

RegEx ile Adres Ayrıştırma

10/11/2020

BLM 4580 Doğal Dil İşlemeye Giriş
Banu Diri

Ahmet Onur Akman
16011059



Giriş

Problem

Text biçiminde verilen veri içerisindeki adresleri ayırtmak amacına uygun bir Regular Expression oluşturmak. Bu adreslerin mahalle, cadde, sokak, kapı numarası, ilçe, şehir bilgilerini ayırtmak.

Veri

Elimizdeki veri text formatındadır ve 6831 adet satır içermektedir. Bu satırların her biri farklı bir adresi barındırmaktadır. Bu adreslerin büyük bir kısmı, Türkiye adres yazım alışkanlıklarına uyumlu olacak şekilde “Mahalle, Cadde, Sokak, Ek Bilgiler, Numara, Ek Bilgiler, İlçe, Şehir” formatında verilmiştir ancak oluşturulacak RegEx ile bu formata uymayan adres bilgilerinin de başarıyla okunup ayrıştırılması hedeflenmektedir.

Yaklaşım

Ele alınan problem bir metnin regular expression ile taranıp ayrıştırılmasını gerektirdiğinden, çalışma ortamının yazılan ifadenin syntax'ını kontrol edebilen, başarısını ölçebilen ve verilen veri üzerindeki taramasını ayrıntılarıyla bize gösterebilen bir ortam olması gerekliliği ortaya çıkmıştır. Bu gereklilik doğrultusunda www.regex101.com sitesi çalışma ortamı olarak seçilmiştir.

Türkiye'deki ev ve ofis adreslerinin yazım formatı başka ülkelere kıyasla daha karmaşık bir formattadır. Bu format doğrultusunda adresi verilecek mekanın konum bilgisinin bütün detayları adrese dahil edilir ve genellikle bu bilgiler büyük yerleşim yerinden daralarak yazılır. Bu ödev kapsamında ele alınmış olan adreslerin geneli de bu şekilde yazılmıştır.

Bazı istisnalar dışında, adreslerin genel yazılış tarzlarının şu ilerleyişe sadık kaldığı görülmektedir:

Mahalle, Cadde, Sokak, Numara, ilçe, Şehir

Bu formatın yanı sıra adres içerisinde site, kompleks, iş hanı veya AVM ismi, şehirler arası yola göre tarif gibi detaylar da eklenmektedir. Bahsedilen bu detayların adreste yerleştirilebileceği her yerin ayrı ayrı taranması ve her ek bilginin neyi ifade ettiğinin çözümlenmeye çalışılması ifadeyi oldukça kalabalıklaştıracığından, adreslerdeki genel yazılış biçimi incelenmiş ve bu ek bilgileri en yüksek doğrulukta yakalamak ve kodu da basitlikten uzaklaştırmamak hedefiyle bir RegEx formatı belirlenmiştir. Buna göre yazılan RegEx'in formatı aşağıdaki gibi planlanmıştır.

Mahalle (Varsa), Cadde (Varsa), Sokak (Varsa), Ek Bilgiler (Varsa), Numara (Varsa), Ek bilgiler (Varsa), İlçe (Zorunlu), Şehir (Zorunlu)

Eldeki adres bilgilerinin tamamı İstanbul içi adresler olmasına rağmen oluşturulan RegEx'te bu durum gözardı edilmiştir.

Adresler arası ele alınması gereken başka bir problem de mahalle, sokak gibi kelimelerin yazılışlarının veya kısaltma biçimlerinin yazan kişinin tercihinine bağlı olarak değişiyor olmasıdır. Yazılan RegEx ile muhtemel bütün kısaltmaların bulunması sağlanmıştır. Buna göre muhtemel olduğu varsayılan yazılışlar ve kısaltmalar şu şekildedir:

MH, MAH, MAHALLE, MAHALLESİ, MAHALLESİ: **Mahalle**

CD, CAD, CADDE, CADDESİ, CADDESİ: **Cadde**

SK, SOK, SOKAK, SOKAĞI, SOKAGI: **Sokak**

N., N:, NO., NO:, NR., NR:: **Numara**

Bu ve bunun gibi adres parçalarının içerebileceği karakterler ve aralarına konulacak boşluk ve noktalama işaretleri düşünülerek her senaryoda başarılı olunması hedeflenmiştir.

