



Doğal Dil İşleme

1. Ödev Raporu

17011044 - Uğur Keskin

Ödevin Tanımı

- Verilen adres dosyası içerisinde, adresleri doğru ve düzenli bir şekilde ayrıştıracak RegEx Pattern'ini bulmak.

Ödev Notları

- **Parçalar ve sıraları:**

Adresleri Mahalle, Cadde, Sokak, Blok, Kat, Bulvar, Site, No, No ismi, İş Hanı, Apartman, İl ve İlçeleri düzenli şekilde ayrıştıracak şekilde bir RegEx yazmayı amaçladım.

Bu parçaların sıraları, gözlemlediğim kadarıyla

Mahalle -> Bulvar ->Asfalt -> Cadde -> Site -> Blok -> Sokak ->No -> İlçe -> İl,

İş Hanı: Sokaktan sonra, İlçeden önce

Apartman İsmi: Sokaktan sonra İlçe'den önce

Kat Tarifi : Bloktan sonra İlçe'den önce yer almaktadır.

No İsmi, (Örn. Dükkan No: 3 dizimindeki “Dükkan” **No** ismidir.), **İş Hanı**, **Apartman İsmi** ve **Kat Tarifi** gibi parçalarla karıştırılabileceği için bu konudaki çözüm regex’i daha da karmaşıktırabileceği için bu gibi şeyleri Sokak ve İlçe arasında aramak yerine Sokak ve No arasında aramayı, No’dan sonraki kısmın ancak ilçe olabileceği varsayıp bunun dışındaki satırları matchlemeyen bir regex oluşturmaya gayret ettim.

Durak, **Önü/Yanı/İçi** gibi tarifler de mahalle ve ilçe arasında bulundukları için bu işi daha da zorlaştırmaktadır. Apartman Önü, Blok Yanı gibi kavramlardan dolayı bu tarifleri yakalamak çok zordur. Bu yüzden bu kısımlar, wildcard kullanarak “Diğer” başlığı altına alınıp kolayca kaçmak **yerine** doğruluğu daha doğru hesaplayabilmek adına göz ardı edilmiştir.

RegEx:

```
^(?:(?<Mahalle>"?.*?)MA?H(?:ALLESİ)?[\\.\\/\,]?)?\s*(?:(?<Bulvar>.*?)BU?LV(?:\.?|(?:(?:ARI)))?)?\s*(?:(?<Asfalt>.*?)ASFALTI)?\s*(?:(?<Cadde>.*?)CA?D(?::(?:DES[İİ]))?[\\.\\/\,]?)?\s*(?:(?<Site>.*?)S[İİ]T(?:\.?|(?:(?:ES[İİ]))?)?\s*(?:(?<Blok>.*?)BL(?:K.|O(?:K|ĞU)\.*)?)?\s*(?:(?<Sokak>.*?)S(?:O?K\.?|OKAK|\.))?\s*(?:(?<Apartman>.*?)APT\.)?\s*(?:(?<is_HANI>.*?)İŞ ?HANI)?\s*(?:(?:N[00]? *[:.])?(?<No>[0-9] *[/\ -A-Z0-9]{0,5}"?))?\s+(?<ilce>[a-zA-ZğüşöçİĞÜŞÖÇ]+)\s*\s*(?<il>[a-zA-ZğüşöçİĞÜŞÖÇ]+)$
```

```
^(?:(?<Mahalle>"?.*?)MA?H(?:ALLESİ)?[\\.\\/\,]?)?\s*(?:(?<Bulvar>.*?)BU?LV(?:\.?|(?:(?:ARI)))?)?\s*(?:(?<Asfalt>.*?)ASFALTI)?\s*(?:(?<Cadde>.*?)CA?D(?::(?:DES[İİ]))?[\\.\\/\,]?)?\s*(?:(?<Site>.*?)S[İİ]T(?:\.?|(?:(?:ES[İİ]))?)?\s*(?:(?<Blok>.*?)BL(?:K.|O(?:K|ĞU)\.*)?)?\s*(?:(?<Sokak>.*?)S(?:O?K\.?|OKAK|\.))?\s*(?:(?<Apartman>.*?)APT\.)?\s*(?:(?<is_HANI>.*?)İŞ ?HANI)?\s*(?:(?:N[00]? *[:.])?(?<No>[0-9] *[/\ -A-Z0-9]{0,5}"?))?\s+(?<ilce>[a-zA-ZğüşöçİĞÜŞÖÇ]+)\s*\s*(?<il>[a-zA-ZğüşöçİĞÜŞÖÇ]+)$
```

