# DA5402: Machine Learning Operations Laboratory

# Assignment 4

Nandhakishore C S (DA24M011)

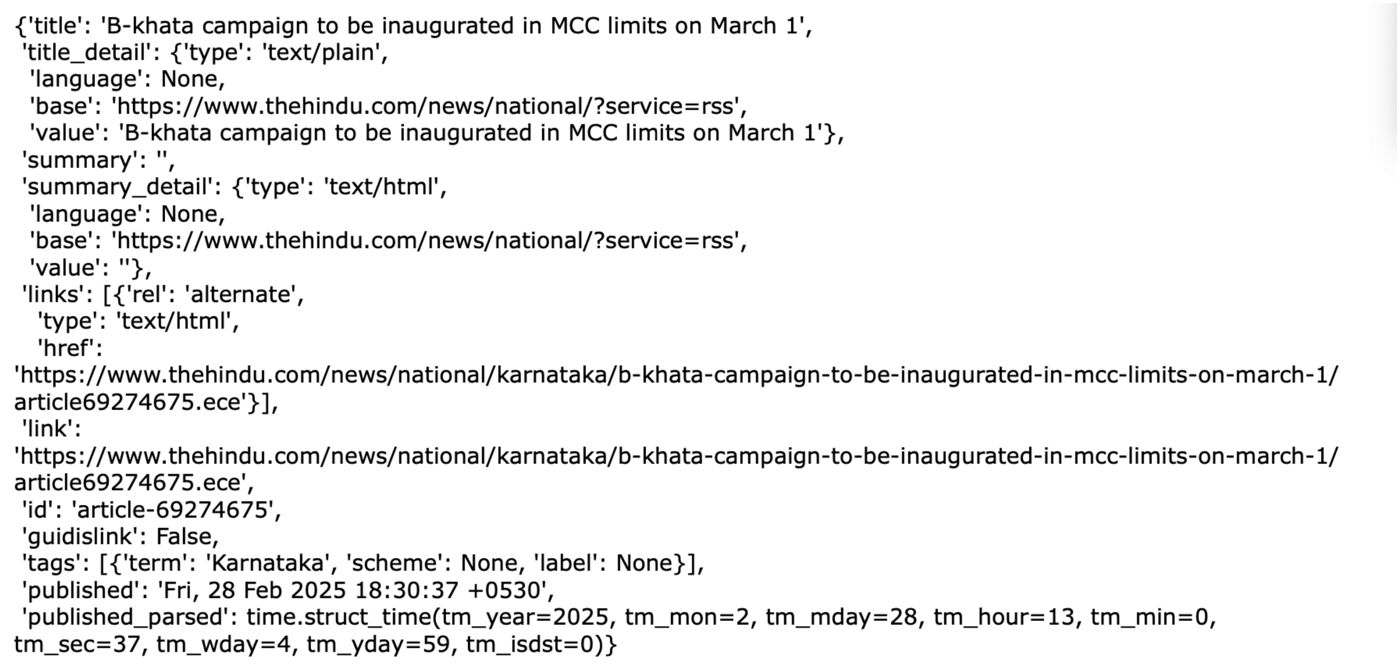
March 14, 2025

## Problem Statement

We have learned about scrapping websites into a dataset of images and captions. Now, let's turn our attention to RSS feeds. RSS feeds are the XML structured data streams that news websites typically expose for easier consumption into a software application.
Python has a 'feedparser' library that can read an RSS feed either from a saved XML file or a web URL. The following code snippet allows one to read the feed directly into a Python dictionary. Here, we have used the RSS feed from The Hindu newspaper.

```python
feed = feedparser.parse('https://www.thehindu.com/news/national/?service=rss')
```

The generated dictionary is organized as entries, where each entry contain the details of a specific news article streamed via the RSS feed. A typical dict structure of an entry without media content.



```
{'title': 'B-khata campaign to be inaugurated in MCC limits on March 1',
 'title_detail': {'type': 'text/plain',
  'language': None,
  'base': 'https://www.thehindu.com/news/national/?service=rss',
  'value': 'B-khata campaign to be inaugurated in MCC limits on March 1'},
 'summary': '',
 'summary_detail': {'type': 'text/html',
  'language': None,
  'base': 'https://www.thehindu.com/news/national/?service=rss',
  'value': ''},
 'links': [{'rel': 'alternate',
   'type': 'text/html',
   'href':
'https://www.thehindu.com/news/national/karnataka/b-khata-campaign-to-be-inaugurated-in-mcc-limits-on-march-1/
article69274675.ece'}],
 'link':
'https://www.thehindu.com/news/national/karnataka/b-khata-campaign-to-be-inaugurated-in-mcc-limits-on-march-1/
article69274675.ece',
 'id': 'article-69274675',
 'guidislink': False,
 'tags': [{'term': 'Karnataka', 'scheme': None, 'label': None}],
 'published': 'Fri, 28 Feb 2025 18:30:37 +0530',
 'published_parsed': time.struct_time(tm_year=2025, tm_mon=2, tm_mday=28, tm_hour=13, tm_min=0,
tm_sec=37, tm_wday=4, tm_yday=59, tm_isdst=0)}
```

