09.05.2024

Web Scraping ve Veri Toplama

Kaan Çardak
Bilgisayar Mühendisliği
20360859087

İçindekiler

1. Giriş

Web Scraping nedir?

Web scraping, web sayfalarından veri toplama sürecidir. Genellikle Bir bilgisayar programı veya bir bot yardımıyla gerçekleştirilen bu işlem, belirli bir web sayfasının HTML yapısını analiz ederek istenilen bilgileri çıkarmayı içerir. Web scraping genellikle veri madenciliği, makine öğrenimi, analiz veya raporlama gibi amaçlarla kullanılır.

2. Temel Kavramlar

HTTP ve HTTPS protokollerinin önemi

Web scraping yaparken kullanılan temel HTTP istekleri şunlardır:

- GET
- POST
- HEAD
- PUT, DELETE, OPTIONS, PATCH vb.

2. Temel Kavramlar

HTML temelleri

- <html>, <head>, <body>: Temel belge yapısını tanımlar
- , : Blok ve satır düzeyinde içerikleri gruplar
- <p>, <h1>-<h6>: Metin içeriğini ve başlıkları tanımlar

3. Web Scraping Araçları

BeautifulSoup, Scrapy ve Selenium

BeautifulSoup, python tabanlı, HTML ve XML analiz kütüphanesidir. Kullanımı kolaydır. Veri çekmek için, BeautifulSoup ile ilgili URL'yi belirtir, belgeyi analiz eder ve istediğimiz verileri çekmek için belirli öğeleri kullanabiliriz.

Scrapy, python tabanlı bir web scraping framework'ü, daha büyük ve karmaşık verileri toplama projeleri için idealdir. Scrapy ile veri toplamak için bir örümcek oluştururuz. Örümcek, belirli URL'leri ziyaret eder ve verileri çeker. Paralel veri indirme ve otomatik tekrar deneme başlıca avantajlarındandır.

3. Web Scraping Araçları

BeautifulSoup, Scrapy ve Selenium

Selenium, asıl amacı web tarayıcılar üzerinde otomatik testler yapmaktır fakat web scraping için de kullanılabilir. Dinamik web sayfalarını otomatik olarak gezerek veri toplamak için kullanılır. Tarayıcılarda tıpkı gerçek bir kullanıcı gibi davranarak veri topladığı için JavaScript tabanlı web sitelerinde daha esnek ve geniş bir kullanım alanı sağlar. Ancak, genelde daha fazla kaynak tüketir ve daha yavaştır.

4. Etik ve Yasal Konular

robots.txt ve sitelerin kullanım koşulları

- **robots.txt dosyası:** Bir web sitesindeki robots.txt dosyası, web tarayıcısı ve botlara hangi sayfaların taranabileceğini ve hangilerinin taranamayacağı hakkında bilgi veren bir dosyadır. Bu dosyadaki yönergeleri izlemek ve kısıtlamalara uymak, web scraping sonrasında yasal bir sorunla karşılaşmamak için önemlidir.

5. Web Scraping Yaygın Kullanım Alanları

- **Haber Sitelerinden Veri Toplama**

  Güncel haberleri takip etmek ve trendleri analiz etmek için kullanılır.

- **E-Ticaret Veri Analizi**

  Piyasadaki verileri analiz ederek, rakip firmalara göre strateji oluşturmada kullanılır.

- **Sosyal Medya Veri İşleme**

  Sosyal medya platformlarından kullanıcı verilerini toplamak ve bu verileri işleyerek kullanıcı profilini analiz etmek için kullanılır.

6. Veri Temizleme ve Analiz

Veri Temizleme Yöntemleri

- Veri Düzenleme
- Veri Temizleme
- Veri Formatlama

Analiz Araçları

- Microsoft Excel
- Pandas
- R programlama dili
- MATLAB

7. Örnek Uygulama

Steam URL'si girilen oyunun piyasaya çıkış tarihini veren basit web scraping uygulaması

```python
import requests
from bs4 import BeautifulSoup

def get_game_release_date(game_url):
    response = requests.get(game_url)
    soup = BeautifulSoup(response.content, 'html.parser')

    release_date_container = soup.find('div', class_='date')
    if release_date_container:
        release_date = release_date_container.text.strip()
        return release_date
    else:
        return "Cikis tarihi bulunamadi."

game_url = input("Steam oyunu URL'sini girin: ")

release_date = get_game_release_date(game_url)
print(f"Oyun cikis tarihi: {release_date}")
```

```
PS C:\Users\user\Desktop\web tabanlı programlama> & C:/Users/user/AppData/Local/Microsoft/WindowsApps/python3.11.exe
Steam oyunu URL'sini girin: https://store.steampowered.com/app/271590/Grand_Theft_Auto_V/
Oyun cikis tarihi: 14 Apr, 2015
```

8. Kaynaklar

- https://medium.com/kaveai/web-scraping-453e96a86195
- https://veribilimiokulu.com
- https://codiasoft.com/blog/web-scraping-nedir-neden-yapilir

Sorularınız

Dinlediğiniz için teşekkürler