Ayrımlar

Mahalle:

```
(?: (?<Mahalle> " ? .* ) MA ? H (?: ALLES İ ) ? [ \ . \ / \ , ] ? ) ? \s *
```

Bulvar:

```
(?: (?<Bulvar> . *) BU ? LV (?: \ . ? | (?: ARI ) ) ? ) ? \s *
```

Asfalt:

```
(?: (?<Asfalt> . *) ASFALTI ) ? \s *
```

Cadde:

```
(?: (?<Cadde> . *) CA ? D (?: (?: DES [ İ İ ] ) ) ? [ \ . \ / \ , ] ? ) ? \s *
```

Site:

```
(?: (?<Site> . *) S [ İ İ ] T (?: \ . | (?: ES [ İ İ ] ) ) ? ) ? \s *
```

Blok:

```
(?: (?<Blok> . *) BL (?: K . | O (?: K | Ğ U ) \ . * ) ? ) ? \s *
```

Sokak:

```
(?: (?<Sokak> . *) S (?: O ? K \ . ? | OKAK [ \ . ] ) ? ) ? \s *
```

Apartman:

```
(?: (?<Apartman> . *) APT \ . ? ) ? \s *
```

İş Hanı:

```
(?: (?<is_HANI> . *) İ Ş ? HANI ) ? \s *
```

No:

```
(?: (?: N [ 0 0 ] ? * [ : . ] ? ) ? (?: <No> [ 0 - 9 ] * [ \ / \ - A - Z 0 - 9 ] { 0 , 5 } " ? ) ) ? \s +
```

ilçe / il:

```
(?: <ilce> [ a - z A - Z ğ ü ş ö ç İ Ğ Ü Ş Ö Ç ] + ) \s * \ / \s * (?: <il> [ a - z A - Z ğ ü ş ö ç İ Ğ Ü Ş Ö Ç ] + ) $
```

Ekran Görüntüleri

/^((?:<Mahalle>.*))MA?H(?:ALLESİ)?[\\.\\/\,]??\s*((?:<Bulvar>.*))BU?LV(?:\.?|(?:(?:ARI)))??

Text Tests

YUKARI DUDULLU.. MAH TAVUKÇUYOLU CADDESİ TANDOĞANAY. SOK 2/B1/1-ÜMRANIYE/İSTANBUL

Tools

| | | |
|------------|-----|------------------|
| 'Mahalle' | n/a | YUKARI DUDULLU.. |
| 'Bulvar' | n/a | <empty> |
| 'Asfalt' | n/a | <empty> |
| 'Cadde' | n/a | TAVUKÇUYOLU |
| 'Site' | n/a | <empty> |
| 'Blok' | n/a | <empty> |
| 'Sokak' | n/a | TANDOĞANAY. |
| 'Apartman' | n/a | <empty> |
| 'is_HANI' | n/a | <empty> |
| 'No' | n/a | 2/B1/1 |
| 'ilce' | n/a | ÜMRANIYE |
| 'il' | n/a | İSTANBUL |

/^((?:<Mahalle>.*))MA?H(?:ALLESİ)?[\\.\\/\,]??\s*((?:<Bulvar>.*))BU?LV(?:\.?|(?:(?:ARI)))??\s*((?:<?<