The structure of the RSS feed differs from one provider to another. You may have to review the dictionary to suite to your needs. Alright, let's get the business now. The objective is to setup an RSS reader which will automatically fetch and process the required news articles into a list of tuples. We shall push these tuples into a database table of our choice to persistent storage. Beware the the RSS feed is a dynamic stream. Depending on when you read, the data received shall be different. Nonetheless, when you read the stream to often, you may get duplicates as expected. The tuple should have the following information, which eventually get stored in the persistent database table.

1. Title (cannot be blank)
2. Publication Timestamp (to be stored as datetime)
3. Weblink to the article (no need to scrape the page; can't be blank)
4. The picture of the article. (the actual image should be downloaded; can be blank)

5. Tags (can be one or more tag terms)
6. Summary (can be blank)

*** Necessary logs should be created for debugging and information needs.**

## Task 1 [20 points]

Create a docker container of your favorite database. Add scripts to initialize your database with the necessary credentials and create the necessary table(s) to save the data with the above-listed columns. So, when the container is up (during the service creation), the database should get created with the supplied credentials, tables are to be created. The database gets created when the docker container is created for the first time. When the container is restarted, there should be a checker that validates the existence of the database and the required tables. If the validation fails, the db table creation should be redone again. If there is some issue with the initialization, the container should not start up. The required configuration settings from setting up the db should be via environment variables.

A corresponding Dockerfile is created for virtualising the database creating process.

```
FROM postgres:15
ENV POSTGRES_USER=news_admin
ENV POSTGRES_PASSWORD=securepassword
ENV POSTGRES_DB=news_db
COPY init.sql /docker-entrypoint-initdb.d/
```

Figure 1: Dockerfile for Task1

A shell script is used to initiate the data using postgres server with required sql queries.

```
#!/bin/bash
set -e
echo "Checking database initialization..."

PGPASSWORD=$POSTGRES_PASSWORD psql -U $POSTGRES_USER -d $POSTGRES_DB -tc "SELECT 1 FROM pg_database
    WHERE datname='$POSTGRES_DB';" | grep -q 1 || {
    echo "Database $POSTGRES_DB does not exist, initializing..."
    psql -U $POSTGRES_USER -d postgres -c "CREATE DATABASE $POSTGRES_DB;"
}

TABLE_EXISTS=$(PGPASSWORD=$POSTGRES_PASSWORD psql -U $POSTGRES_USER -d $POSTGRES_DB -tc "SELECT
    to_regclass('public.news_articles');")

if [[ "$TABLE_EXISTS" != "news_articles" ]]; then
    echo "Table does not exist. Creating..."
    psql -U $POSTGRES_USER -d $POSTGRES_DB -f /docker-entrypoint-initdb.d/init.sql
else
    echo "Table exists. Skipping initialization."
fi

echo "Database is ready."
exec docker-entrypoint.sh postgres
```

Figure 2: Shell script to initialise database

A set of SQL queries are send to create a table with the necessary aatributes.

```
CREATE TABLE IF NOT EXISTS news_articles (
    id SERIAL PRIMARY KEY,
    headline TEXT NOT NULL,
    url TEXT NOT NULL UNIQUE,
    published_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP,
    thumbnail BYTEA,  -- Binary storage for images
    summary TEXT
);
```

