Midterm Project Report
Advanced Computer Programming

Web Scraping with Python

Student Name : SIYABONGA NHLEKO

Student ID : 112021176

Teacher : DINH-TRUNG VU

2024-04

• Introduction

1. Github

- **Personal Github Account**: https://github.com/Siya0198
- **Group Github Account**: https://github.com/niican/College-Dropouts
- **Group Project Repository**: https://github.com/niican/College-Dropouts
- **List of submitted file**s:

  Mid-project: main.py

  supporters.py

• Topic

Build a web scraper to extract information about the latest finance news from www.moneycontrol.com

• Project Overview

Brief description of advanced language features and libraries used in your work as well as results you have archieved. For example:

The data class, pprint, pattern matching, regulation expresion, and Beautiful Soup have used in my program. My program have extracted information about URL, About, Language, Number of Commits of repositories on the page

re (Regular Expressions):

This library is used for working with regular expressions. In the code, it's utilized to extract the domain part from a URL using the re.match() function. This enables the code to extract relevant information from URLs by specifying patterns to match against them.

requests:

The requests library is used to send HTTP requests to web servers and retrieve web pages. In the code, it's primarily used to fetch the HTML content of web pages using the requests.get() function. This allows the code to access the content of web pages programmatically, which is essential for web scraping.

BeautifulSoup:

BeautifulSoup is a library for parsing HTML and XML documents. In the code, it's used to parse the HTML content fetched using requests and extract relevant information from it. BeautifulSoup provides a convenient interface for navigating and searching HTML documents, making it easier to extract specific elements or data from web pages.

sys:

The sys module provides access to some variables used or maintained by the Python interpreter and to functions that interact strongly with the interpreter. In the code, it's used for accessing information about exceptions (sys.exc_info()) to handle errors more effectively during web scraping.

https://github.com/Siya0198.

• Implementation

• Class 1

Class; ArticlePreview

Attributes:

title: A string representing the title of an article.

link: A string representing the URL link to the article.

preview_content: A string representing the preview content of the article.

- **Fields**

Fields refer to the attributes or variables associated with an object of a class. The code uses the following function parameters and local variables to store and manipulate data

1. Function Parameters:

- Function parameters are used to pass data into functions. In the provided code, parameters such as `full_url`, `article_link`, `preview_target`, `results`, `both_belong_area`, `indi_part`, `title_target`, `img_target`, `container`, and `preview_results` are used to provide inputs to the functions.

2. Local Variables:

- Local variables are variables defined within the scope of a function and are only accessible within that function. In the provided code, variables like `rs`, `drilling_site`, `soup_pot`, `preview`, `s`, `titles`, `imgs`, `clean_title`, `link`, and `img_link` are local variables used for data processing and manipulation within the functions.

These parameters and local variables act as the "fields" in the context of the functions in which they are used. They store and manipulate data during the execution of the functions, enabling the code to perform its intended tasks such as web scraping and data extraction.

- **Methods**

Constructor method to initialize the attributes.
Methods to access and manipulate the attributes.

- **Functions**

I used the following functions in my code;

1. `re.match(pattern, string)`: This function from the `re` module is used in the `get_domain()` function to match a regular expression pattern (`"https://(.?)/"`) against a string (`full_url`). It returns a match object if the pattern is found in the string.

2. `requests.get(url)`: This function from the `requests` module is used to send an HTTP GET request to the specified URL (`article_link` or `full_url`). It returns a Response object.

3. `BeautifulSoup(text, parser)`: This function from the `bs4` (Beautiful Soup) module is used to create a BeautifulSoup object from the HTML content (`drilling_site.text`). It takes the HTML content and a parser type (in this case, `"html.parser"`) as arguments.

4. `soup.select(selector)`: This method of the BeautifulSoup object is used to find all elements in the HTML document that match the given CSS selector (`preview_target`, `both_belong_area + " " + indi_part + " " + title_target`,

`both_belong_area + " " + indi_part + " " + img_target`, etc.).

5. `print()`: This built-in Python function is used to output text or variables to the console for debugging purposes. It's used in exception handling blocks to print error messages.

6. `len()`: This built-in Python function is used to get the length of a sequence (e.g., a list or string). It's used to check if a list of elements returned by `soup.select()` is empty or not.

