Web Scraping News from Different Countries using RSS Feeds

Title: Web Scraping and Data Extraction

Selection competitive Key creterias: Selected intern would have demonstrated competitive edge making the code as comprehensive, generic and complete as time permits.

Objective

The purpose of this assignment is to evaluate candidates' skills in **web scraping**, **data extraction**, and **automation** using **RSS feeds**. By the end of this assignment, candidate will:

- demonstrate how to extract structured news data from different countries.
- Work with RSS feed parsing in Python.
- Store and process the extracted data.
- Handle common web scraping challenges such as encoding, rate limits, and missing data.

Task Overview

You will develop a **Python program** to scrape news headlines and summaries from **at least twenty different countries** (the more you do, the better) using publicly available RSS feeds from major news agencies. The extracted data should be **stored in a structured format** (CSV, JSON, or a database). **Extract historical data as well for atleast past one year.**

Requirements

1. Data Collection

- Identify and select **at least twenty** different **news RSS feeds** from different countries (e.g., BBC for the UK, CNN for the US, Al Jazeera for the Middle East, NHK for Japan, India, Singapore, Malaysia, Indonesia, Korea, China, USA and other countries popular news papers – find the list of them all first). **Extract historical data as well for atleast past one year. If you want to include other countries, feel free. Think of all the different news papers, articiles in a country and try to structure it as comprehensive as possible.**

- Find as much as information as you can but aim for at least the below fields from each RSS feed:
  - **News Title**
  - **Publication Date**
  - **Source (News Agency)**
  - **Country**
  - **Summary/Description**
  - **News URL (Link to Full Article)**

2. Data Storage & Formatting

- Store the extracted data in **CSV** or **JSON** format. Or any other consistent format.
- Ensure proper encoding to handle non-English characters (UTF-8).
- Remove duplicate or irrelevant entries.

3. Code Implementation

- Use Python with **BeautifulSoup** or **feedparser** to parse RSS feeds. Or feel free to use others if you so wish.
- Implement error handling (e.g., handling missing fields, network failures).
- Modularize code using **functions** and **classes** where appropriate.
- Add comments explaining each section of the code.
- Feel free to use chatgpt and Claude ai or any other LLM.

4. Advanced Features

You demonstrate your enhanced skills by implementing any of the following:

- Store data in **SQLite/PostgreSQL** database or any other format if you so wish, instead of CSV/JSON.
- Create **a simple Flask/Django API** to serve the news data. Or any other way if you so wish.
- Add a **language detection** feature to categorize news by language.
- Implement **scheduling (cron jobs)** to update the feed automatically.

Technical Guidelines

1. **Programming Language:** Python
2. **Required Libraries:**
   - `feedparser` (for RSS parsing)
   - `BeautifulSoup4` (for HTML parsing, if needed)
   - `pandas` (for data storage in CSV/JSON)
   - `sqlite3` (for optional database storage)
   - `requests` (for handling network requests)
   - any other libraries of your choice
3. **Development Environment:** Jupyter Notebook / VS Code / PyCharm
4. **Submission Format:**
   - Python script (`.py` file) or Jupyter Notebook (`.ipynb`).
   - Sample output file (CSV/JSON).
   - A short **README.md** explaining how to run the script.

Example RSS Feeds

Here are some example RSS feeds you can use:

- **BBC News (UK):** `http://feeds.bbci.co.uk/news/rss.xml`
- **CNN (US):** `http://rss.cnn.com/rss/edition.rss`
- **Al Jazeera (Middle East):** `https://www.aljazeera.com/xml/rss/all.xml`
- **NHK (Japan):** `https://www3.nhk.or.jp/rss/news/cat0.xml`

Submission Instructions

1. Upload your **Python script or Jupyter Notebook** to your **GitHub**.
2. Attach your **CSV/JSON output file**.
3. Include a **README.md** explaining:
   - How to install dependencies.
   - How to run the script.
   - Any issues encountered.
4. If you implemented **bonus features**, highlight them in your README.

Example of summary information

Country	News Agency	Total articles downloaded	Total historical data
Country 1	A, B, C	1m	Since 2021
2			
3			
..			
n			