Figure 3: SQL script for creating tables

## Task 2 [20 points]

Build an application (a python runnable script here) that would fetch the news articles from the configured RSS feed and push the required fields into the containerized database. The URL of the RSS feed, the dictionary paths of the required fields are to be configurable via environment variables. This way, we can configure the application to fetch from any RSS feed from source. For now, we are hanging on to The Hindu. But the expectation is to make it work for others such as, TOI, Indian Express, NDTV, India Today, News18, etc. Dockerize the application. The application should poll the RSS feed for updates every 10 mins (configurable) and fetch the feed when the data has changed. Mind that data change does not mean no-duplicates.
For creating the database in a docker, the psycopg library is used with required credentials (db_name, password, port number, etc). There is a user defined function which creates the database and checks for existing database when the docker is run again. A.env file for configuring the docker file with database credentials is created to make the process easy.

```
# RSS Feed Configuration
RSS_FEED_URL=https://www.thehindu.com/news/national/feeder/default.rss
FETCH_INTERVAL=600  # Time interval in seconds

# PostgreSQL Configuration
POSTGRES_HOST=database  # Matches Docker service name
POSTGRES_PORT=5432
POSTGRES_DB=news_db
POSTGRES_USER=news_admin
POSTGRES_PASSWORD=securepassword
```

Figure 4: Database credentials for Task1

A separate Dockerfile is created to virtualise task2 (fetching RSS separately)

```
FROM python:3.10
WORKDIR /app
COPY fetch_news.py requirements.txt ./
RUN pip install --no-cache-dir -r requirements.txt
CMD ["python", "fetch_news.py"]
```

Figure 5: DockerFile for task2

The project directory is structured as follows:

## Task 3 [10 points]

We got two containerized applications. Use docker-compose to create a multi-application (services) dockerization such that the containerized RSS reader app could push the data into the containerized DB. Upon docker-compose up, both services should be created and applications should be started for RSS reading and db storage. Docker-compose down should stop services and remove them. Likewise, docker-compose start and stop should have the necessary functionalities.

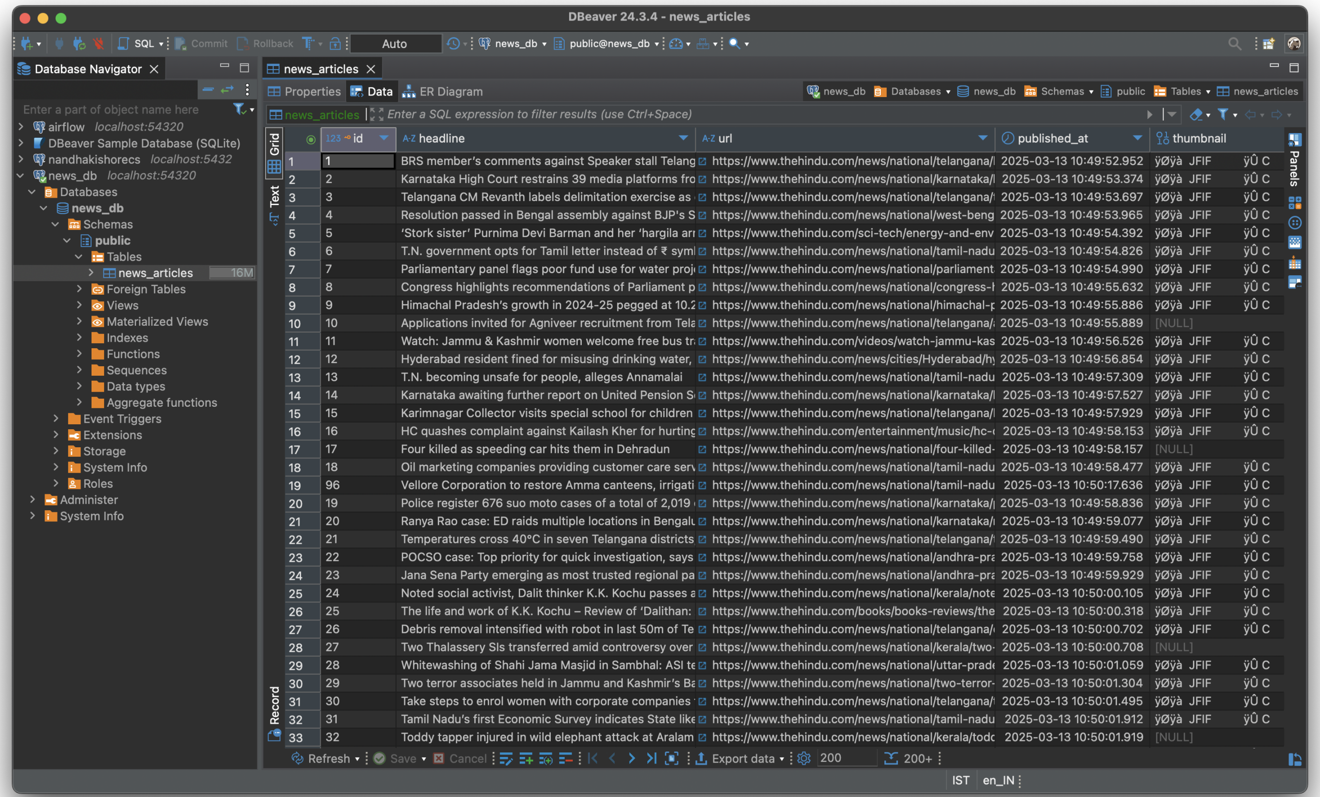
For multi application services, a common docker-compose file is created and both the docker container's details are read. This file initiates the docker and creates two containers and make them work in tandem by passing the information from one to other.