7. `append()`: This method is used to add an element to the end of a list. It's used to add cleaned-up preview content (`s`) or scraped data (`[clean_title, link, img_link]`) to lists.

- ## Class 2

Class; WebPage

Attributes:

url: A string representing the URL of the web page.

content: A string representing the HTML content of the web page.

- ## Method/Function 1

Methods to fetch and parse the web page content.

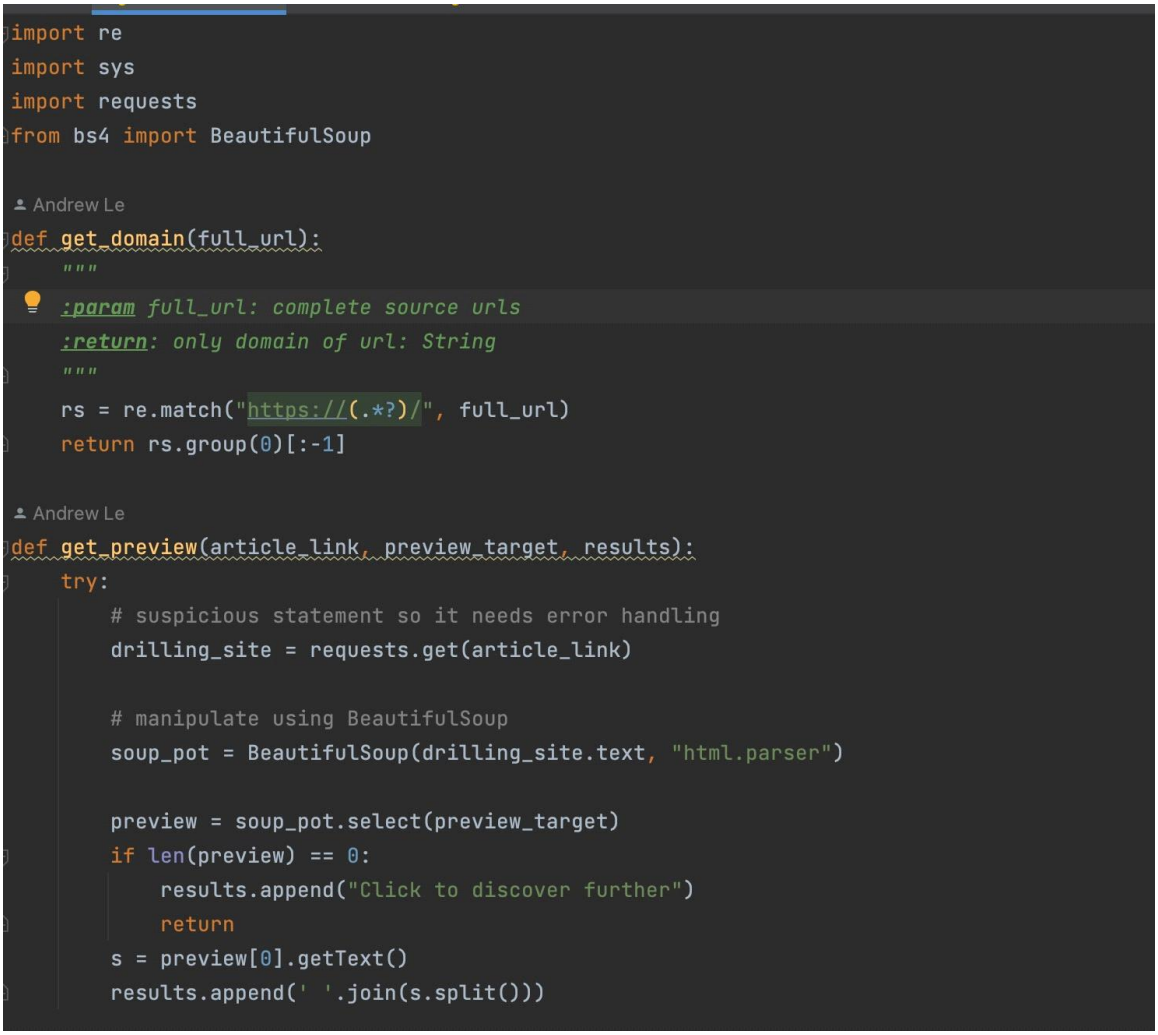
- ## Method/Function 2

Methods to handle errors during web page retrieval.

