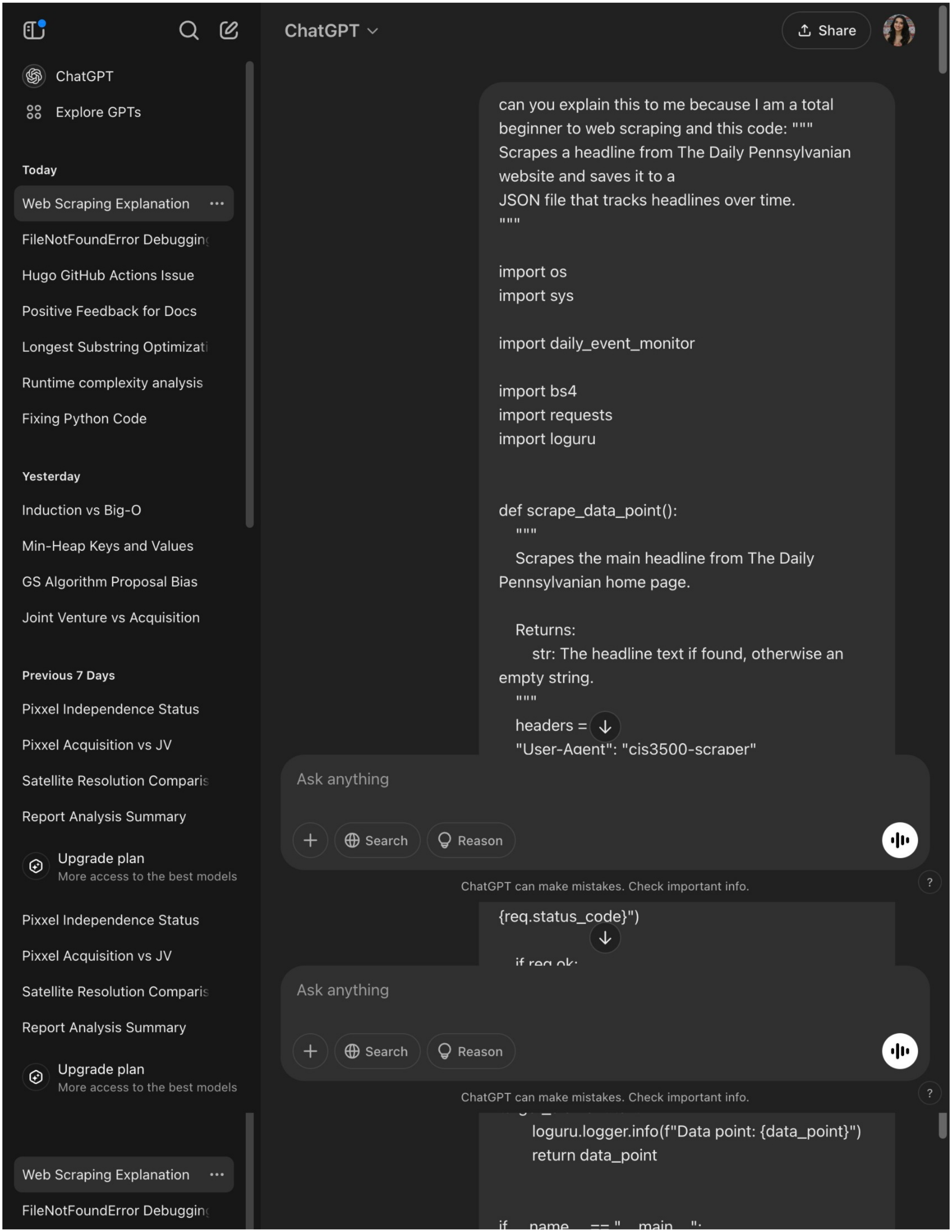

ChatGPT
Explore GPTs
Today
Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Yesterday
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Previous 7 Days
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Web Scraping Explanation
FileNotFoundError Debuggin
ChatGPT
Share

can you explain this to me because I am a total beginner to web scraping and this code: """
Scrapes a headline from The Daily Pennsylvanian website and saves it to a
JSON file that tracks headlines over time.
"""