Oluşturulan RegEx ve Başarı Oranı

Regular Expression

Yaklaşım bölümünde anlatılmış olan faktörler göz önünde bulundurularak aşağıdaki RegEx oluşturulmuştur:

a) Renklendirilmiş Format

```
/ ^(\\"?((?P<Mahalle>[^\n]+)(MAHALLE|MAHALLES[İİ]|MA?H)[\.,\s]+)?((?P<Cadde>[^\n]+)(CADDES[İİ]|CADDE|CA?D)[\.,\s]+)?((?P<Sokak>[^\n]+)(\s|\.)+(SOKA[GĞ]I|SOKAK|SO?K)[\.,\s]+)?(?P<Ekler1>[\w\sĞÜŞİÖÇÂ()\.\"-\",]+[\.,\s]+)?((NO|NR|N)?(\s*)(\s|\s*):)?(\s*)(?P<No>[\wĞÜŞİÖÇ\.\s\"-:]{1,7})[\.,\s]+)?(?P<Ekler2>[^\n]+[\.,\s]+)?((?P<Semt>[^\n\s]+\s)/)(\s*)(?P<Sehir>[^\d\s]+))$
```

b) Ayırıştırılmış Format

^(\\"?

((?P<Mahalle>[^\n]+)(MAHALLE|MAHALLES[İİ]|MA?H)[\.,\s]+)?

((?P<Cadde>[^\n]+)(CADDES[İİ]|CADDE|CA?D)[\.,\s]+)?

((?P<Sokak>[^\n]+)(\s|\.)+(SOKA[GĞ]I|SOKAK|SO?K)[\.,\s]+)?

(?P<Ekler1>[\w\sĞÜŞİÖÇÂ()\.\"-\",]+[\.,\s]+)?

((NO|NR|N)?(\s*)(\s|\s*):)?(\s*)(?P<No>[\wĞÜŞİÖÇ\.\s\"-:]{1,7})[\.,\s]+)?

(?P<Ekler2>[^\n]+[\.,\s]+)?

((?P<Semt>[^\n\s]+\s)/)(\s*)

(?P<Sehir>[^\d\s]+)

)\$

c) Regex101 Linki

regex101.com/r/TXOF5X/2

Başarı Durumu

Oluşturulan RegEx, verilen adres bilgilerinin tamamını ayrı ayrı olacak şekilde algılamayı başarmıştır. RegEx ile veri üzerinde yapılan tarama sonucunda, il ve ilçe bilgileri dahil olmak üzere en az 3 adres parçasına sahip **6831** adet adres tespit edilmiştir.

Kullanılan ortamın şartları sebebiyle, yapılan tarama sonucu bulunan mahalle, cadde, sokak, no, ilçe, şehir ve ek bilgiler ayrı ayrı olacak şekilde bir veri yapısında toplanamamıştır. Modelin doğruluk oranını daha iyi gözlemlemek için önceki sayfada verilen RegEx101 linkini ziyaret edebilirsiniz.

Muhtemel olarak değerlendirilen formatların dışında yazılmış olan, yazım hataları içeren ve diğer adreslere kıyasla beklenmedik bilgiler içeren adreslerin okumalarında yüksek bir başarı oranı elde edildiyse de, aşağıdaki 10 adet örneği verilen bazı adreslerin ayrıştırılmasında karışıklıklar meydana gelmiştir:

- 1) **KUYUMCU KENT ATÖLYE DURAĞI, 29 EKİM CAD. LADİN SOK. K:1, SK:9, NO:6 BAHÇELİEVLER/ İSTANBUL**
“Kuyumcu Kent Atölye Durağı” bilgisi normalde ek bilgi olarak kabul edilmektedir. Burada beklenenin aksine adresin en başına yazıldığı için ek bilgi olarak değerlendirilmemiş, cadde ismi ile birleştirilmiştir.
- 2) **TARLABAŞI BULVARI NO:190/B BEYOĞLU/ İSTANBUL**
“Tarlabaşı Bulvarı” ifadesi, “bulvar” bilgisinin nadiren veriliyor olması ve verildiği çoğu zaman “xxx bulvarı caddesi” şeklinde veriliyor olması sebepleriyle “Bulvar” olarak değil “Ek bilgi” olarak sınıflandırılmıştır.
- 3) **SİNAN PAŞA MAH. BARBAROS BULVARI KÖY İÇİ CAD. GİRİŞİ NO:1/7 BEŞİKTAŞ/ İSTANBUL**
“CAD. GİRİŞİ” ifadesi nadir kullanılan bir ifade olduğu için muhtemel bir senaryo olarak değerlendirilmemiştir. Bu sebeple RegEx bu adreste “Girişi” ifadesinin adresin öncesinden ayırarak “Ek bilgi” olarak sınıflandırmıştır.
- 4) **AKSEMSETTİN MAH FATİH BULVARI NO : 505 SULTANBEYLİ/ İSTANBUL**
Burada RegEx “NO :” ifadesini önceki parçadan ayıramamıştır. Bu sebeple “Fatih Bulvarı No :” ifadesi “Ek bilgi”, “505” verisi “No” olarak sınıflandırılmıştır.