Text Tests

29 EKİM CAD. KUYUMÇUKENT KOMPLEKSİ ATOLYE BLOĞU ZEMİN KAT 8. SOKAK NO:18 BAĞÇELİEVLER/İSTANBUL

Tools

| | | |
|------------|-----|------------------------------|
| 'Mahalle' | n/a | <empty> |
| 'Bulvar' | n/a | <empty> |
| 'Asfalt' | n/a | <empty> |
| 'Cadde' | n/a | 29 EKİM |
| 'Site' | n/a | <empty> |
| 'Blok' | n/a | KUYUMÇUKENT KOMPLEKSİ ATOLYE |
| 'Sokak' | n/a | ZEMİN KAT 8. |
| 'Apartman' | n/a | <empty> |
| 'is_HANI' | n/a | <empty> |
| 'No' | n/a | 18 |
| 'ilce' | n/a | BAĞÇELİEVLER |
| 'il' | n/a | İSTANBUL |

/^((?:(<Mahalle>.*))MA?H(?:ALLESİ)?[\\.\\/\\,]?)?\\s*((?:(<Bulvar>.*))BU?LV(

Text

Tests

AKSARAY MAH.TİRYAKİ HASANPAŞA SOK.VALDE İŞ HANI 39C FATİH/İSTANBUL

Tools

| | | |
|------------|-----|-------------------|
| 'Mahalle' | n/a | AKSARAY |
| 'Bulvar' | n/a | <empty> |
| 'Asfalt' | n/a | <empty> |
| 'Cadde' | n/a | <empty> |
| 'Site' | n/a | <empty> |
| 'Blok' | n/a | <empty> |
| 'Sokak' | n/a | TİRYAKİ HASANPAŞA |
| 'Apartman' | n/a | <empty> |
| 'is_HANI' | n/a | VALDE |
| 'No' | n/a | 39C |
| 'ilce' | n/a | FATİH |
| 'il' | n/a | İSTANBUL |

Performans

RegEx Testing
From Dan's Tools

Web Dev

Regular Expression

JavaScript flags

/^((?<Mahalle>.*MA?H(?:ALLESİ)?[.\\/\s,]?)*(?:(?<Bulvar>.*BU?LV(?:\s|(?<ARI))?)|(?<Asfalt>.*A

5834 matches

Test String

"MEHMET AKİF MAH. FATİH BULVARI NO: 147/B " SULTANBEYLİ/ İSTANBUL
"NAMIK KEMAL MAH. SÜTÇÜ İMAM CAD." ÜMRANİYE/ İSTANBUL
KORDONBOYU MAH. ANKARA CAD. TOGAY APT. ANT SİTESİ A BLOK 175/1 KARTAL/ İSTANBUL
"KAZİM KARABEKİR MAH. ERZURUM CAD. NO:113/1" ÜMRANİYE/ İSTANBUL
"EVLIYA ÇELEBİ MAH. SEMT FERASET SOK. 6A" TUZLA/ İSTANBUL
BALAT MAH. MANYAS ZADE CAD. NO: 43 FATİH/ İSTANBUL
"SULTANİYE MAH. BALIK YOLU CAD. NO:2/C" ESENYURT/ İSTANBUL
"FETİH TEPE MAH. FATİH SULTAN CAD. 87/B" BEYOĞLU/ İSTANBUL
"PINAR MAH. BALABAN DERE CAD. NO:16" SARIYER/ İSTANBUL
"SANCAR TEPE MAH. BİRLİK CAD. NO:40/B" BAĞCILAR/ İSTANBUL
"HAVAALANI MAH. TAŞOCAĞI CAD. NO:12/C" ESENLER/ İSTANBUL
"PINAR MAH. 1271 SOK. NO: 50/2" ESENYURT/ İSTANBUL
KARTALTEPE MAH. ÇUKUROVA CAD. NO: 1A BAYRAMPAŞA/ İSTANBUL
AYDINEVLER MAH.SELÇUKBEY CAD. NO:43/B MALTEPE/ İSTANBUL
İHLAMURKUYU MAH. ALEMDAĞ CAD. NO:263/B ÜMRANİYE/ İSTANBUL
FATİH MAH. SEMT YAKACIK CAD 74.A SANCARTEPE/ İSTANBUL
KAVACIK MAH. NÜVE SOK. NO:24 D.1 BEYKOZ/ İSTANBUL
MİMAR SİNAN MAH. İSTANBUL CAD. NO:64 ESENLER/ İSTANBUL
ATAKENT MAH. 221. SOK. NO:3/C ROTA OFFICE NO:28 KÜÇÜKÇEKMECE/ İSTANBUL

RegEx sonucunda 5849 tane satır ile eşleşilmiştir. Bu da toplam satırın 6832 olduğunu düşünürsek adreslerin **%85,5'**ü ile eşleştirilmiştir, yazılan regEx doğru ayırım yapılamayan satırları eşleştirilmemek üzere tasarlandığı için ve "Diğer" gibi wildcard bir parça bulunmadığı için, eşleştirilen kısmın çok büyük oranının adresi doğru parçalarına ayırdığını varsayabiliriz.

