AI601-Data Engineering for AI Systems

Assignment 1: Project Data Mosaic

Due Date: 11:59 PM 16th February

Overview & Scenario

You have been hired by the **Data Mosaic Initiative**, an organization aiming to gather **multi-faceted insights** on emerging topics. Your mission is to **collect data from multiple sources of different types (structured, unstructured, semi-structured)**, store or upload it to a repository, and produce a short report that demonstrates your pipeline design and addresses theoretical considerations.

Data Sources to Integrate:

1. **Reddit** (via praw API)
2. **Yahoo Finance API** (using yfinance)
3. **Data dumps** (Kaggle, Government Websites, Open Data Portals, etc.)

Part 1: Choose Your Topic

Pick **one** of the following themes to focus your data collection. Your pipeline should revolve around gathering data related to this theme from each source.

1. **Green Energy**
    - Reddit discussions around renewable energy.
    - Public datasets on global power consumption.
    - Financial trends for “energy,” etc.
2. **Remote Work**
    - Reddit discussions on r/RemoteWork or r/WorkFromHome.
    - Public datasets on employment trends or labor statistics.
    - Financial trends for “technology” etc.
3. **Electric Vehicles (EVs)**

- Reddit discussions in r/ElectricVehicles or r/TeslaMotors.
- Public dataset on vehicle registrations or alternative fuels.
- Financial trends for “energy,” etc.

4. **Telehealth / Online Healthcare**
   - Reddit posts on r/Telemedicine or r/AskDocs.
   - Public healthcare-related datasets (hospital admissions, telemedicine usage if available).
   - Financial trends for “healthcare” etc.
5. **Cryptocurrencies**
   - Reddit communities like r/CryptoCurrency or r/Bitcoin.
   - Public crypto datasets (on-chain data, trading volumes).
   - Financial trends for “financial services” etc.
6. **Sports**
   - Reddit discussions on cricket/football match outcomes
   - Public datasets on sports statistics (e.g., MLB, soccer)
   - Financial trends for “consumer cyclical” etc.
7. **Public Sentiment on Upcoming Elections**
   - Reddit posts in political subreddits tracking candidate mentions
   - Public datasets on voter turnout or election results
   - Financial trends for “communication services” etc.

Part 2: Data Collection Requirements

1. **Reddit**
   - Collect a small dataset (e.g., ~100–200 posts or comments) containing your keywords (e.g., “electric vehicle,” “remote work,” etc.).
   - Use an *official API* (Reddit’s praw) or minimal scraping with caution, respecting Reddit’s TOS.
   - Fields to include: title, post text, author, date, upvotes, subreddit name.
   - Use csv library to write to a CSV file.
2. **Public Datasets**
   - Find at least **one** relevant public dataset
   - Export the query results as a CSV or JSON file.
3. **Yahoo Finance API**
   - Use yfinance to extract data from the finance.yahoo.com site on the sector of your choice.
   - Define a list of major companies belonging to your selected sector or ETFs (e.g `"XLE", "CVX", etc.`)
   - Fetch at least 2 years of data. Keep the 'Close' position of the stock only [Hint: use Ticker class to fetch the stock and use history method to get the previous data].
   - Use pandas library to process the data.

Part 3: Technical Deliverables

1. **Data Collection Scripts**
   - **Reddit**: Python script retrieving data via praw or an equivalent approach.
   - **Yahoo Finance**: Script using yfinance.
   - For public data, include the link to datasets and approach in the report. If you use any programmatic way to gather data, submit that script as well.
   - Note: Please follow the structure of code we used in lab1 i.e. dividing the code into functions (fetch_data, save_to_csv, clean, summarize).
2. **Dataset Storage**
   - Save raw data files (CSV, JSON) in a structured folder (e.g., /datasets/raw/reddit_posts.csv, /datasets/raw/yfinance.csv, etc.).
3. **Pipeline Diagram**
   - A simple flowchart that **illustrates** your multi-source pipeline:
     - Input: (APIs, Public Data, Yahoo Finance)
     - Processing: (scripts for each)
     - Output: (structured CSV/JSON)
4. **GitHub Submission**
   - Create a private repository containing below folders and share the link in the PDF report:
     - **Report**: assignment1/report.pdf
     - **Code**: 2 scripts (assignment1/scripts/reddit.py, assignment1/scripts/yfinance.py), neatly labeled.
     - **Datasets**: At least 3 datasets, one for each source (assignment1/datasets/reddit_posts.csv, etc.) Up to a feasible size—if too large, provide a subset or instructions to regenerate.
     - Add TAs as collaborators in your github repository.
   - Please upload **ONLY** the PDF file to LMS. It should have the github link.

Part 4: Reporting & Theoretical Questions

1. Write your group number, student ids, and summarize contributions of both students in the report.
2. **Overview of Your Topic**: Why did you choose it? What data do you expect to see?
3. **Data Collection Process**: Summarize the steps you took for each source and any challenges (API rate limits, incomplete data, TOS constraints).
4. **Initial Observations**: Generate a summary of the datasets using pandas. Add the screenshot of the console output of pandas DataFrame in the document.
5. What AI product will you make using this data?
6. Which terms of service constraints or privacy issues might arise when collecting data from Reddit and Yahoo? Consider limitations on storing or redistributing user-generated content.
7. How does collecting from multiple sources help or hinder data quality? What conflicts or discrepancies might you face?
8. Can you think of ways to store and combine all of this data?
9. (Optional) Provide at least one table or chart per dataset. Any format is okay. For instance:

- **Reddit**: A word frequency chart or average upvotes over time.
- **Public Data**: Basic descriptive stats (count, mean, min, max of relevant fields).
- **Yahoo Finance**: A line chart of interest over time for your keywords.