Data Science Assignment 2

Kerui Wu

Data Collection

Goal/mode

The record of the class times and participation policy in each department at RPI in the past 5 years, which in turn helps to learn each department and major's features from the time perspective. By analyzing each class's time schedule in each department, students can have a better understanding of each department and major's requirements and teaching styles, which, in turn, can help them to choose the major that fits them.

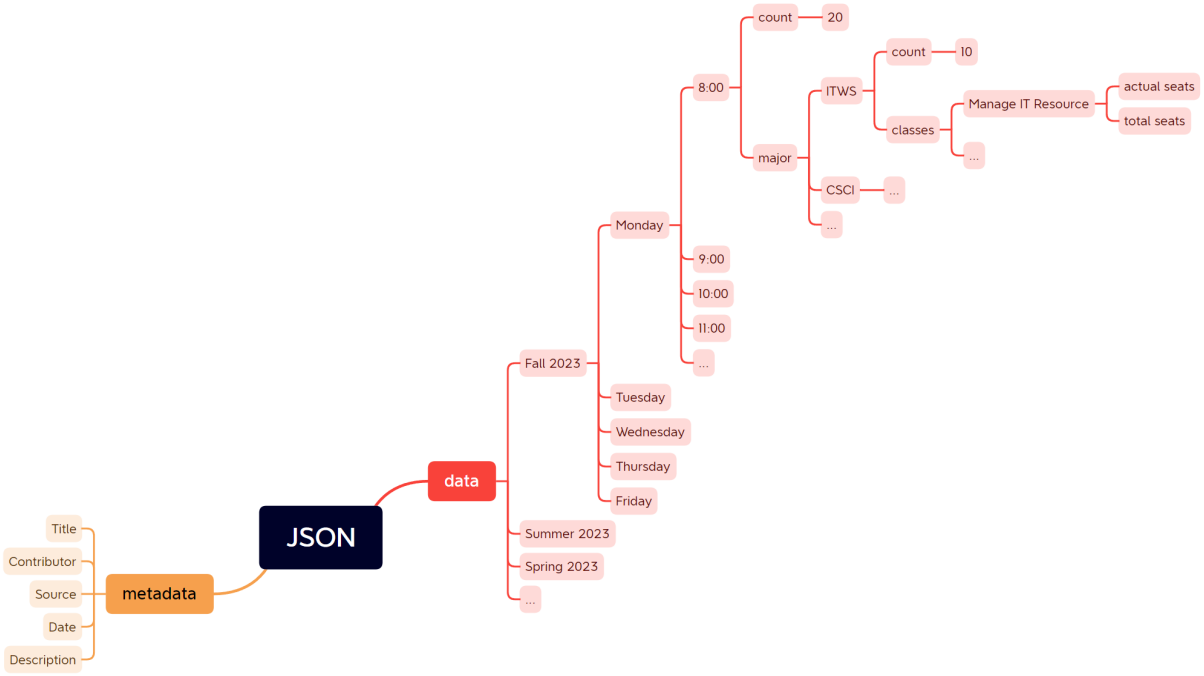
Process

The data was scraped with the use of Python script. To be more specific, Selenium(a Python package that is used to control web browsers and perform browser automation) is used to access Quacs(an authority class scheduling website that has all of the information about RPI's class schedules) read, and parse data. The collected data are exported into a JSON file.

Format

The JSON file consists of 2 keys: data and metadata

```
"""
result = {
    "metadata": {
        "Title": "...",
        "Contributors": ["..."],
        "Source": "...",
        "Date": "...",
        "Description": "..."
    }
    "data":{
        "Fall_2023": {
            "1": {  # Monday
                "8": {  # 8:00
                    "count": 0,
                    "majors": {
                        "ITWS": {
                            "count": 0,
                            "courses": {}
                        },
                        ...
                    }
                },
                ...
            },
            ...
        },
        ...
    }
}
"""
```



Store

On behalf of Open Source Spirit, the JSON data was published on GitHub with a GPL license for the public to use. More than that, the JSON file was imported into MongoDB as a Disaster Recovery solution.

Data Management

Logical Collection

Because of the JSON format's characteristics, the collected data has been parsed and categorized into different groups based on the semester and time that align with the project's goal.

Physical Data Handling

Physical data handling was not involved in the process since the source of the data is from the online website.

Security Support

With the knowledge of risk transference, the collected data was stored in the third-party platforms, GitHub and MongoDB, to take advantage of industry-standard security protection.

Data Ownership

As mentioned above, the collected data was stored in the GitHub repository with the license GPL, which means that everyone is free to make use of the data with the use of GPL license in their projects.

Metadata Management

The metadata was stored as a part of a JSON file, which can be updated easily to ensure the information's correctness.

Improvement

There are several points to improve during the collection process:

- With the use of HDF5, metadata can be stored in a separate table instead of as a part of the data's main body.
- Metadata's format can be more formal by taking advantage of some popular metadata standards like the Dublin Core Metadata Initiative(DCMI)
- Another script could be implemented to automate the update process if Quacs updated the incorrect data or published the new semester's schedule