Çilem AFACAN - 502431015 - Doğal Dil İşlemede İstatistiksel Yöntemler Dersi 1. Ödev Raporu

Metin Ön İşleme: Neden Önemlidir?

Doğal Dil İşleme (NLP) projelerinde ham metin verisi, çoğu zaman doğrudan analiz veya modelleme için kullanılamaz. Metinler; düzensiz karakterler, noktalama işaretleri, gereksiz kelimeler ve biçimsel farklılıklar içerebilir. Bu nedenle, metinlerin doğru ve etkili bir şekilde işlenebilmesi için ön işleme adımları uygulanmalıdır.

Metin ön işleme, veriyi daha temiz, tutarlı ve işlenebilir hale getirmek için yapılan bir dizi dönüşüm sürecini kapsar. Bu süreç, kullanılan modele ve uygulama alanına bağlı olarak değişebilir ancak genel olarak şu nedenlerle kritik bir öneme sahiptir:

Gürültüyü Azaltma: Gereksiz karakterler, semboller ve durak (stop) kelimelerinin kaldırılması, veriyi daha anlamlı hale getirir.
Veri Tutarlılığını Artırma: Farklı biçimlerde yazılmış kelimeleri normalize etmek, metnin daha tutarlı olmasını sağlar.
Hesaplama Maliyetini Azaltma: Gereksiz veriyi eleyerek modelin daha verimli çalışmasına yardımcı olur.
Genelleştirme Yeteneğini Artırma: Kelime köklerine indirgeme (stemming/lemmatization) gibi işlemler, modelin farklı bağlamlarda da iyi performans göstermesini sağlar.

Metin Ön İşleme Aşamaları

Metin ön işleme süreci, çeşitli dönüşümleri içeren birden fazla adımdan oluşur:

Büyük/Küçük Harf Dönüşümü (Lowercasing): Metindeki tüm harflerin küçük harfe çevrilmesi.
Noktalama İşaretlerinin Kaldırılması (Removing Punctuation): Noktalama işaretlerinin silinmesi veya gerektiğinde değiştirilmesi.
Sayıların Kaldırılması (Removing Numbers): Sayıların metinden çıkarılması (bazı durumlarda korunabilir).
Durak Kelimelerin Kaldırılması (Removing Stopwords): "ve", "bu", "bir" gibi bilgi taşımayan yaygın kelimelerin çıkarılması.
Özel Karakterlerin ve Emojilerin Kaldırılması (Removing Special Characters & Emojis): Metindeki özel sembollerin ve emojilerin temizlenmesi.
Metin Normalizasyonu (Text Normalization): Kelimelerin standart bir forma dönüştürülmesi (örneğin, "olmıyacak" → "olmayacak", "tlfn" → "telefon").
Kök veya Gövdeye İndirgeme (Stemming & Lemmatization): Kelimeleri kök haline getirme (stemming) veya köküne en yakın hâle getirme (lemmatization).
Kelime Tokenizasyonu (Tokenization): Metni kelime veya cümlelere ayırma.