PYTHON PROJECT

Web scraping

```
import requests
from bs4 import BeautifulSoup
import csv

# Web scraping for the first website
url1 = 'https://builtin.com/companies/tech/aws-companies?page={}'
base_url1 = 'https://builtin.com/companies/tech/aws-companies?page={}'
company_data1 = []

for page_number in range(1, 44):
    current_url = base_url1.format(page_number)
    r1 = requests.get(current_url)
    soup1 = BeautifulSoup(r1.text, 'html.parser')
    company_list1 = soup1.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
    company_data1.extend([company.get_text(strip=True) for company in
company_list1])

# Web scraping for the second website
url2 = 'https://www.forbes.com/lists/cloud100/?sh=1a9a50077d9c'
r2 = requests.get(url2)
soup2 = BeautifulSoup(r2.text, 'html.parser')
company_list2 = soup2.find_all("div", class_="organizationName second
table-cell company")
company_data2 = [company.get_text(strip=True) for company in
company_list2]

# Web scraping for the third website
url3 = 'https://www.builtinla.com/companies/tech/aws-companies'
r3 = requests.get(url3)
soup3 = BeautifulSoup(r3.text, 'html.parser')
company_list3 = soup2.find_all("h2", class_="fw-extrabold fs-xl hover-
underline d-inline-block company-title-clamp mb-0")
company_data3 = [company.get_text(strip=True) for company in
company_list3]

# Combine the company data from both sources
all_company_data = company_data1 + company_data2

# Create a CSV file and write the data
```

```
csv_file_path = 'combined_companies.csv'
with open(csv_file_path, 'w', newline='') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name'])

    # Write the combined company data to the CSV file
    for company in all_company_data:
        csv_writer.writerow([company])

print(f"Data has been successfully written to {csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company names from a given URL
def scrape_company_names(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
    return [company.get_text(strip=True) for company in company_list]

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store company names
company_names = []

# Iterate over each page and scrape company names
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    company_names.extend(scrape_company_names(current_url))

# Save company names to CSV file
csv_file_path = 'company_names.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
```

```
    csv_writer.writerow(['Company Name'])

    # Write the company names to the CSV file
    for company_name in company_names:
        csv_writer.writerow([company_name])

print(f"Company names have been successfully written to
{csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape links and company names from a given URL
def scrape_data(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scrape links
    link_list = soup.find_all("a", class_="info-item company-website-
link")  # Adjust the class based on your HTML structure
    links = [link['href'] for link in link_list]

    # Scrape company names
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
    company_names = [company.get_text(strip=True) for company in
company_list]

    return list(zip(company_names, links))

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store companies and their links
companies_data = []

# Iterate over each page and scrape data
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    companies_data.extend(scrape_data(current_url))
```

```
# Print companies and their links
for company, link in companies_data:
    print(f"Company: {company}\nLink: {link}\n")

# Save companies and their links to CSV file
csv_file_path = 'companies_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name', 'Link'])

    # Write the companies and links to the CSV file
    csv_writer.writerows(companies_data)

print(f"Companies and their links have been successfully written to
{csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company names from a given URL
def scrape_company_names(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Adjust the class based on your HTML structure
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")

    return [company.get_text(strip=True) for company in company_list]

def scrape_company_links(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Adjust the class based on your HTML structure
    company_link = soup.find_all("a", class_="btn btn-lg btn-outline-
primary w-100")

    return [company.get_text(strip=True) for company in company_link]

# Replace this URL with the actual URL of the page you want to scrape
url = 'https://www.builtinla.com/companies?page=1'
```

```
# Scrape company names and links
company_names = scrape_company_names(url)
company_links = scrape_company_links(url)

# Save company data to CSV file
csv_file_path = 'company_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name', 'Company Link'])

    # Write the company data to the CSV file
    for company_name, company_link in zip(company_names,
company_links):
        csv_writer.writerow([company_name, company_link])

print(f"Company names and links have been successfully written to
{csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company names and links from a given URL
def scrape_company_links(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Adjust the class based on your HTML structure
    company_links = soup.find_all("a", class_="btn btn-lg btn-outline-
primary w-100")

    return [company.get("href") for company in company_links]

# Replace this URL with the actual URL of the page you want to scrape
url = 'https://www.builtinla.com/companies?page=1'

# Scrape company links
company_links = scrape_company_links(url)

# Save company links to CSV file
csv_file_path = 'company_links.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
```

```
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Link'])

    # Write the company links to the CSV file
    for link in company_links:
        csv_writer.writerow([link])

print(f"Company links have been successfully written to
{csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company names and links from a given URL
def scrape_company_data(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company names
    company_names = []
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
    company_names = [company.get_text(strip=True) for company in
company_list]

    # Scraping company links
    company_links = []
    company_elements = soup.find_all("a", class_="btn btn-lg btn-
outline-primary w-100")
    company_links = [company.get("href") for company in
company_elements]

    return list(zip(company_names, company_links))

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store company data (names and links)
all_companies_data = []
```

```
# Iterate over each page and scrape company data
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    companies_data = scrape_company_data(current_url)
    all_companies_data.extend(companies_data)

# Save all company data to CSV file
csv_file_path = 'all_companies_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name', 'Company Link'])

    # Write the company data to the CSV file
    for company in all_companies_data:
        csv_writer.writerow(company)

print(f"All company data (names and links) have been successfully
written to {csv_file_path}.")
