Doğal Dil İşlemeye Giriş Ödevi

Dersin Yürütücüsü: *Banu Diri*

Öğrencinin;

Adı: Muhammet Kayra

Soyadı: Bulut

Numarası: 20011901

Giriş

Verilen ödevde REGEX ile adres verisi içerisinde adres bloklarının ayrıştırılması işlemi yapılmıştır.

Veri

600 (599) satırdan oluşan veri her satırda farklı şekillerde yazılmış adreslerden oluşmaktadır.

Veride en çok tekrar eden özellik her satırın sonunda il ve ilçelerin "/" karakteriyle birbirlerinden ayrılmış olmasıdır. Bunun satırdan satıra değişmekle birlikte cadde, sokak ve no, mahalle gibi bilgileri içermektedir. Nadir olsa da bazı satırlarda iş hanı, avm, bulvar gibi kısımlar da bulunmaktadır.

Problem

Temel sorunumuz bu verilerin hangilerinde hangi özelliklerin(Adres bloklarının) bulunup bulunmadığını bulmaktır. Bunu çözdükten sonra da verinin hangi kısmının bu özelliği gösterdiğini bulmaktır. Örneğin eğer bir satır içinde "mah." bloğu geçiyorsa bu satırda mahalle bilgisi bulunuyor diyebiliriz. Bunun ardından da mahalleden önceki öbeği alarak bu öbeğin mahalle bilgisini içerdiğini söyleyebiliriz. Örneğin "23 Nisan Mah." bu öbekte "Mah."tan önce yer alan "23 Nisan" buradaki mahallenin adını belirtmektedir.

Yöntem

Regular expression, "https://regex101.com/" yardımıyla yazılmıştır. Elimizdeki veri çok fazla sayıda kombinasyon içerdiğinden verilerin bir kısmından feragat etmek durumunda kalıyoruz.

Bu kaybı minimuma indirmek için veri seti incelenip en çok tekrar eden veri gruplarını kapsayacak regular expression yazılmaya çalışılmıştır. Her satırda en sonda il ve ilçe blokları bulunmaktadır. Yazılan regex her satır için il ve ilçe bilgisini ayrıştırmaktadır. Satır başında ise en çok rastlanan üç seçenek bulunmaktadır bunlar:

1. Mahalle

2. Cadde

3. Sokak

Satırlar okunurken eğer bloklar bulunuyorsa bu sırada bulunacağı varsayılmıştır. Farklı sıralarda girilen adresler hata vermektedir. Bu bilgiler ayrıştırılırken satır,

verilen sırayla regex'te belirtilen keyworde kadar okunmuş ve gelinen kısım ilgili seçeneğe ait sayılmıştır. Bu üç özelliğin ardından Genellikle no komponenti gelmektedir. No komponenti bu üç özelliğin tersine bilgisini kendinden sonra gelen genelde sayılardan oluşan kısımda içermektedir. Bu yüzden de diğerlerinin tersine etiketi No bulunduktan sonrasını kapsamaktadır

İlk olarak ifadelerdeki mahalleleri bulmak için "(([A-z0-9ŞÜÂĞİÖşçğiöÇü"\.\-]*[Mm][Aa]?[Hh][Aa]?[Ll]?[Ll]?[Ee]?[Ss]?[İIıi]?,? ?\.?))" ifadesini ekledim.

Sonrasında bulvarları bulmak için "([A-z0-9ŞÜÂĞİÖşçğiöÇü"\.\-]* BULVARI)" ifadesini ekledim.

Sonrasında caddeleri bulmak için "([\."A-zÜÂçği0-9-öÇĞİÖŞşü,]*CA?DD?E?S?[İI]?\.?)" ifadesini ekledim.

Sonrasında sokakları ve blokları bulmak için **"([\."A-zĞİÖŞÜÂçğ0-9\/ iöÇşü]* SO?[Kk]A?K?Ğ?I?\.?)"** ifadesini ekledim.

Sonrasında apartman isimlerini bulmak için **"([A-zÜÂçği0-9-öÇĞİÖŞşü]* APT\.?)"** ifadesini ekledim.

Sonrasında numaraları bulmak için **"(?.*NO ?:? ?[0-9]+[A-z\-]?)"** ifadesini ekledim.

Sonrasında daireleri bulmak için **"(\/?[0-9]*\/?[A-z]?)"** ifadesini ekledim.

Sonrasında ilçeleri bulmak için **"(.*)"** ifadesini ekledim.

Sonrasında posta kodlarını bulmak için **"([0-9]{5})"** ifadesini ekledim.

Sonrasında şehirleri bulmak için "**"([A-zÇŞÜÂçğiĞİÖöşü]*)"** ifadesini ekledim.

REGEX'imin son haliyse aşağıdaki gibi oldu.

(([A-z0-9ŞÜÂĞİÖşçğiöÇü\.\-]*[Mm][Aa]?[Hh][Aa]?[Ll]?[Ll]?[Ee]?[Ss]?[İIıi]?,? ?\.?))?([A-z0-9ŞÜÂĞİÖşçğiöÇü"\.\-]* [Bb]U?[Ll][Vv]A?R?I?\.?)?([\."A-zÜÂçği0-9-öÇĞİÖŞşü,]*CA?DD?E?S?[İI]?\.?)?([\."A-zĞİÖŞÜÂçğ0-9\/ ,iöÇşü]* SO?[Kk]A?K?Ğ?I?\.?)?(.*BLOK)?([A-zÜÂçği0-9-öÇĞİÖŞşü]* APT\.?)?(?.*NO ?:? ?[0-9]+[A-z\-]?)?(\/?[0-9]*\/?[A-z]?)?(.*)?(?[A-zÇĞİÖçğiöŞÜÂşü]*)?(\/ ?)([0-9]{5})?([A-zÇŞÜÂçğiĞİÖöşü]*)

	Mah	Cad	Bulvar	Sok	Blok	Apt	No	Daire	PK	Semt	Şehir
Kaç adet	513	434	22	187	4	6	494	323	5	598	598
Bulunan	502	425	19	151	3	5	456	307	1	589	597
Başarı(%)	97,9	97,9	86,37	80,07	75	83,33	92,31	95,05	20	98,5	99,83

Bulunamayan adresler;

1-)Fenerbahçe, Iğrıp Sk. No:13, 34726 İstanbul

2-)KAPTAN PAŞA MAH. YAY GEÇİDİ SOK. NO:6/C BEYOĞLU

3-)Ziya Gökalp, Mahallesi, Eski Turgut Özal Cd. No:29 D:Kat.2, 34000 İkitelli Osb/Başakşehir/İstanbul

4-)Mehmet Nesih Özmen, Fatih Cad. & Karadal Sokagı , Merter Keresteciler Sitesi No:20, 34173 Güngören/İstanbul