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.

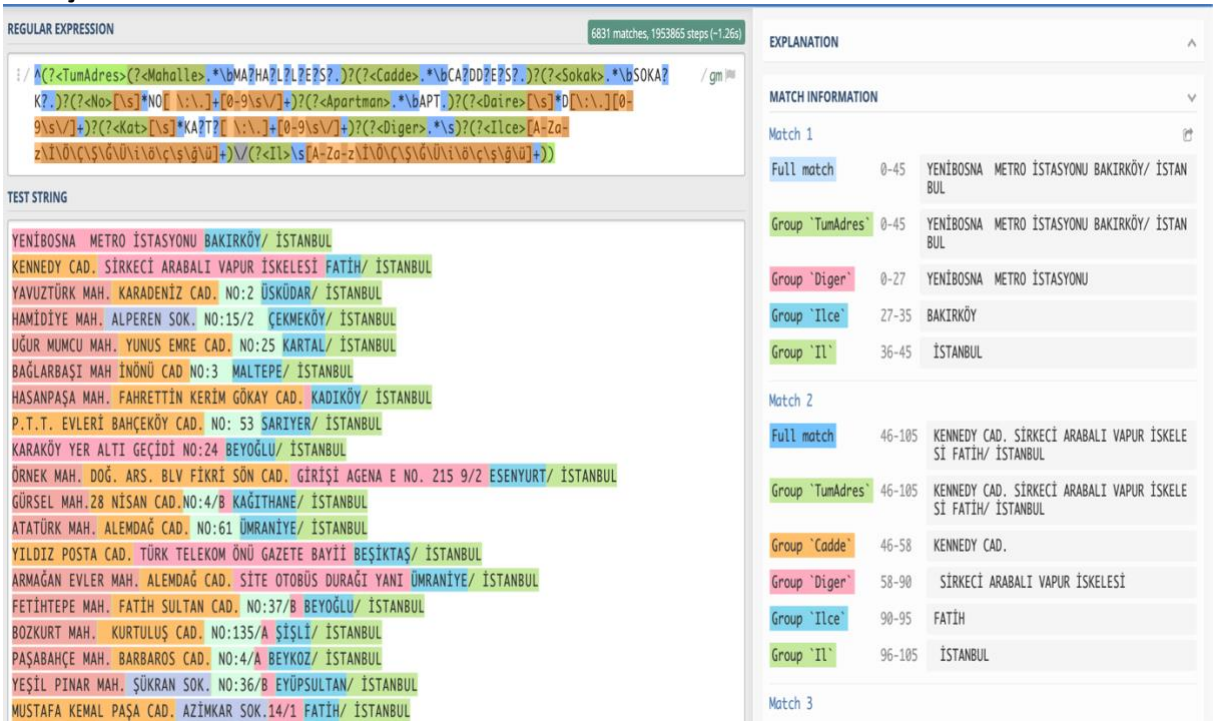
Regullar Expression:

