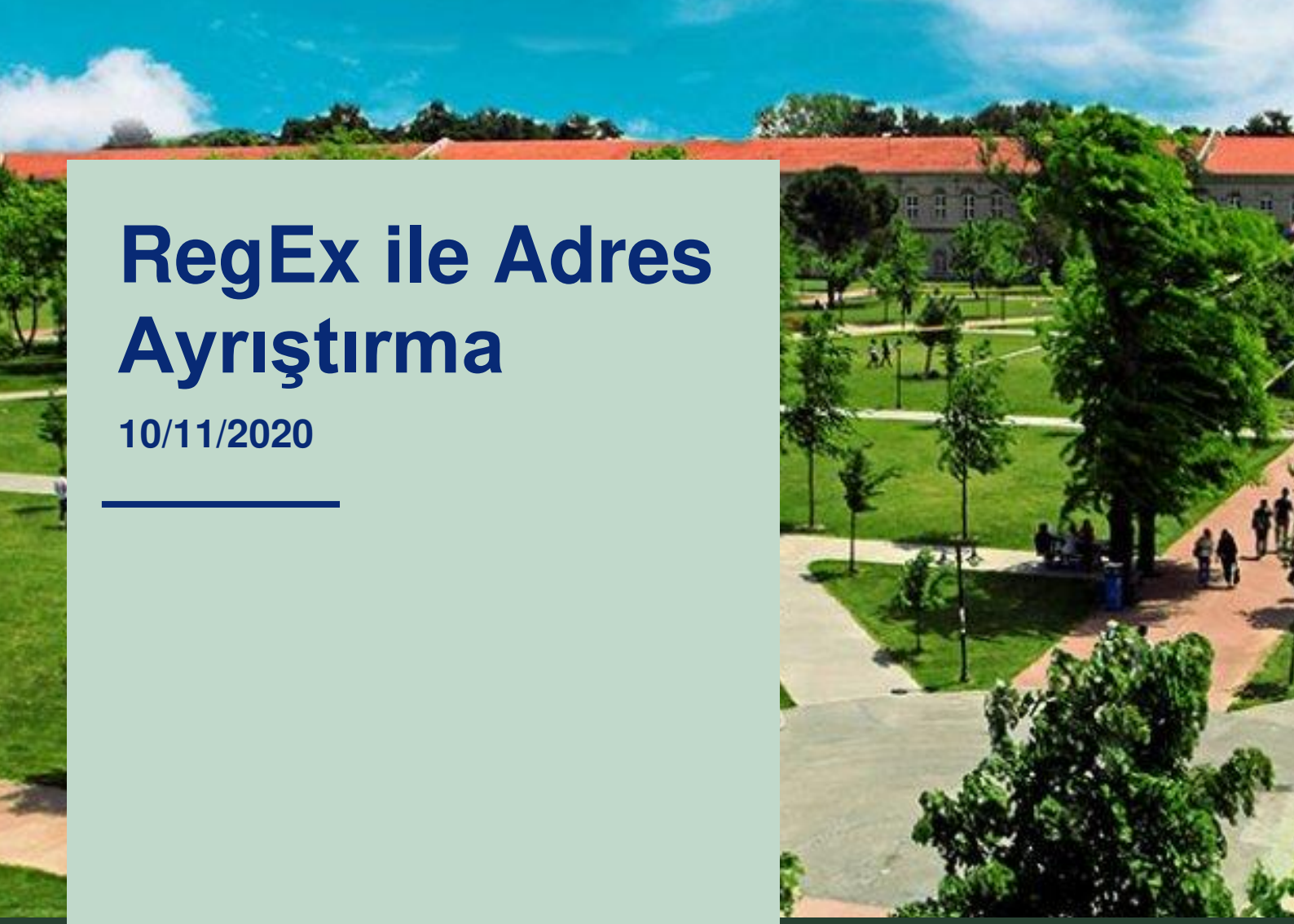


RegEx ile Adres Ayrıştırma

10/11/2020

BLM 4580 Doğal Dil İşlemeye Giriş

Banu Diri

Ahmet Onur Akman

16011059

YILDIZ TEKNİK ÜNİVERSİTESİ

1911

Giriş

Problem

Text biçiminde verilen veri içerisindeki adresleri ayrıştırmak amacına uygun bir Regular Expression oluşturmak. Bu adreslerin mahalle, cadde, sokak, kapı numarası, ilçe, şehir bilgilerini ayrıştırmak.

