



**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*[V|-]*\s*[A-Z1-9]{0,4}\s|  
(\s?[A-Z,Ç,Ğ,İ,Ö,Ş,Ü]+\s*[V])| ([^\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*[V|-]*\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*[V])
```

- 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 tanımda satır olarak tüm satırları tanımaktaki başarısı 6831 adres içerisinde 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	İÇ KAPI	NO:1 KADIKÖY/ İSTANBUL
KİRAZLITEPE MAH.	TOMRUK CAD.	NO:A/1	ÜSKÜDAR/	İSTANBUL
VELİEFENDİ MAH	75 SOK	NO 25/A	ZEYTİNBURNU/	İSTANBUL
FETİH MAH	ADNAN MENDERES CAD	NO:3	ATAŞEHİR/	İSTANBUL

Başarılı sınıflandırmalara örnek olarak,

REGULAR EXPRESSION v10 31656 matches, 8867567 steps (~4.44s)

```
:/ (. * \s M[A]{0,1} H[. | \s | A]{0,1} ) ( (. * \s C[A]{0,1} D[ \s | . | D]{0,1} ) |  
(. * \s S[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 ] )
```

TEST STRING

ERTUĞRUL GAZİ MAH. ATIF EFENDİ CAD. NO:41 PENDİK/ İSTANBUL  
GÜZELYALI MAH. YANYOL CAD. NO:3 PENDİK/ İSTANBUL  
YENİSAHRA MAH. YAVUZER CAD. NO:32/F ATAŞEHİR/ İSTANBUL  
KAYIŞDAĞI MAH. EĞEMENLİK SOK. NO:1/A ATAŞEHİR/ İSTANBUL  
HEKİMBAŞI MAH. KÜÇÜKSU CAD. NO:371 ÜMRANIYE/ İSTANBUL  
ATATÜRK CAD. YAYLA SOK. NO:3/A GÜVEN İŞ MERKEZİ ÖNÜ MALTEPE/ İSTANBUL  
AYAZAĞA MAH. ATATÜRK CAD. NO:57/B SARIYER/ İSTANBUL  
HAMİDİYE MAH. GİRNE CAD. NO:2/B KAĞITHANE/ İSTANBUL  
ERENKÖY MAH. FAHRETTİN KERİM GÖKAY CAD. NO:292 KADIKÖY/ İSTANBUL  
"YENİKÖY MAH. GÜZELCE ALİ PAŞA CAD. NO:67/A " SARIYER/ İSTANBUL  
PİYALEPAŞA MAH. FATİH SULTAN CAD. NO:298/4 BEYOĞLU/ İSTANBUL  
EMİRGAN MAH. BALTALIMANI CAD. NO: 74 SARIYER/ İSTANBUL  
ERGENEKON CAD. NO:1/1 ŞİŞLİ/ İSTANBUL  
ÇELİKTEPE MAH. İNÖNÜ CAD. NO: 53A/A KAĞITHANE/ İSTANBUL  
EMİRGAN CADDESİ NO:122 SARIYER/ İSTANBUL  
ALT KAYNARCA MAH. YANYOL CAD. NO:166/B PENDİK/ İSTANBUL  
PİYADE ER YAŞAR YEŞİM KIR SOKAK NO:31 ZEYTİNBURNU/ İSTANBUL  
İSMET PAŞA MAH. ÇİMEN SOK. NO:24 BAYRAMPAŞA/ İSTANBUL  
TUNA MAH. 65 SOK NO:10 ESENLER/ İSTANBUL  
NURİ PAŞA CAD. NO: 80/G SARIYER/ İSTANBUL  
ABDİ İPEKÇİ CAD. NO:56 GÜNGÖREN/ İSTANBUL  
MEHMET AKİF ERSOY MAH. FATİH BULVARI NO:131 SULTANBEYLİ/ İSTANBUL  
YENİ MAH. ATATÜRK CAD. NO:78/B KARTAL/ İSTANBUL  
BALABAN AĞA MAH. VEZNEÇİLER CAD. FATİH/ İSTANBUL

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

YENİBOSNA METRO İSTASYONU BAKIRKÖY/ İSTANBUL  
KARAKÖY YER ALTI GEÇİDİ NO:24 BEYOĞLU/ İSTANBUL  
YENİ DOĞAN MAH. KIŞLA CAD. NUR SANAYİ SİTESİ NO: 69 BAYRAMPAŞA/ İSTANBUL  
DENİZ KÖŞKLER MAH. ESKİ LONDRA ASFALTI NO: 22/3 AVCILAR/ İSTANBUL  
ORTAÇEŞME SONDURAK BEYKOZ/ İSTANBUL  
MEHMET AKİF ERSOY MAH. NATO YOLU BOSNA BULVARI NO: 115 ÜSKÜDAR/ İSTANBUL  
DERBENT MAH. DEREİCİ OTOBÜS SON DURAK SARIYER/ İSTANBUL

## Kaynak

Geliřtirme ortamı olarak regex101.com kullanılmıřtır.

Baęlantı: <https://regex101.com/r/QY86yX/10>

Not : grup bazlı başarı ölçülememiřtir.