```

```
!pip install selenium
chrome_driver_path = '/content/drive/MyDrive/chromedriver'
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape official link from the company detail page
def scrape_official_link(company_detail_url):
    # Add your logic to extract the official link based on the
structure of the company detail page
    # This is a placeholder, modify it according to the actual
structure
    official_link = "Not available"

    # Example:
    # r = requests.get(company_detail_url)
    # soup = BeautifulSoup(r.text, 'html.parser')
    # official_link = soup.find("a", class_="official-link-
class").get("href")

    return official_link
```

```
# Function to scrape company names and links from a given URL
def scrape_company_data(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company names
    company_names = []
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
    company_names = [company.get_text(strip=True) for company in
company_list]

    # Scraping company links
    company_links = []
    company_elements = soup.find_all("a", class_="btn btn-lg btn-
outline-primary w-100")
    company_links = [company.get("href") for company in
company_elements]

    # Combine names and links
    companies_data = list(zip(company_names, company_links))

    # Scraping official links from each company's detail page
    for i, (company_name, company_detail_url) in
enumerate(companies_data):
        print(f"Scraping {i + 1}/{len(companies_data)}:
{company_name}")
        official_link = scrape_official_link(company_detail_url)
        companies_data[i] += (official_link,)

    return companies_data

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store company data (names, links, and official links)
all_companies_data = []

# Iterate over each page and scrape company data
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    companies_data = scrape_company_data(current_url)
    all_companies_data.extend(companies_data)
```

```
# Save all company data to CSV file
csv_file_path = 'all_companies_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name', 'Company Link', 'Official
Link'])

    # Write the company data to the CSV file
    for company in all_companies_data:
        csv_writer.writerow(company)

print(f"All company data (names, links, and official links) have been
successfully written to {csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company names, links, and official links from a
given URL
def scrape_company_data(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company names and links
    company_names = []
    company_links = []
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
    for company in company_list:
        company_name = company.get_text(strip=True)
        company_link_element = company.find("a")
        company_link = company_link_element["href"] if
company_link_element else None
        company_names.append(company_name)
        company_links.append(company_link)

    # Scraping official links from company detail pages
    official_links = []
    for company_link in company_links:
        if company_link:
            company_detail_page = requests.get(company_link)
            company_detail_soup =
BeautifulSoup(company_detail_page.text, 'html.parser')
```

```
            official_link_element = company_detail_soup.find("a",
text="Official Website")
            official_link = official_link_element["href"] if
official_link_element else "Not available"
            official_links.append(official_link)
        else:
            official_links.append("Not available")

    # Combine names, links, and official links
    companies_data = list(zip(company_names, company_links,
official_links))

    return companies_data

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store company data (names, links, and official links)
all_companies_data = []

# Iterate over each page and scrape company data
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    companies_data = scrape_company_data(current_url)
    all_companies_data.extend(companies_data)

# Save all company data to CSV file
csv_file_path = 'all_companies_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name', 'Company Link', 'Official
Link'])

    # Write the company data to the CSV file
    for company in all_companies_data:
        csv_writer.writerow(company)

print(f"All company data (names, links, and official links) have been
successfully written to {csv_file_path}.")
