AI601-Data Engineering for AI Systems

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

LINK TO REPO https://github.com/Heuscartist/ai601-assignments/tree/main/assignment1

1. Write your group number, student ids, and summarize contributions of both students in the report.

Ans 1.

24280007: Code for Reddit and Kaggle Dataset

24280006: Code for Wikipedia and yahoo finance data

2. **Overview of Your Topic**: Why did you choose it? What data do you expect to see?

Ans 2.

Our topic was Electric Vehicles. We chose this because this is something we were personally interested in. We expected to see reviews on different electric vehicles, comparisons between non-electric and EVs, and also information on the different electric vehicle makes and models available in the market.

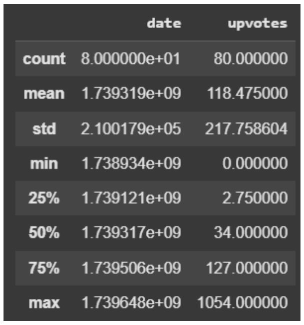
3. **Data Collection Process**: Summarize the steps you took for each source and any challenges (API rate limits, incomplete data, TOS constraints).

Ans 3.

We first looked into the api itself and its usage from its documentation. Some APIs couldn't be used directly and needed API keys. As for challenges did not face any issues for the particular topic we choose.

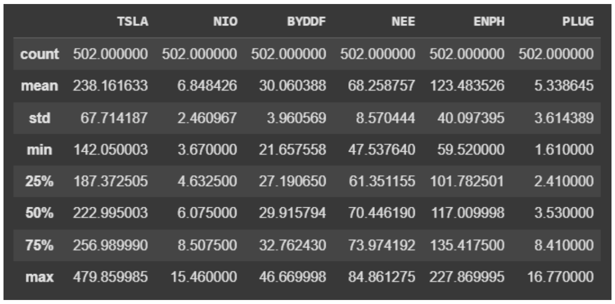
4. **Initial Observations**: Generate a summary of the datasets using pandas. Add the screenshot of the console output of pandas DataFrame in the document.

   Ans 4.



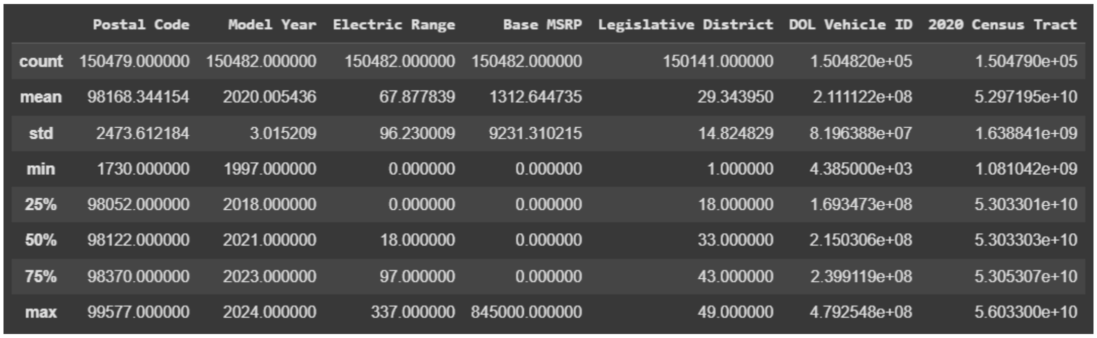
	date	upvotes
count	8.000000e+01	80.000000
mean	1.739319e+09	118.475000
std	2.100179e+05	217.758604
min	1.738934e+09	0.000000
25%	1.739121e+09	2.750000
50%	1.739317e+09	34.000000
75%	1.739506e+09	127.000000
max	1.739648e+09	1054.000000

Reddit Data



	TSLA	NIO	BYDDF	NEE	ENPH	PLUG
count	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000
mean	238.161633	6.848426	30.060388	68.258757	123.483526	5.338645
std	67.714187	2.460967	3.960569	8.570444	40.097395	3.614389
min	142.050003	3.670000	21.657558	47.537640	59.520000	1.610000
25%	187.372505	4.632500	27.190650	61.351155	101.782501	2.410000
50%	222.995003	6.075000	29.915794	70.446190	117.009998	3.530000
75%	256.989990	8.507500	32.762430	73.974192	135.417500	8.410000
max	479.859985	15.460000	46.669998	84.861275	227.869995	16.770000