Veri

Elimizdeki veri text formatındadır ve 6831 adet satır içermektedir. Bu satırların her biri farklı bir adresi barındırmaktadır. Bu adreslerin büyük bir kısmı, Türkiye adres yazım alışkanlıklarına uyumlu olacak şekilde "Mahalle, Cadde, Sokak, Ek Bilgiler, Numara, Ek Bilgiler, İlçe, Şehir" formatında verilmiştir ancak oluşturulacak RegEx ile bu formata uymayan adres bilgilerinin de başarıyla okunup ayrıştırılması hedeflenmektedir.

Yaklaşım

Ele alınan problem bir metinin regular expression ile taranıp ayrıştırılmasını gerektirdiğinden, çalışma ortamının yazılan ifadenin syntax'ını kontrol edebilen, başarısını ölçebilen ve verilen veri üzerindeki taramasını ayrıntılarıyla bize gösterebilen bir ortam olması gerekliliği ortaya çıkmıştır. Bu gereklilik doğrultusunda www.regex101.com sitesi çalışma ortamı olarak seçilmiştir.

Türkiye'deki ev ve ofis adreslerinin yazım formatı başka ülkelere kıyasla daha karmaşık bir formattadır. Bu format doğrultusunda adresi verilecek mekanın konum bilgisinin bütün detayları adrese dahil edilir ve genellikle bu bilgiler büyük yerleşim yerinden daralarak yazılır. Bu ödev kapsamında ele alınmış olan adreslerin geneli de bu şekilde yazılmıştır.

Bazı istisnalar dışında, adreslerin genel yazılış tarzlarının şu ilerleyişe sadık kaldığı görülmektedir:
Mahalle, Cadde, Sokak, Numara, ilçe, Şehir

Bu formatın yanı sıra adres içerisinde site, compleks, iş hanı veya AVM ismi, şehirler arası yola göre tarif gibi detaylar da eklenmektedir. Bahsedilen bu detayların adreste yerleştirilebileceği her yerin ayrı ayrı taranması ve her ek bilginin neyi ifade ettiğinin çözümlenmeye çalışılması ifadeyi oldukça kalabalıklaştıracağından, adreslerdeki genel yazılış biçimi incelenmiş ve bu ek bilgileri en yüksek doğrulukta yakalamak ve kodu da basitlikten uzaklaştırmamak hedefiyle bir RegEx formatı belirlenmiştir. Buna göre yazılan RegEx'in formatı aşağıdaki gibi planlanmıştır.

Mahalle (Varsa), Cadde (Varsa), Sokak (Varsa), Ek Bilgiler (Varsa), Numara (Varsa), Ek bilgiler (Varsa), İlçe (Zorunlu), Şehir (Zorunlu)

Eldeki adres bilgilerinin tamamı İstanbul içi adresler olmasına rağmen oluşturulan RegEx'te bu durum gözardı edilmiştir.

