Modül 1 - Day 2

Modül-1: Data Collection

İçerik: API'lar, loglama, sensory data, web scraping.

Anahtar sözcükler: JSON, XML, HTTP, HTML, DOM, grep, RegExp.

Araçlar: Postman, log4j, python-logging, BeautifulSoup, Jsoup, Selenium

HTTP

“http”, bilginin sunucudan kullanıcıya nasıl ve ne şekilde aktarılacağını gösteren protokoldür. İnternet ağında sunucular ve kullanıcılar arasında nasıl bir veri alışverişi olacağı hakkında kurallar vardır. Bu düzeni sağlayan protokol de HTTP’dir.internet sitesine girmek için adres çubuğuna sitenin adresini yazdığınız vakit HTTP ile sunucuya bir istek gönderilir ve sunucu bu isteğe cevap verdiği vakit internet sitesinin verileri size gelir.

🏛️ Restful Mimarisi

RESTful servisler veri iletiminde farklı HTTP metotlarını kullanmaktadır. Yapılan HTTP request’i için çağrılan URL ile beraber HTTP method bilgisi bahsi geçen 4 metottan biri olarak seçilir ve sunucu yapılan talebin kayıt üzerine nasıl etki edeceğini buna göre belirler.

> 👉 **GET:** Veri listeleme - veri görüntülemek için kullanılır.

> 👉 **POST:** Veri eklemek için kullanılır.

Diğer Metodlar

- **PUT:** Veriyi Güncelleme isteği olarak kullanılır.

- **PATCH:** Verinin sadece bir parçasını güncellemek için kullanılır. Örneğin bir issue'nun durumunun aktiften çözüldü haline getirilmesi.
- **DELETE:** Veriyi silmek için kullanılır.
- **OPTIONS:** Bir api urline Options isteği yapıldığında o url in hangi istekleri kabul ettiği bilgisi verilir.

 httpbin.org sitesinden bu denemeler yapılabilir

HTTP Status kodları

Kodlara linkten ulaşılabilir. https://www.argenova.com.tr/http-durum-ve-hata-kodlari

En sık karşılaşılan hata kodları

- **HTTP 200 (OK):** yanıtın başarılı olduğunu gösterir. Yani, istemci ile sunucu arasındaki iletişim herhangi bir hata olmadan sorunsuz bir şekilde yürütülmüştür.
- **HTTP 404 (Not Found):** istenen kaynağın sunucu tarafından bulunamayacağı anlamına gelir. Bu, geçici bir aksaklıktan kaynaklanıyor olabilir ve gelecekte başka bir istekte bulunulursa kaynak kullanılabilir olabilir. Çoğunlukla, 404'e götüren bağlantılara genellikle bozuk bağlantılar denir.
- **HTTP 502 (Bad Gateway):** sunucunun proxy olarak hareket ederken istekte bulunurken sunucudan geçersiz bir yanıt aldığını gösterir.

🥌 CURL

Çoğu Unix bazlı sistemde mevcut olan bir komuttur ve “Client URL”nin kısaltılmışıdır. Curl komutları URL’lerin bağlanabilirliğini kontrol etmek ve veri transferi için harika bir araç olarak kullanılmak için üretilmiştir.

Basit Curl Command Sözdizimi

Hadi Curl komutlarını nasıl kullanacağımızı öğrenelim. Curl’ün basit sözdizimi böyledir:

```
curl [OPTIONS] [URL]
```

Curl'ün en basit kullanımı bir sayfanın içeriklerini göstermektir. Aşağıdaki örnek testalanadi.com'un ana sayfasının içeriklerini gösterecektir:

```
curl testalanadi.com
```

Derste kullanılan örnekler

```
curl -x GET "http://httpbin.org/get" -H "accept: application/json"
```

 ipify.org

https://api.ipify.org?format=json&param=2

URL içerisindeki özel karakterler

- **"?":** "sorgu parametreleri" olarak adlandırılır. Sunucu, bu ek bilgilere dinamik olarak cevap verebilir ve göreceğiniz sayfayı bu parametrelere göre oluşturabilir. Bu, sayfadaki bir alana otomatik olarak yazılmış bir bilgi veya bir arama motorundaki arama sorgunuz olabilir.
- **"%":** "Escaping" olarak adlandırılan bu işlem, URL'deki boşluk karakterinin soruna yol açmaması için alternatif bir biçime dönüştürülmesidir. Örneğin + yazdığınızda bu karakter, %3F ile değiştirilir.
- **"=":** anahtar-değer çiftlerini temsil eder. Anahtar-değer çiftine bir örnek, sayfa=4 olabilir. Burada "sayfa" anahtar, "4" ise değerdir.

