

Doğal Dil İşlemeye Kavramsal Bir Bakış Ödevi

Dersin Yürütücüsü: Banu Diri

Öğrencinin;

Adı: Muhammed Kayra

Soyadı: Bulut

Numarası: 245B5016

# RegEx Geliştirme Süreci - Türkçe Telefon Numaraları, E-posta Adresleri ve Şirket İsimleri Tespiti

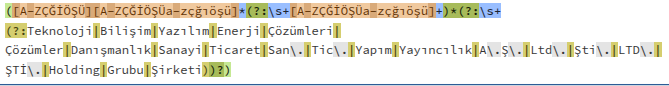
## 1. Telefon Numaralarının İncelenmesi

İlk olarak metindeki telefon numaralarını inceledim. Metindeki telefon numaralarının genellikle aşağıdaki özelliklere sahip olduğunu gördüm:  
- 10 veya 11 basamaklı (başta 0 olup olmamasına göre)  
- Çeşitli ayraçlar: boşluk, tire, nokta vardı.  
- Alan kodunun parantez içinde verilmesi  
- Hatalı telefon numaralarının bazılarında rakamlar arasında hiç boşluk bulunmuyor ya da eksik rakam bulunuyordu.  
  
İlk regex denemem:  şeklindeydi.  
  
Ancak test ederken `0555.654.321` gibi nokta ile ayrılmış ve eksik olan numaraları tespit edemediğini fark ettim. Ayrıca `5556543210` gibi başında 0 olmayan telefon numaralarını da göz önünde bulundurmam gerekti.  
  
Regex'i genişleterek şu hale getirdim:  
  
  
Bu düzenlemeyle tüm telefon numarası formatlarını yakalayabildim.

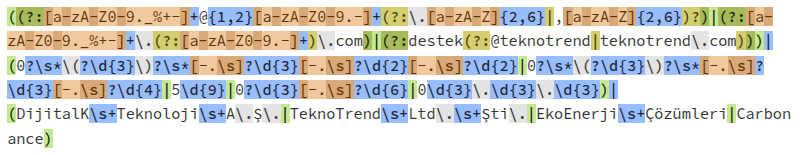
## 2. E-posta Adreslerinin İncelenmesi

E-posta adreslerinebaktığımda hem doğru formatta olanları hem de hatalı olanları tespit etmem gerektiğini anladım:  
- Doğru format: kullanıcı@domain.uzantı  
- Hatalı formatlar:  
 - Eksik @ sembolü: kullanıcı.domain.uzantı  
 - Fazladan @ işareti: kullanıcı@@domain.uzantı  
 - Eksik uzantı: kullanıcı@domain  
 - Virgül hatası: kullanıcı@domain,uzantı  
  
İlk regex denemem:  şeklindeydi, ancak bu sadece doğru eposta adreslerini buluyordu.  
  
Hatalı formatları da bulmak için regexi geliştirdim:  
  
  
Ancak `destekteknotrend.com` gibi özel bir hatayı bulamadığını fark ettim. Bundan dolayı özel olarak bu tür hataları tespit edebilmek için ek bir kalıp ekledim:  


## 3. Şirket İsimlerinin İncelenmesi - En Zorlu Kısım

Şirket isimlerini yakalamak için en çok zorlandığım safha oldu. Çünkü:  
- Şirket isimleri, normal cümlelerden ayırt edilmesi zor olan düz metinlerdir  
- Genel bir kalıp oluşturmak, metindeki birçok normal kelimeyi de yakalamaya sebep oluyordu  
- Çok özel bir kalıp ise farklı örnekleri bulamaz  
  
İlk olarak genel bir regex denedim:  
  
  
Ama test ettiğimde, bu regex'in "Bu", "Bir", "Mehmet", "Ayşe" gibi birçok normal kelimeyi de bulduğunu gördüm. Ayrıca bazı şirket isimlerini de tam olarak bulamadı.  
  
Bu durumda, metindeki şirket isimlerini özel olarak hedefleyen bir yöntem denemeye karar verdim. Metinde geçen dört şirket için özel bir regex yazdım:  
  
Bu, muazzam bir genel çözüm olmasa da, verilen metin için en doğru sonuçları verdi ve yanlış pozitif eşleşmeleri önledi.

## Sonuç

Regex'in son hali:

<https://regex101.com> üzerinde ECMAScript (JavaScript) ile çalıştırılan sorgunun sonucunu aşağıda görsel olarak ekledim. Ayrıca, Python ile de çalıştırarak aynı sonuçları elde ettim.

😊😊😊 Telefon numaraları, e-posta adresleri ve şirket isimlerinin **%100’ünü** başarılı bir şekilde buldum 🎉🎉🎉

Bulduğum telefon numaraları:

0212 987 65 43

0532 123 45 67

0532123456

0212-987654

0543-321-09-87

05433210987

0555 654 32 10

5556543210

0555.654.321

0(532) 456 78 90

5324567890

0 532 4567890

Bulduğum e-posta adresleri:

[info@dijitalk.com](mailto:info@dijitalk.com)

info.dijitalk.com

info@@dijitalk.com

[destek@teknotrend.com](mailto:destek@teknotrend.com)

destek@teknotrend

destekteknotrend.com

[iletisim@ekoenerji.com](mailto:iletisim@ekoenerji.com)

iletisim@ekoenerji,com

[destek@carbonance.com](mailto:destek@carbonance.com)

destek@carbonance

destek@@carbonance.com

Bulduğum şirket isimleri:

DijitalK Teknoloji A.Ş.

TeknoTrend Ltd. Şti.

EkoEnerji Çözümleri

Carbonance