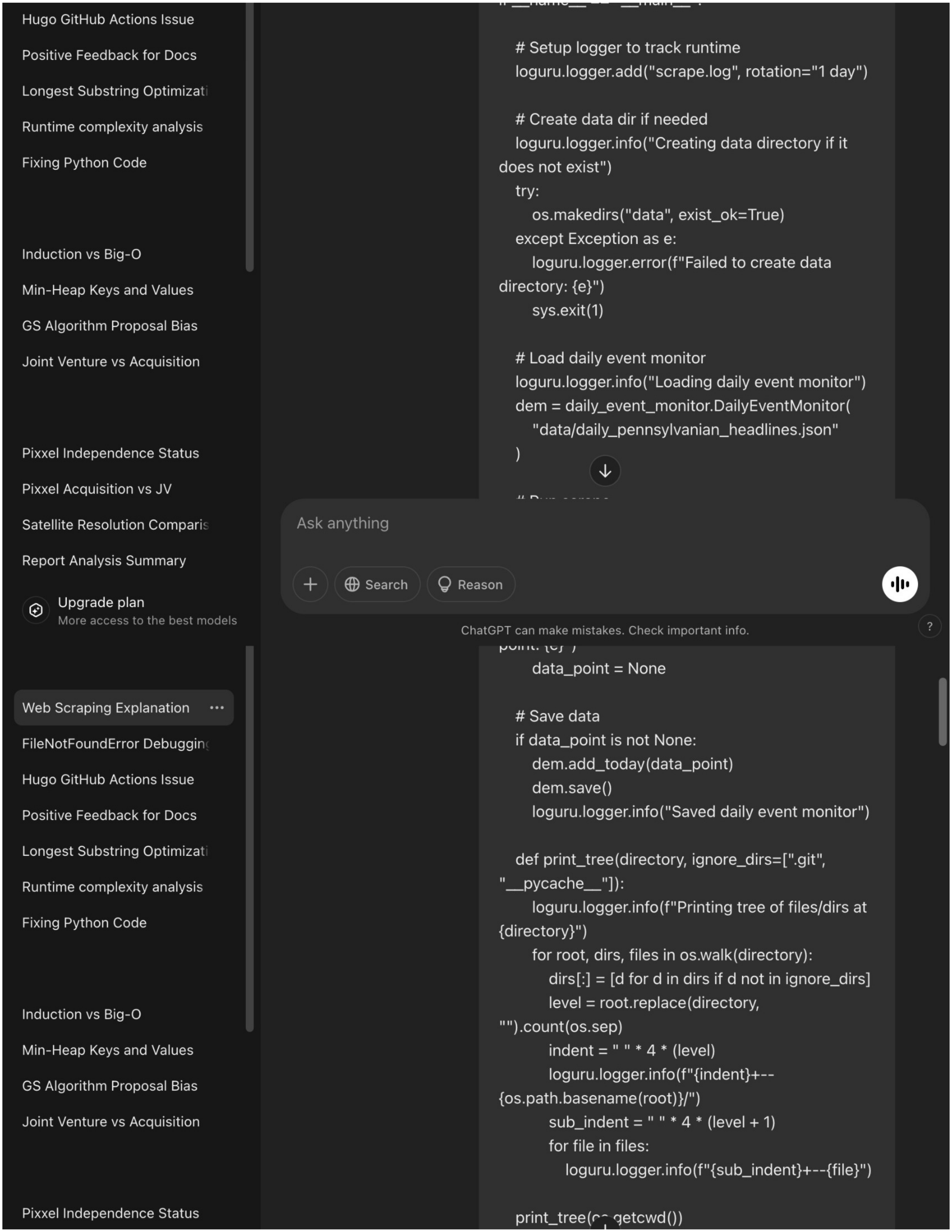
```
import os
import sys

import daily_event_monitor

import bs4
import requests
import loguru


def scrape_data_point():
   """
   Scrapes the main headline from The Daily Pennsylvanian home page.

   Returns:
      str: The headline text if found, otherwise an empty string.
   """
   headers =
   "User-Agent": "cis3500-scraper"
```

ChatGPT can make mistakes. Check important info.

```
{req.status_code}")

   if req.ok:
```

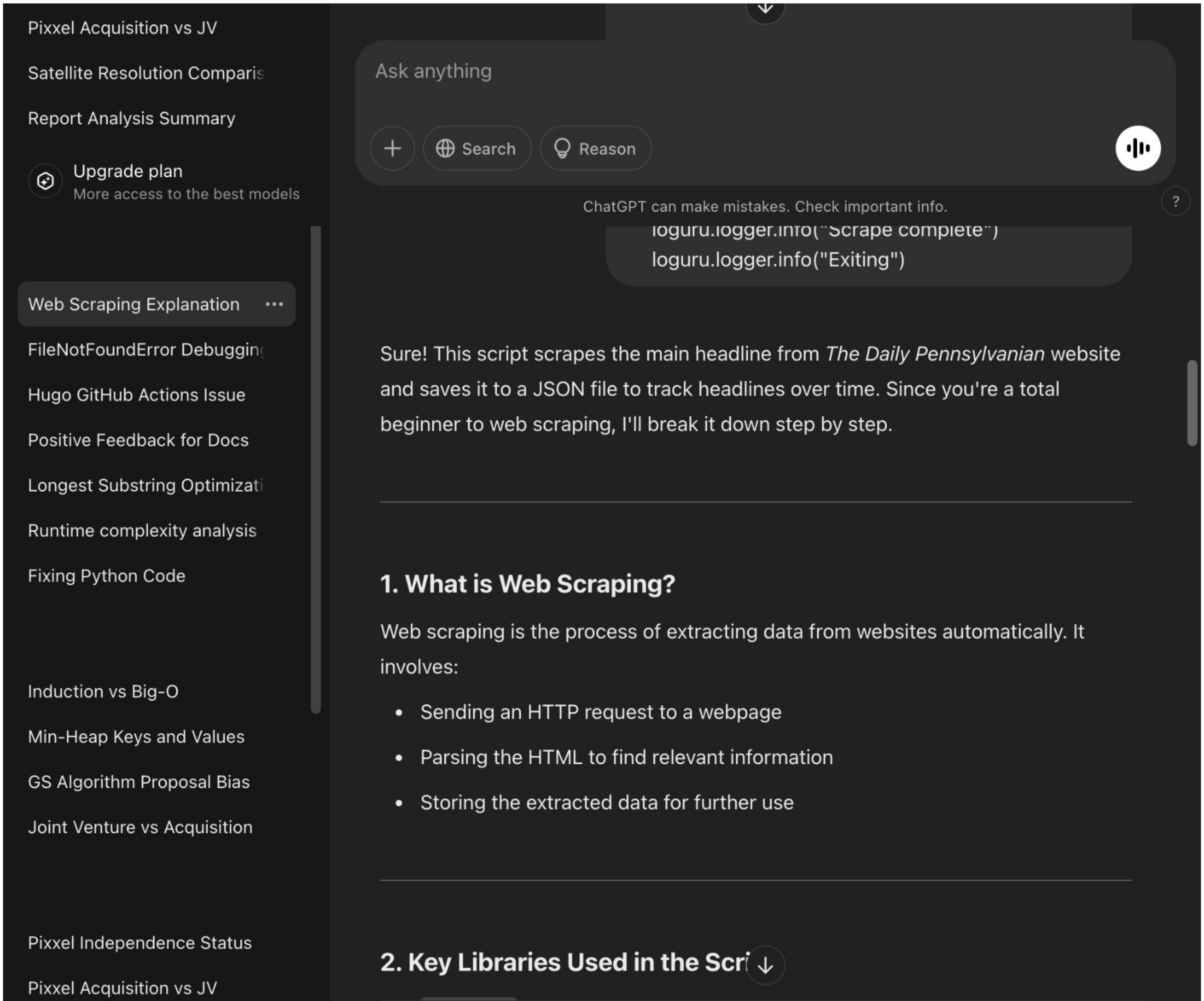
ChatGPT can make mistakes. Check important info.

```
      loguru.logger.info(f"Data point: {data_point}")
      return data_point


if __name__ == "__main__":
```


Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Ask anything
Search
Reason
ChatGPT can make mistakes. Check important info.
Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status

```
    # Setup logger to track runtime
    loguru.logger.add("scrape.log", rotation="1 day")

    # Create data dir if needed
    loguru.logger.info("Creating data directory if it
does not exist")
    try:
        os.makedirs("data", exist_ok=True)
    except Exception as e:
        loguru.logger.error(f"Failed to create data
directory: {e}")
        sys.exit(1)

    # Load daily event monitor
    loguru.logger.info("Loading daily event monitor")
    dem = daily_event_monitor.DailyEventMonitor(
        "data/daily_pennsylvanian_headlines.json"
    )
```

```
point: {e}")
        data_point = None

    # Save data
    if data_point is not None:
        dem.add_today(data_point)
        dem.save()
        loguru.logger.info("Saved daily event monitor")

    def print_tree(directory, ignore_dirs=[".git",
"__pycache__"]):
        loguru.logger.info(f"Printing tree of files/dirs at
{directory}")
        for root, dirs, files in os.walk(directory):
            dirs[:] = [d for d in dirs if d not in ignore_dirs]
            level = root.replace(directory,
"").count(os.sep)
            indent = " " * 4 * (level)
            loguru.logger.info(f"{indent}+--
{os.path.basename(root)}/")
            sub_indent = " " * 4 * (level + 1)
            for file in files:
                loguru.logger.info(f"{sub_indent}+--{file}")

    print_tree(os.getcwd())
```


Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Web Scraping Explanation
FileNotFoundError Debugging
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Ask anything
Search
Reason
ChatGPT can make mistakes. Check important info.

```
loguru.logger.info("Scrape complete")
loguru.logger.info("Exiting")
```

Sure! This script scrapes the main headline from *The Daily Pennsylvanian* website and saves it to a JSON file to track headlines over time. Since you're a total beginner to web scraping, I'll break it down step by step.

---

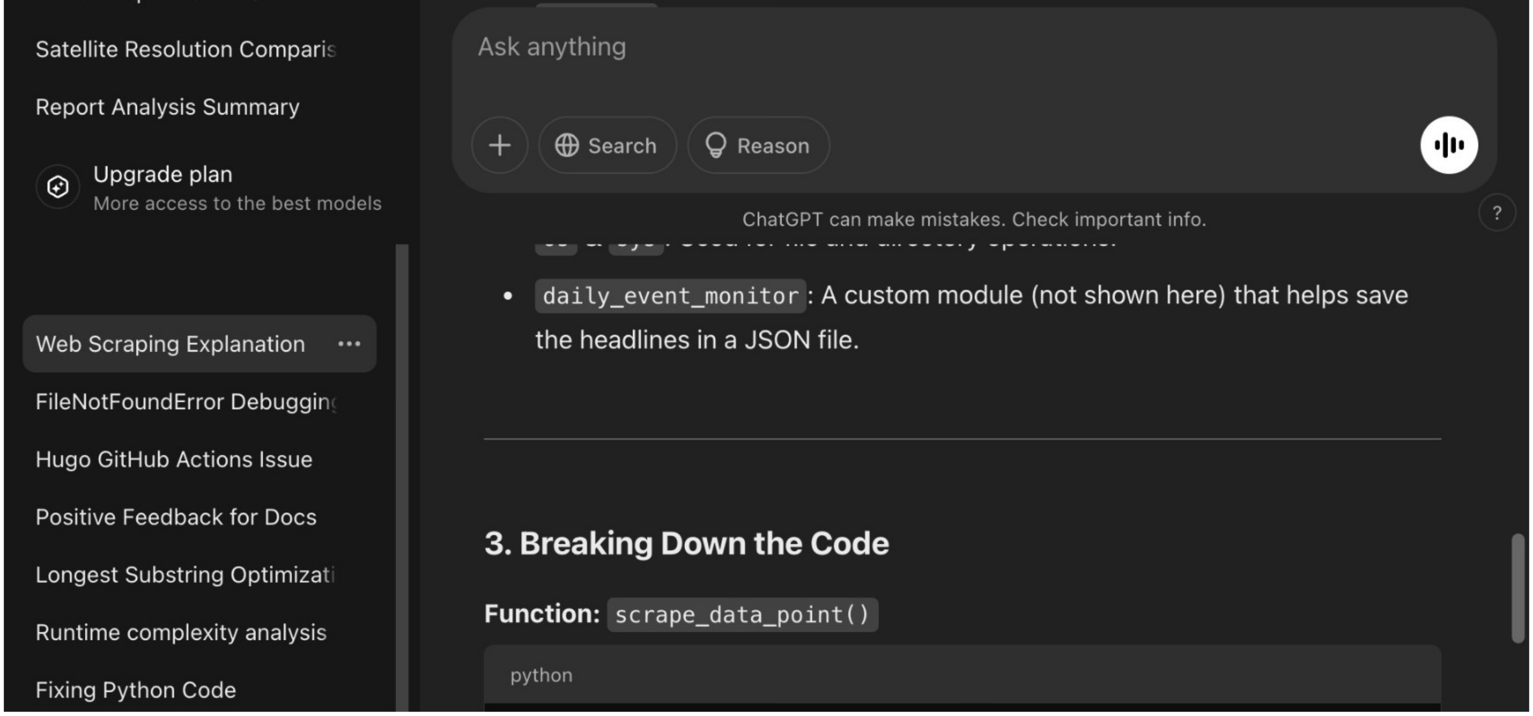
1. What is Web Scraping?