Bulunamayan Adresler

Eşlenmemiş 983 tane adres vardır. Bunların eşlenmemesinin nedeni **en az** bir parçasının (**Durak, Kat Numarası** vs.) ayırt edilemiyor olması yani gruplanamıyor olmasıdır.

Bu adresler:

- Numara ve İlçe arasında başka kısımlar olan adresler
- Yanı/Önü/Karşısı/Kesişimi/İçi gibi tarif ifadeleri bulunan adresler
- Durak, Bayi, AVM, Merkez, Büfe gibi ifadeleri içeren adresler
- En az bir parçası gruplandıramayan tüm adresler (Eğer sadece bir parçası gruplandırılmış bir adres olsaydı bu adresin tüm parçalarının doğru bir şekilde gruplanıp gruplanmadığından emin olamayacağımız/ bunu test edemeyeceğimiz için bu adresleri eşlemeyecek bir regex yazdım)

Bulunamayan adreslerin birkaçı:

- YENİBOSNA METRO İSTASYONU BAKIRKÖY/ İSTANBUL
- KENNEDY CAD. SİRKEÇİ ARABALI VAPUR İSKELESİ FATİH/ 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
- YILDIZ POSTA CAD. TÜRK TELEKOM ÖNÜ GAZETE BAYİİ BEŞİKTAŞ/ İSTANBUL
- ARMAĞAN EVLER MAH. ALEMDAĞ CAD. SİTE OTOBÜS DURAĞI YANI ÜMRANIYE/ İSTANBUL
- ORTAÇEŞME SONDURAK BEYKOZ/ İSTANBUL
- DERBENT MAH. DEREİCİ OTOBÜS SON DURAK SARIYER/ İSTANBUL
- BÜYÜKDERE CAD. NİMET ABLA CAMİ YANI ŞİŞLİ/ İSTANBUL
- TOPKAPI TİCARET MERKEZİ KARŞISI ÜSTGEÇİT ALTI ZEYTİNBURNU/ İSTANBUL
- MİLLET CAD. YUSUF PAŞA ZİRAAT BANKASI ÖNÜ FATİH/ İSTANBUL
- MALTEPE MAH. MATBAACILAR SİTESİ MALTEPE METROBÜS ÇIKIŞI NO:133 BAYRAMPAŞA/ İSTANBUL
- MEHMET AKİF MAH. TEPE ÜSTÜ GÜR DEMİREL C29 KÜÇÜKÇEKMECE/ İSTANBUL
- HÜRRİYET MAH. KUYU K.K NO:2 E BAĞCILAR/ İSTANBUL
- ERENKÖY MAH. ŞEMSETTİN GÜNALTAY CAD. ÖĞRETMEN HAYRULLAH DURAĞI KADIKÖY/ İSTANBUL
- SAHİP MOLLA CAD. NO:37 PAŞABAHÇE BEYKOZ/ İSTANBUL
- FİKİRTEPE MAH FAHRETTİN KERİM GÖKAY CAD. ABDİBEY SOK. KESİŞİMİ KADIKÖY/ İSTANBUL
- KADIKÖY DENİZ OTOBÜSÜ İSKELE ÖNÜ RASİMPAŞA KADIKÖY/ İSTANBUL
- E5 KARAYOLU ÜZERİ GÖZTEPE OTOBÜS DURAGI YANI KADIKÖY/ İSTANBUL
- ÇAĞLAYAN MAH. ÇAĞLAYAN BENZİN İSTASYONU YANI KAĞITHANE/ İSTANBUL