Yahoo Finance Data



	Postal Code	Model Year	Electric Range	Base MSRP	Legislative District	DOL Vehicle ID	2020 Census Tract
count	150479.000000	150482.000000	150482.000000	150482.000000	150141.000000	1.504820e+05	1.504790e+05
mean	98168.344154	2020.005436	67.877839	1312.644735	29.343950	2.111122e+08	5.297195e+10
std	2473.612184	3.015209	96.230009	9231.310215	14.824829	8.196388e+07	1.638841e+09
min	1730.000000	1997.000000	0.000000	0.000000	1.000000	4.385000e+03	1.081042e+09
25%	98052.000000	2018.000000	0.000000	0.000000	18.000000	1.693473e+08	5.303301e+10
50%	98122.000000	2021.000000	18.000000	0.000000	33.000000	2.150306e+08	5.303303e+10
75%	98370.000000	2023.000000	97.000000	0.000000	43.000000	2.399119e+08	5.305307e+10
max	99577.000000	2024.000000	337.000000	845000.000000	49.000000	4.792548e+08	5.603300e+10

Kaggle Data

5. What AI product will you make using this data?

   Ans 5.

   There could be a stock forecasting model that would predict whether a stock would go up or not depending on the trending topics on reddit and the stock information from yahoo finance.

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.

   Ans 6.

   Privacy issues may include people's personal information in the post such as name, address, numbers that might be in the post that would need processing. There could also be some sort of biased posts that might affect our data quality. Then there could be some user generated content which we dont have a right to share within the posts as well. All of these need to be considered.

7. How does collecting from multiple sources help or hinder data quality? What conflicts or discrepancies might you face?

   Ans 7.

   This depends on the use case. In case we want to build something like the finance prediction model, multiple sources will help by giving us more insight into the latest trend and user activity / interest regarding a particular topic. This could improve our model in its forecasting. On the other hand this might cause problems in trying to store and process data in different formats to make it usable. Different data formats and mix of structured and unstructured data would be more difficult to process compared to getting data from a single source.

8. Can you think of ways to store and combine all of this data?

Ans 8.

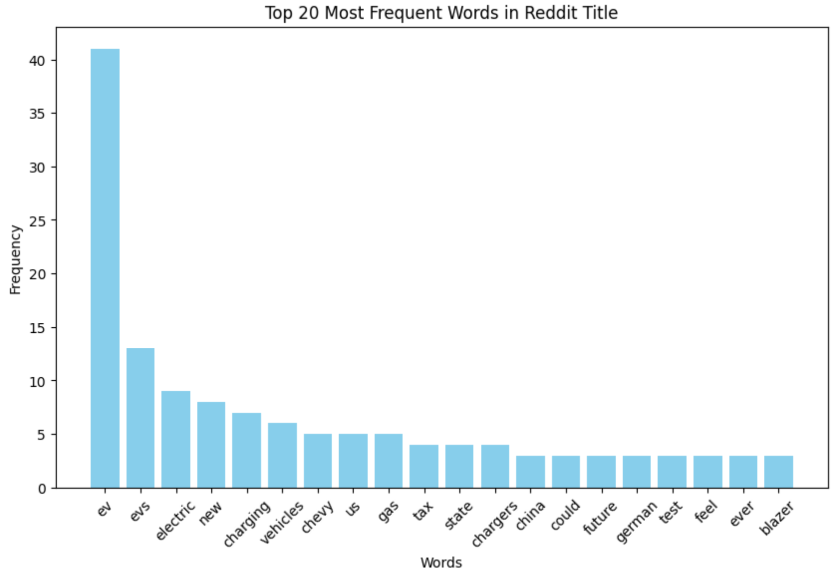
The data can be combined. For example we can append data from the reddit posts such as upvotes and user semantics into the daily financial trend dataset to get more information on whether a stock price would go up or down. This can then be stored into a single structured format like a csv file.