Web scraping is the process of extracting data from websites automatically. It involves:

- Sending an HTTP request to a webpage
- Parsing the HTML to find relevant information
- Storing the extracted data for further use

---

2. Key Libraries Used in the Scri


Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Web Scraping Explanation
FileNotFoundError Debugging
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Ask anything
Search
Reason
ChatGPT can make mistakes. Check important info.

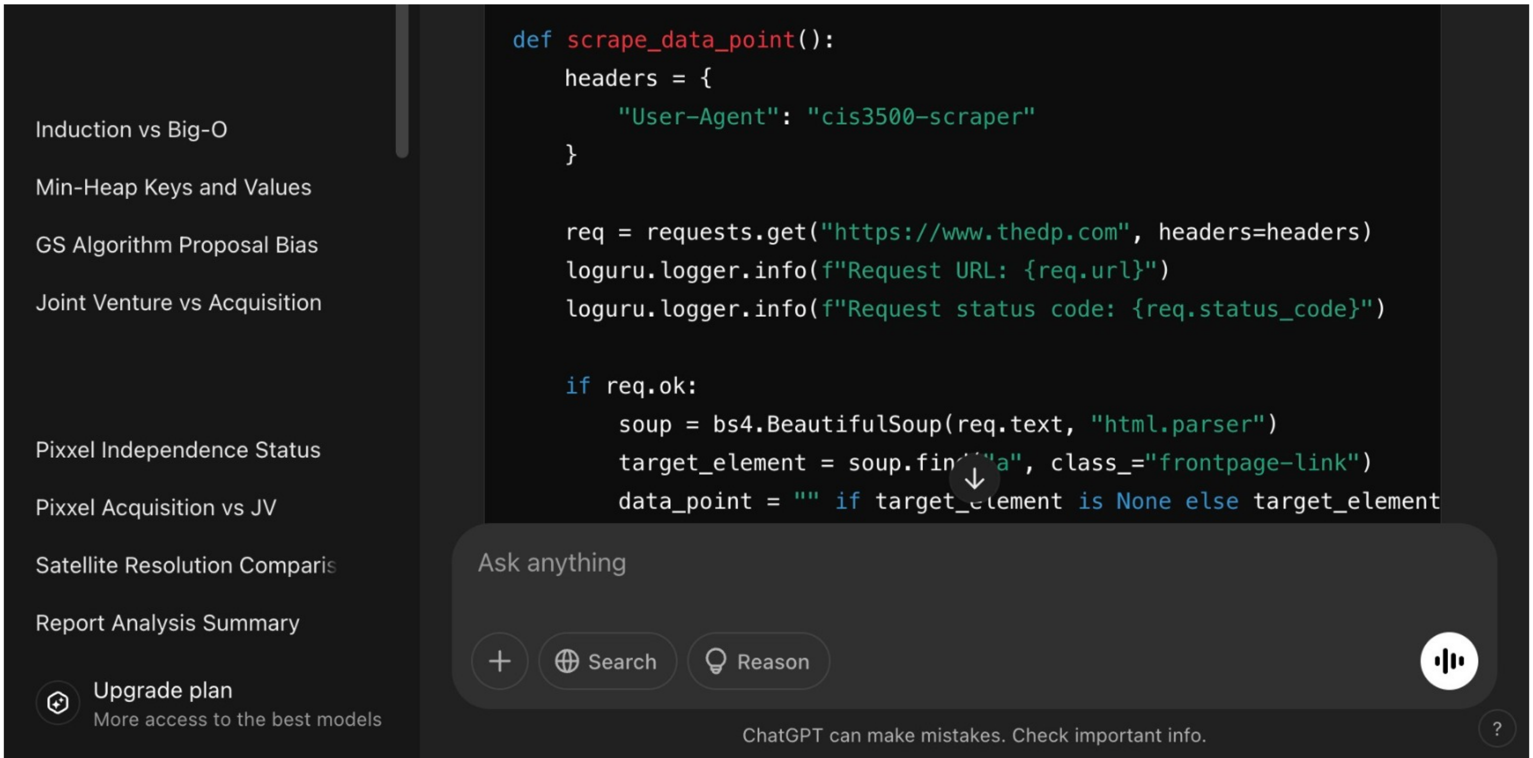
- `daily_event_monitor`: A custom module (not shown here) that helps save the headlines in a JSON file.

