkIiGhLfZdaIANiUAYxwYFYvOdGLyn4A/UY070l…
newspaper : "BusinessLine"
```

Figure 5: Scraped data in the MongoDB database

Module 5

Before adding data to the database, it is important to check whether the same, or similar, articles are already present in the database. Two articles may have semantically similar headlines or visually similar thumbnails. Having such articles in our image-caption dataset is redundant and not useful.

To compare the similarities of the headlines and thumbnails of articles, we use embeddings from OpenAI's CLIP model. Both the headline embedding and the thumbnail embedding (vectors of size 512) of each article are stored in the MongoDB database. If the headline embeddings or thumbnail embeddings of two articles have a cosine similarity above a threshold that is configured in `config.json`, they are deemed duplicates. If one of these articles is already present in the database, the second article is not added to the database.

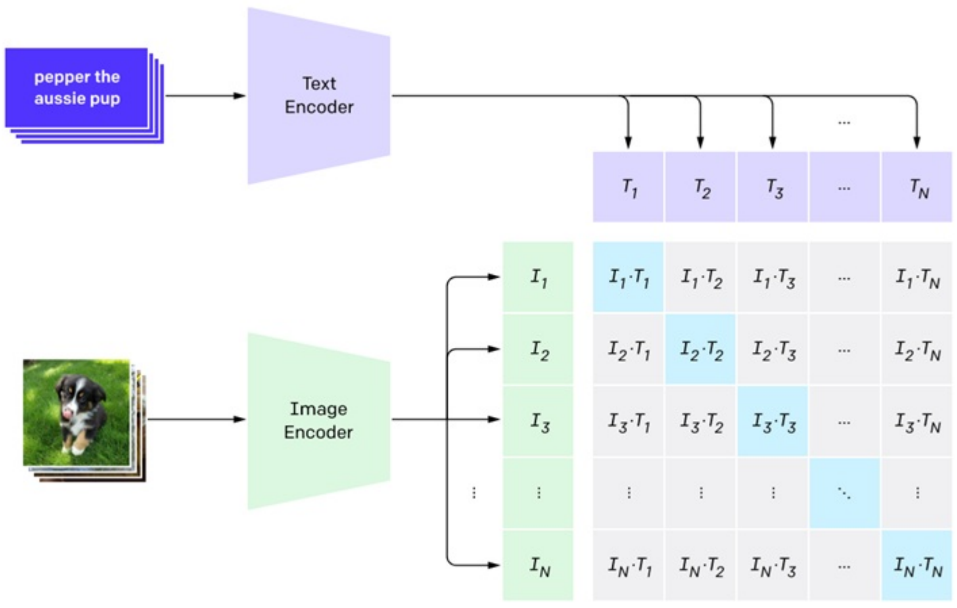

pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Figure 6: CLIP embeddings

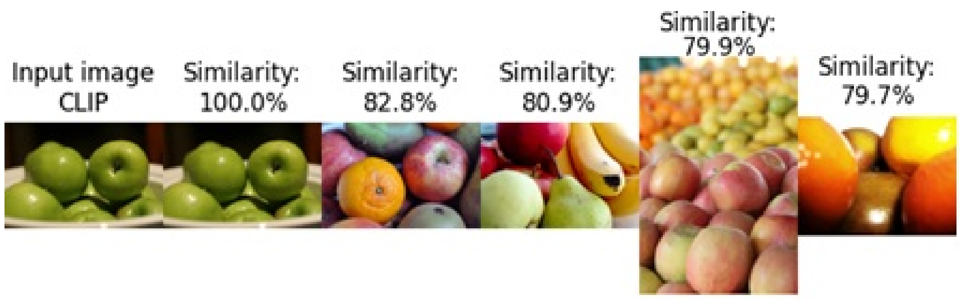


Figure 7: Comparing image similarity using CLIP

Module 6

All modules are orchestrated in a cascaded manner in the `main.py` script.

