DATA COLLECTION AND INTEGRATION REPORT

AI601-Data Engineering for AI Systems

Ahmed Hannan
Hammad Javed

Contents

Contributions

- **Student 1 (ID: 23030026):**
  - Configured and executed data collection from Reddit using PRAW.
  - Developed the pipeline for scraping posts and comments, ensuring proper field extraction.
  - Collected and processed the public dataset using the **Electric Vehicle Charging Dataset** from open data sources.
  - Developed the Google Trends extraction process using `pytrends`.
  - Contributed to the discussion on API challenges and ethical considerations.
  - Generated initial dataset summaries and contributed to the report.
- **Student 2 (ID: 23030010):**
  - Collected and processed the public dataset using the **Electric Vehicle Charging Dataset** from open data sources.
  - Developed the Google Trends extraction process using `pytrends`.
  - Generated initial dataset summaries and contributed to the report on data integration and privacy issues.

1 Overview of the Topic

Topic: Electric Vehicles (EVs)

Why EVs?
We chose Electric Vehicles because they are at the forefront of sustainable transportation, combining technological innovation with environmental responsibility. The EV market is rapidly expanding, and analyzing related data can provide valuable insights into market trends, user sentiment, and charging infrastructure developments.

Expected Data:

- **Reddit:**
  User discussions and opinions on EVs from subreddits such as `r/TeslaMotors` and `r/ElectricVehicles`. Collected fields include post title, text, author, timestamp, upvotes, and subreddit name.

- **Public Dataset:**
  The **Electric Vehicle Charging Dataset** provides structured data on charging station usage, installation locations, and other metrics related to EV charging infrastructure.
- **Google Trends:**
  Time-series data over 12 months for keywords like “electric vehicle,” “tesla,” and “ev charging,” including date/time, interest score, and regional interest data.

2 Data Collection Process

2.1 Reddit Collection

Method:
We used Reddit’s official API (`PRAW`) to search for posts in targeted subreddits using EV-related keywords.

Steps:

1. Configured API credentials (`client_id`, `client_secret`, `user_agent`).
2. Queried subreddits such as `r/ElectricVehicles` and `r/TeslaMotors` with keywords like “electric vehicle,” “tesla,” and “ev charging.”
3. Extracted fields including title, post text, author, date (timestamp), upvotes, and subreddit name.
4. Saved the results into a CSV file using Python’s `csv` library.

Challenges:

- Managing API rate limits and asynchronous operation warnings.
- Adhering to Reddit’s Terms of Service regarding user-generated content.

2.2 Public Dataset Collection

Dataset: Electric Vehicle Charging Dataset

Method:

Downloaded the Electric Vehicle Charging Dataset from a public open data

portal or Kaggle using the Kaggle API.

Steps:

1. Configured Kaggle API credentials by uploading the `kaggle.json` file.
2. Downloaded the dataset via the Kaggle command-line interface.
3. Unzipped the dataset and converted it into a CSV file for further analysis.

Challenges:

- Ensuring the dataset is current and accurately represents EV charging infrastructure.
- Handling potential missing data or format discrepancies in the downloaded dataset.

2.3 Google Trends Collection

Method:

Employed the `pytrends` library to fetch public interest data over a 12-month period.
Steps:

1. Configured `pytrends` with relevant parameters.
2. Collected interest-over-time data and, optionally, regional interest data for keywords such as “electric vehicle,” “tesla,” and “ev charging.”
3. Transformed the data to include fields for date/time, keyword, and interest score (and region for the regional analysis).
4. Saved the resulting data into CSV files.

Challenges:

- Limited granularity of trends data and ensuring consistent formatting.
- Converting aggregated time-series data into a format suitable for detailed analysis.

3 Future AI Product

AI-Powered EV Insights Platform:

We plan to develop a platform that leverages AI to provide comprehensive insights into the Electric Vehicle market. The platform will:

- Analyze sentiment from Reddit data to understand public perception of EVs and related charging infrastructure.
- Combine insights from the Electric Vehicle Charging Dataset to gauge the real-world availability and usage of charging stations.
- Integrate Google Trends data to monitor public interest and predict market trends.
- Offer dashboards and predictive analytics to assist stakeholders in making informed decisions regarding EV adoption and infrastructure investment.

4 Terms of Service Constraints and Privacy Issues

Reddit:

- **API Constraints:** Reddit's API usage limits and guidelines restrict the frequency and volume of requests.
- **Privacy Considerations:** Redistribution of raw user-generated content may violate privacy policies unless properly anonymized.

Google Trends:

- **Data Usage:** Google Trends data is aggregated and anonymized; however, its commercial use may be subject to specific licensing or usage guidelines.

General Considerations:

- **Data Redistribution:** Verify that the usage and redistribution of data conform to the terms set by each data source.

5 Data Integration: Benefits and Challenges

Benefits:

- **Holistic View:** Integrating data from Reddit, public EV charging data, and Google Trends provides multiple perspectives, enabling richer analysis.
- **Cross-Validation:** Multiple sources allow us to validate findings, e.g., comparing public sentiment with actual charging usage statistics and search interest trends.

Challenges:

- **Data Quality and Consistency:** Variations in data collection methods, formats, and update frequencies can lead to inconsistencies.
- **Conflicting Signals:** Social media sentiment may not always align with quantitative data (charging usage or search trends), requiring careful analysis to reconcile differences.
- **Technical Integration:** Merging structured and unstructured data requires robust ETL pipelines and may necessitate additional data cleaning steps.

6 Data Storage and Integration Strategy

Data Storage:

- **Database Solutions:** Store structured data in relational databases (e.g., PostgreSQL) and unstructured data in NoSQL databases (e.g., MongoDB) or cloud storage services.
- **Cloud Storage:** Utilize services like AWS S3 or Google Cloud Storage to maintain raw data files (CSV, JSON) and enable scalable access.

Data Integration:

- **ETL Pipelines:** Develop Extract, Transform, Load (ETL) pipelines using tools like Apache Airflow or custom Python scripts to regularly update and merge data.
- **Data Warehouse:** Consider using a data warehouse to combine and analyze datasets from different sources efficiently.
- **APIs:** Implement internal APIs to allow real-time access to integrated data for downstream applications or dashboards.