---

3. Breaking Down the Code

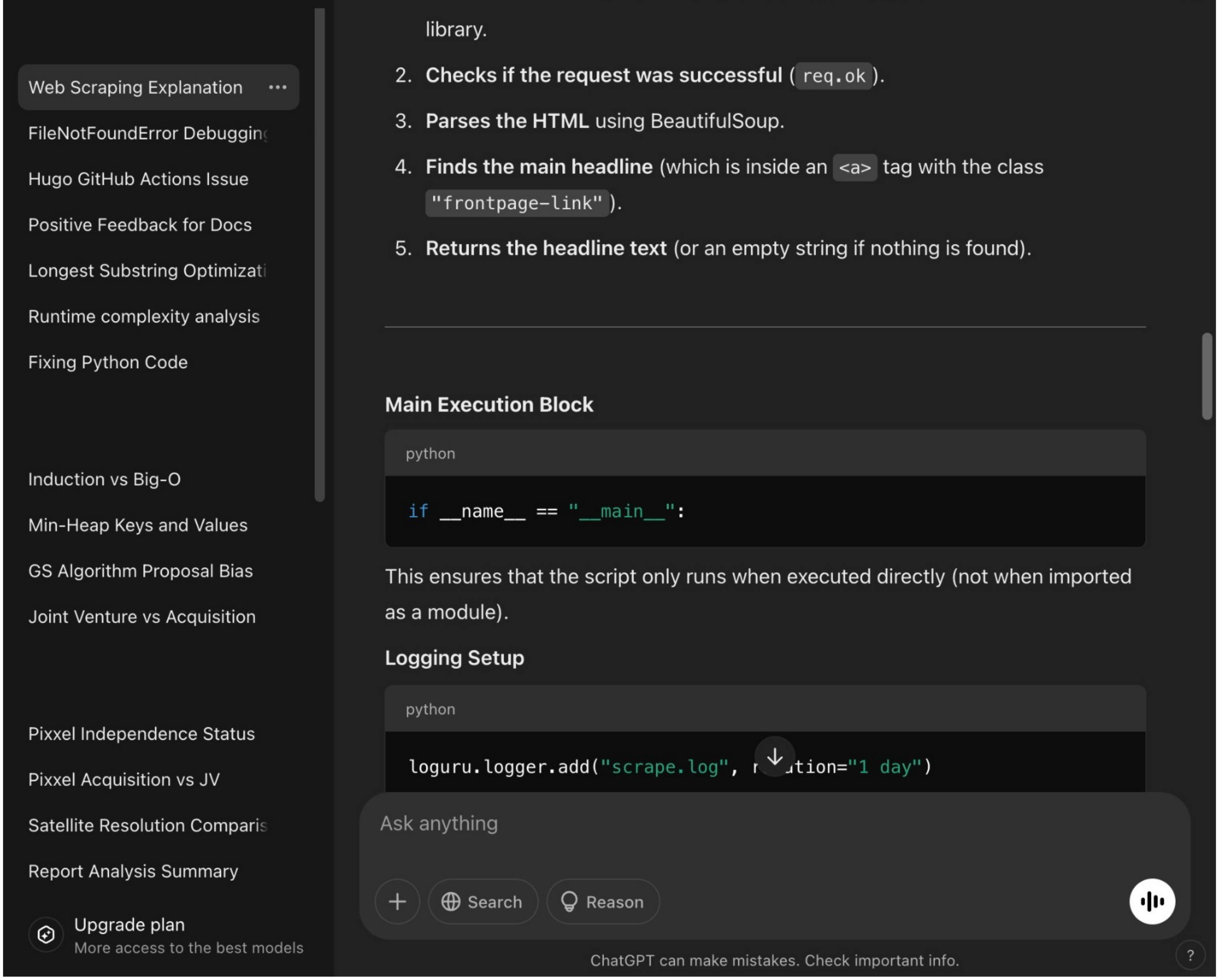
Function: `scrape_data_point()`

python


Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Ask anything
Search
Reason

```python
def scrape_data_point():
    headers = {
        "User-Agent": "cis3500-scraper"
    }

    req = requests.get("https://www.thedp.com", headers=headers)
    loguru.logger.info(f"Request URL: {req.url}")
    loguru.logger.info(f"Request status code: {req.status_code}")

    if req.ok:
        soup = bs4.BeautifulSoup(req.text, "html.parser")
        target_element = soup.find("a", class_="frontpage-link")
        data_point = "" if target_element is None else target_element
```

ChatGPT can make mistakes. Check important info.


Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Ask anything
Search
Reason

library.

2. **Checks if the request was successful** (`req.ok`).
3. **Parses the HTML** using BeautifulSoup.
4. **Finds the main headline** (which is inside an `<a>` tag with the class `"frontpage-link"`).
5. **Returns the headline text** (or an empty string if nothing is found).

