**Internship Assignment: RSS Feed Parser**

**OBJECTIVES:**

**1. Feed Parser and Data Extraction:**

* Parsing RSS Feeds: Used the *feedparser* library available in Python to parse each feed from the given URLs and extract relevant information such as title, content, publication date, and source URL.
* Handling Duplicates: To avoid storing duplicate articles from the same feed, stored a set of existing URLs which check for duplicates before insertion.

**2. Database Storage:**

* Database: Used a MySQL database through the mysql-connector library to store the extracted news article data. The schema includes fields for title, content, publication date, source URL, and category.
* Insertion Logic: Articles are inserted into the database using a Python function insert\_news\_article. Before insertion, the function checks for duplicates based on the source URL and inserts only if the article is not already present.

3. Task Queue and News Processing:

* Celery: Although the objective mentioned setting up a Celery queue, I was not able to implement it due to a lack of prior knowledge. But I'm learning about it now and will update the code accordingly.
* Article Processing Logic: Developed a function to classify each article into predefined categories using NLTK and updated the database with the assigned category for each article.

4. Logging and Error Handling:

* Logging Configuration: Logging is implemented using Python's logging module. Logs are stored in a file named rss\_feed.log with timestamps, log levels, and messages.