

DOĞAL DİL İŞLEME

1. ÖDEVİ

ALP BİNTUĞ UZUN
17011066

Yazılan RegEx:

Yazılan Regular Expression'da 7 adet grup bulunmaktadır. Bu gruplar "Mahalle", "Cadde", "Sokak", "Yer Bilgisi", "No", "Kalan Adres" ve "İl" gruplarıdır.

"Yer Bilgisi" grubu, herhangi bir anahtar kelime (CAD., MAH. vb.) girilmeden girilmiş olan adres bilgilerini ya da bu bilgilerden sonra kalan adres bilgilerini belirtir.

"Kalan Adres" grubu, "No" bilgisi girildiyse, bu bilgiden sonra ve "İl" grubundan önce kalan, genellikle ilçe bilgilerini belirtir.

RegEx metninde açılan ilk parantez ana grubu belirtir, bu grup içerisinde "Mahalle, Cadde ya da Sokak" bilgileri bulunan adresler ayrıştırılmıştır. Bu bilgilerin bulunabilmesi için hem tam adları hem de kısaltmaları göz önünde bulundurulmuştur. Bu gruba ait olmayan adres satırları içinde ayrıştırma yapmak için numara ve il bilgisini belirten "/" işareti kullanılmıştır. Herhangi bir numara bilgisi, öncesinde "no" yazmasa bile "İl" bilgisi ile karıştırılmaması için il ve kalan adres grubunda sadece Türkçe karakterler kabul edilmiştir, böylece "/" işaretinden önce ve sonra gelen sayıların "ilçe/il" olarak sınıflandırılmaları önlenmiştir.

REGULAR EXPRESSION

6830 matches, 17130540 steps (~7.08s)

```
// / (^(\"?(?P<mahalle>[\\w\\d[ÜİĞŞÇÖÄ0-9 \\.\\-]*([MAHALLES[İİİİ]MAH[İİİİ],?))?(\\\"?(?P<cadde>[\\w\\d[ÜİĞŞÇÖÄ0-9 \\.\\-]*([. ]CADD[İİİİ]([. ]CA[Dİİİ]?[.\\s, ]*))?(\\\"?(?P<sokak>[\\w\\d[ÜİĞŞÇÖÄ0-9 \\.\\-]*([SOKAK[İİİİ]SOKAĞI[İİİİ],?))?(\\\"?(?P<yerBilgisi>[\\w\\d[ÜİĞŞÇÖÄ0-9 \\.\\-]*)(?P<no>[\\d\\-\\/]*)?(NO[İİİİ]?[.\\s, ]*)?(?P<kalanAdres>[\\w\\d[ÜİĞŞÇÖÄ0-9 \\.\\-]*)(?P<il>[\\w\\d[ÜİĞŞÇÖÄ0-9 \\.\\-]*)?)?)?)? / gm
```

TEST STRING

SOĞUKSU MAH. M.AKİF ERSOY CD. NO:141/C BEYKOZ/ İSTANBUL
19 MAYIS MAH. İNÖNÜ CAD. NO:48/B KADIKÖY/ İSTANBUL
DUMLUPINAR MAH. ANADOLU CAD. NO:53/B ÜMRANİYE/ İSTANBUL
İSTİKLAL MAH. HACI RÜSTEMOĞLU SOK. NO:3/A ÜMRANİYE/ İSTANBUL
İHASANİYE MAH. HAREM İSKELE CAD. NO:6/A ÜSKÜDAR/ İSTANBUL
VALİDEATİK MAH. NUHKUYUSU CAD. NO:69/A ÜSKÜDAR/ İSTANBUL
GÜZELTEPE MAH. KÜÇÜK NAMAZGAH CAD. NO:43/A ÜSKÜDAR/ İSTANBUL
ÇATALMEŞE MAH. REŞADİYE CAD. NO:2/B ÇEKMEKÖY/ İSTANBUL
GÜLTEPE MAH. ŞEHİT ÖZCAN DEMİRCİ CAD. ŞAHİNLER SOK KÜÇÜKÇEKMECE/ İSTANBUL
ÖMERLİ MAH. TERME SOK. A.V.M ÇARŞI NO:13 ARNAVUTKÖY/ İSTANBUL
CERRAH PAŞA MAH. KOCAMUSTAFA PAŞA CAD. NO:59A FATİH/ İSTANBUL
ORDU CAD. MARMARA ÇARŞISI ÖNÜ GAZETE BAYİ 35/1 FATİH/ İSTANBUL
MUSTAFA KEMAL PAŞA MAH. BEYOĞLU CAD. NO:33/B AVCILAR/ İSTANBUL
2.KÖY MAH. TATAR SOK. 83 SARIYER/ İSTANBUL
SİNAN PAŞA MAH. DOLMABAĞÇE CAD. NO :11 BEŞİKTAŞ/ İSTANBUL
DERVİŞ ALİ MAH. FEVZİ PAŞA CAD. NO.150/1 FATİH/ İSTANBUL
KAZIM KARABEKİR MAH. ATIŞ ALANI CAD. NO:97-A ESENLER/ İSTANBUL
YEŞİLYURT MAH. İSTASYON CAD. NO:1 BAKIRKÖY/ İSTANBUL
KARADENİZ MAH. ESKİ EDİRNE ASFALTI NO:360/B GAZİOSMANPAŞA/ İSTANBUL

Analiz:

Analiz için Python dili kullanılmıştır. Verilen “Adresler.txt” dosyası satır satır okunmuş ve her bir satırda ilgili RegEx kullanılarak eşleşen elemanlar ayrıştırılmıştır.

Yapılan analizler sonucunda şu bilgilere ulaşılmıştır:

Mahalle sayısı	5841
Cadde sayısı	4873
Sokak sayısı	1438
Yer bilgisi sayısı	2190
No sayısı	431
Kalan adres sayısı	5687
İl sayısı	6829

Yukarıdaki ayırım kriterleri göz önüne alındığında bir satırda bulunan grup sayıları şöyledir;

Hiçbir grup bulunamayan satır sayısı	0
Sadece bir grup bulunan satır sayısı	0
İki grup bulunan satır sayısı	258
Üç grup bulunan satır sayısı	859
Dört grup bulunan satır sayısı	4765
Beş grup bulunan satır sayısı	573
Altı grup bulunan satır sayısı	361
Tüm grupların bulunduğu satır sayısı	15

Yukarıdaki bilgiler göz önüne alındığında programın tüm satırlarda en azından iki grup bulduğu görülmektedir. Satır tanımadaki başarı oranı %100'dür.

Yazılan RegEx'in başarı oranı gruplanmış olan kelimelerin, satır içerisinde bulunan kelimelerin oranına göre yapıldığında en başarılı ayırmada %100, en başarısız ayırmada %15 başarıya ulaştığı görülmekte, genel başarı oranı ise %87.6 olmaktadır.

RegEx başarı oranı kelime yerine harfler üzerinden hesaplanırsa en başarılı ayırmada %100, en başarısız ayırmada %12.7, genel başarı oranı %95.11 olmaktadır.