

**Türkiye Cumhuriyeti**  
**Yıldız Teknik Üniversitesi**  
**Bilgisayar Mühendisliği Bölümü**



**Doğal Dil İşlemeye Giriş**

**1. Ödev**

**Regex Komutu Kullanarak Adres Bilgisinin Bulunması**

**19011604 - İlker Tınkır**

**Kasım 2020**

## Regex Komutu:

```
/^(.*(MAH\.|MH\.|MAH))*(.*(CD\.|CAD\.|CAD))*(.*(SK\.|SOK\.))*(.NO[.:0-9]+[\/\-\\s0-9]*[A-Z0-9]{1})+\s(.*[A-ZğüşöçİĞÜŞÖÇ\ ])*$
```

Kullandığım regex komutunu genel hatlarıyla açıklamak gerekirse: veriseti üzerindeki her bir adres satırına tek tek bakıp, o satır içerisindeki “Mahalle”, “Cadde” ve “Sokak” bilgilerinin \* işareti yardımıyla ayrı ayrı işaretlenmesini sağladım. Verisetindeki bu 3 bilgi adresin girişinde karmaşık sırada 0 veya 1 kez geçecek şekilde olduğu için \* operatörü bu bilgileri çoğunlukla doğru şekilde ayırt edebildi. Ayrıca bu bilgileri verisetinde en çok geçen yazım şekillerini kapsayacak şekilde “or” operatörü ile birleştirerek en fazla sayıda adresin işaretlenmesini hedefledim. Adresin “Numara” bilgisi verisetinde çok fazla farklı şekilde tutulduğu için tüm numara bilgisi kombinasyonlarını kaydetmemeye çalışarak komutun daha fazla hafıza harcamamasını sağlamaya çalışarak. Yazdığım komutta “No” ibaresiyle başlayan numaralar daire numarası, blok numarası gibi bilgileri birlikte başarıyla işaretlenebilirken bu ibareyi içermeyen adresler bulunmamaktadır. Komutun en son kısmında şehir ve ilçe bilgileri aynı anda çekilerek kaydedilmiştir.

Yapılan tüm test işlemleri <https://regex101.com/> websitesi üzerinden PHP regex yapısı kullanılarak gerçekleştirilmiştir. Testler sonucunda:

**Verisetindeki toplam adres sayısı: 6831**

**Doğru sonuç veren adres sayısı: 5128**

**Yanlış sonuç veren adres sayısı: 1703**

**Doğruluk oranı: %75.06**

Olarak hesaplanmıştır.

## Çıkarımlar:

Yapılan test işlemleri sonucunda tüm satırları içerisinde regex komutuna göre doğru ayrılamayan adreslerin çoğunlukla bozuk Türkçeyle yazılmış, hatalı adresler olduğu görülmüştür. Bunun dışında fazla hafıza harcamamak adına regex komutunun Numara bulan kısmının bazı adreslerde yanlış sonuç göstermesine göz yumulmuştur. Verisetindeki adreslerin büyük çoğunluğunun kısaltma kullanılarak yazılması göz önünde bulundurularak en çok kullanılan 2, 3 kısaltma regex komutuna eklendiği için yine bu kısımda da tam hali yazılmadığı için bulunamayan adresler olmuştur.

### Doğru bulunan adres örnekleri:

SÜMER MAH. 18.SOK.NO:17/A ZEYTİNBURNU/ İSTANBUL  
YILDIRIM MAH. ŞEHİT KAMİL BALKAN CD. NO:55/A BAYRAMPAŞA/ İSTANBUL  
M. N. ÖZMEN MH. NADİDE SK. NO:39 GÜNGÖREN/ İSTANBUL  
GÖKTÜRK MAH. GÖKTÜRK CAD. NO:51/D FATİH/ İSTANBUL  
GÖKTÜRK MERKEZ MAH. BELEDİYE CAD. NO:7 B EYÜPSULTAN/ İSTANBUL  
M.N ÖZMEN MAH ALAYBEY SOK MUTLUOĞLU İŞHANI NO:10/2 GÜNGÖREN/ İSTANBUL  
CEVATPAŞA MAH. MİLLET CAD. NO:221/A BAYRAMPAŞA/ İSTANBUL  
CEVATPAŞA MH. MEHMET AKİF CD. NO:7/1 BAYRAMPAŞA/ İSTANBUL  
OSMANİYE MAH.CEMİYET SK.NO:45 /B BAKIRKÖY/ İSTANBUL  
ABDİ İPEKÇİ CAD. NO:1 GÜNGÖREN/ İSTANBUL  
ÜÇEVLER MAH.URFALILAR CAD.NO:56/B ESENYURT/ İSTANBUL  
SULTANİYE MAH.DOĞAN ARASLI BULVARI NO:168/A KENT İŞ MERKEZİ ESENYURT/ İSTANBUL  
İNÖNÜ MAH. DOĞAN ARASLI BULVARI NO: 94/1/23 ESENYURT/ İSTANBUL  
TAYAKADIN MH. KÖY İÇİ CD. NO:7/2 NO:1 ARNAVUTKÖY/ İSTANBUL  
YEŞİLKENT MH. BALIKYOLU CD. NO:47/B AVCILAR/ İSTANBUL  
GÜZELYURT MAH. YILDIRIM BEYAZIT CD. 2118 SK. NO:12 DÜKKAN:1 ESENYURT/ İSTANBUL  
YEŞİLKENT MAH.AMASYALILAR CAD. NO:53/B AVCILAR/ İSTANBUL  
ADNAN MENDERES MH. CEREN SK. NO:17 ARNAVUTKÖY/ İSTANBUL  
MUSTAFA KEMAL PAŞA MAH. FATİH CAD. NO:113/A ARNAVUTKÖY/ İSTANBUL

### Bulunamayan adres örnekleri:

ESENKENT MAH. SÜLEYMAN DEMİREL CAD. KİLER AVM İÇİ ESENYURT/ İSTANBUL

#### Numara bilgisi içermeyen adres

AYDINLI MAH KONAŞLI MEVKİİ SÜHEYLA SOK NO 6 TUZLA/ İSTANBUL

#### Yanlış formatta Numara içeren adres

NAMIK KEMAL MAH ABDİ İPEKÇİ 32 1 ESENYURT/ İSTANBUL

#### Sokak, cadde ve yanlış formatta Numara içeren bir adres

ÇAMÇEŞME MH MARMARA CD 64B PENDİK/ İSTANBUL  
NEOMARIN AVM KEMİKLİ DERE MEVKİİ ZAMİN KAT:H-17 PENDİK/ İSTANBUL

#### Regex komutunda bulunmayan bir mahalle bilgisi içeren adres

#### Regex komutuna uymayan bir adres

OSMANGZİ MH GENÇ OSMAN CD 69B SANCAKTEPE/ İSTANBUL  
OSMANGAZİ MH MİMAR SİNAN CD 97/1 SANCAKTEPE/ İSTANBUL

#### Verisinde yaygın olmayan kısaltmalar içeren adresler