Part 4: Reporting & Theoretical Questions

- Group ID 19
- Student ID 24280051, 24280057

Member Contribution

Topic Selection and Coding: 24280051
Code Review and Report: 24280057

Topic Overview

Our ecosystem is experiencing severe climate changes that negatively impact us. Traditional energy production methods come with drawbacks such as pollution and environmental damage. As a result, the world is shifting toward modern energy sources like solar and wind, which help preserve the environment by reducing pollution and other harmful effects. Following are the expected insights:

1. **World View**

   What are people discussing and debating about green energy? Are there concerns about the cost and effectiveness of green energy adoption?

2. **Global Power Data**

   Renewable vs. non-renewable energy usage over time, regional variations.

3. **Google Trends**

   Seasonal spikes in interest, geographic variations in search behavior.

Data Collection Process

The assignment involved collecting data from three different sources. The first data source was Reddit. The Python package to scrap this data source was Praw which was easily installed in Pycharm virtual environment without any dependency conflict. Furthermore, we did not face API rate limits in our case. One thing was that there was no content for some posts only titles and images. So in the case of performing any statistical analysis or modeling such data points won't contribute much. The second source was Kaggle from where the .csv file of the data was directly downloaded. Thirdly, for Google Search Trends, we did face an API limit issue for which it is included to attempt scrapping five times in the script.

Initial Observations

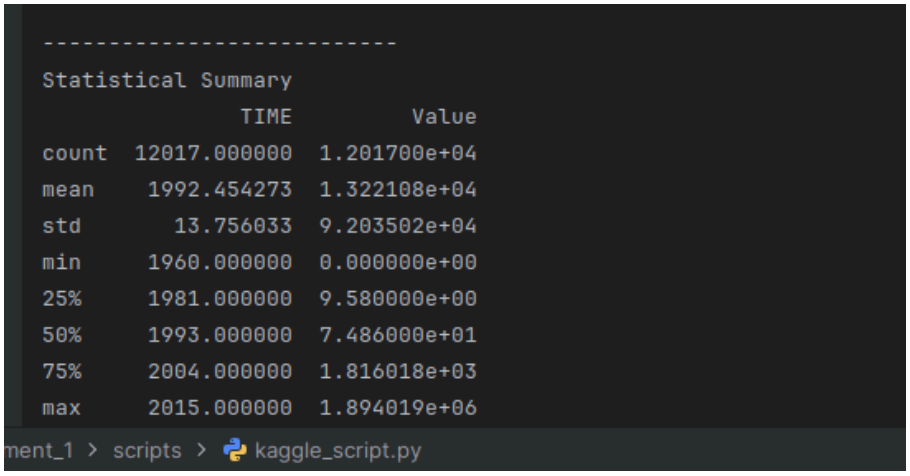
Reddit Summary

```
/home/user/PycharmProjects/assignment_1/.venv/bin/python /home/user/PycharmProjects/assignment_1/script
Statistical Summary of Data:
       Unnamed: 0       upvotes
count  100.000000    100.000000
mean    49.500000  23054.420000
std     29.011492  22262.195028
min      0.000000      0.000000
25%     24.750000   4378.250000
50%     49.500000  15859.500000
75%     74.250000  32446.750000
max     99.000000  86094.000000
Categorical Summaries:
          author   subreddit
count        100         100
unique        75          30
top     Wagamaga  Futurology
freq           9          19

Process finished with exit code 0
```

Google Trends

```
Statistical Summary of Data:
       interest_score
count      162.000000
mean        61.845679
std         24.536592
min         16.000000
25%         34.000000
50%         74.000000
75%         81.000000
max        100.000000

Process finished with exit code 0
```

Kaggle Summary



```
--------------------------
Statistical Summary
               TIME         Value
count  12017.000000  1.201700e+04
mean    1992.454273  1.322108e+04
std       13.756033  9.203502e+04
min     1960.000000  0.000000e+00
25%     1981.000000  9.580000e+00
50%     1993.000000  7.486000e+01
75%     2004.000000  1.816018e+03
max     2015.000000  1.894019e+06
ment_1 > scripts > kaggle_script.py
```

AI Products

1. **Investment Advisory Tool**
   A tool that advises investors based on renewable energy search trends.
2. **Content Strategy Assistant**
   The data scrapped in this assignment can be used to build a content strategy assistant for journalists, social media managers, and news anchors discussing the topic of renewable energy.

ToS Constraints and Privacy Issues

Reddit

1. Reddit API has restrictions on how data can be stored, used, and shared. Some of the content like images, also require explicit permission.
2. Scraping deleted content or private posts violates Reddit ToS.
3. Collecting names, emails, and other personal information without consent can lead to privacy breaches.

Google

1. Google Trends and other search-based data sources are subject to licensing and cannot be stored and republished with proper consent.
2. Using APIs can lead to data rate limits and commercial use restrictions.

Benefits of Multi-Source Data Collection

1. Combining data sources provides a complete picture of user behavior and opinions.

2. Cross-referencing data from multiple sources improves the accuracy and readability of insights.
3. Different data sources contain different perspectives which reduces bias.

Challenges & Conflicts in Multi-Source Data Collection

1. Google search data is structured and quantitative, while Reddit discussions are unstructured and qualitative, requiring different processing techniques.
2. Each platform has its own demographic and behavioral biases.
3. Aggregating data from multiple sources may introduce repetition, requiring additional filtering.

Combine and Store

To effectively store and integrate data from Reddit, Google, and other sources, we can use the following approaches:

1. Relational Databases can be used to store Google search trends and metadata from Reddit.
2. NoSQL can be used to store posts

Methods for Combining Data

1. Keyword-based matching can be used to combine data from Reddit posts and Google trends searches.
2. Use NLP models like BERT to detect topic similarity between sources.

(Graphs for each data source are added in the Github repo.)