🚀 GREP

Elimizdeki loglarda buradan terminal ekranında hızlıca sorgularımız getirebiliriz. Logların güzel şekilde tutulması işimizi kolaylaştırıyor. Pipe ile çok daha etkin kullanım yöntemleri mevcut.

💡 grep "42.236.10.125" * —color | wc -lw

📚 Ders Örnekleri:

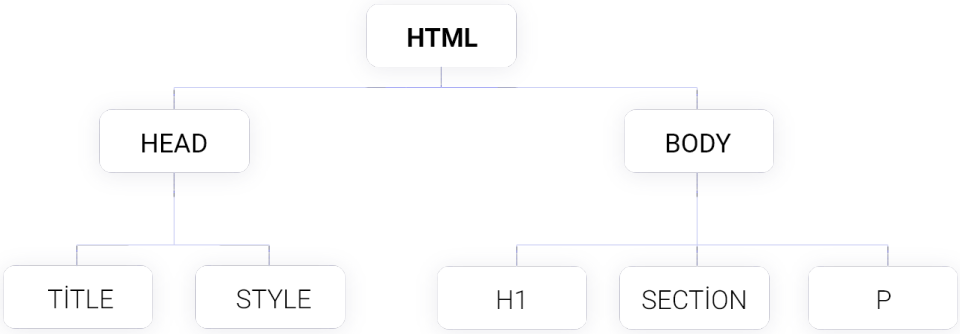
grep "GET" * —color

grep "42.236.10.125" * —color | wc -lw

grep "mozilla" * —color

🌲 DOM

HTML için kullanılan doküman nesne modelidir. HTML Elementlerini objeler olarak, HTML elementlerinin tüm özelliklerini, HTML elementlerine erişmek için metotları, tüm HTML elementleri için olayları tanımlar. Diğer bir deyişle HTML DOM yeni elementler eklemek, elementleri değiştirmek veya silmek için kullanılır.




```
<html>
<head>
    <title></title>
    <style></style>
</head>
```

```
<body>
   <div>
      <p>Cobanov</p>
   </div>
</body>
</html>
```

Python Web Scraping

Logging

Kaynak: https://medium.com/@umut.boz/python-logging-a8fdd36fee7

Loglama, bir sistemdeki hareketliliği kaydetmek için kullanılan yapıdır. Python standart kütüphanesi içinde loglama için çok güçlü bir kütüphane barındırır. Bu kütüphane ile geliştirdiğimiz programlarda hata ayıklamak aynı zamanda ifadeleri yazdırmak için loglama kullanabiliriz.

Requests

Kaynak: https://medium.com/python/python-requests-modülü-4af79ebdb8c8

Python, standart modüllerinin yanında harici yüzlerce kullanışlı modül ile birlikte çok güçlü bir dil. Bu gücü veren harika modüller var bunlardan biri de **Requests modülü.**

Bu modül ile web üzerindeki isteklerinizi yöneteceksiniz. Mesela bu modül ile API entpointlerine PUT, DELETE, POST gibi istekler atabilirsiniz.

BeautifulSoup4

Kaynak: https://medium.com/@nuriyavuz2.71/python-beautifulsoup-modülü-689ef499ee16

BeautifulSoup, HTML veya XML dosyalarını işlemek için oluşturulmuş güçlü ve hızlı bir kütüphanedir.

Bu modül ile bir kaynak içerisindeki HTML kodlarını parse edip,botlar yazabiliriz.

Selenium

Kaynak: https://www.sinanerdinc.com/python-selenium-modulu-kullanimi-1

Selenium, bilgisayarınıza yükleyeceğiniz bir driver yardımı ile ekrana chrome, firefox gibi bir tarayıcı açarak, gerçek bir insan gibi istediğiniz tüm işlemleri programlama dili yardımıyla çalıştırmanızı sağlayan bir araçtır.

☝️ **Homework:** Vatan bilgisayardaki ürünlerin görsellerin veya isimleri ile ücretlerini scrap edebilirsin.

💡 **Ödev Linki:** https://github.com/cobanov/dataeng-bootcamp/blob/main/homeworks/scraping_homework.py

🔗 GISTLER:

- **BS4**: https://gist.github.com/sirin/695abacaa207ad7af20f18c946d19858
- **Selenium**: https://gist.github.com/sirin/0e1491163b8f485a476e0991ad228b86

End of the second day!