BLM 4580 Doğal Dil İşleme Giriş Dersi

1.Ödev

Konu : Adresler için Regex hazırlanması

Hazırlayan: 16011002 Aykut AKDENİZ

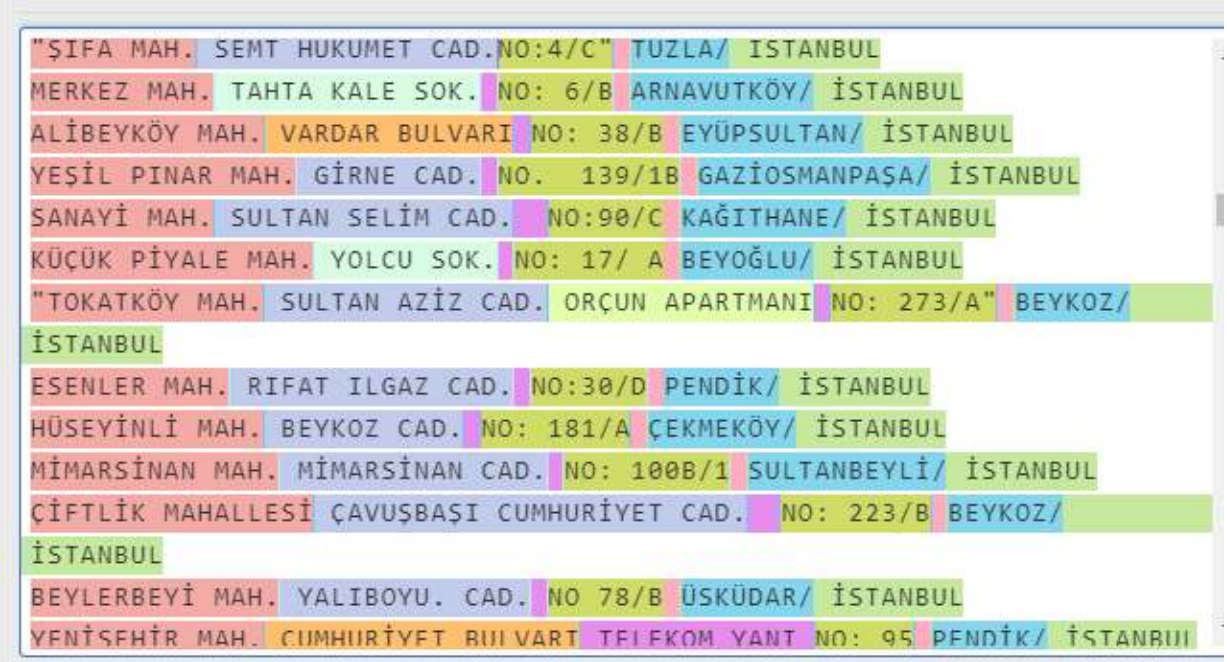
Adreslerin formatında bazı örneklerde değişse de genel olarak sırayla mahalle-bulvar-cadde-sokak-apartman-no-ilçe-il olarak gittiği için buna göre regex hazırlanmıştır.(https://regex101.com/ adresinde PCRE(PHP) ile yapılmıştır.)(İl olarak [A-ZĞÜŞİÖÇ]{3-14} çalışma süresini çok fazla arttırdığı için il yerine İstanbul yazılmıştır.)

^(?<Adres>(?<Mahalle>.[^\/.^\n]*M\.|.*MA?H\s?[\.|\s]|.*MAHALLES?[İI]?)|(?<Bulvar>.*BULVARI|.*B\.?U?L\.?V\.?|.*BUL\.)|(?<Cadde>.*C\.|.*[\s\.]CA?D\s?[\s\.]|.*CADDES?[İI]?\s)|(?<Sokak>.*S\.|.*[\s\.]SO?K\s?[\.\s]|.*SOKA[KĞG]I?)|(?<Apartman>.*\sAPT\.|.*APARTMANI?)|(?<Bilinmeyen>.*)(?<No>[\.\s]?NO\s?[.:\s]*\s*[0-9A-Z]+[\/-]?[A-Z0-9]*[\s\/\-\"]?[A-Z0-9]*[\s\.\"]?)|(?<Tanim_cumlesi>.*\s)(?<Ilce>\p{L}+\/)(?<Il>\sİSTANBUL))+$

Regex adres dosyasındaki tüm adresleri grup içerisine alabiliyor ancak istasyon, bayii, camii, durak, büfe önü, bayii vb. örneklerin grup olarak aynı grup içerisine alıp tam olarak ismini tanımlayamıyor(Tanım cümlesi başlığında toplanmışlardır). Aşağıda verilen başlıkta bulunan durumlarda yanlış gruplandırsa da adres bilgilerinde bulunan 6831 adresi de kapsamaktadır.

Regex ile başarılı bulunan adres sayısı: 6788

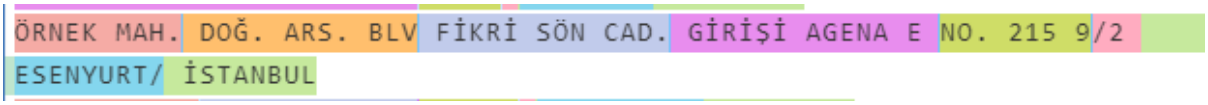
Regex'in başarı oranı: %99.37



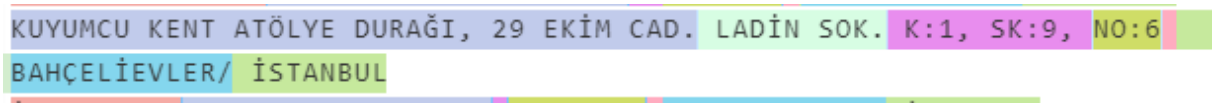
Yanlış Gruplandırılan Örnekler

Genel olarak durağı, camii, iskele gibi kesin olmayan tarif içeren adreslerde belirli bir gruplama yapılamamıştır ve bu kısımlar Tanım_cumlesi grubuna alınmıştır. Sitesi, Apartmanı, iş merkezi gibi bilgiler için de ifade uzunluğu artacağı için "Bilinmeyen" olarak

gruplandırılmıştır. Başına No eklenmemiş rastgele numara bilgileri de belirsizlikten dolayı eklenememiştir.



Yukarıdaki örnekte "/2" kısmını no grubuna dahil edememiştir.



Yukarıdaki örnekte durak bilgisini tanımadığı için cadde bilgisine eklemiştir.

Yukarıdaki örnekte m. Kısmından dolayı mahalle bilgisini ön kısmından ayıramamıştır.

Yukarıdaki örnekte "Mihrimah" kısmındaki mah kısmından dolayı mahalle grubuna dahil edilmiştir.

Yukarıdaki örnekte ise diğer adres bilgilerinde bulvar caddeden önce gelmesine rağmen burada sırası değiştiği için bulvar grubunu caddeyi de içerecek şekilde bulmuştur.

Yukarıdaki örnekte mahalle, caddeden sonra geldiği için mahalle kısmını olduğundan uzun bulmuştur.

Uzun ve karmaşık yapıda "No" bilgilerinin genelini kapsayacak bir ifade yazıldığı için bu örnekteki gibi yanlış harfleri de dahil edebilmektedir.