Doğal Dil İşleme Dersi
1. Ödev Raporu

Düzenli ifade (regular expresion) kullanarak adres ayrıştırılması

Uygar Köroğlu 16011052

10.11.2020

Bu ödevde insan eliyle yazılmış adres satırlarından mahalle, cadde, sokak, il, ilçe, numara ve varsa ekstra bilginin düzenli ifade kullanılarak ayrıştırılması yapılmıştır.

Veri 6831 satırlık ve her bir satırda bir adres tanımı bulunan bir text dosyası olarak verilmiştir. Bu adreslerde mutlak düzenli bir yapı söz konusu değildir fakat incelenen veriden yapılan bazı çıkarımlar şu şekildedir:

- Mahalle bilgisi cadde bilgisinden önce gelir, mahalleyi ifade etmek için MAH. MH. MAHALLE MAHALLESİ MAH VE MH kullanılmıştır
- Cadde bilgisi sokak bilgisinden önce gelir, caddeyi ifade etmek için CAD. CD. CADDESİ CADDE CAD ve CD kullanılmıştır
- Sokak bilgisi numara bilgisinden önce gelir, sokağı ifade etmek için SOK. SK. SOKAĞI SOKAK SOK ve SK kullanılmıştır
- No bilgisi mahalle, cadde ve sokaktan farklı olarak değerini yazıda kendinden sonra barındırır (örneğin: x MAH., NO: 12/A), numaranın ilk değeri kesinlikle bir sayıdır, ayrıca numara bilgisinden önce veya sonra ekstra veri bulundurabilir
- İlçe ve il içerilerinde boşluk barındırmaz ve araların "/" işareti vardır, sıralamada en sonda yer alırlar

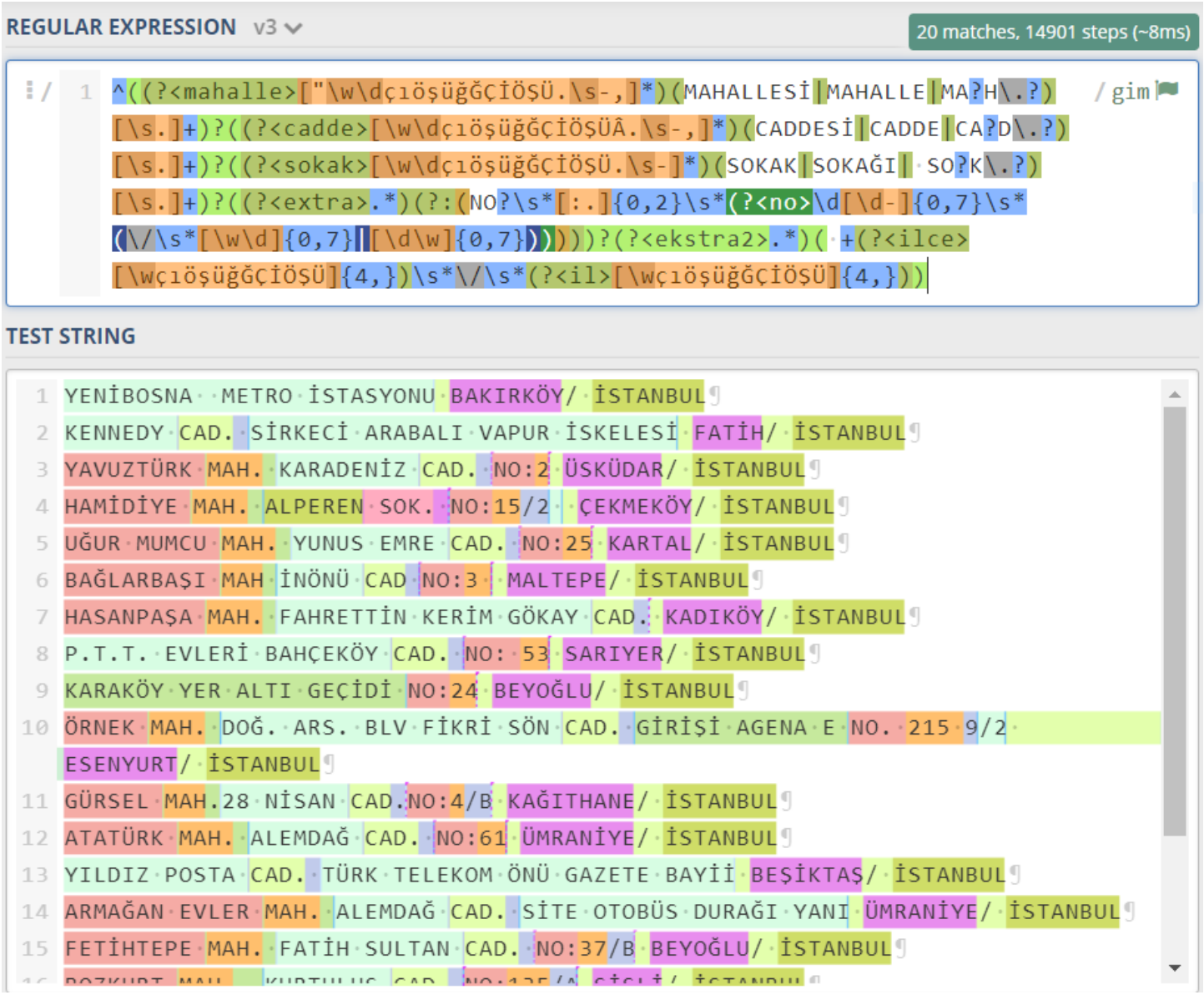
Bu tespitler göz önünde bulundurularak aşağıdaki regex komutu regex101 sitesinde oluşturulmuştur. PCRE(PHP) formatındadır. Regex içerisindeki "?<label>" şeklindeki ifadeler içinde bulunduğu grubu adlandırmak ve kod içinde daha rahat kullanmak için eklenmiştir.

