Handbook: Web Scraping 102

Last week, we covered the fundamentals of web scraping. This week, we will continue to scrape news websites.

Today, we will cover:

1. How to webscrape
2. How to clean the scraped data..
3. How to contribute a dataset to AI Characters.

Before we start, if you need to do some revision in our previous session (Web scraping 101), kindly go to the link below:
https://shendai.notion.site/Workshop-Web-Scraping-101-94e3792315ea4a28875da7133f5e0331?pvs=4

How to Web Scrape News?

Step 1: Choose your AI Character to Contribute

Character list: Web Scraping 102

Step 2: Search for News articles for the Character Chosen.

1. Find a SINGLE News source, means the news for the character should be coming only from the Media Company. I strongly suggest you can take directly from:
   - BBC
   - CNN
   - CNBC
2. Search for news articles that have long paragraphs, try to avoid video news sources.
3. Find 10-15 news, that are related to the character.

Step 3: Open Google Collab file, Make a Copy

1. Open the link below: Web Scraping 102 (Google Collab)
   https://colab.research.google.com/drive/1xAnv2Km9XUMlLQQIPiSLt8QFbTGdKarL
2. Make a copy of the Collab notebook (File > Save a copy in Drive)
3. Go to the new copy of notebook to start scraping

Step 4: Scrape your News

1. Insert the news source (link) into the List of URLs.

   The example format to put in:
   Urls = [
   ‘News Link 1 ’ ,
   ‘News Link 2 ’ ,
   ‘News Link 3 ’ ,
   …
   ‘News Link 15 ’,
   ]

```
            content = '\n'.join(content)

            articles_data.append({'URL': url, 'Title': title, 'Content': co
        return articles_data

# List of URLs
urls = [
    'https://www.bbc.com/news/articles/ce448zzwp2go',
    'https://www.bbc.com/news/articles/cjqqkjy41zno',
    'https://www.bbc.com/news/articles/c977njnvq2do',
]

# Scrape article data
articles_data = scrape_article_data(urls)

# Define the CSV file path
csv_file_path = 'scraped_data2.csv'
```

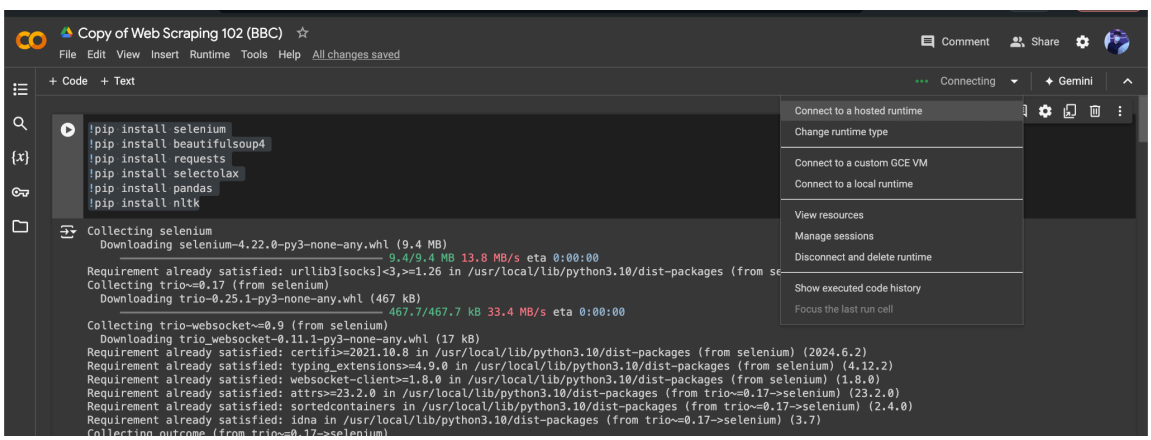
2. Replace the line of code based on which media company source you are taking from:

```
if title_tag:
    title = title_tag.text
else:
    title = 'N/A'

# Extract text from all paragraphs within the main article content
# If news source is BCC:
paragraphs = soup.find_all('div', attrs={'data-component': 'text-block', 'class': 'sc-43e6b7ba-0 bWSguZ'})
#If News source is CNN, replace 'paragraphs' by using code below:
'''
paragraphs = soup.select('div[class*="article"] p')
'''
#If News source is CNBC, replace 'paragraphs' by using code below:
'''
paragraphs = soup.find_all('div', class_ = 'group')
'''


content = []
for div in paragraphs:
    content.extend([p.text for p in div.find_all('p')])
```

3. Connect to hosted runtime (Connect > Connect to a hosted runtime)



4. Run from the first set of code (Installing Package), until the last one

Note:
- When using Google Collab Notebook, everytime when you have restarted the process, you have to start from the beginning (first set).
- The web scraping has been divided into a few sets of codes. This is because:
    1. Easier for participants to understand
    2. Lesser overwhelmed runtime, shorter the duration.
    3. Easier to check which dataset goes wrong during the process.

5. Once you have gone through all the process, Congrats you have successfully done Web Scraping Media articles ✅

   And you will need to download the csv file, rename it to “<your character>-scraped”
   For example: “JohnCena-scraped”

Step 5. Train the AI Agent with the dataset prepared

1. In order to train your AI character, you will need to log in with your account. You can log in with either: (a) email, (b) crypto wallet.
2. Now once you have logged in an account, head to Virtual Protocol to train your character with the prepared character card.
   Link: https://app.virtuals.io/contribution
3. Go to ‘Search Virtuals’ > Type in your AI Agent chosen > Contribute > Cognitive Core > Dataset
4. Follow the instruction below:

Type	Text
Package Name (must be same as your csv file name)	
Description:	This is the Dataset for AI Agent - John Cena . I have performed web scraping the information from 100+ of internet sources. Contributed by: Shuenrui

After filing in the information, click on “Submit”.
Congratulations 🎇, you have submitted the character card for your AI Agent. The character card will be used to train the agent, after the voting is passed.