News Article Categorization Application

# Task Overview

The objective of this project is to collect news articles from various RSS feeds, store them in a PostgreSQL database, and categorize each article into predefined categories using Natural Language Processing (NLP). The four main categories for classification are:  
1. Terrorism / Protest / Political Unrest / Riot  
2. Positive/Uplifting  
3. Natural Disasters  
4. Others  
The project also utilizes Celery to handle task scheduling and asynchronous article processing, ensuring efficient and scalable data ingestion.

# Tech Stack

- Programming Language: Python  
- RSS Feed Parsing: Feedparser  
- Database: PostgreSQL with SQLAlchemy ORM  
- Task Queue: Celery with Redis as the message broker  
- NLP (Text Classification): spaCy  
- Logging: Python's logging library  
- Other Libraries:   
 - datetime for date and time handling  
 - sqlalchemy for database operations  
 - celery for background task processing

# Code Explanation

## 1. RSS Feed Fetching

The fetch\_rss\_feed() function utilizes the feedparser library to parse the list of RSS feeds.  
It extracts the title, content, publication date, and source URL from each article entry.  
It handles errors gracefully, logging them for further analysis.

## 2. Database Setup

The SQLAlchemy ORM is used to model and interact with the PostgreSQL database.  
The NewsArticle class represents the news articles in the database, with fields like id, title, content, published\_date, source\_url, and category.  
The store\_article() function checks for duplicates and stores new articles in the database.

## 3. NLP-Based Categorization

The classify\_text() function uses the spaCy NLP model to categorize articles based on their content.  
Simple keyword-based classification is employed for this task, mapping content containing terms like 'protest,' 'riot,' or 'earthquake' to predefined categories.

## 4. Celery Task Queue

The project uses Celery for task scheduling. The process\_articles() function is a Celery task that fetches RSS feed articles asynchronously, classifies them, and stores them in the database.  
Redis is configured as the message broker for Celery.

## 5. Logging and Error Handling

The logging library is used throughout the application to capture important events and errors during feed parsing, article storage, and task execution.

# Code Execution

To execute this project, follow these steps:

## 1. Clone the Repository

git clone https://github.com/yourusername/news-article-categorization.git  
cd news-article-categorization

## 2. Install Dependencies

pip install -r requirements.txt

## 3. Set Up PostgreSQL

Install and start PostgreSQL.  
Create a database named news\_db and update the DATABASE\_URL in the script with your credentials:  
DATABASE\_URL = "postgresql://username:password@localhost:5432/news\_db"

## 4. Run Database Migrations

python initialize\_db.py

## 5. Set Up Redis and Celery

Install and start Redis on your system:  
redis-server  
  
Start the Celery worker to process articles:  
celery -A app worker --loglevel=info

## 6. Run the Application

python app.py  
  
This will start fetching news articles from the RSS feeds, classify them using NLP, and store them in the database under appropriate categories.