5) **METROGARDEN AVM NECİP FAZIL MH. DUDULLU YOLU ÜZERİ, MABEYN CAD. NO:3-A**
ÜMRANIYE/ İSTANBUL

Bu adreste verilen ek bilgilerin yerleri diğer adreslere kıyasla sıradışı bir şekilde verildiği için, altı çizili bloklar sırasıyla “Mahalle” ve “Cadde” olarak sınıflandırılmıştır.

6) **ÜSKÜDAR CADDESİ. YUKARI MH. BAYRAMLAR İŞ MERKEZİ. NO:9 KARTAL/ İSTANBUL**

Bu adreste kullanılan format öngörülenin aksine cadde -> mahalle olarak ilerlemiştir. Adreste ilk olarak mahalle bilgisinin varlığını sorgulayan RegEx, altı çizili alanı “Mahalle” olarak sınıflandırmıştır.

7) **BATTALGAZİ MAH BOSNA BULV 86A SULTANBEYLİ/ İSTANBUL**

Burada numara bilgisinden önce “NO:” veya diğer alternatif ifadeler kullanılmadığı ve “86A” ifadesinin “Ek bilgiler” sınıflandırması yazım kuralları gerekliliklerini karşılaması sebebiyle altı çizili blok “Ek bilgi” olarak sınıflandırılmıştır.

8) **CAMİ MAH. BALIKÇILAR SOK. - SİT. C BLOK APT. NO: 20 C / 96 TUZLA/ İSTANBUL**

Bu adreste “No” bilgisi beklenenden fazla karakter sayısı ile ifade edildiği için altı çizili alan “Ek bilgi” olarak sınıflandırılmıştır.

9) **VIAPORT AVM YENİSEHİR MAH.DEDEPASA CAD.NO:19 PENDİK/ İSTANBUL**

Normalde “Ek bilgi” olarak algılanması gereken “Viaport AVM” ifadesi adresin en başına yazıldığı için kendisinden sonra gelen ifadeden ayrıştırılamamış, altı çizili bölgenin tamamı “Mahalle” olarak sınıflandırılmıştır.

10) **EMNİYETEVLERİ M. ÇELEBİ MEHMET S. NO:12/ KAĞITHANE/ İSTANBUL**

Mahalle ve Sokak ifadeleri beklenen ifade ve kısaltma formatları dışında kısaltıldığı için altı çizili alanın tamamı “Ek Bilgi” olarak tanımlanmıştır.

Ekran Görüntüleri

REGULAR EXPRESSION v2 ▾ 6831 matches, 5054163 steps (~4.42s)

```
/ / \((?P<Mahalle>[^\n]+)(MAHALLE|MAHALLES|I|I|MAH|H)[^\.,\s]+)?((?P<Cadde>[^\n]+)(CADES|I|I|CADE|C|A|D)[^\.,\s]+)?((?P<Sokak>[^\n]+)(\s|\.)+|SOKA|GG|I|SOKAK|SO|K)[^\.,\s]+)?(?P<Eklr1>[^\w$GÜSİÖÇ\(\)\.\-\,\,]+[^\.,\s]+)?((NO|NR|N)?(?P<S>)(\s|\.)+)?(?P<No>[^\w$GÜSİÖÇ\(\)\.\-\,\,]+[1,7])(\s|\.)+)?(?P<Eklr2>[^\n]+[^\.,\s]+)?((?P<Semt>[^\n\s]+)\/)(\s*)(?P<Sehir>[^\d\n\s]+))$
```