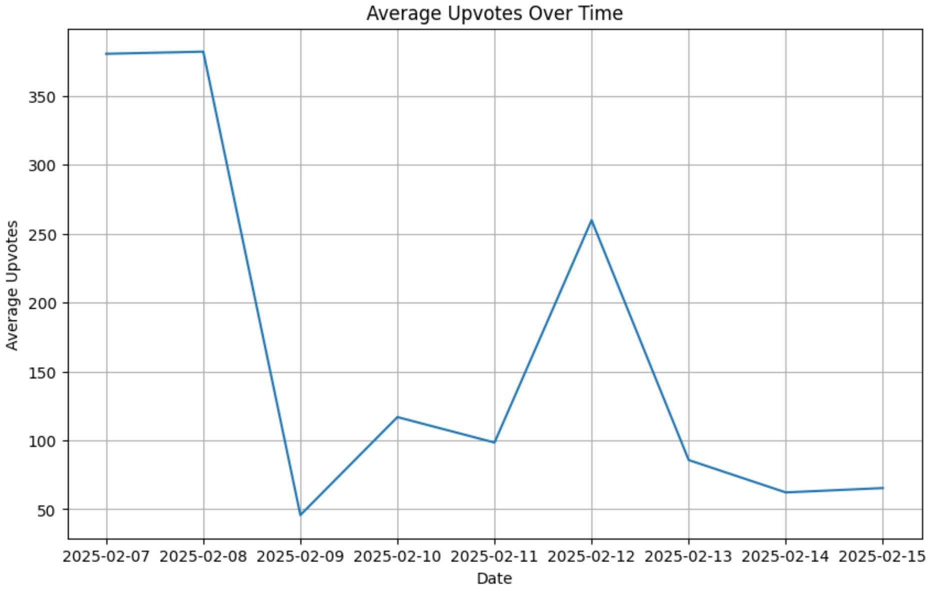
9. (Optional) Provide at least one table or chart per dataset. Any format is okay. For instance:

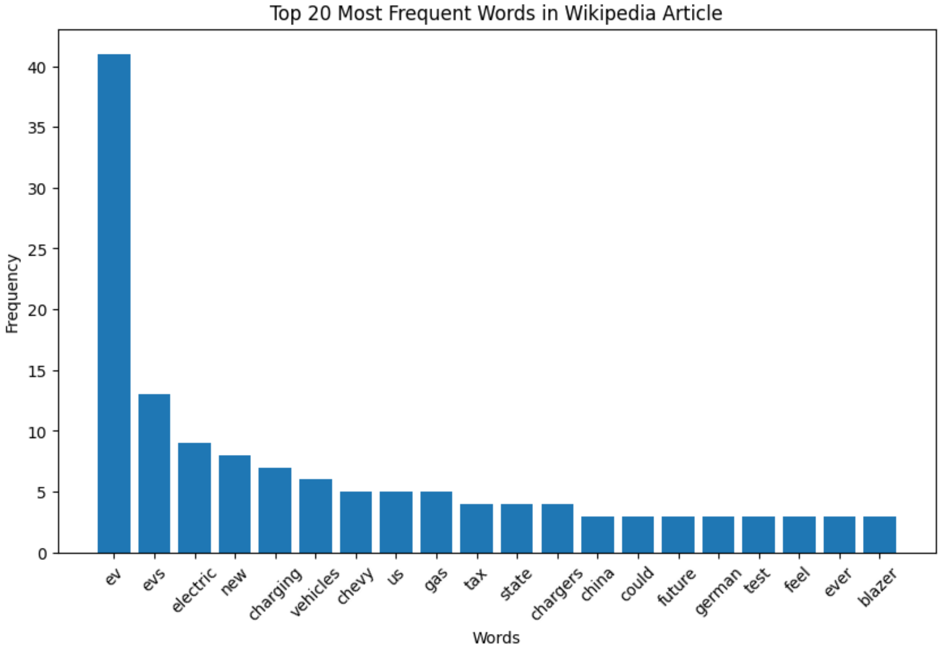
○ **Reddit**: A word frequency chart or average upvotes over time.

