Assignment 1: Project Data Mosaic

Due Date: 11:59 PM 15th 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. **Google Search Trends** (using [pytrends](#))
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.
   ○ Google Trends for "solar power," "wind energy," etc.
2. **Remote Work**
   ○ Reddit discussions on r/RemoteWork or r/WorkFromHome.
   ○ Public datasets on employment trends or labor statistics.
   ○ Google Trends for "remote jobs," "hybrid work," etc.
3. **Electric Vehicles (EVs)**

- ○ Reddit discussions in r/ElectricVehicles or r/TeslaMotors.
- ○ Public dataset on vehicle registrations or alternative fuels.
- ○ Google Trends for "electric vehicle," "Tesla," "EV charging."

4. **Telehealth / Online Healthcare**
   - ○ Reddit posts on r/Telemedicine or r/AskDocs.
   - ○ Public healthcare-related datasets (hospital admissions, telemedicine usage if available).
   - ○ Google Trends for "telehealth," "telemedicine," "online doctor."

5. **Cryptocurrencies**
   - ○ Reddit communities like r/CryptoCurrency or r/Bitcoin.
   - ○ Public crypto datasets (on-chain data, trading volumes).
   - ○ Google Trends for "Bitcoin price," "crypto regulation," etc.

6. **Sports**
   - ○ Reddit discussions on cricket/football match outcomes
   - ○ Public datasets on sports statistics (e.g., MLB, soccer)
   - ○ Google Trends for "india vs pak," "champions league," etc.

7. **Public Sentiment on Upcoming Elections**
   - ○ Reddit posts in political subreddits tracking candidate mentions
   - ○ Public datasets on voter turnout or election results
   - ○ Google Trends for "early voting," "candidate debates," 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. **Google Search Trends**
   - ○ Use pytrends to extract data from the Google Trends site on the topic of your choice.
   - ○ Collect at least **6–12 months** of interest data for a set of **2–3 keywords** related to your topic.
   - ○ Fields to include: keyword, date/time, interest score, and (optionally) region if you are doing a region-based analysis.
   - ○ Use pandas library to write to a CSV file.

Part 3: Technical Deliverables

1. **Data Collection Scripts**
   - **Reddit**: Python script retrieving data via praw or an equivalent approach.
   - **Google Trends**: Script using pytrends.
   - 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/pytrends.csv, etc.).
3. **Pipeline Diagram**
   - A simple flowchart that **illustrates** your multi-source pipeline:
     - Input: (APIs, Public Data, Google Trends)
     - 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/google_trends.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 Google? 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:

AI601-Data Engineering for AI Systems

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