## Brownie [10 points]

Create a simple containerized web application to read the data from the database with suitable date filters (default is today's news). The web application should show the Title, Image  Summary from the database for every record in the table. Additionally, upon clicking the news title or image, you should open another tab with the actual news using the URL that you got from the db table. The web application should also become a part of the docker-compose environment, which will now have three services. Up, down, start and stop commands of docker-compose should have the necessary functionalities.
A simple RESTAPI based web application is created in python with a standard HTML, CSS layer.

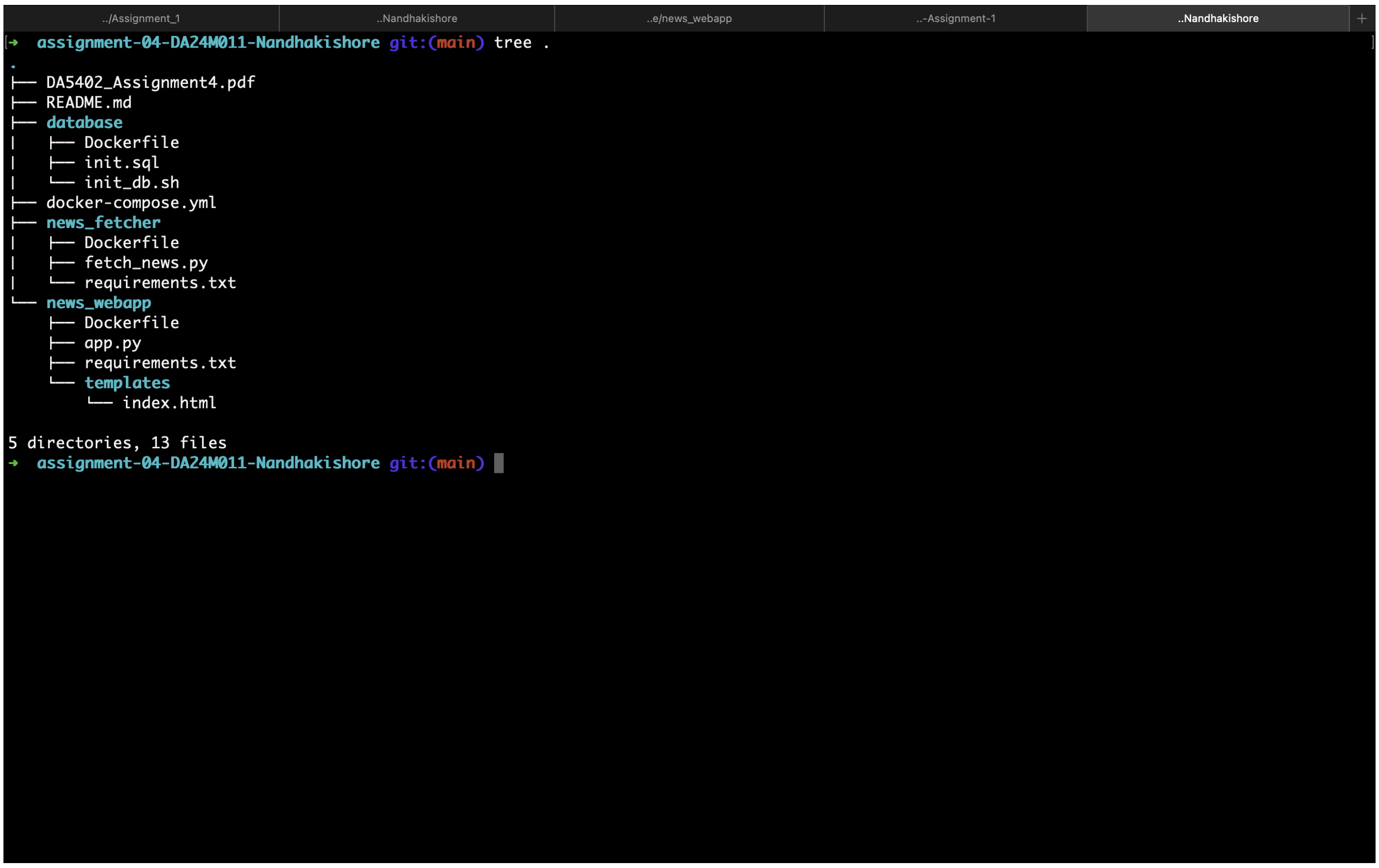
```python
# Load environment variables
RSS_FEED_URL = os.getenv("RSS_FEED_URL", "https://www.thehindu.com/news/national/feeder/default.rss")
FETCH_INTERVAL = int(os.getenv("FETCH_INTERVAL", 600))  # Default: 10 minutes
DB_CONFIG = {
    "dbname": os.getenv("POSTGRES_DB", "news_db"),
    "user": os.getenv("POSTGRES_USER", "news_admin"),
    "password": os.getenv("POSTGRES_PASSWORD", "securepassword"),
    "host": os.getenv("POSTGRES_HOST", "database"),  # Updated to match service name
    "port": os.getenv("POSTGRES_PORT", "5432"),
}

LAST_FEED_CONTENT = None  # Stores last feed content for change detection

def connect_db():
    """Connect to PostgreSQL database."""
    return psycopg2.connect(**DB_CONFIG)

def fetch_rss_feed():
    """Fetch and parse the RSS feed."""
    global LAST_FEED_CONTENT
    print(f"Fetching RSS feed from {RSS_FEED_URL}...")
    feed = feedparser.parse(RSS_FEED_URL)

    if not feed.entries:
        print("No new entries found.")
        return []

    feed_content = [(entry.title, entry.link) for entry in feed.entries]

    # Check if content has changed
    if feed_content == LAST_FEED_CONTENT:
        print("Feed has not changed since the last fetch. Skipping...")
        return []

    LAST_FEED_CONTENT = feed_content  # Update last fetched content
    return feed.entries

def process_and_store_news(entries):
    """Process and store news articles in PostgreSQL."""
    conn = connect_db()
    cur = conn.cursor()

    for entry in entries:
        headline = entry.title
        url = entry.link
        summary = entry.get("summary", "")

        # Fetch the first image (if available)
        thumbnail_data = None
        if "media_content" in entry:
            image_url = entry.media_content[0]['url']
            try:
                response = requests.get(image_url)
                if response.status_code == 200:
                    img = Image.open(BytesIO(response.content))
                    img = img.convert("RGB")
                    img_io = BytesIO()
                    img.save(img_io, format="JPEG")
                    thumbnail_data = img_io.getvalue()
            except Exception as e:
                print(f"Could not fetch image: {e}")

        try:
            cur.execute("""
                INSERT INTO news_articles (headline, url, summary, thumbnail)
                VALUES (%s, %s, %s, %s)
                ON CONFLICT (url) DO NOTHING;
            """, (headline, url, summary, psycopg2.Binary(thumbnail_data) if thumbnail_data else None)
    )

            conn.commit()
            print(f"Inserted: {headline}")
        except psycopg2.IntegrityError:
            conn.rollback()
            print(f"Skipping duplicate: {headline}")

    cur.close()
    conn.close()
```