Main Execution Block

python

```python
if __name__ == "__main__":
```

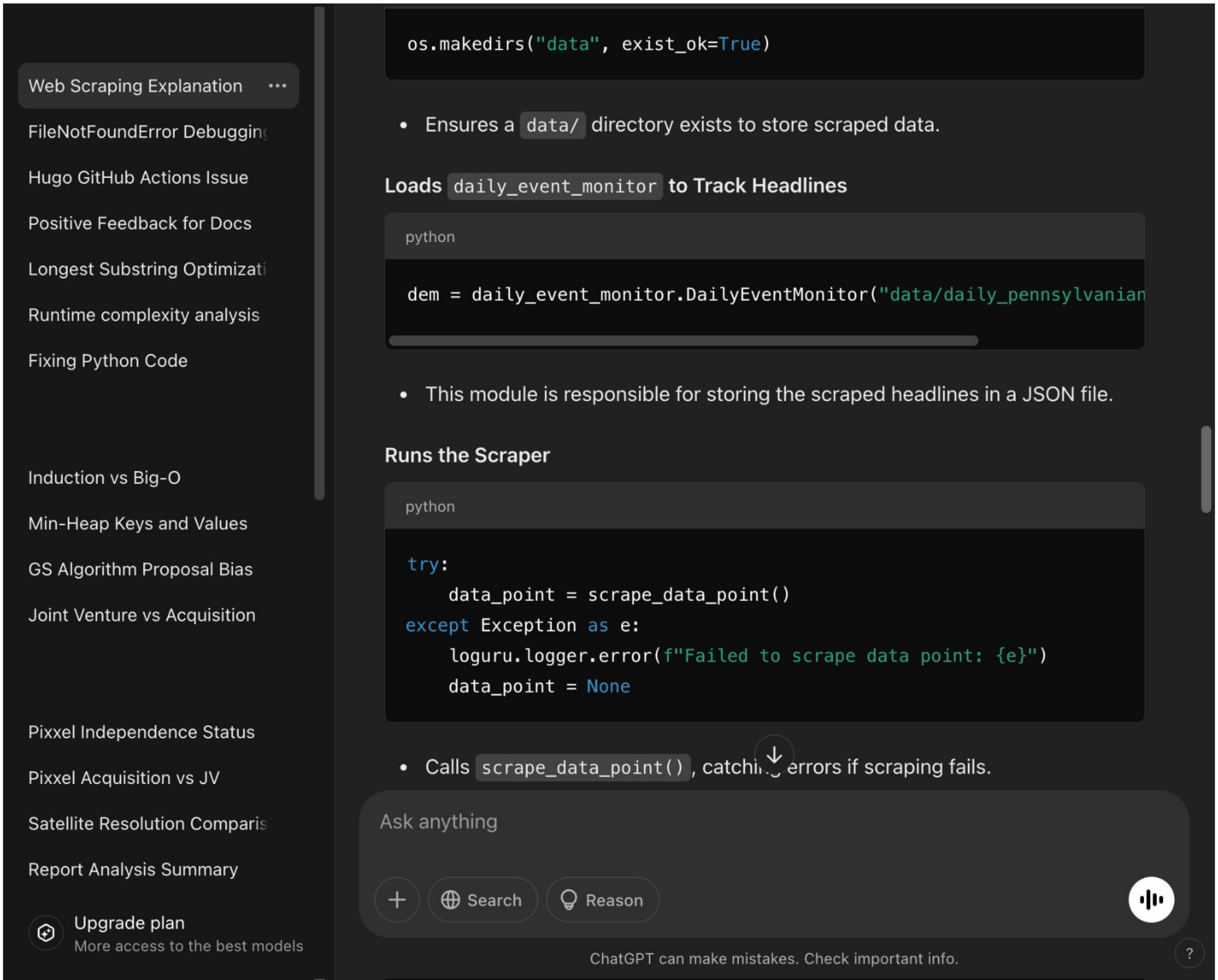
This ensures that the script only runs when executed directly (not when imported as a module).

Logging Setup

python

```python
loguru.logger.add("scrape.log", rotation="1 day")
```

ChatGPT can make mistakes. Check important info.


Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Ask anything
Search
Reason
?

```
os.makedirs("data", exist_ok=True)
```

- Ensures a `data/` directory exists to store scraped data.

Loads `daily_event_monitor` to Track Headlines

python

```
dem = daily_event_monitor.DailyEventMonitor("data/daily_pennsylvanian
```

- This module is responsible for storing the scraped headlines in a JSON file.

