BİTİRME PROJESİ

Haftalık Rapor – 18.03.2022

18 MART 2022

KIRIKKALE ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ İÖ

AHMET MUNGAN – 160255081

İÇİNDEKİLER

ÖZET

Metin madenciliği sürecinde kilit rol oynayan veri seti oluşturma işlemlerinin başlangıcı yapılmıştır. Veri seti neden oluşturulmalı, ne gibi şartları sağlamalı, oluştururken dikkat edilmesi gerekenler ve veriye yönelik rıza nelerdir gibi sorulara cevap verilmiştir. Veri seti oluşturma işlemi başlanıp, ilk işlemler Python aracılığı ile yapılmıştır. Çıktılar gözlenip, veri seti oluşturma işlemlerinde program koduyla hedef ve amaçların nasıl sağlanacağı belirlenmiştir.

VERİ SETİ OLUŞTURMA İŞLEMLERİ

Metin madenciliğinde yapılandırılmış veri seti bulmak mümkündür. Fakat beklenmedik durumların da oluşabileceği, her zaman yapılandırılmış veri seti bulmanın zorluğu, piyasadaki veri setlerinin büyük bir kısmının yapılandırılmamış veri halinde bulunduğu gibi sebeplerden ötürü hazır veri setleri büyük bir anlam ifade etmemektedir. Dolayısıyla sürecin tamamını anlamak açısından, tabiri caizse sıfırdan sisteme kadar olan süreçleri deneyimlemek açısından veri setini bu proje kapsamında veriyi toplayarak elde etmek avantajlı olacaktır. Bu sayede Python dilinin getirdiği avantajları ve veri toplarken kaynakların sınırsızlığının esnekliği kullanılarak bu süreci yönetmek zor olacaktır. Fakat bu zorluk projenin her noktasına da bir o kadar hakim olunacağı gerçeğini değiştirmemekle birlikte, sürecin performans ve verim gibi başarı ölçütlerine etkisi de yüksek olması beklenmektedir.

Veri Toplama

Dünyadaki en büyük e-ticaret platformu olan Amazon'un verileri toplanacaktır [1]. Amazon'un seçilmesinde kapsamı geniş ve farklı kalemde ürün çeşitlerinin olmasıdır. Fakat tüm ürün kalemleri ele alınacak olursa ve sadece ürünler incelenecek olursa metin madenciliği bu işin arka planında kalacaktır. Dolayısıyla belirli bir ürün kalemine gelen kullanıcı/müşteri yorumlarının üzerinde durulacaktır. Bu sayede yorumlara bakılarak ürünlerden belirli çıkarımlar ile madencilik sürecine katkı sağlaması beklenen verilerin toplanma aşaması gerçekleştirilecektir.

İlk etapta; Amazon'un en çok satan elektronik ürünlerinden olan telefonlar üzerinde durulacaktır [2]. Telefonlar birçok kapsamda karşılaştırılarak belirli formatlarda bir çıktı elde edilebilir fakat burada telefonların özellikleri genel itibariyle yardımcı öznitelikler olacaktır. Bu sayede metin madenciliği ile hedeflenen kısımlara daha fazla odaklanılmış olacaktır. Ayrıca Amazon üzerinde bu kısmın seçilmesinin bir diğer sebebi ise, Amazon her alt sitesine erişimi serbest kılmamıştır. Dolayısıyla erişime açık, yani Amazon'un rızası olan kısımların kullanılabileceği unutulmamalıdır. Aksi takdirde *503 Service Unavailable* hatası alınabilir.

Code 1

```
>>> import requests
>>> r = requests.get("https://www.amazon.com.tr/s/ref=nb_sb_noss?__mk_tr
_TR=%C3%85M%C3%85%C5% ... C1384")
>>> r
<Response [503]>
```

Code 1'de[1] Amazon'un özel bir alt sitesine istek gönderildiğinde cevap olarak 503 kodlu bir çıktı alındığı görülmektedir. Code 1'deki örnek işte tam bu sebepten dolayı bazı alt sitelere erişimi engellediği için, izin verilen ölçüde veri toplama gerçekleştirilmelidir.