Figure 6: Code for creating database and fetching RSS for task2



|  | id | headline | url | published_at | thumbnail |
|---|---|---|---|---|---|
| 1 | 1 | BRS member's comments against Speaker stall Telang | https://www.thehindu.com/news/national/telangana/l | 2025-03-13 10:49:52.952 | ÿØÿà JFIF ÿÛ C |
| 2 | 2 | Karnataka High Court restrains 39 media platforms fro | https://www.thehindu.com/news/national/karnataka/l | 2025-03-13 10:49:53.374 | ÿØÿà JFIF ÿÛ C |
| 3 | 3 | Telangana CM Revanth labels delimitation exercise as | https://www.thehindu.com/news/national/telangana/t | 2025-03-13 10:49:53.697 | ÿØÿà JFIF ÿÛ C |
| 4 | 4 | Resolution passed in Bengal assembly against BJP's S | https://www.thehindu.com/news/national/west-beng | 2025-03-13 10:49:53.965 | ÿØÿà JFIF ÿÛ C |
| 5 | 5 | 'Stork sister' Purnima Devi Barman and her 'hargila arı | https://www.thehindu.com/sci-tech/energy-and-env | 2025-03-13 10:49:54.392 | ÿØÿà JFIF ÿÛ C |
| 6 | 6 | T.N. government opts for Tamil letter instead of ₹ syml | https://www.thehindu.com/news/national/tamil-nadu | 2025-03-13 10:49:54.826 | ÿØÿà JFIF ÿÛ C |
| 7 | 7 | Parliamentary panel flags poor fund use for water proj | https://www.thehindu.com/news/national/parliament | 2025-03-13 10:49:54.990 | ÿØÿà JFIF ÿÛ C |
| 8 | 8 | Congress highlights recommendations of Parliament p | https://www.thehindu.com/news/national/congress-l | 2025-03-13 10:49:55.632 | ÿØÿà JFIF ÿÛ C |
| 9 | 9 | Himachal Pradesh's growth in 2024-25 pegged at 10.2 | https://www.thehindu.com/news/national/himachal-p | 2025-03-13 10:49:55.886 | ÿØÿà JFIF ÿÛ C |
| 10 | 10 | Applications invited for Agniveer recruitment from Tela | https://www.thehindu.com/news/national/telangana/ | 2025-03-13 10:49:55.889 | [NULL] |
| 11 | 11 | Watch: Jammu & Kashmir women welcome free bus tr | https://www.thehindu.com/videos/watch-jammu-kas | 2025-03-13 10:49:56.526 | ÿØÿà JFIF ÿÛ C |
| 12 | 12 | Hyderabad resident fined for misusing drinking water, | https://www.thehindu.com/news/cities/Hyderabad/hy | 2025-03-13 10:49:56.854 | ÿØÿà JFIF ÿÛ C |
| 13 | 13 | T.N. becoming unsafe for people, alleges Annamalai | https://www.thehindu.com/news/national/tamil-nadu | 2025-03-13 10:49:57.309 | ÿØÿà JFIF ÿÛ C |
| 14 | 14 | Karnataka awaiting further report on United Pension S | https://www.thehindu.com/news/national/karnataka/l | 2025-03-13 10:49:57.527 | ÿØÿà JFIF ÿÛ C |
| 15 | 15 | Karimnagar Collector visits special school for children | https://www.thehindu.com/news/national/telangana/l | 2025-03-13 10:49:57.929 | ÿØÿà JFIF ÿÛ C |
| 16 | 16 | HC quashes complaint against Kailash Kher for hurting | https://www.thehindu.com/entertainment/music/hc-c | 2025-03-13 10:49:58.153 | ÿØÿà JFIF ÿÛ C |
