Doğal Dil işleme Raporu

1. **Giriş**

Normalizasyon, genel anlamda, verileri bir standart veya norma göre düzenleme ve optimize etme sürecini ifade eder.

Bir doğal dil kaynağını normalize ettiğimizde, onu önceden belirlenmiş bir "standarta" daha yakın hale getirerek içindeki rastgeleliği azaltmayı amaçlarız. Bu, bilgisayarın ele alması gereken farklı bilgi miktarını azaltmaya yardımcı olur ve böylece verimliliği artırır.

1. **Kullanım Alanları**

Metin normalizasyonu, metin madenciliği, duygu analizi, makine çevirisi, sohbet botları ve diğer pek çok NLP uygulamasında temel bir adımdır. Bu süreç, verileri makine öğrenimi modelleri için daha uygun hale getirerek, daha doğru ve etkili sonuçlar elde edilmesini sağlar.

Örnek olarak sohbet botlarında, Kullanıcıların girdiğini anlamak ve uygun yanıtlar üretmek için metin normalizasyonu kullanılır. Bu, botların doğal dildeki çeşitlilikle daha iyi başa çıkmasını sağlar. Duygu analizi yapılırken gürültülü verilerin temizlenmesi analizin doğruluğunun artırılmasında önemli rol oynar.

1. **Temel Adımlar**

* Tokenization
* Durma Kelimelerinin Kaldırılması (Stop Words),
* Boşluk ve Satır Düzenlemeleri
* Küçük/Büyük Harf Dönüşümü (Case Normalization)
* Noktalama İşaretlerinin Kaldırılması
* Sayıların ve Özel Karakterlerin Kaldırılması
* Kök Bulma (Stemming) ve Ek Ayrıştırma (Lemmatization)
* Dil Özgü İşlemler

**3.1. Kök Bulma (Stemming)**

Kelimelerin köklerini ya da gövdelerini bulmayı ve bu şekilde kelimeleri basitleştirmeyi amaçlar. Temelde, bir kelimenin türevlerini (örneğin, çekimlenmiş, zaman veya kip ekleri eklenmiş hallerini) alıp kelimenin temel formuna indirger.

Avantajları,

* Hız ve Verimlilik
* Basitleştirilmiş Karşılaştırma

Dezavantajları

* Az veya aşırı kesme sonucu anlam kaybı

**3.2. Ek Ayrıştırma (Lemmatization)**

Lemmatization, çekimlenmiş kelimeleri kök kelimesine indirgemek için kullanılan başka bir tekniktir. Bu, çekimlenmiş bir kelimenin "lemma"sını (sözlük formunu), kelimenin amaçlanan anlamına dayanarak belirleyen algoritmik bir süreci tanımlar.

Uygulanması Stemming’e göre daha fazla zaman gerektirmesine karşı doğruluk oranı daha yüksektir.