Web Scraping: A Key Tool in Data Science

Estimated Effort: 5 mins

Introduction

Web scraping, also known as web harvesting or web data extraction, is a technique used to extract large amounts of data from websites. The data on websites is unstructured, and web scraping enables us to convert it into a structured form.

Importance of Web Scraping in Data Science

In the field of data science, web scraping plays an integral role. It is used for various purposes such as:

1. **Data Collection:** Web scraping is a primary method of collecting data from the internet. This data can be used for analysis, research, etc.
2. **Real-time Application:** Web scraping is used for real-time applications like weather updates, price comparison, etc.
3. **Machine Learning:** Web scraping provides the data needed to train machine learning models.

Web Scraping with Python

Python provides several libraries for web scraping. Here are some of them:

1. **BeautifulSoup:** BeautifulSoup is a Python library used for web scraping purposes to pull the data out of HTML and XML files. It creates a parse tree from page source code that can be used to extract data in a hierarchical and more readable manner.

```
from bs4 import BeautifulSoup
import requests
URL = "http://www.example.com"
page = requests.get(URL)
soup = BeautifulSoup(page.content, "html.parser")
```

2. **Scrapy:** Scrapy is an open-source and collaborative web crawling framework for Python. It is used to extract the data from the website.

```
import scrapy
class QuotesSpider(scrapy.Spider):
    name = "quotes"
    start_urls = ['http://quotes.toscrape.com/tag/humor/',]
    def parse(self, response):
        for quote in response.css('div.quote'):
            yield {'quote': quote.css('span.text::text').get()}
```

3. **Selenium:** Selenium is a tool used for controlling web browsers through programs and automating browser tasks.

```
from selenium import webdriver
driver = webdriver.Firefox()
driver.get("http://www.example.com")
```

Applications of Web Scraping

Web scraping is used in various fields and has many applications:

1. **Price Comparison:** Services such as ParseHub use web scraping to collect data from online shopping websites and use it to compare the prices of products.

2. **Email address gathering:** Many companies that use email as a medium for marketing, use web scraping to collect email ID and then send bulk emails.

3. **Social Media Scraping:** Web scraping is used to collect data from Social Media websites such as Twitter to find out what's trending.

Conclusion

Web scraping is an essential skill in the fast-growing world of data science. It provides the ability to turn the web into a source of data that can be analyzed, processed, and used for a variety of applications. However, it's important to remember that one should use web scraping responsibly and ethically, respecting the terms of use or robots.txt files of the websites being scraped.

Author(s)

Abhishek Gagneja

Skills Network