| 17 | 17 | Four killed as speeding car hits them in Dehradun | https://www.thehindu.com/news/national/four-killed- | 2025-03-13 10:49:58.157 | [NULL] |
| 18 | 18 | Oil marketing companies providing customer care serv | https://www.thehindu.com/news/national/tamil-nadu | 2025-03-13 10:49:58.477 | ÿØÿà JFIF ÿÛ C |
| 19 | 96 | Vellore Corporation to restore Amma canteens, irrigati | https://www.thehindu.com/news/national/tamil-nadu | 2025-03-13 10:50:17.636 | ÿØÿà JFIF ÿÛ C |
| 20 | 19 | Police register 676 suo moto cases of a total of 2,019 | https://www.thehindu.com/news/national/karnataka/ | 2025-03-13 10:49:58.836 | ÿØÿà JFIF ÿÛ C |
| 21 | 20 | Ranya Rao case: ED raids multiple locations in Bengalu | https://www.thehindu.com/news/national/karnataka/ | 2025-03-13 10:49:59.077 | ÿØÿà JFIF ÿÛ C |
| 22 | 21 | Temperatures cross 40°C in seven Telangana districts | https://www.thehindu.com/news/national/telangana/t | 2025-03-13 10:49:59.490 | ÿØÿà JFIF ÿÛ C |
| 23 | 22 | POCSO case: Top priority for quick investigation, says | https://www.thehindu.com/news/national/andhra-pra | 2025-03-13 10:49:59.758 | ÿØÿà JFIF ÿÛ C |
| 24 | 23 | Jana Sena Party emerging as most trusted regional pa | https://www.thehindu.com/news/national/andhra-pra | 2025-03-13 10:49:59.929 | ÿØÿà JFIF ÿÛ C |
| 25 | 24 | Noted social activist, Dalit thinker K.K. Kochu passes a | https://www.thehindu.com/news/national/kerala/note | 2025-03-13 10:50:00.105 | ÿØÿà JFIF ÿÛ C |
| 26 | 25 | The life and work of K.K. Kochu – Review of 'Dalithan: | https://www.thehindu.com/books/books-reviews/the | 2025-03-13 10:50:00.318 | ÿØÿà JFIF ÿÛ C |
| 27 | 26 | Debris removal intensified with robot in last 50m of Te | https://www.thehindu.com/news/national/telangana/ | 2025-03-13 10:50:00.702 | ÿØÿà JFIF ÿÛ C |
| 28 | 27 | Two Thalassery SIs transferred amid controversy over | https://www.thehindu.com/news/national/kerala/two- | 2025-03-13 10:50:00.708 | [NULL] |
| 29 | 28 | Whitewashing of Shahi Jama Masjid in Sambhal: ASI te | https://www.thehindu.com/news/national/uttar-prade | 2025-03-13 10:50:01.059 | ÿØÿà JFIF ÿÛ C |
| 30 | 29 | Two terror associates held in Jammu and Kashmir's Ba | https://www.thehindu.com/news/national/two-terror- | 2025-03-13 10:50:01.304 | ÿØÿà JFIF ÿÛ C |
| 31 | 30 | Take steps to enrol women with corporate companies | https://www.thehindu.com/news/national/telangana/t | 2025-03-13 10:50:01.495 | ÿØÿà JFIF ÿÛ C |
| 32 | 31 | Tamil Nadu's first Economic Survey indicates State like | https://www.thehindu.com/news/national/tamil-nadu | 2025-03-13 10:50:01.912 | ÿØÿà JFIF ÿÛ C |
| 33 | 32 | Toddy tapper injured in wild elephant attack at Aralam | https://www.thehindu.com/news/national/kerala/todc | 2025-03-13 10:50:01.919 | [NULL] |