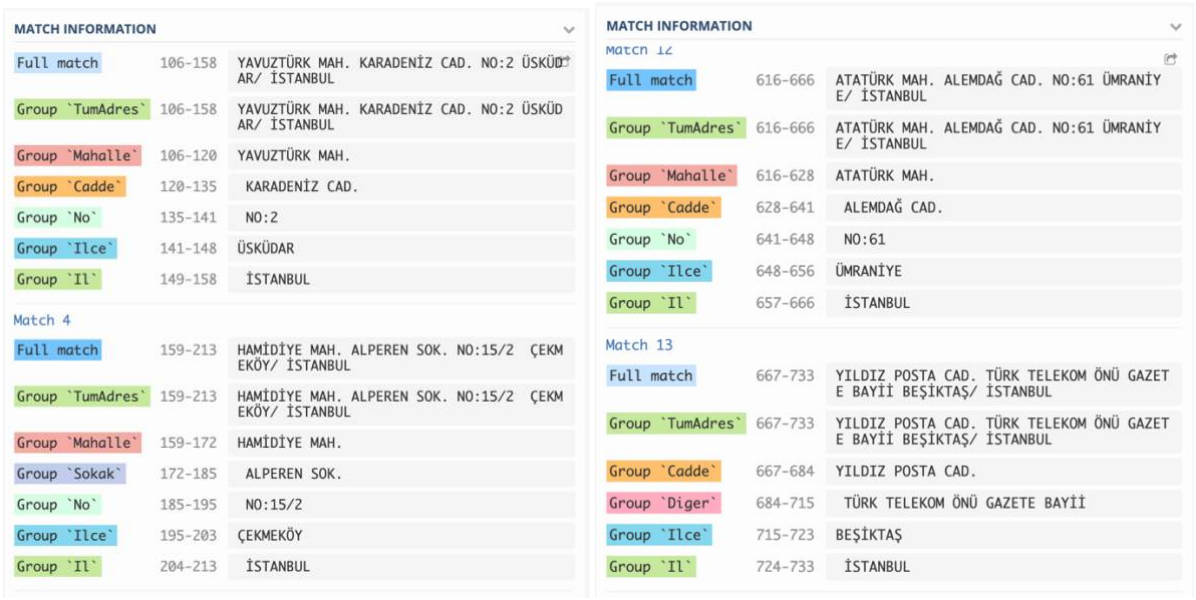
REGULAR EXPRESSION

6831 matches, 1953865 steps (~1.26s)

```
/ ^(?<TumAdres>(?<Mahalle>.*\bMA?HA?L?L?E?S?.)?(?<Cadde>.*\bCA?DD?E?S?.)?(?<Sokak>.*\bSOKA?K?.)?(?<No>[\s]*NO[ \:\.]+[0-9\s\/]+)?(?<Apartman>.*\bAPT.)?(?<Daire>[\s]*D[\:\.][0-9\s\/]+)?(?<Kat>[\s]*KA?T?[ \:\.]+[0-9\s\/]+)?(?<Diger>.*\s)?(?<Ilce>[A-Za-z\İ\Ö\Ç\Ş\Ğ\Ü\i\ö\ç\ş\ğ\ü]+)\/(?<Il>\s[A-Za-z\İ\Ö\Ç\Ş\Ğ\Ü\i\ö\ç\ş\ğ\ü]+))
/ gm
```

Sonuç ekranı:





MATCH INFORMATION

Full match	106-158	YAVUZTÜRK MAH. KARADENİZ CAD. NO:2 ÜSKÜDAR/ İSTANBUL
Group `TumAdres`	106-158	YAVUZTÜRK MAH. KARADENİZ CAD. NO:2 ÜSKÜDAR/ İSTANBUL
Group `Mahalle`	106-120	YAVUZTÜRK MAH.
Group `Cadde`	120-135	KARADENİZ CAD.
Group `No`	135-141	NO:2
Group `Ilce`	141-148	ÜSKÜDAR
Group `Il`	149-158	İSTANBUL
Match 4		
Full match	159-213	HAMİDİYE MAH. ALPEREN SOK. NO:15/2 ÇEKMEKÖY/ İSTANBUL
Group `TumAdres`	159-213	HAMİDİYE MAH. ALPEREN SOK. NO:15/2 ÇEKMEKÖY/ İSTANBUL
Group `Mahalle`	159-172	HAMİDİYE MAH.
Group `Sokak`	172-185	ALPEREN SOK.
Group `No`	185-195	NO:15/2
Group `Ilce`	195-203	ÇEKMEKÖY
Group `Il`	204-213	İSTANBUL

