











- **There are 6 Python scripts in this folder**
 - **module_1.py** - Scrapes Google News homepage HTML code.
 - **module_2.py** - Extracts 'Top stories' page link using the HTML code
 - **module_3.py** - Gets thumbnail URLs, captions/titles, article links, date, and time of all the articles on the top stories page at the time of pipeline execution.
 - **module_4.py** - Downloads all thumbnails locally in a 'thumbnails' folder and stores the module_3 data in a local MongoDB database inside the 'articles' collection. The thumbnail URL is replaced with the image file path of that thumbnail before being stored in the database collection.
 - **module_5.py** - It contains a function to check duplicated articles every time the pipeline runs so that the same article is not stored more than once. module_4 uses this function while storing it in the database collection.
 - **module_6.py** - This is the orchestrator file that executes the above modules in a cascaded style. It logs the time stamps of invocation, error statuses, and other relevant details for debugging later into a file called '**orchestrator.log**'.
- **Sample output from the **orchestrator.log** file once module_6.py is executed:**

```
2025-02-02 18:20:25,398 - INFO - Pipeline started
2025-02-02 18:20:25,858 - INFO - Module 1 executed
successfully
2025-02-02 18:20:26,305 - INFO - Module 2 executed
successfully
2025-02-02 18:20:27,600 - INFO - Module 3 executed
successfully
2025-02-02 18:20:30,933 - INFO - Module 4 executed
successfully using Module 5
2025-02-02 18:20:30,933 - INFO - Pipeline completed in 5.53
seconds
```

- This pipeline can be easily set up as a **CronJob** in Linux. I've used **Windows Task Scheduler** to run this pipeline every day at **6.20 pm** (You see down below)

Name	Status	Triggers	Next Run Time	Last Run Time
 NvTmRep_Cr...	Ready	At 12:25 every day	03-02-2025 12:25:24	01-02-2025 12:25:25
 NvTmRep_Cr...	Ready	At 18:25 every day	03-02-2025 18:25:24	02-02-2025 18:25:25
 NvTmRep_Cr...	Ready	At 00:25 every day	03-02-2025 00:25:24	02-02-2025 00:25:25
 NvTmRep_Cr...	Ready	At 06:25 every day	03-02-2025 06:25:24	01-02-2025 10:20:16
 OneDrive Per...	Ready	At 20:00 on 01-05-1992 - After triggered, repeat every 1.00:00:00 indefinitely.	02-02-2025 21:54:25	01-02-2025 21:02:55
 OneDrive Re...	Ready	At 21:03 on 17-10-2022 - After triggered, repeat every 1.00:00:00 indefinitely.	02-02-2025 21:03:40	01-02-2025 21:03:41
 Opera sched...	Ready	Multiple triggers defined	03-02-2025 15:31:24	02-02-2025 15:31:25
 Opera sched...	Ready	Multiple triggers defined	03-02-2025 15:31:43	02-02-2025 15:31:44
 Overwolf Up...	Ready	At 16:41 on 19-12-2022 - After triggered, repeat every 04:00:00 indefinitely.	02-02-2025 20:41:13	02-02-2025 17:04:26
 SecurityScan...	Running	At log on of any user		31-01-2025 12:58:47
 TopStoriesPi...	Ready	At 18:20 every day	03-02-2025 18:20:43	02-02-2025 18:20:43

- A Snippet of my MongoDB compass

MongoDB Compass - Pranav/Assignment_1.articles

Connections Edit View Collection Help

Compass

{ } My Queries

CONNECTIONS (1)

Search connections

▼ Pranav

▼ Assignment_1

articles

admin

config

local

articles

Pranav > Assignment_1 > articles

Documents 45 Aggregations Schema Indexes 1 Validation

Type a query: { field: 'value' } or [Generate query](#)

Explain Reset Find </> Options

ADD DATA EXPORT DATA UPDATE DELETE

100 1 - 45 of 45

articles

	_id ObjectId	title String	link String	date String	time String	image_l
1	ObjectId('679f68f54d96925...	"New Income Tax Slabs Exp...	"https://news.google.com/...	"2025-02-02"	"14:02:00"	"thumbna / / /
2	ObjectId('679f68f54d96925...	"Donald Trump excludes In...	"https://news.google.com/...	"2025-02-02"	"16:47:30"	"thumbna / / /
3	ObjectId('679f68f64d96925...	"Delhi Assembly election ...	"https://news.google.com/...	"2025-02-02"	"10:10:00"	"thumbna / / /
4	ObjectId('679f68f64d96925...	"Complaint filed against ...	"https://news.google.com/...	"2025-02-02"	"09:44:00"	"thumbna / / /
5	ObjectId('679f68f64d96925...	"Union Budget 2025: Keral...	"https://news.google.com/...	"2025-02-01"	"14:10:00"	"thumbna / / /
6	ObjectId('679f68f74d96925...	"'I failed, I will resign...	"https://news.google.com/...	"2025-02-02"	"15:38:00"	"thumbna / / /
7	ObjectId('679f68f74d96925...	"After Bihar budget bonan...	"https://news.google.com/...	"2025-02-02"	"12:16:39"	"thumbna / / /
8	ObjectId('679f68f84d96925...	"What did the Budget say ...	"https://news.google.com/...	"2025-02-02"	"11:19:00"	"thumbna / / /
9	ObjectId('679f68f84d96925...	"EAM Jaishankar feels 'as...	"https://news.google.com/...	"2025-02-02"	"08:45:51"	"thumbna / / /
10	ObjectId('679f68f84d96925...	"Vasant Panchami 2025: Ho...	"https://news.google.com/...	"2025-02-02"	"07:00:07"	"thumbna / / /
11	ObjectId('679f68f94d96925...	"India clinch back-to-bac...	"https://news.google.com/...	"2025-02-02"	"14:52:30"	"thumbna / / /
12	ObjectId('679f68f94d96925...	"Union Budget 2025: Railw...	"https://news.google.com/...	"2025-02-01"	"20:47:00"	"thumbna / / /
13	ObjectId('679f68f94d96925...	"US Army finally reveals ...	"https://news.google.com/...	"2025-02-02"	"05:53:07"	"thumbna / / /
14	ObjectId('679f68fa4d96925...	"How will Budget 2025 imp...	"https://news.google.com/...	"2025-02-01"	"16:12:25"	"thumbna / / /
15	ObjectId('679f68fa4d96925...	"Union Budget 2025 'Are...	"https://news.google.com/...	"2025-02-01"	"20:54:43"	"thumbna / / /
16	ObjectId('679f68fb4d96925...	"Welfare Measures For Gig...	"https://news.google.com/...	"2025-02-01"	"13:51:45"	"thumbna / / /