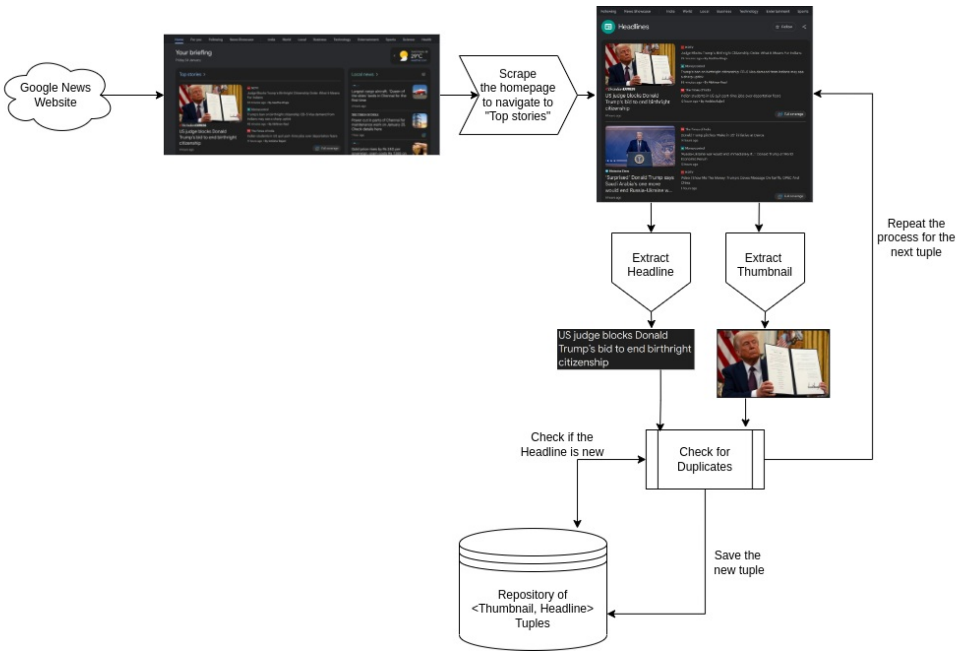
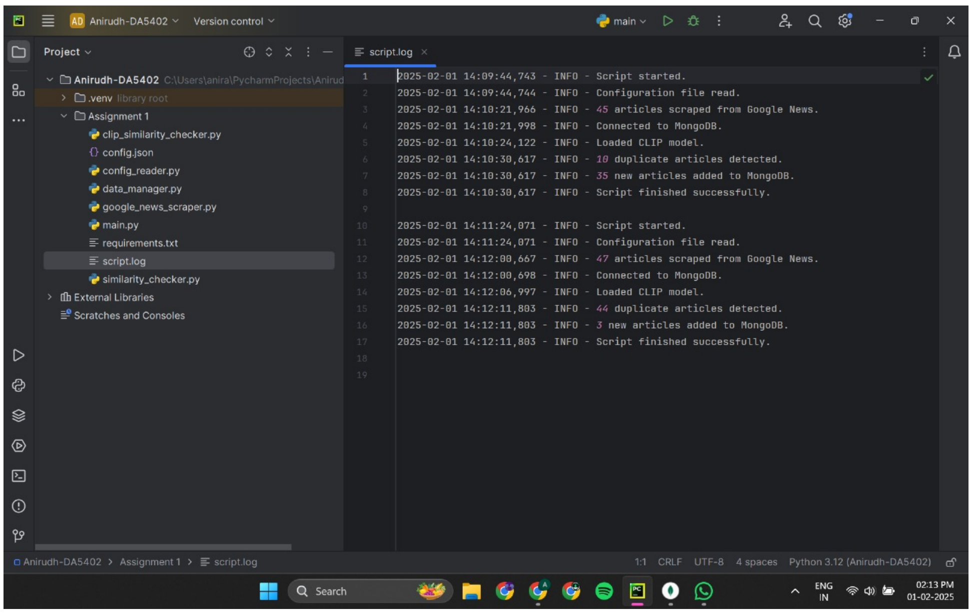

Google News Website
Scrape the homepage to navigate to "Top stories"
Extract Headline
Extract Thumbnail
US judge blocks Donald Trump's bid to end birthright citizenship
Repeat the process for the next tuple
Check if the Headline is new
Check for Duplicates
Save the new tuple
Repository of <Thumbnail, Headline> Tuples

Figure 8: The flow of the script

The `logging` library is used to log time stamps of invocation, error statuses, and other relevant details in the `script.log` file. The script takes about 10 minutes to run completely.