Figure 7: Database viewed with Dbeaver



```
→ assignment-04-DA24M011-Nandhakishore git:(main) tree .
.
├── DA5402_Assignment4.pdf
├── README.md
├── database
|   ├── Dockerfile
|   ├── init.sql
|   └── init_db.sh
├── docker-compose.yml
├── news_fetcher
|   ├── Dockerfile
|   ├── fetch_news.py
|   └── requirements.txt
└── news_webapp
    ├── Dockerfile
    ├── app.py
    ├── requirements.txt
    └── templates
        └── index.html

5 directories, 13 files
→ assignment-04-DA24M011-Nandhakishore git:(main)
```

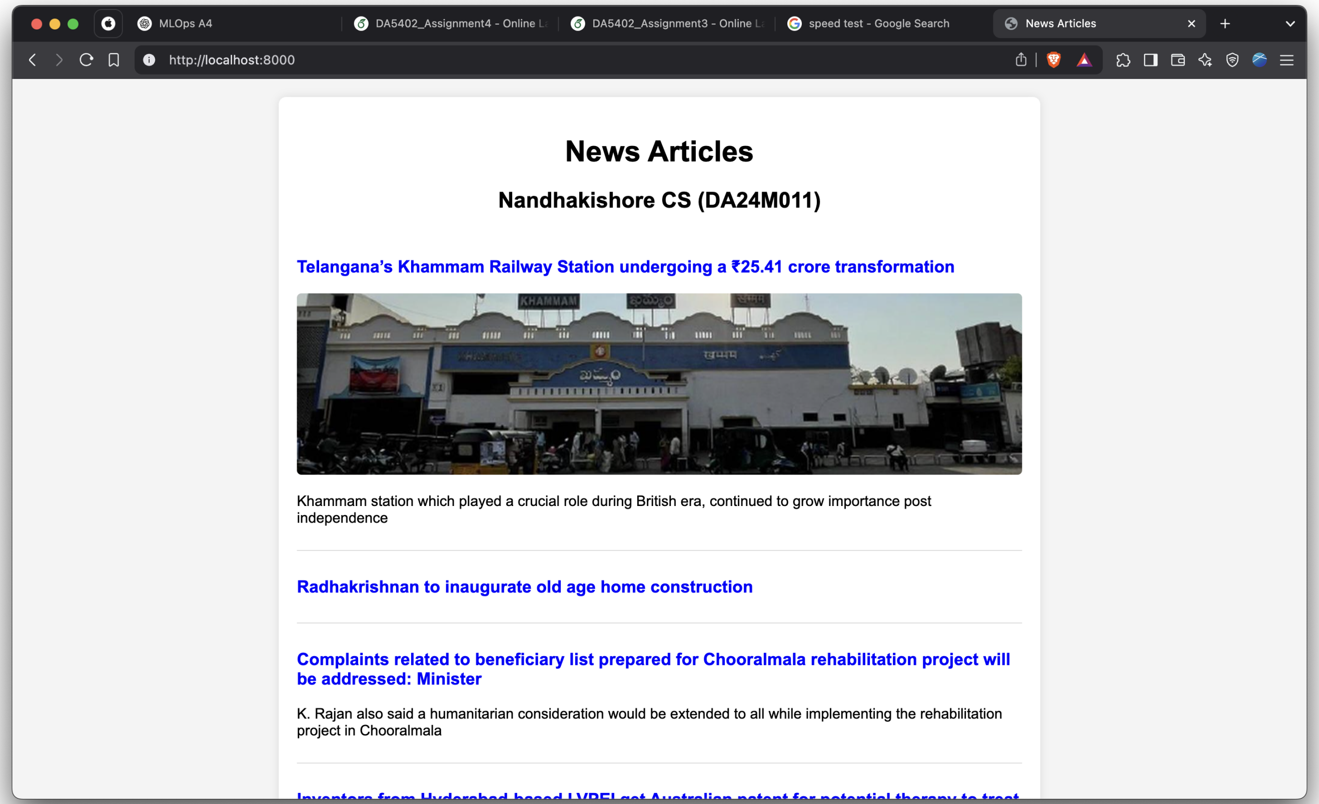
Figure 8: Directory structure for assignment 4

```yaml
services:
  database:
    build: ./database
    container_name: news_postgres
    restart: unless-stopped
    environment:
      POSTGRES_USER: news_admin
      POSTGRES_PASSWORD: securepassword
      POSTGRES_DB: news_db
    ports:
      - "54320:5432"
    volumes:
      - pgdata:/var/lib/postgresql/data

  news_fetcher:
    build: ./news_fetcher
    container_name: news_fetcher
    restart: unless-stopped
    env_file:
      - ./news_fetcher/.env
    depends_on:
      - database

  news_webapp:
    build: ./news_webapp
    container_name: news_webapp
    restart: unless-stopped
    ports:
      - "8000:8000"
    env_file:
      - ./news_webapp/.env
    depends_on:
      - database

volumes:
  pgdata:
```

