Cristelle Andrea Aguilar

Extracting and Analyzing Data Using Regular Expressions

```
!pip install requests
!pip install beautifulsoup4
!pip install pandas
!pip install selenium
!pip install chromedriver_autoinstaller
!pip install chromedriver-binary
```

[1] ✓ 1m 52.2s Python

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: requests in c:\users\laptop\appdata\local\packages\pythonsoftwarefoundation.python.3.12_qbz5n2kfra8p0\localcache\local-packages\pyt
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\laptop\appdata\local\packages\pythonsoftwarefoundation.python.3.12_qbz5n2kfra8p0\localcache\lo
Requirement already satisfied: idna<4,>=2.5 in c:\users\laptop\appdata\local\packages\pythonsoftwarefoundation.python.3.12_qbz5n2kfra8p0\localcache\local-packages
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\laptop\appdata\local\packages\pythonsoftwarefoundation.python.3.12_qbz5n2kfra8p0\localcache\local-pa
Requirement already satisfied: certifi>=2017.4.17 in c:\users\laptop\appdata\local\packages\pythonsoftwarefoundation.python.3.12_qbz5n2kfra8p0\localcache\local-pa
Defaulting to user installation because normal site-packages is not writeable
Collecting beautifulsoup4
  Downloading beautifulsoup4-4.12.3-py3-none-any.whl.metadata (3.8 kB)
Collecting soupsieve>1.2 (from beautifulsoup4)
  Downloading soupsieve-2.5-py3-none-any.whl.metadata (4.7 kB)
Downloading beautifulsoup4-4.12.3-py3-none-any.whl (147 kB)
   ---------------------------------------- 0.0/147.9 kB ? eta -:--:--
   -------- ------------------------------- 30.7/147.9 kB ? eta -:--:--
   ------------------------------ --------- 112.6/147.9 kB 1.3 MB/s eta 0:00:01
   ---------------------------------------- 147.9/147.9 kB 1.3 MB/s eta 0:00:00
Downloading soupsieve-2.5-py3-none-any.whl (36 kB)
Installing collected packages: soupsieve, beautifulsoup4
Successfully installed beautifulsoup4-4.12.3 soupsieve-2.5
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pandas in c:\users\laptop\appdata\local\packages\pythonsoftwarefoundation.python.3.12_qbz5n2kfra8p0\localcache\local-packages\pytho
```

```
from bs4 import BeautifulSoup
import pandas as pd
from selenium import webdriver
from selenium.webdriver.firefox.options import Options
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import re
# Set up Selenium
options = Options()
options.set_preference("general.useragent.override", "Mozilla/5.0 (Windows
NT 10.0; Win64; x64; rv:58.0) Gecko/20100101 Firefox/58.0")

driver = webdriver.Firefox(options=options)

# URLs
faq_url = "https://www.packtpub.com/en-us/help/faqs"
terms_url = "https://www.packtpub.com/en-us/help/terms-and-conditions"

# Fetch and parse the FAQ page
driver.get(faq_url)

# Wait for the FAQs to load
```

```
WebDriverWait(driver,
10).until(EC.presence_of_element_located((By.TAG_NAME, 'html')))

# Get the HTML content
faq_html = driver.page_source

# Parse the HTML content using BeautifulSoup
faq_soup = BeautifulSoup(faq_html, 'html.parser')

# Extract FAQs and their answers
def extract_faqs(soup):
    faqs = []
    cards = soup.find_all('div', class_='card')
    for card in cards:
        question_button = card.find('button', class_='btn btn-info
collapsed', type='button')
        if question_button:
            question = question_button.text.strip()
            answer_p = card.find('p', class_='card-text')
            if answer_p:
                answer = answer_p.text.strip()
                faqs.append({'Question': question, 'Answer': answer})
    return faqs

faqs = extract_faqs(faq_soup)

# Create a DataFrame for FAQs
faq_df = pd.DataFrame(faqs)
print("FAQs and their answers:")
print(faq_df)

# Fetch and parse the terms and conditions page
driver.get(terms_url)
terms_html = WebDriverWait(driver,
10).until(EC.presence_of_element_located((By.TAG_NAME,
'html'))).get_attribute('outerHTML')
terms_soup = BeautifulSoup(terms_html, 'html.parser')

# Extract email addresses and phone numbers using regular expressions
terms_text = terms_soup.get_text()
```

```
emails = re.findall(r'\b[A-Za-z0-9._%+-]+@packt+\.[A-Z|a-z]{2,}\b',
terms_text)
phone_numbers = re.findall(r'\+?\d[\d -]{8,}\d', terms_text)

print("\nEmail addresses found:")
for email in emails:
    print(email)

print("\nPhone numbers found:")
for phone_number in phone_numbers:
    print(phone_number)

# Close the Selenium driver
driver.quit()
```

```
FAQs and their answers:
                                        Question  \
0                How can I cancel my subscription?
1  What happens when I cancel my Packt subscription?
2                               What are credits?
3  How can I download a video package for offline...
4                 How can I extract my video file?
5  How can I get help and support around my video...
6            Why can't I download my video package?
7  What happens if an Early Access Course is canc...
8  Where can I send feedback about an Early Acces...
9  Can I download the code files for Early Access...

                                              Answer
0  To cancel your subscription with us simply go ...
1  If you cancel your subscription, you lose your...
2  Credits can be earned from reading 40 section ...
3  Login to your account at Packtpub.com.Click on...
4  All modern operating systems ship with ZIP fil...
5  If your video course doesn't give you what you...
6  In the even that you are having issues downloa...
7  Projects are rarely cancelled, but sometimes i...
8  If you have any feedback about the product you...
9  We try to ensure that all books in Early Acces...
...

Phone numbers found:
121 265 6484
121 212 1419
```