N-Gram Modelleri

N-gram modelleri, doğal dil işleme (NLP) alanında sıklıkla kullanılan istatistiksel dil modelleridir. Bir dildeki kelime veya karakter dizilimlerini olasılıksal olarak modellemek için kullanılırlar.

N-Gram Tanımı

Bir N-gram, metin içerisindeki ardışık N kelimelik veya karakterlik bir diziyi ifade eder. Örneğin:

- **Unigram (1-gram)**: "merhaba"
- **Bigram (2-gram)**: "merhaba dünya"
- **Trigram (3-gram)**: "merhaba dünya bugün"

N-Gram Modellerinin Kullanımı

1. **Dil Modelleme**: N-gramlar, bir cümlenin olasılığını tahmin etmek için kullanılır.
2. **Metin Tahmini**: Otomatik metin tamamlama ve klavye tahmin sistemlerinde kullanılır.
3. **Makine Çevirisi**: İstatistiksel çeviri modellerinde kelime dizilimlerini analiz eder.
4. **Yazım Denetimi**: Yazım hatalarını düzeltmek için yaygın kelime kombinasyonlarını belirler.

N-Gram Olasılık Hesaplama

Bir cümlenin olasılığı aşağıdaki gibi hesaplanır:

[P(w_1, w_2, ..., w_n) \approx \prod_{i=1}^{N} P(w_i | w_{i-(N-1)}, ..., w_{i-1})]

Burada, bir kelimenin olasılığı, önceki N-1 kelimeye bağlı olarak hesaplanır.

Avantajları ve Dezavantajları

Avantajlar:

- Hesaplama açısından basit ve hızlıdır.
- Küçük veri setlerinde etkili sonuçlar verebilir.

Dezavantajlar:

- Büyük N değerlerinde veri kıtlığı problemi yaşanır.
- Bağlamı tam olarak anlayamaz ve uzun mesafeli bağımlılıkları göz ardı eder.

N-gram modelleri, modern derin öğrenme tabanlı dil modellerine kıyasla daha sınırlı olsa da, halen birçok NLP uygulamasında temel bir yöntem olarak kullanılmaktadır.