Figure 9: common docker-compose.yaml file

```python
from fastapi import FastAPI, Request, Query
from fastapi.templating import Jinja2Templates
import psycopg2
import os
import base64
from datetime import date
from dotenv import load_dotenv

# Load environment variables
load_dotenv()

app = FastAPI()
templates = Jinja2Templates(directory="templates")

# Database connection
def get_db_connection():
    return psycopg2.connect(
        dbname=os.getenv("POSTGRES_DB"),
        user=os.getenv("POSTGRES_USER"),
        password=os.getenv("POSTGRES_PASSWORD"),
        host=os.getenv("POSTGRES_HOST"),
        port=os.getenv("POSTGRES_PORT")
    )

@app.get("/")
def home(request: Request):
    conn = get_db_connection()
    cur = conn.cursor()
    cur.execute("""
        SELECT headline, url, summary, thumbnail
        FROM news_articles
        WHERE date(published_at) = CURRENT_DATE
        ORDER BY published_at DESC
    """)
    articles = cur.fetchall()
    conn.close()

    # Format articles properly
    formatted_articles = []
    for article in articles:
        headline, url, summary, thumbnail = article
        if isinstance(thumbnail, memoryview):  # Convert binary images to Base64
            thumbnail = f"data:image/png;base64,{base64.b64encode(thumbnail.tobytes()).decode()}"
        formatted_articles.append((headline, url, summary, thumbnail))

    # Pass your name and roll number to the template
    return templates.TemplateResponse("index.html", {
        "request": request,
        "articles": formatted_articles,
        "name": "Nandhakishore CS",  # Replace with your actual name
        "roll_number": "DA24M011"    # Replace with your actual roll number
    })
```

Figure 10: RESTAPI based webapp


MLOps A4
DA5402_Assignment4 - Online L
DA5402_Assignment3 - Online L
speed test - Google Search
News Articles
http://localhost:8000

News Articles

Nandhakishore CS (DA24M011)

Telangana's Khammam Railway Station undergoing a ₹25.41 crore transformation

Khammam station which played a crucial role during British era, continued to grow importance post independence

Radhakrishnan to inaugurate old age home construction

Complaints related to beneficiary list prepared for Chooralmala rehabilitation project will be addressed: Minister

K. Rajan also said a humanitarian consideration would be extended to all while implementing the rehabilitation project in Chooralmala

Inventors from Hyderabad-based LVPEI get Australian patent for potential therapy to treat

Figure 11: Screenshot of the webapp