- **Results**

- **Result 1**

.



```
import re
import sys
import requests
from bs4 import BeautifulSoup

Andrew Le
def get_domain(full_url):
    """
    :param full_url: complete source urls
    :return: only domain of url: String
    """
    rs = re.match("https://(.*?)/", full_url)
    return rs.group(0)[:-1]

Andrew Le
def get_preview(article_link, preview_target, results):
    try:
        # suspicious statement so it needs error handling
        drilling_site = requests.get(article_link)

        # manipulate using BeautifulSoup
        soup_pot = BeautifulSoup(drilling_site.text, "html.parser")

        preview = soup_pot.select(preview_target)
        if len(preview) == 0:
            results.append("Click to discover further")
            return
        s = preview[0].getText()
        results.append(' '.join(s.split()))
```

```python
import concurrent
import time
from supporters import *
from concurrent import futures

# main program
if __name__ == "__main__":

    # sources and their scrapping areas
    # src 1
    money_control = "https://www.moneycontrol.com/news/tags/currency.html/news/"
    money_news = "#cagetory"
    money_indi_1 = ""
    money_indi_2 = ".clearfix"
    money_img = "a img"  # data-src
    money_title = "h2 a"
    money_preview = ".article_desc"
```

- **Result 2**

```
 the previous session. The currency traded in a narrow 83.2450 to 83.27 range in the spot session.']
['Forex reserves decline by $2.36 billion to $583.53 billion', 'https://www.moneycontrol
 .com/news/business/forex-reserves-decline-by-2-36-billion-to-583-53-billion-11614361.html', 'https://images.moneycontrol
 .com/static-mcnews/2022/08/Forex-Reserves-770x433.png?impolicy=website&width=135&height=80', 'The Indian currency came close to testing its record low of 83.29, hit in
  October 2022, towards the end of the spot trading session.']
['Rupee falls 6 paise to end at 83.23 against US dollar', 'https://www.moneycontrol.com/news/currency/rupee-falls-6-paise-to-end-at-83-23-against-us-dollar-11604691
 .html', 'https://images.moneycontrol.com/static-mcnews/2023/10/rupee_dollar-770x433.jpg?impolicy=website&width=135&height=80', 'Truth be told, we weren't supposed to
 stay around for a long while. We were a stop-gap arrangement to speed up the remonetization of the economy after the 2016 demonetization exercise']
['How will the surge in US bond yields rub-off on other asset classes?', 'https://www.moneycontrol
 .com/news/business/markets/how-will-the-surge-in-us-bond-yields-rub-off-on-other-asset-classes-11599771.html', 'https://images.moneycontrol
 .com/static-mcnews/2023/10/bonds-mc-770x433.jpg?impolicy=website&width=135&height=80', 'The rupee rose by 13 paise on December 14 but lost ground due to the ongoing
 dollar buying pressure']
['Rupee falls 6 paise to 83.18 against dollar amid strengthening US bond yields', 'https://www.moneycontrol
 .com/news/currency/rupee-falls-6-paise-to-83-18-against-dollar-amid-strengthening-us-bond-yields-11587581.html', 'https://images.moneycontrol
 .com/static-mcnews/2023/10/rupee_dollar-1-770x433.jpg?impolicy=website&width=135&height=80', 'Indian markets gained for fifth straight sessions with gaining around 2.5
  percent during this period. So far this year, both Sensex and Nifty gained around 17 percent each.']
['Rupee ends flattish, likely aggressive intervention helps avert record low', 'https://www.moneycontrol
 .com/news/business/markets/rupee-ends-flattish-likely-aggressive-intervention-helps-avert-record-low-11564651.html', 'https://images.moneycontrol
 .com/static-mcnews/2023/10/rupee_dollar-770x433.jpg?impolicy=website&width=135&height=80', 'The weakening of the Yen, especially to a 34-year low, can have significant
  impacts on various sectors in India, particularly the automotive industry']
Confirm number: 24

Checkmark: 24 articles in 7.74 secs (included previews)

Process finished with exit code 0
```

.

- **Conclusions**

In conclusion, studying the provided code provides valuable insights into web scraping techniques, error handling practices, modularity, library utilization, data extraction strategies, and code readability in Python programming. These insights can be applied to develop effective web scraping solutions and improve overall programming skills.