

Doğal Dil İşlemeye Kavramsal Bir Bakış Ödevi

Dersin Yürütücüsü: Banu Diri

Öğrencinin;

Adı: Muhammed Kayra

Soyadı: Bulut

Numarası: 23501901

Öncelikle bütün adreslere bakıp genel niteliklerini incelediğimde, çoğunun sonunda /İstanbul olduğunu fark ettim.

Bu ifadeleri **(.\*[/]\s\*[İi](stanbul|STANBUL))** regex kalıbı ile bir araya getirdim. Ayrıca **(T.\*N0:.\*)** i ifadesini de başka bir durum olarak ele aldım.

Başlangıç ifadesi aşağıdaki gibidir. **(T.\*N0:.\*)|(.\*[\/]\s\*[İi](stanbul|STANBUL))**

Sonuçlara bakarken, ilk çözümümdeki regex'in yetersiz olduğunu fark ettim. **'adresteki işler tamamlandıktan sonra Yıldırım Mah. Kemal Paşa Sok. No:8 BAYRAMPAŞA/İSTANBUL'** cümlesinde, ifadenin tamamı adres olarak algılanıyordu. Ayrıca, adreslerin **'YENİBOSNA'** gibi başlangıç kısımları da eşleşmiyordu. Bu nedenle, özelleştirme yapmaya karar verdim. Bu ve benzeri cümleleri incelediğimde, adreslerin büyük harfle başlaması gerektiğine karar verdim. İfadenin başındaki .\* regex'in sorunlu olduğunu gördüğüm için değiştirmeye başladım. Büyük harfle başlayan kelimeleri (tek harfliler dahil) kapsayan regex'i [A-ZĞÜİÖÇŞ 0-9] olarak yazdım. Bunların tekrar edebilmesi için, sonunda boşluk veya tab karakterleri de olabileceğini düşünerek regex'i genişlettim. Sorgu bu haliyle aşağıdaki haline ulaştı.

**([A-ZĞÜİÖÇŞ 0-9][ \t ]\*)+**

Daha sonra, adreslerin içinde **., :, -, /** gibi karakterlerin de bulunabileceğini göz önünde bulundurarak, regex'in ikinci kısmı olan **boşluk** bölümünü **[ .:-/]** şeklinde genişlettim. Sorgu bu haliyle aşağıdaki haline ulaştı.

**(T.\*N0:.\*)|([A-ZĞÜİÖÇŞ0-9][ \.:\-\/]\*[\t ]\*)+([\/]\s\*[İi](stanbul|STANBUL))**

Örnekleri incelediğimde, **'Yıldırım Mah. Kemal Paşa Sok. No:8 BAYRAMPAŞA/İSTANBUL'** gibi adresleri tam olarak tespit edemediğimi gördüm. Bu nedenle, **([A-ZĞÜİÖÇŞ0-9][ \.:\-\/]\*[\t ]\*)** sorgusunu genişleterek **([0-9\.:\-\/A-ZĞÜİÖÇŞ]\*[\t ]\*)(([A-ZĞÜİÖÇŞ0-9]|:)[0-9\.:\-\/a-zA-ZĞÜİÖÇŞğüıöçş]\*[\t ]\*)** şekline dönüştürdüm. Sorgu bu haliyle aşağıdaki haline ulaştı.

**(T.\*N0:.\*)|([0-9\.:\-\/A-ZĞÜİÖÇŞ]\*[\t ]\*)(([A-ZĞÜİÖÇŞ0-9]|:)[0-9\.:\-\/a-zA-ZĞÜİÖÇŞğüıöçş]\*[\t ]\*)+([\/]\s\*[İi](stanbul|STANBUL))**

Bu aşamadan sonra, **'Kaıköy Deniz Otobüsü İskelesi önü Rasimpaşa Kadıköy/İstanbul'** gibi adreslerin aslında tespit edilemediğini fark ettim. Bu tür uç durumlar için **'önü'** ifadesini regex'ime ekledim. Sorgu, bu değişiklikle aşağıdaki şekle geldi.

**(T.\*N0:.\*)|([0-9\.:\-\/A-ZĞÜİÖÇŞ]\*[\t ]\*)(([A-ZĞÜİÖÇŞ0-9]|önü|:)[0-9\.:\-\/a-zA-ZĞÜİÖÇŞğüıöçş]\*[\t ]\*)+([\/]\s\*[İi](stanbul|STANBUL))**

Bu şekliyle, sorgunun tespit edemediği bir adres kalmadı. Bu durum beni oldukça memnun etti. Regex'in son halini, okunabilirliği artırmak amacıyla renklendirdim.

**Sorgunun Son Hali:**

**(T.\*N0:.\*)|([0-9\.:\-\/A-ZĞÜİÖÇŞ]\*[\t ]\*)(([A-ZĞÜİÖÇŞ0-9]|önü|:)[0-9\.:\-\/a-zA-ZĞÜİÖÇŞğüıöçş]\*[\t ]\*)+([\/]\s\*[İi](stanbul|STANBUL))**

<https://regex101.com> üzerinde ECMAScript (JavaScript) ile çalıştırılan sorgunun sonucu aşağıda gösterilmiştir. Ayrıca, Python ile de çalıştırarak aynı sonuçları elde ettim. Sonuç görselini aşağıya ekledim. 😊😊😊 Adreslerin **%100’ünü** buldum 🎉🎉🎉

