AD: Serdar

SOYAD: Yıldız

NO: 16011004

DERS: Doğal Dil İşleme

ÖDEV NO: 1

Problem

Verilen adres veri setine uygun regex komutunu yazarak adresleri sokak, mahalle gibi parçalara ayırmaktır.

Regex Komutu

Tüm Komut:

(.*\sM[A]{0,1}H[.|\s|A]{0,1})| (.*\sC[A]{0,1}D[\s|.|D]{0,1})| (.*\sS[O]*K[\s|.|A])| (NO[\s|:|.]*\s*[\w]*\s*[\/|-]*\s*[A-Z1-9]{0,4}\s)| (\s?[A-Z,Ç,Ğ,İ,Ö,Ş,Ü]+\s*[\/])| ([^\s]*[\r\n|\r|\n])

Komut Açıklamaları:

Mahalle

Mahalle bilgisini ayırmak için kullanılan regex komutu,

(.*\sM[A]{0,1}H[.|\s|A]{0,1})

- M harfinden önce boşluk olması gerekmektedir.
- A harfi 1 kere olabilir veya hiç olamaz.
- H harfi olmalıdır.
- H harfinden sonra nokta boşluk veya A harfi gelebilir.

Cadde

Cadde bilgisini ayırmak için kullanılan regex komutu,

(.*\sC[A]{0,1}D[\s|.|D]{0,1})

- C harfinden önce boşluk olması beklenmektedir. İçinde CD geçen kelimelerin alınmaması için bu gereklidir.
- C harfinden sonra A harfi olabilir.
- A harfinden sonra veya C harfinden sonra D harfi bulunmalıdır.
- D harfinden sonra boşluk, nokra veya D harfi gelmelidir.

Sokak

Sokak bilgisini ayırmak için kullanılan regex komutu,

(.*\sS[O]*K[\s|.|A])

- S harfinden önce boşluk gelmesi beklenir.
- S harfi olmak zorundadır.
- O harfi olabilir ya da olmaya bilir.
- K harfi olmak zorundadır.
- K harfinden sonra boşluk, nokta veya A harfi gelebilir.

No

Numara bilgisini almak için kullanılan regex komutu,

```
(NO[\s|:|.]*\s*[\w]*\s*[\/|-]*\s*[A-Z1-9]{0,4}\s)
```

- N ve O harfleri olmak zorundadır.
- O harfinden sonra boşluk, iki nokta üst üste ve noktadan bulunabilir veya hiç olmayabilir.
- Daha sonrasında boşluklar atlanır
- Boşluk gelene kadar tüm karakterler alınır.
- Boşluklar atlanır.
- "/" veya "-" gelme durumu kontrol edilir.
- Boşluklar atlanır.
- Daha sonrasında gelebilecek karakterler alınır ve bunların maksimum boyutu 4 digit olarak kabul edilir.
- Bitiş olarak boşluk olması beklenir.

İlçe

İlçe bilgisini almak için kullanılan regex komutu,

```
(\s?[A-Z,Ç,Ğ,İ,Ö,Ş,Ü]+\s*[\/])
```

- İl ile ayrılabilmesi için en son karakterinin "/" olması beklenir.
- "\" karakterinden önce boşluk bulunabilir.
- Boşluklardan önce Türkçe karakterlerin de dahil olduğu bir kelime bulunmak zorundadır.