```

```
from bs4 import BeautifulSoup


# Parse the HTML content
soup = BeautifulSoup(html, 'html.parser')

# Find the anchor tag with the specified class
company_link_element = soup.find('a', class_='info-item company-
website-link')

# Extract the href attribute (company link)
company_link = company_link_element['href'] if company_link_element
else None

# Print the company link
print("Company Link:", company_link)
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company links from a given URL
def scrape_company_links(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company links
    company_links = []
    company_elements = soup.find_all("a", class_="btn btn-lg btn-
outline-primary w-100")
    company_links = [company.get("href") for company in
company_elements]

    return company_links

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store company links
all_company_links = []

# Iterate over each page and scrape company links
```

```
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    company_links = scrape_company_links(current_url)
    all_company_links.extend(company_links)

# Print all company links
for i, link in enumerate(all_company_links, start=1):
    print(f"Company Link {i}: {link}")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company links from a given URL
def scrape_company_links(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company links
    company_links = []
    company_elements = soup.find_all("a", class_="btn btn-lg btn-
outline-primary w-100")

    for company in company_elements:
        try:
            company_link = company.get("href")
            company_links.append(company_link)
        except TypeError:
            print(f"Failed to retrieve data from {url}")
            company_links.append(None)

    return company_links

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 382

# List to store company links
all_company_links = []

# Iterate over each page and scrape company links
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
```

```
    company_links = scrape_company_links(current_url)
    all_company_links.extend(company_links)

# Print all company links
for i, link in enumerate(all_company_links, start=1):
    print(f"Company Link {i}: {link}")
```

```
import requests
from bs4 import BeautifulSoup
import csv

def scrape_company_links(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company links
    company_links = []
    company_elements = soup.find_all("a", class_="btn btn-lg btn-
outline-primary w-100")
    company_links = [company.get("href") for company in
company_elements]

    return company_links

def scrape_official_link(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Find the anchor tag with the specified class
    company_link_element = soup.find('a', class_='info-item company-
website-link')

    # Extract the href attribute (company link)
    company_link = company_link_element['href'] if company_link_element
else None

    return company_link

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (382 in this case)
total_pages = 3

# List to store company links
```

```
all_company_links = []

# Iterate over each page and scrape company links
for page_number in range(1, total_pages + 1):
    current_url = base_url.format(page_number)
    company_links = scrape_company_links(current_url)
    all_company_links.extend(company_links)

# List to store company data (names, links, and official links)
all_companies_data = []

# Iterate through the list of company links and scrape official company
links
for company_link in all_company_links:
    official_link = scrape_official_link(company_link)
    all_companies_data.append((company_link, official_link))

# Save all company data to CSV file
csv_file_path = 'all_companies_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Link', 'Official Link'])

    # Write the company data to the CSV file
    for company_data in all_companies_data:
        csv_writer.writerow(company_data)

print(f"All company data (links and official links) have been
successfully written to {csv_file_path}.")
```

```
import requests
from bs4 import BeautifulSoup
import csv

# Function to scrape company names and links from a given URL
def scrape_company_data(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'html.parser')

    # Scraping company names
    company_names = []
    company_list = soup.find_all("h2", class_="fw-extrabold fs-xl
hover-underline d-inline-block company-title-clamp mb-0")
```

```
    company_names = [company.get_text(strip=True) for company in
company_list]

    # Scraping company links
    company_links = []
    company_elements = soup.find_all("a", class_="btn btn-lg btn-
outline-primary w-100")
    company_links = [company.get("href") for company in
company_elements]

    return list(zip(company_names, company_links))

# Base URL
base_url = 'https://www.builtinla.com/companies?page={}'

# Number of pages to scrape (let's scrape the first 10 pages, each page
contains around 100 companies)
pages_to_scrape = 10

# List to store company data (names and links)
all_companies_data = []

# Iterate over each page and scrape company data
for page_number in range(1, pages_to_scrape + 1):
    current_url = base_url.format(page_number)
    companies_data = scrape_company_data(current_url)
    all_companies_data.extend(companies_data)

# Save all company data to CSV file
csv_file_path = 'all_companies_data.csv'
with open(csv_file_path, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)

    # Write the header row
    csv_writer.writerow(['Company Name', 'Company Link'])

    # Write the company data to the CSV file
    for company in all_companies_data:
        csv_writer.writerow(company)

print(f"The first 1000 companies' names and links have been
successfully written to {csv_file_path}.")
```