DA5402: Assignment 4

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.

```
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)}
```


A typical dict structure of an entry without media content.


```
{'title': 'Several colleges, research institutions celebrate National Science Day in Tiruchi',
 'title_detail': {'type': 'text/plain',
  'language': None,
  'base': 'https://www.thehindu.com/news/national/?service=rss',
  'value': 'Several colleges, research institutions celebrate National Science Day in Tiruchi'},
 'summary': 'Workshops, science exhibitions mark the occasion in many colleges; Union Minister Mansukh Mandaviya
inaugurates the programme via video conferencing at Shrimati Indira Gandhi College',
 'summary_detail': {'type': 'text/html',
  'language': None,
  'base': 'https://www.thehindu.com/news/national/?service=rss',
  'value': 'Workshops, science exhibitions mark the occasion in many colleges; Union Minister Mansukh Mandaviya
inaugurates the programme via video conferencing at Shrimati Indira Gandhi College'},
 'links': [{'rel': 'alternate',
   'type': 'text/html',
   'href':
'https://www.thehindu.com/news/cities/Tiruchirapalli/several-colleges-research-institutions-celebrate-national-science-
day-in-tiruchi/article69275084.ece'}],
 'link':
'https://www.thehindu.com/news/cities/Tiruchirapalli/several-colleges-research-institutions-celebrate-national-science-
day-in-tiruchi/article69275084.ece',
 'id': 'article-69275084',
 'guidislink': False,
 'tags': [{'term': 'Tiruchirapalli', 'scheme': None, 'label': None}],
```

'published': 'Fri, 28 Feb 2025 19:00:30 +0530',
'published_parsed': time.struct_time(tm_year=2025, tm_mon=2, tm_mday=28, tm_hour=13, tm_min=30, tm_sec=30, tm_wday=4, tm_yday=59, tm_isdst=0),
'media_content': [{'height': '675',
'medium': 'image',
'url':
'https://th-i.thgim.com/public/incoming/xm5svw/article69275230.ece/alternates/LANDSCAPE_1200/10295_28_2_2025_18_36_22_2_NRCB.JPG',
'width': '1200'}]}

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.

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 https://rss.feedspot.com/indian_news_rss_feeds/. 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.

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.

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.