İl

İl bilgisini almak için kullanılan regex komutu,

```
([^\s]*[\r\n|\r|\n])
```

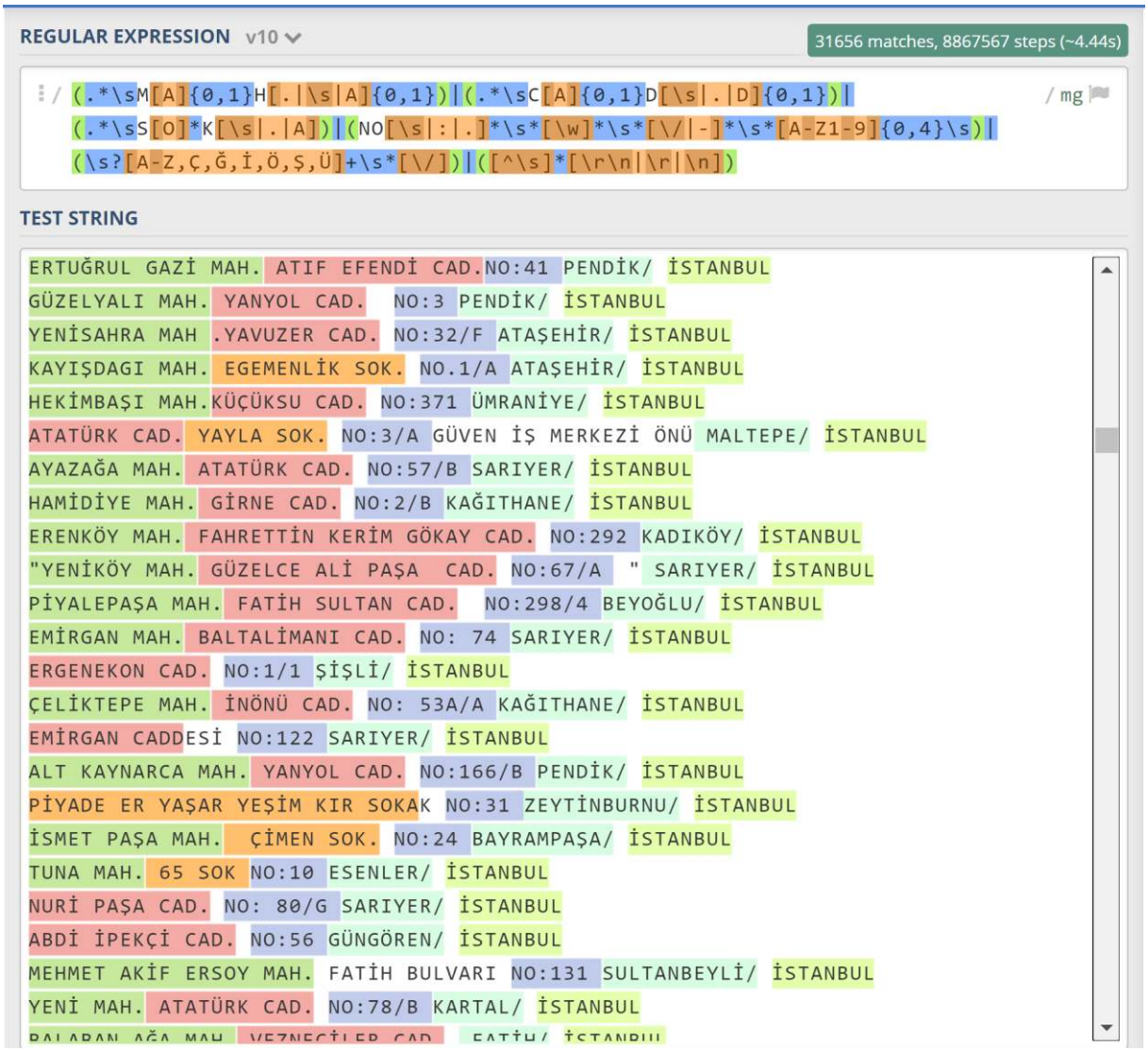
- Son kelimenin il olduğu kabul edilmiştir. Bu sebeple satır sonunun en sonda olması koşulu eklenmiştir.
- Satır sonundan boşluk olana kadar tüm karakterler dahil edilmiştir.

Başarı

Adres tanımada satır olarak tüm satırları tanımaktaki başarısı 6831 adres içerisinden 36 adet adresi bir alt satırdaki adres ile birleştirerek yanlış tanıdığı için %99,4 tür. Bu duruma örnek aşağıdaki görsel verilebilir.

```
AYAZAGA MAH MUSTAFA KEMAL ATATÜRK CAD NO 24/B SARIYER/ İSTANBUL
BOZHANE MAH. ÇARŞI CAD. NO:2 BEYKOZ/ İSTANBUL
EĞİTİM MAH. ABDİBEY SK NO:49/A KADIKÖY/ İSTANBUL
FINDIKLI MAH ÇINAR CAD NO 37/B MALTEPE/ İSTANBUL
CADDEBOSTAN MAH. PROF. DRÇ HULUSİ BEHÇET CAD. NO:2 IÇ KAPI NO:1 KADIKÖY/ İSTANBUL
KİRAZLITEPE MAH. TOMRUK CAD. NO:A/1 ÜSKÜDAR/ İSTANBUL
VELİEFENDİ MAH 75 SOK NO 25/A ZEYTİNBURNU/ İSTANBUL
FETIH MAH ADNAN MENDERES CAD NO:3 ATAŞEHİR/ İSTANBUL
```

Başarılı sınıflandırmalara örnek olarak,



Mahalle sokak isimleri bulunmayan adresler için başarısız sonuçlar elde edilmiştir. Örnek olarak,

Kaynak

Geliştirme ortamı olarak regex101.com kullanılmıştır.

Bağlantı: https://regex101.com/r/QY86yX/10

Not : grup bazlı başarı ölçülememiştir.