```
2025-02-01 14:09:44,743 - INFO - Script started.
2025-02-01 14:09:44,744 - INFO - Configuration file read.
2025-02-01 14:10:21,966 - INFO - 45 articles scraped from Google News.
2025-02-01 14:10:21,998 - INFO - Connected to MongoDB.
2025-02-01 14:10:24,122 - INFO - Loaded CLIP model.
2025-02-01 14:10:30,617 - INFO - 10 duplicate articles detected.
2025-02-01 14:10:30,617 - INFO - 35 new articles added to MongoDB.
2025-02-01 14:10:30,617 - INFO - Script finished successfully.

2025-02-01 14:11:24,071 - INFO - Script started.
2025-02-01 14:11:24,071 - INFO - Configuration file read.
2025-02-01 14:12:00,667 - INFO - 47 articles scraped from Google News.
2025-02-01 14:12:00,698 - INFO - Connected to MongoDB.
2025-02-01 14:12:06,997 - INFO - Loaded CLIP model.
2025-02-01 14:12:11,803 - INFO - 44 duplicate articles detected.
2025-02-01 14:12:11,803 - INFO - 3 new articles added to MongoDB.
2025-02-01 14:12:11,803 - INFO - Script finished successfully.
```

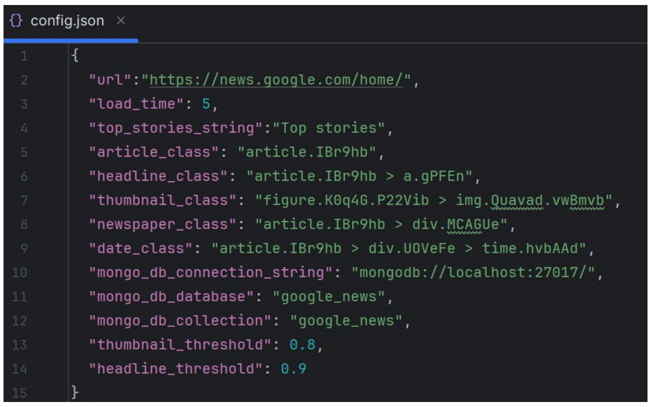
Figure 9: The `script.log` file after running the script twice

A `requirements.txt` file is also provided to install the libraries necessary to run this script.

The script is made friendly to be executed as a CronJob.

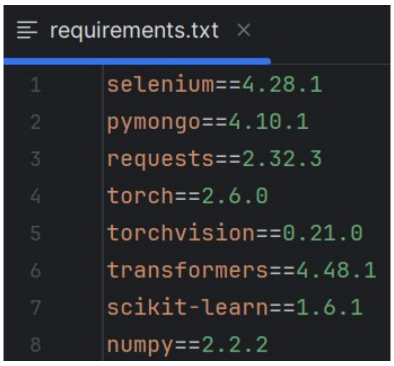
- There are no `print` statements in the script.
- Logging and error handling are performed diligently.
- The parameters are configurable through `config.json`.
- The shebang line is provided at the top of `main.py` for the Python interpreter.

The script can be scheduled as a CronJob using the `crontab` after granting the necessary permissions.



```json
{
  "url":"https://news.google.com/home/",
  "load_time": 5,
  "top_stories_string":"Top stories",
  "article_class": "article.IBr9hb",
  "headline_class": "article.IBr9hb > a.gPFEn",
  "thumbnail_class": "figure.K0q4G.P22Vib > img.Quavad.vwBmvb",
  "newspaper_class": "article.IBr9hb > div.MCAGUe",
  "date_class": "article.IBr9hb > div.UOVeFe > time.hvbAAd",
  "mongo_db_connection_string": "mongodb://localhost:27017/",
  "mongo_db_database": "google_news",
  "mongo_db_collection": "google_news",
  "thumbnail_threshold": 0.8,
  "headline_threshold": 0.9
}
```

Figure 10: The `config.json` file



```
selenium==4.28.1
pymongo==4.10.1
requests==2.32.3
torch==2.6.0
torchvision==0.21.0
transformers==4.48.1
scikit-learn==1.6.1
numpy==2.2.2
```

Figure 11: The `requirements.txt` file