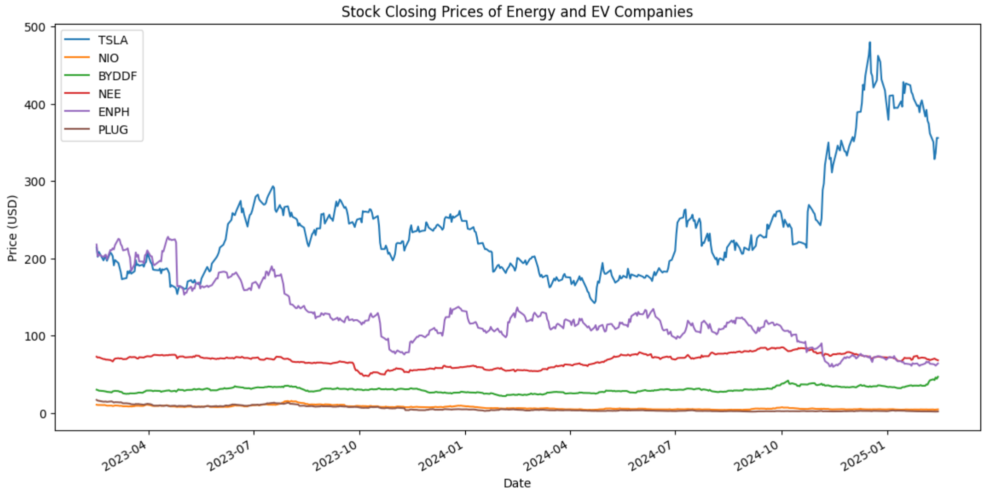
○ **Public Data**: Basic descriptive stats (count, mean, min, max of relevant fields).

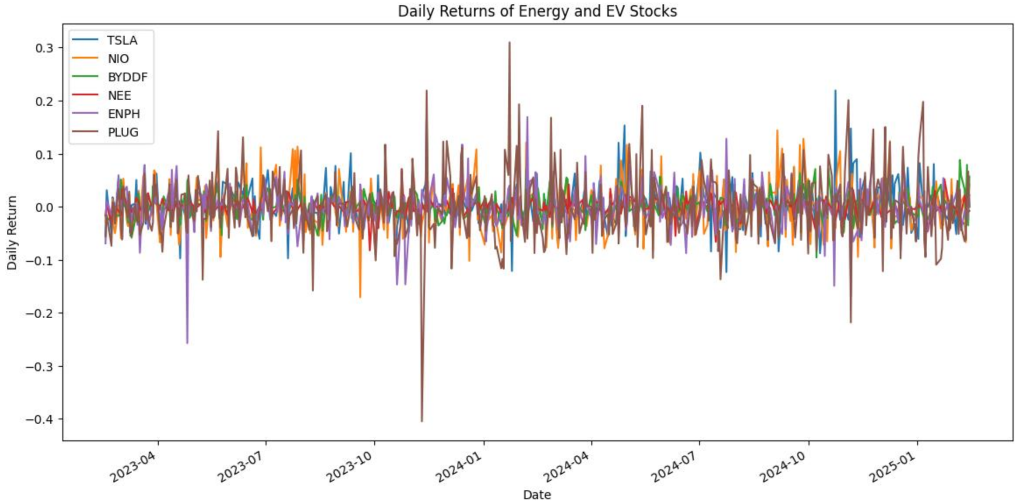
○ **Google Trends**: A line chart of interest over time for your keywords.




Average Upvotes Over Time
Average Upvotes
350
300
250
200
150
100
50
2025-02-07
2025-02-08
2025-02-09
2025-02-10
2025-02-11
2025-02-12
2025-02-13
2025-02-14
2025-02-15
Date


Top 20 Most Frequent Words in Wikipedia Article
Frequency
40
35
30
25
20
15
10
5
0
ev
evs
electric
new
charging
vehicles
chevy
us
gas
tax
state
chargers
china
could
future
german
test
feel
ever
blazer
Words


Stock Closing Prices of Energy and EV Companies
TSLA
NIO
BYDDF
NEE
ENPH
PLUG
Price (USD)
0
100
200
300
400
500
2023-04
2023-07
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
Date


Daily Returns of Energy and EV Stocks
TSLA
NIO
BYDDF
NEE
ENPH
PLUG
Daily Return
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
2023-04
2023-07
2023-10
2024-01
2024-04
2024-07
2024-10
2025-01
Date