TEST STRING

MEHMET AKİF MAH. AYDIN MENDERES CAD. NO: 73/A ÇEKMEKÖY/ İSTANBUL
MUSTAFA KEMAL PAŞA MAH. FİDAN SOK. NO: 105A/4 AVCILAR/ İSTANBUL
MENDERES MAH. 377 SOK. NO: 29 ESENLER/ İSTANBUL
KEMER MAH. BÜLBÜL SOK. NO: 17/A SARIYER/ İSTANBUL
MİMAR SİNAN MAH. MİMAR SİNAN CAD. ÇALTEPE SOK. NO: 13/B ÇEKMEKÖY/ İSTANBUL
NEHEHATUN MAH. AZİZİYE CQ. NO: 50/A ESENLER/ İSTANBUL
TURGUT ÖZAL CAD. OKTAY RIFAT CAD. NO: 90/92 ESENYURT/ İSTANBUL
ESENKENT MAH. YILMAZ SOK. NO: 150/A MALTEPE/ İSTANBUL
MERKEZ MAH. USKÜDAR CAD. NO: 23 ÇEKMEKÖY/ İSTANBUL
ESENŞEHİR MAH. MAREŞAL FEVZİ ÇAKMAK CAD. NO: 37/A ÜMRANİYE/ İSTANBUL
ÇİFTEHAVUZLAR MAH. DAVUTPAŞA KAMPÜS İCİ ESENLER/ İSTANBUL
MİMAR SİNAN MAH. RUMELİ CAD. NO: 8/A ESENLER/ İSTANBUL
YILDIRIM MAH. MİLLET CAD. NO: 94/96A BAYRAMPAŞA/ İSTANBUL
AŞAĞI DUDULLU MAH. YENİ SOK. NO: 25/A ÜMRANİYE/ İSTANBUL
ALTINTEPSİ MAH. MALAZGİRT SOK. NO: 33/A-B BAYRAMPAŞA/ İSTANBUL
AŞAĞI DUDULLU CAMGAZ MEVKİİ ÇOBANÇESME CAD. NO: 50 ÜMRANİYE/ İSTANBUL
İNCİRTEPE MAH. 229 SOK. NO: 20/A ESENYURT/ İSTANBUL
SÜMER MAH. ŞEHİT ER YAVUZ BAHA SOK. NO: 17 ZEYTİNBURNU/ İSTANBUL
AŞAĞI DUDULLU HUZUR MAH. ÇAMLICA CAD. NO: 65 ÜMRANİYE/ İSTANBUL
SÜMER MAH. MERV CAD. NO: 101 ZEYTİNBURNU/ İSTANBUL
KOCATEPE MAH. 32 SOK. NO: 35/B BAYRAMPAŞA/ İSTANBUL
MEVLANA MAH. KUŞKONMAZ SOK. NO: 19/B ATAŞEHİR/ İSTANBUL
TAYAKADIN MAH. TAYAKADIN YASSİÖREN CAD. NO: 66A ARNAVUTKÖY/ İSTANBUL
ÇİRPİCİ MAH. SEYİTNİZAM CAD. NO: 108 ZEYTİNBURNU/ İSTANBUL
NURİPAŞA MAH. MERV CAD. NO: 12 ZEYTİNBURNU/ İSTANBUL
KARABURUN: KARABURUN KÖYÜ SOK. NO: 13/B ARNAVUTKÖY/ İSTANBUL
NURİPAŞA MAH. 14 SOK. NO: 17 ZEYTİNBURNU/ İSTANBUL
TELSİZ MAH. 72 SOK. NO: 13 ZEYTİNBURNU/ İSTANBUL
GÖKALP MAH. 39/4 SOK. NO: 26/40A ZEYTİNBURNU/ İSTANBUL
PARSELLER MAH. TAŞKÖPRÜ CAD. NO: 19/A ÜMRANİYE/ İSTANBUL
CEHİL MERİÇ MAH. ÇAYIRONU CAD. NO: 156A/A ÜMRANİYE/ İSTANBUL
YUKARI DUDULLU MAH. GÜVERCİN SOK. NO: 8A/A ÜMRANİYE/ İSTANBUL

REGULAR EXPRESSION v2 ▾ 6831 matches, 5054163 steps (~4.42s)

```
/ / \((?P<Mahalle>[^\n]+)(MAHALLE|MAHALLES|I|I|MAH|H)[^\.,\s]+)?((?P<Cadde>[^\n]+)(CADES|I|I|CADE|C|A|D)[^\.,\s]+)?((?P<Sokak>[^\n]+)(\s|\.)+|SOKA|GG|I|SOKAK|SO|K)[^\.,\s]+)?(?P<Eklr1>[^\w$GÜSİÖÇ\(\)\.\-\,\,]+[^\.,\s]+)?((NO|NR|N)?(?P<S>)(\s|\.)+)?(?P<No>[^\w$GÜSİÖÇ\(\)\.\-\,\,]+[1,7])(\s|\.)+)?(?P<Eklr2>[^\n]+[^\.,\s]+)?((?P<Semt>[^\n\s]+)\/)(\s*)(?P<Sehir>[^\d\n\s]+))$
```