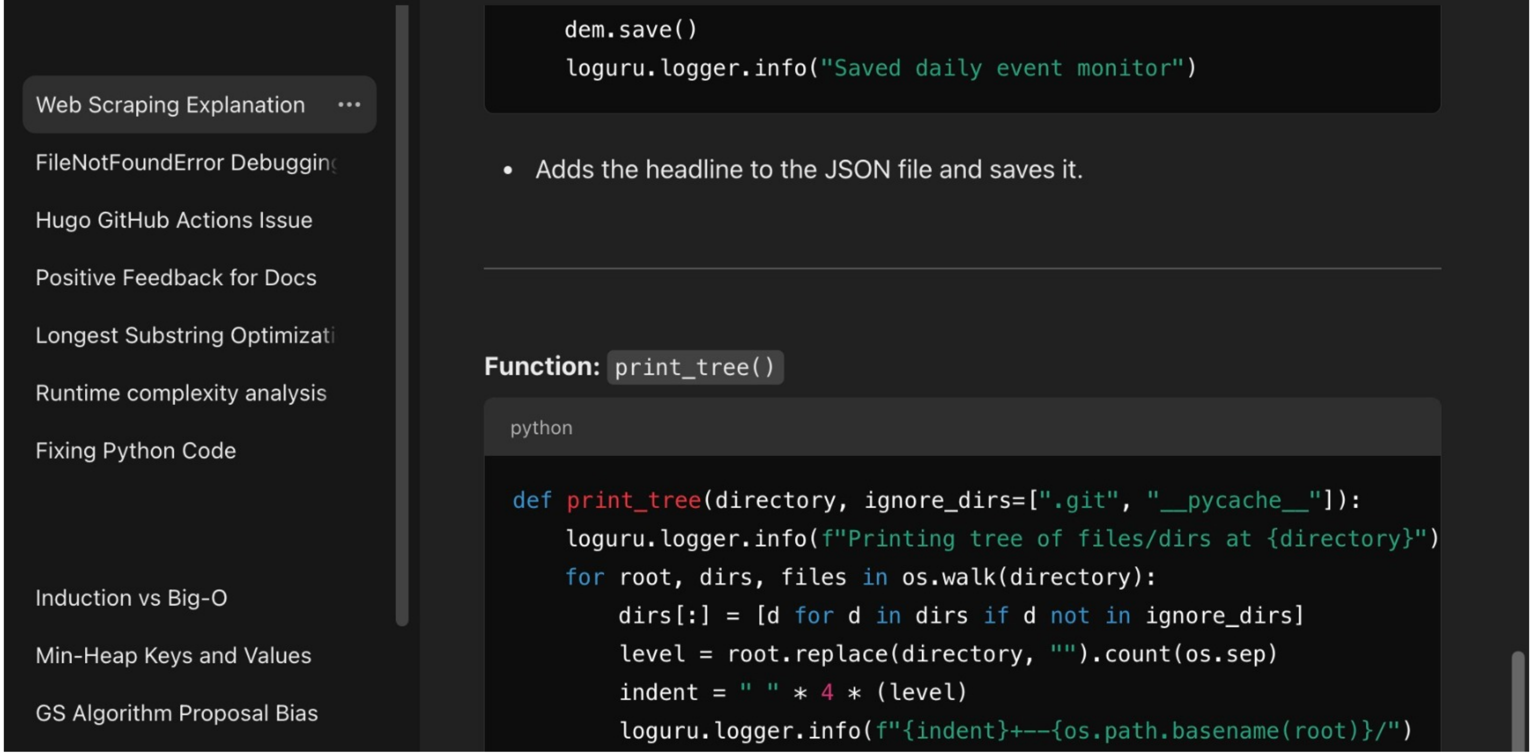
Runs the Scraper

python

```
try:
    data_point = scrape_data_point()
except Exception as e:
    loguru.logger.error(f"Failed to scrape data point: {e}")
    data_point = None
```

- Calls `scrape_data_point()`, catchi errors if scraping fails.

ChatGPT can make mistakes. Check important info.


Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias

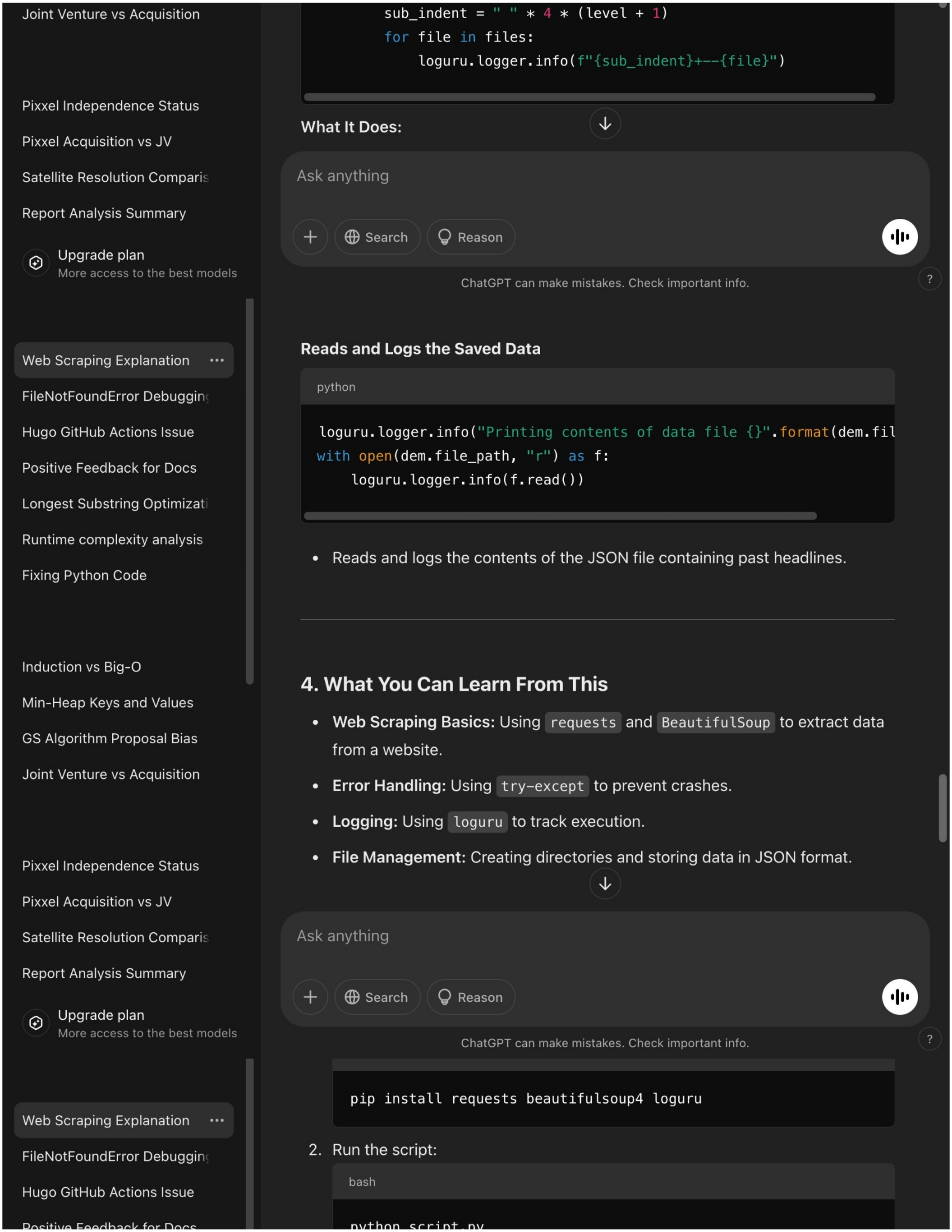
```
    dem.save()
    loguru.logger.info("Saved daily event monitor")
```

- Adds the headline to the JSON file and saves it.