Adresler arası ele alınması gereken başka bir problem de mahalle, sokak gibi kelimelerin yazılışlarının veya kısaltma biçimlerinin yazan kişinin tercihine bağlı olarak değişiyor olmasıdır. Yazılan RegEx ile muhtemel bütün kısaltmaların bulunması sağlanmıştır. Buna göre muhtemel olduğu varsayılan yazılışlar ve kısaltmalar şu şekildedir:

MH, MAH, MAHALLE, MAHALLESİ, MAHALLESI: **Mahalle**
CD, CAD, CADDE, CADDESİ, CADDESI: **Cadde**
SK, SOK, SOKAK, SOKAĞI, SOKAGI: **Sokak**
N., N:, NO., NO:, NR., NR:: **Numara**

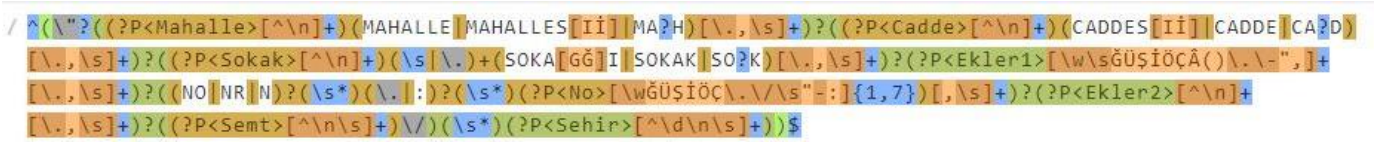
Bu ve bunun gibi adres parçalarının içerebileceği karakterler ve aralarına konulacak boşluk ve noktalama işaretleri düşünülerek her senaryoda başarılı olunması hedeflenmiştir.

Oluşturulan RegEx ve Başarı Oranı

Regular Expression

Yaklaşım bölümünde anlatılmış olan faktörler göz önünde bulundurularak aşağıdaki RegEx oluşturulmuştur:

a) Renklendirilmiş Format



b) Ayrıştırılmış Format

```
^(\"?
    ((?P<Mahalle>[^\n]+)(MAHALLE|MAHALLES[Iİ]|MA?H)[\.,\s]+)?
    ((?P<Cadde>[^\n]+)(CADDES[Iİ]|CADDE|CA?D)[\.,\s]+)?
    ((?P<Sokak>[^\n]+)(\s|\.)+(SOKA[GĞ]I|SOKAK|SO?K)[\.,\s]+)?
    (?P<Ekler1>[\w\sĞÜŞİÖÇÂ()\.\-",]+[\.,\s]+)?
    ((NO|NR|N)?(\s*)(\.|:)?(\s*)(?P<No>[\wĞÜŞİÖÇ\.\/\s"-:]{1,7})[,\s]+)?
    (?P<Ekler2>[^\n]+[\.,\s]+)?
    ((?P<Semt>[^\n\s]+)\/)(\s*)
    (?P<Sehir>[^\d\n\s]+)
)$
```

c) Regex101 Linki

regex101.com/r/TXOF5X/2

Başarı Durumu

Oluşturulan RegEx, verilen adres bilgilerinin tamamını ayrı ayrı olacak şekilde algılamayı başarmıştır. RegEx ile veri üzerinde yapılan tarama sonucunda, il ve ilçe bilgileri dahil olmak üzere en az 3 adres parçasına sahip **6831** adet adres tespit edilmiştir.

Kullanılan ortamın şartları sebebiyle, yapılan tarama sonucu bulunan mahalle, cadde, sokak, no, ilçe, şehir ve ek bilgiler ayrı ayrı olacak şekilde bir veri yapısında toplanamamıştır. Modelin doğruluk oranını daha iyi gözlemlemek için önceki sayfada verilen RegEx101 linkini ziyaret edebilirsiniz.

Muhtemel olarak değerlendirilen formatların dışında yazılmış olan, yazım hataları içeren ve diğer adreslere kıyasla beklenmedik bilgiler içeren adreslerin okumalarında yüksek bir başarı oranı elde edildiyse de, aşağıdaki 10 adet örneği verilen bazı adreslerin ayrıştırılmasında karışıklıklar meydana gelmiştir:

1) **KUYUMCU KENT ATÖLYE DURAĞI, 29 EKİM CAD. LADİN SOK. K:1, SK:9, NO:6 BAHÇELİEVLER/ İSTANBUL**
   "Kuyumcu Kent Atölye Durağı" bilgisi normalde ek bilgi olarak kabul edilmektedir. Burada beklenenin aksine adresin en başına yazıldığı için ek bilgi olarak değerlendirilmemiş, cadde ismi ile birleştirilmiştir.