Yapılan denemeler sonucu su sonuçlar elde edilmiştir. Toplamda 6831 adresten, 5927 tane mahalle, 4952 tane cadde, 1337 tane sokak, 5963 tane numara, 6831 tane ilçe, 6831 tane il tespit edilmiştir.

Mahalle, cadde, sokak, numara, ilçe, il değerlerinden hiçbirini bulamadığı adres sayısı 0, sadece 1 tanesini bulabildiği adres sayısı 0, sadece 2 tanesini bulabildiği adres sayısı 125, sadece 3 tanesini bulabildiği adres sayısı 351, sadece 4 tanesini bulabildiği adres sayısı 1442, sadece 5 tanesini bulabildiği adres sayısı 4708, hepsini bulabildiği adres sayısı 205 şeklindedir.

Regex Komutu:

```
^((?<mahalle>["\w\dçıöşüğĞÇİÖŞÜ.\s-,]*)(MAHALLESİ|MAHALLE|MA?H\.?)[\s.]+)?((?<cadde>[\w\dçıöşüğĞÇİÖŞÜÂ.\s-,]*)(CADDESİ|CADDE|CA?D\.?)[\s.]+)?((?<sokak>[\w\dçıöşüğĞÇİÖŞÜ.\s-]*)(SOKAK|SOKAĞI| SO?K\.?)[\s.]+)?((?<extra>.*)(?:(NO?\s*[:.]{0,2}\s*(?<no>\d[\d-]{0,7}\s*(\/\s*[\w\d]{0,7}|[\d\w]{0,7})))))?(?<ekstra2>.*)(
+(?<ilce>[\wçıöşüğĞÇİÖŞÜ]{4,})\s*\/\s*(?<il>[\wçıöşüğĞÇİÖŞÜ]{4,}))
```



https://regex101.com/r/ZJTcfx/3

https://larcis.github.io/addressRegex/

Sonuç:

Yukarıdaki regex101.com sitesinden alınan sonuçta da görüldüğü gibi adreslerden saf mahalle, cadde, sokak, no, ilçe, il bilgileri beklenildiği gibi elde ediliyor. Hatalı ya da eksik sonuç aldığım durumlar regex in yanlışlığından ziyade regexi tanımlarken yukarıda yapmış olduğum tespit ve varsayımlardan kaynaklanıyor. Örnek vermek gerekirse mahallenin her zaman cadde ve sokaktan önce yazıldığını varsaydım (ki genel olarak veri setinde durum bu) fakat veri setindeki bazı adres girdilerinde cadde, mahalleden önce yazılmış durumda ki bu da caddeyi bulamama ve mahallenin saf bilgisini yanlış çıkarmama neden olmaktadır. Regexi yazarken adres bileşenlerinin belli bir sırada yazıldığını varsaymak performansı yüksek oranda etkiliyor ve veri setinde yanlış sayılabilecek şekilde yazılmış girdi sayısı göz ardı edilecek kadar az.

Başarı oranı: %99.41

Başarılı bulunan adres sayısı: 6791