Function: `print_tree()`

python

```
def print_tree(directory, ignore_dirs=[".git", "__pycache__"]):
    loguru.logger.info(f"Printing tree of files/dirs at {directory}")
    for root, dirs, files in os.walk(directory):
        dirs[:] = [d for d in dirs if d not in ignore_dirs]
        level = root.replace(directory, "").count(os.sep)
        indent = " " * 4 * (level)
        loguru.logger.info(f"{indent}+--{os.path.basename(root)}/")
```



```
        sub_indent = " " * 4 * (level + 1)
        for file in files:
            loguru.logger.info(f"{sub_indent}+--{file}")
```

What It Does:

Reads and Logs the Saved Data

```python
loguru.logger.info("Printing contents of data file {}".format(dem.fil
with open(dem.file_path, "r") as f:
    loguru.logger.info(f.read())
```

- Reads and logs the contents of the JSON file containing past headlines.

---

4. What You Can Learn From This

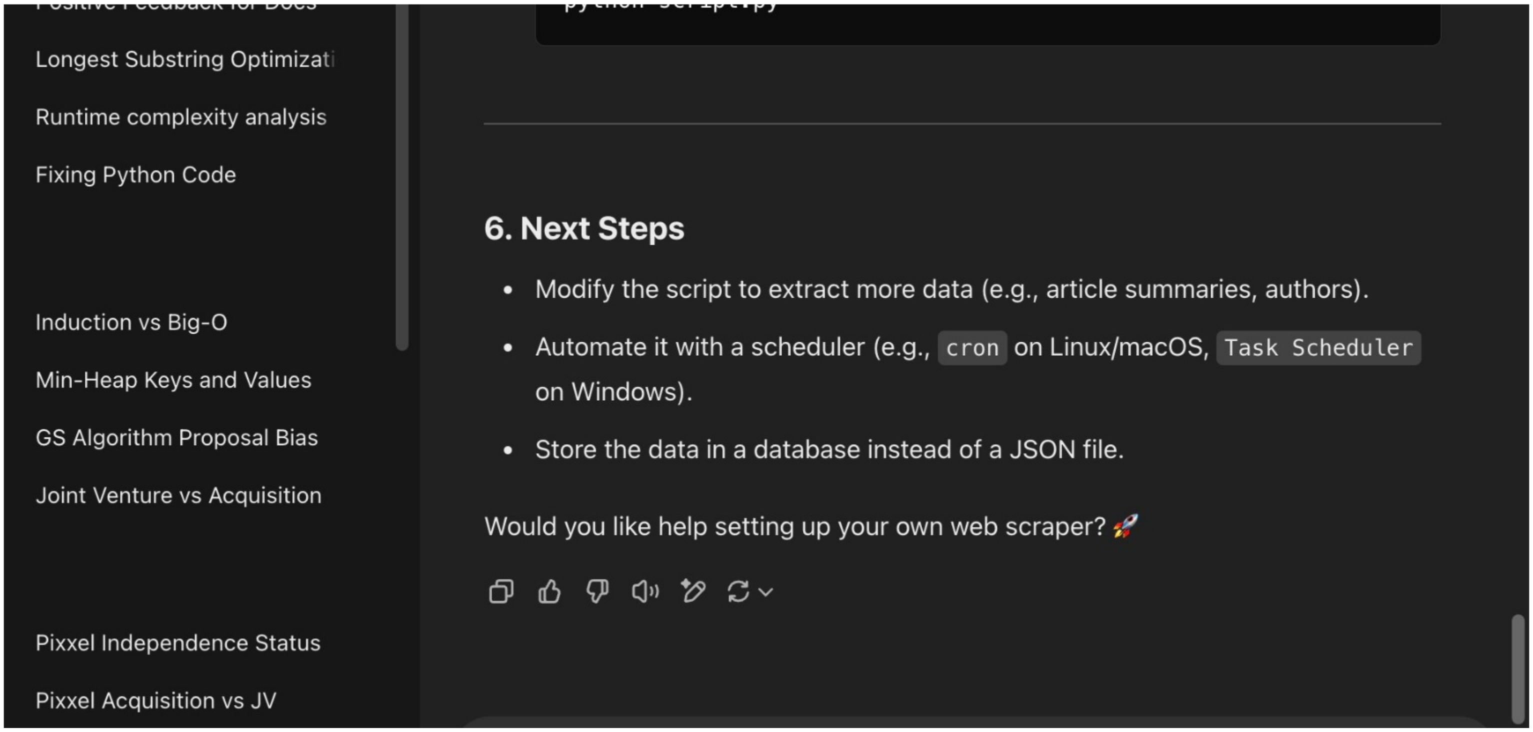
- **Web Scraping Basics:** Using `requests` and `BeautifulSoup` to extract data from a website.
- **Error Handling:** Using `try-except` to prevent crashes.
- **Logging:** Using `loguru` to track execution.
- **File Management:** Creating directories and storing data in JSON format.

```
pip install requests beautifulsoup4 loguru
```

2. Run the script:

```bash
python script.py
```

Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV
Satellite Resolution Comparis
Report Analysis Summary
Upgrade plan
More access to the best models
Web Scraping Explanation
FileNotFoundError Debuggin
Hugo GitHub Actions Issue
Positive Feedback for Docs


Positive Feedback for Docs
Longest Substring Optimizati
Runtime complexity analysis
Fixing Python Code
Induction vs Big-O
Min-Heap Keys and Values
GS Algorithm Proposal Bias
Joint Venture vs Acquisition
Pixxel Independence Status
Pixxel Acquisition vs JV

```
python script.py
```

---

6. Next Steps

- Modify the script to extract more data (e.g., article summaries, authors).
- Automate it with a scheduler (e.g., `cron` on Linux/macOS, `Task Scheduler` on Windows).
- Store the data in a database instead of a JSON file.

Would you like help setting up your own web scraper? 🚀