TEST STRING

AS İZMİR MAH. ÇAYIRONU CAD. NO: 156A/A ÜMRANİYE/ İSTANBUL
CUMHURİYET MAH. DR SADIK AHMET CAD. NO: 15/1 KÜÇÜKÇEKMECE/ İSTANBUL
MEVLANAKAPI MAH. KALBURCU MEHMET CAMİİ SOK. OTOPARK 8/10 FATİH/ İSTANBUL
KAYASEHİR 24. BÖLGE CAD. DUKKANLARI 46/A BAŞAKŞEHİR/ İSTANBUL
SARIDEMİR MAH. RAGİP GÜMÜŞPALA CAD. HALIÇ METRO ÇIKIŞI İEET DURAĞI YANI FATİH/ İSTANBUL
GÜNESTEPE MAH. BAĞLAR CAD. NO: 77/A GÜNGÖREN/ İSTANBUL
KEMER MAH. ATİŞLALANI CAD. NO: 212/A ESENLER/ İSTANBUL
ATIKALI MAH. YAVUZSELMİ CAD. NO: 23/A FATİH/ İSTANBUL
SULTANCIĞI MAH. ESKİ EDİRNE ASFALTI CAD. 617/B SULTANGAZİ/ İSTANBUL
ALEMDAĞ CAD. NO: 490/C ÜMRANİYE/ İSTANBUL
BATI MAH. BURHAN TOPRAK CAD. 20/E PENDİK/ İSTANBUL
ZÜHRÜTEVLER MAH. CANDAN CAD. 91/A MALTEPE/ İSTANBUL
NAMIK KEHAL MAH. SÜTCÜ İMAH CAD. NO: 26/23 ÜMRANİYE/ İSTANBUL
BAŞAKŞEHİR MAH. ÇİHANGİR SOK. KIPTAŞ ÇARŞISI NO: 13C/2.18 BAŞAKŞEHİR/ İSTANBUL
MECİDİYE MAH. MECİDİYE CAD. NO: 43 SULTANBEYLİ/ İSTANBUL
KAYABAŞI KAYASEHİR BULVARI 6-3 BULVARI EMLAK KONUT EVLERİ NO: 45 BAŞAKŞEHİR/ İSTANBUL
SEHREMENİ MAH. BÜYÜK SARAY MEYDANI CAD. NO: 8A FATİH/ İSTANBUL
İSLAMBAY MAH. İSLAMBAY CAD. NO: 124/A EYÜPSULTAN/ İSTANBUL
HASANPAŞA MAH. SOĞUTLUCESME CAD. KADIKÖY/ İSTANBUL
KAYABAŞI MAH. BOTANİK SOK. BAŞAKŞEHİR/ İSTANBUL
TUNA MAH. ŞEHİT MUSTAFA GÜNES CAD. NO: 12/A ESENLER/ İSTANBUL
ATATÜRK MAH. ATAŞEHİR BULVARI NO: 12/B ATAŞEHİR/ İSTANBUL
FATİH MAH. ÇİNCİN DERE CAD. NO: 72B ESENLER ESENLER/ İSTANBUL
ZEYNEP KAMİL MAH. NUHKUYUSU CAD. NO: 5 İÇ KAPI USKÜDAR/ İSTANBUL
TEPEÜSTÜ MAH. ALEMDAĞ CAD. 570/A ÜMRANİYE/ İSTANBUL
ÇAMÇESME MAH. MİSAKİ MİLLİ CAD. NO: 60/A PENDİK/ İSTANBUL
GÜLLÜ BAĞLAR MAH. TANDOĞAN CAD. NO: 60/B PENDİK/ İSTANBUL
OSMANİYE İNCİRLİ BAKIRKÖY METRO ÖNÜ NO: 22 İÇ KAPI NO: 3-A BAKIRKÖY/ İSTANBUL
ARMAGANEVLER MAH. MİTHAT PAŞA CAD. NO: 231/ A ÜMRANİYE/ İSTANBUL
AŞAĞI DUDULLU MAH. ALEMDAĞ CAD. NO: 465/A ÜMRANİYE/ İSTANBUL
SAADETDERE MAH. ABDİ İPEKÇİ CAD. NO: 27 ESENYURT/ İSTANBUL
NAMIK KEHAL MAH. FEVZİ ÇAKMAK CAD. NO: 58/A-1 ESENYURT/ İSTANBUL
RAİTKÖYLÜ MAH. RAİTKÖYLÜ CAD. NO: 132/ 1 ESENYURT/ İSTANBUL