Web scrapping

Web scraping is the process of collecting and parsing raw data from the Web. The Internet hosts perhaps the greatest source of information on the planet. Many disciplines, such as data science, business intelligence, and investigative reporting, can benefit enormously from collecting and analyzing data from websites.

```
Scrape and parse text from any website and show the below:
```

```
from urllib.request import urlopen
import re
from bs4 import BeautifulSoup as bs
import csv

url = url = "https://www.reddit.com/"

page = urlopen(url)

html_bytes = page.read()
html = html_bytes.decode("utf-8")
```

1. Extract Text From HTML With String Methods

```
# Getting the title and style of the webpage

title = html.find("<title>") + len("<title>")
title_end = html.find("</title>")
lang = html.find('<html lang="') + len('<html lang="')
lang_end = html.find('">')

title_text = html[title : title_end]
lang_text = html[lang : lang_end]

print(f"Title: \n{title_text}")
print(f"\nLang: \n{lang_text}")
```

```
Title:
Reddit - Dive into anything

Lang:
en-US" class="theme-beta theme-light
```

2. Extract Text From HTML With Regular Expressions

```
pattern = "<h2>.*?</h2>"
matches = re.findall(pattern, html)

print(f"The sub headings in {url} are: ")
```

```
for title in matches:
    print(re.sub("<.*?>", "", title))
```

```
The sub headings in https://www.reddit.com/ are:
```

```
import re
from urllib.request import urlopen

url = "https://www.reddit.com/"
page = urlopen(url)
html = page.read().decode("utf-8")

pattern = "<title.*?>.*?</title.*?>"
match_results = re.search(pattern, html, re.IGNORECASE)
title = match_results.group()
title = re.sub("<.*?>", "", title) # Remove HTML tags

print(title)
```

```
Reddit - Dive into anything
```

3. Use an HTML Parser for Web Scraping in Python

```
soup = bs(html, "html.parser")

# Title
print(f"Title: \n{soup.title.string}", end="\n\n\n")

# Subheading
subs = soup.find_all("h2")
print("All subheadings: ")
for sub in subs:
    print(f"{sub.string}")
print("\n\n")

# Hyperlink titles
links = soup.find_all("a")
print("All hyperlink titles: ")
for link in links:
    if link.string != None:
        print(f"{link.string}")
```

```
Title:
Reddit - Dive into anything


All subheadings:



All hyperlink titles:
Log In
Gaming
Valheim
Genshin Impact
```

```
Minecraft
Pokimane
Halo Infinite
Call of Duty: Warzone
Path of Exile
Hollow Knight: Silksong
Escape from Tarkov
Watch Dogs: Legion
Sports
NFL
NBA
Megan Anderson
Atlanta Hawks
Los Angeles Lakers
Boston Celtics
Arsenal F.C.
Philadelphia 76ers
Premier League
UFC
Business, Economics, and Finance
GameStop
Moderna
Pfizer
Johnson & Johnson
AstraZeneca
Walgreens
Best Buy
Novavax
SpaceX
Tesla
Crypto
Cardano
Dogecoin
Algorand
Bitcoin
Litecoin
Basic Attention Token
Bitcoin Cash
Television
The Real Housewives of Atlanta
The Bachelor
Sister Wives
90 Day Fiance
Wife Swap
```

4. Save the scrapped text to a text file

```
a = [i.string + "\n" for i in links if i.string != None]
b = [i.string + "\n" for i in subs]

with open("output.txt", "w") as file:
    file.writelines(a)
    # file.write("\n")
    file.writelines(b)
    # file.write("\n")
    file.write(soup.title.string)

print("Successfully written")
```

```
Successfully written
```

5. Save the scrapped text to a csv file

```
a = [i.string for i in links if i.string != None]
b = [i.string for i in subs]

print(a)

with open("csvout.csv", "w") as file:
    csvwriter = csv.writer(file)
    csvwriter.writerow(a)
    csvwriter.writerow(b)
```

```
['Log In', 'Gaming', 'Valheim', 'Genshin Impact', 'Minecraft', 'Pokimane', 'Halo Infinite', 'Call of Duty: Warzone', 'Path of Exile', 'Hollow Knight: Silksong', 'Escape from Tarkov', 'Watch Dogs: L
```