MATCH INFORMATION

Match 12		
Full match	616-666	ATATÜRK MAH. ALEMDAĞ CAD. NO:61 ÜMRANİYE/ İSTANBUL
Group `TumAdres`	616-666	ATATÜRK MAH. ALEMDAĞ CAD. NO:61 ÜMRANİYE/ İSTANBUL
Group `Mahalle`	616-628	ATATÜRK MAH.
Group `Cadde`	628-641	ALEMDAĞ CAD.
Group `No`	641-648	NO:61
Group `Ilce`	648-656	ÜMRANİYE
Group `Il`	657-666	İSTANBUL
Match 13		
Full match	667-733	YILDIZ POSTA CAD. TÜRK TELEKOM ÖNÜ GAZETE BAYİİ BEŞİKTAŞ/ İSTANBUL
Group `TumAdres`	667-733	YILDIZ POSTA CAD. TÜRK TELEKOM ÖNÜ GAZETE BAYİİ BEŞİKTAŞ/ İSTANBUL
Group `Cadde`	667-684	YILDIZ POSTA CAD.
Group `Diger`	684-715	TÜRK TELEKOM ÖNÜ GAZETE BAYİİ
Group `Ilce`	715-723	BEŞİKTAŞ
Group `Il`	724-733	İSTANBUL

Dosya içerisinde adres dışında yazıların da olduğu örnek:

REGULAR EXPRESSION

15 matches, 5150 steps (~4ms)

```
/ ^(?<TumAdres>(?<Mahalle>.*\bMA?HA?L?L?E?S?.)?(?<Cadde>.*\bCA?DD?E?S?.)?(?<Sokak>.*\bSOKA?K?.)?(?<No>[\s]*NO[ \:\.]+[0-9\s\/]+)?(?<Apartman>.*\bAPT.)?(?<Daire>[\s]*D[\:\.][0-9\s\/]+)?(?<Kat>[\s]*KA?T?[ \:\.]+[0-9\s\/]+)?(?<Diger>.*\s)?(?<Ilce>[A-Za-z\İ\Ö\Ç\Ş\Ğ\Ü\i\ö\ç\ş\ğ\ü]+)\/(?<Il>\s[A-Za-z\İ\Ö\Ç\Ş\Ğ\Ü\i\ö\ç\ş\ğ\ü]+))   / gm
```

TEST STRING

```
YENİBOSNA  METRO İSTASYONU BAKIRKÖY/ İSTANBUL
KENNEDY CAD. SİRKECİ ARABALI VAPUR İSKELESİ FATİH/ İSTANBUL
Prof.Dr. Banu Diri
YAVUZTÜRK MAH. KARADENİZ CAD. NO:2 ÜSKÜDAR/ İSTANBUL
HAMİDİYE MAH. ALPEREN SOK. NO:15/2  ÇEKMEKÖY/ İSTANBUL
UĞUR MUMCU MAH. YUNUS EMRE CAD. NO:25 KARTAL/ İSTANBUL
2020-2021 Doğal Dil İşleme Dersi
BAĞLARBAŞI MAH İNÖNÜ CAD NO:3  MALTEPE/ İSTANBUL
HASANPAŞA MAH. FAHRETTİN KERİM GÖKAY CAD. KADIKÖY/ İSTANBUL
Güz Dönemi
P.T.T. EVLERİ BAHÇEKÖY CAD. NO: 53 SARIYER/ İSTANBUL
KARAKÖY YER ALTI GEÇİDİ NO:24 BEYOĞLU/ İSTANBUL
ÖRNEK MAH. DOĞ. ARS. BLV FİKRİ SÖN CAD. GİRİŞİ AGENA E NO. 215 9/2 ESENYURT/ İSTANBUL
GÜRSEL MAH.28 NİSAN CAD.NO:4/B KAĞITHANE/ İSTANBUL
Ayşe Hilal Doğan
ATATÜRK MAH. ALEMDAĞ CAD. NO:61 ÜMRANİYE/ İSTANBUL
YILDIZ POSTA CAD. TÜRK TELEKOM ÖNÜ GAZETE BAYİİ BEŞİKTAŞ/ İSTANBUL
ARMAĞAN EVLER MAH. ALEMDAĞ CAD. SİTE OTOBÜS DURAĞI YANI ÜMRANİYE/ İSTANBUL

AKŞEMSETTİN  MAH.   CENGİZ TOPEL CAD.  NO:68/2 EYÜPSULTAN/ İSTANBUL
```

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/