Bu durumlar göz önüne alınarak en çok satan telefonlar üzerinde yoğunlaşılırsa veri toplama işlemi başarılı olabilir.

Code 2

```
>>> from bs4 import BeautifulSoup
>>> import requests
>>> istek =
requests.get("https://www.amazon.com.tr/gp/bestsellers/electronics/13709
907031?ref ... wg=c6bmH")
>>> istek
<Response [200]>
>>> soup = BeautifulSoup(istek.content, "lxml")
>>> soup.prettify()
'<!DOCTYPE html>\n<html class="a-no-js" data-19ax5a9jf="dingo" lang="tr-
tr"> ... </html>
```

Code 2'de[2] görüldüğü üzere bs4 kütüphanesinden tüm bu işleri kolaylaştıracak yardımcı modül BeautifulSoup kullanılmıştır. Ayrıca Amazon'dan alınan ve erişime izin verilen URL adresinin girilmesiyle ve *200 OK* onay cevabıyla birlikte siteye ait kodların sorunsuz bir şekilde elde ediliği görülmektedir.

---

[1] Code 1'de yazılan URL rapor sayfasında fazla yer kaplamaması açısından orta noktasından kesilerek üç nokta (…) ile ifade edilmiştir.
[2] Code 2'de yazılan URL ve web sitesine ait çıktılar, uzunluğu sebebiyle kesilerek üç nokta (…) ile ifade edilmiştir.

Code 3

```
>>> telefonlar = soup.find_all("div", attrs = {"id":"gridItemRoot"})
>>> telefonlar
[<div class="a-column a-span12 a-text-center _p13n-zg-list-grid-desktop
... </div></div></div></div></div>]
```

Code 3'te[3] görüldüğü üzere sitedeki ürünlerin genel gösterimli div bölgesine sınıfı vasıtasıyla erişilmiştir. Bu sınıfın adı sitenin kodu incelendiğinde bulunabildiği gibi, tarayıcıların web siteleri üzerinde tüm nesne ve dışarıdan bakılınca nesne gibi gözüken kısımlarına sağ tıklayıp incele seçeneğinde detaylar bulunabilir.

Code 4

```
>>> type(telefonlar)
bs4.element.ResultSet
```

Code 3'te yaratılan telefonlar değişkeninin içerisinde sadece kod barındırdığı düşünüldüğünde, üzerinde belli başlı oynamalar yapabilmek ve Python'un kullanışlı yapısından faydalanmak için bu değişkenin string tipinde olduğu öngörülebilir. Fakat Bu değişken BeautifulSoup modülünden kalıtım aldığından Code 4'te görüldüğü üzere değişken tipi modülün sonuç göstermeye yarayan özel bir listesi halinde olduğu ortadadır. Dolayısıyla elde edilen ürün kodlarının yani bu iş özelinde telefonlar değişkeninin string veya list tipine type casting[4] edilmesi mümkün değildir.

Code 3 ile elde edilen site kodlarının hiçbir anlam ifade etmediği ortadadır. Dolayısıyla burada bu kodların yardımıyla telefonlara ait clickable (tıklanabilir) varlıklar yakalanmalıdır. Daha sonra tıklanabilir varlıklara tıklandığında telefonlara ait yorumlar kısmına erişilip, bu kısımların işlenebilecek türden değişken tipleri ile muhafazası sağlanmalıdır.

[3] Code 3'te telefonlara ait kod kısmı uzunluğu sebebiyle üç nokta (…) ile ifade edilmiştir.
[4] Type casting; bir değişkenin zoraki kodlar ile tipinin değiştirilmesi olarak kastedilmiştir.

REFERANS VE KAYNAKÇA

[1] Amazon’un Türkiye Resmi Web Sitesi. Link için tıklayınız. (Güvenlidir.)

[2] amazon.com.tr alt sitelerinden olan bestseller/electronic/phones sayfası. Link için tıklayınız. (Güvenlidir.)

EKLER

Bitirme Projesi 2'ye ait doküman, haftalık rapor ve ek bilgilerin paylaşıldığı github linki için tıklayınız.