2) **TARLABAŞI BULVARI NO:190/B BEYOĞLU/ İSTANBUL**
   "Tarlabaşı Bulvarı" ifadesi, "bulvar" bilgisinin nadiren veriliyor olması ve verildiği çoğu zaman "xxx bulvarı caddesi" şeklinde veriliyor olması sebepleriyle "Bulvar" olarak değil "Ek bilgi" olarak sınıflandırılmıştır.

3) **SİNAN PAŞA MAH. BARBAROS BULVARI KÖY İÇİ CAD. GİRİŞİ NO:1/7 BEŞİKTAŞ/ İSTANBUL**
   "CAD. GİRİŞİ" ifadesi nadir kullanılan bir ifade olduğu için muhtemel bir senaryo olarak değerlendirilmemiştir. Bu sebeple RegEx bu adreste "Girişi" ifadesinin adresin öncesinden ayırarak "Ek bilgi" olarak sınıflandırmıştır.

4) **AKSEMSETTIN MAH FATIH BULVARI NO : 505 SULTANBEYLİ/ İSTANBUL**
   Burada RegEx "NO :" ifadesini önceki parçadan ayıramamıştır. Bu sebeple "Fatih Bulvarı No :" ifadesi "Ek bilgi", "505" verisi "No" olarak sınıflandırılmıştır.

5) **METROGARDEN AVM NECİP FAZIL MH. DUDULLU YOLU ÜZERİ, MABEYN CAD. NO:3-A ÜMRANİYE/ İSTANBUL**
Bu adreste verilen ek bilgilerin yerleri diğer adreslere kıyasla sıradışı bir şekilde verildiği için, altı çizili bloklar sırasıyla "Mahalle" ve "Cadde" olarak sınıflandırılmıştır.

6) **ÜSKÜDAR CADDESİ. YUKARI MH. BAYRAMLAR İŞ MERKEZİ. NO:9 KARTAL/ İSTANBUL**
Bu adreste kullanılan format öngörülenin aksine cadde -> mahalle olarak ilerlemiştir. Adreste ilk olarak mahalle bilgisinin varlığını sorgulayan RegEx, altı çizili alanı "Mahalle" olarak sınıflandırmıştır.

7) **BATTALGAZİ MAH BOSNA BULV 86A SULTANBEYLİ/ İSTANBUL**
Burada numara bilgisinden önce "NO:" veya diğer alternatif ifadeler kullanılmadığı ve "86A" ifadesinin "Ek bilgiler" sınıflandırması yazım kuralları gerekliliklerini karşılaması sebebiyle altı çizili blok "Ek bilgi" olarak sınıflandırılmıştır.

8) **CAMİ MAH. BALIKÇILAR SOK. - SİT. C BLOK APT. NO: 20 C / 96 TUZLA/ İSTANBUL**
Bu adreste "No" bilgisi beklenenden fazla karakter sayısı ile ifade edildiği için altı çizili alan "Ek bilgi" olarak sınıflandırılmıştır.

9) **VIAPORT AVM YENISEHIR MAH.DEDEPASA CAD.NO:19 PENDİK/ İSTANBUL**
Normalde "Ek bilgi" olarak algılanması gereken "Viaport AVM" ifadesi adresin en başına yazıldığı için kendisinden sonra gelen ifadeden ayrıştırılamamış, altı çizili bölgenin tamamı "Mahalle" olarak sınıflandırılmıştır.

10) **EMNİYETEVLERİ M. ÇELEBİ MEHMET S. NO:12/ KAĞITHANE/ İSTANBUL**
Mahalle ve Sokak ifadeleri beklenen ifade ve kısaltma formatları dışında kısaltıldığı için altı çizili alanın tamamı "Ek Bilgi" olarak tanımlanmıştır.

Ekran Görüntüleri