

YTU BİLGİSAYAR MÜHENDİSLİĞİ DOĞAL DİL İŞLEME

2020-2021 Güz Dönemi ÖDEV – 1

AYŞE HİLAL DOĞAN 17011907

REGEX KOMUTU İLE TEXT İÇERİSİNDEKİ ADRESLERİ TESPİT ETME

Tanım: Oluşturulan RegEx komutu, verilen bir text teki adresleri bulur ve mahalle, cadde, sokak, apartman, daire, no, ilçe, il şeklinde ayrıştırır.

Yöntem: Adresler bazen doğru formatta olmayabilir. Bu ihtimaller düşünülerek örneğin mahalle için MAH., MH, MAHALLESİ, MAH şeklinde farklı yazım şekillerine uygun olacak bir RegEx yazılmıştır. Fakat bazı durumlarda adres içerisinde açıklaması olmayan kapı numarası sokak numarası gibi ifadeler yer almaktadır. Bunlar ayrıştırılamamış, <diğer> kısmına koyulmuştur.

Örnek: BÜYÜKŞEHİR MH. ENVER ADAKAN CD. NO:A2 B 20/18 BEYLİKDÜZÜ/ İSTANBUL Bu örnekte "B 20/18 "kısmı ne olduğu belirtilmediği için gruplanırken <diğer> kısmına koyulmuştur.

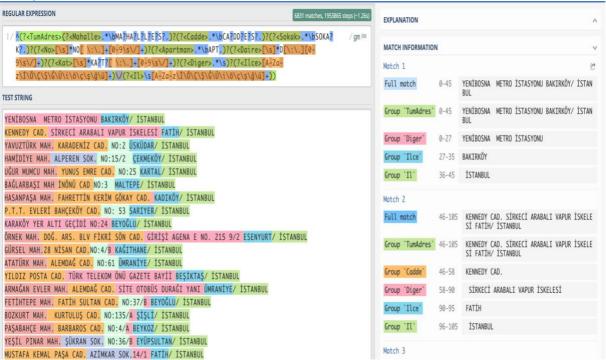
- Üzerinde çalışma yapılan adres dosyasında 6831 adet adres bilgisi bulunmaktadır.
- Yazılan RegEx komutu ile 6831 tanesinin de adres olduğu bilgisi bulunmaktadır. Bazı adreslerde mah., cad. gibi bilgiler bulunmasa da hepsinde ilçe ve il bilgisi bulunduğu için bu bilgi bulunan veriler adres olarak belirlenmektedir.

Örnek: YENİBOSNA METRO İSTASYONU BAKIRKÖY/ İSTANBUL

- 6831 adresten 6831'i de bulunmuş olup , herhangi bir açıklama yapılmayarak veya çok yanlış formatta yazılan adres bilgilerini tam doğru olarak ayrıştıramadığı için başarı oranı %99 olarak belirlenmiştir.

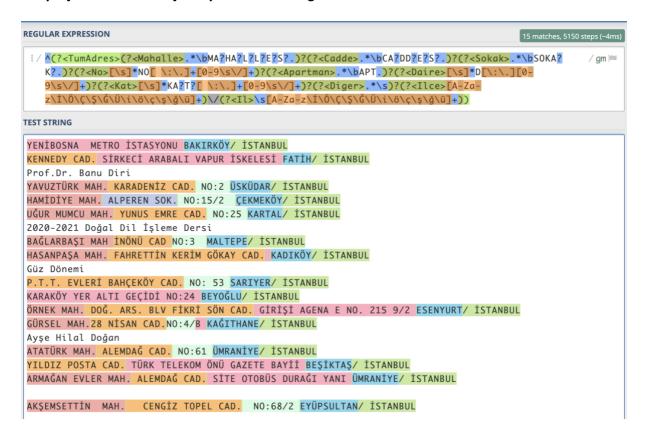
Regullar Expression:

Sonuç ekranı:





Dosya içerisinde adres dışında yazıların da olduğu örnek:



Görüldüğü üzere adres dışındaki yazılar RegEx komutuna uymadığı için adres olarak işaretlenmemiştir.

Yanlış format veya eksik bilgi sebebiyle sokak no, cadde ismi gibi doğru kategorilerde ayrılamayan adreslere örnek:

5. BÖLGE G-5 CAD. BÜFE 2 BAŞAKŞEHİR/ İSTANBUL KARADENİZ MAH. 1176 SOK. C YAN CEPHE 2/3 GAZİOSMANPAŞA/ İSTANBUL MARMARA MAH. HÜRRİYET BULVARI 233-C BEYLİKDÜZÜ/ İSTANBUL

Kaynakça:

RegEx komutunu yazmak ve test etmek için kullanılan site : https://regex101.com/