TF-IDF ve Cosine Similarity: Temel Kavramlar ve Örnekler

TF-IDF (Term Frequency - Inverse Document Frequency) ve Cosine Similarity, metin madenciliği ve doğal dil işleme (NLP) alanında sıklıkla kullanılan iki temel tekniktir. Bu rapor, bu iki yöntemin hem sezgisel hem de matematiksel temellerini açıklamakta ve örneklerle pekiştirmektedir.

Cosine Similarity Nedir?

Cosine Similarity, iki vektör arasındaki açının kosinüsünü ölçerek benzerlik derecesini hesaplar. Matematiksel formülü şu şekildedir:

$$\cos(\theta) = (A \cdot B) / (||A|| \times ||B||)$$

Bu değer -1 ile 1 arasında değişir. 1'e yakınsa yüksek benzerlik, 0 ise ilgisizlik, -1 ise zıt yönleri ifade eder.

TF-IDF Nedir?

TF-IDF, bir kelimenin bir belgede ne kadar önemli olduğunu ölçmek için kullanılır. İki bileşenden oluşur: Term Frequency (TF) ve Inverse Document Frequency (IDF).

1. Term Frequency (TF):
TF(t, d) = (Kelimenin belgede geçiş sayısı) / (Toplam kelime sayısı)

2. Inverse Document Frequency (IDF):
IDF(t) = log(N / (1 + df(t)))
Burada N toplam belge sayısını, df(t) ise kelimenin geçtiği belge sayısını ifade eder.

Örnek:

D1: "kedi köpek"
D2: "kedi balık"
D3: "balık köpek köpek"

Adım 1 : TF Hesaplama

Belge	Kelime	TF
D1	kedi, köpek	0.5, 0.5
D2	kedi, balık	0.5, 0.5
D3	balık, köpek	1/3, 2/3

Adım2 : IDF Hesaplama

Kelime	df(t)	N=3	IDF
kedi	2	3	log(3 / (1 + 2)) = log(1) = **0**
köpek	2	3	log(3 / (1 + 2)) = **0**
balık	2	3	log(3 / (1 + 2)) = **0**

Bu kelimeler her belgede geçtiği için ayırt edici değil. Her bir belgenin birbiriyle olan kosinüs benzerliği değeri matematiksel olarak ya tanımsız olur ya da 0 olarak belirtilir.

D1: "uzay roket mars"
D2: "uzay dünya"
D3: "dünya savaş barış"

Bu örnekte 'roket', 'mars', 'savaş' ve 'barış' gibi kelimeler daha ayırt edicidir ve TF-IDF değerleri yüksektir.

Sonuç

TF-IDF yöntemi ile metinlerde önemli kelimeler öne çıkarılırken, cosine similarity yardımıyla belgeler arasında benzerlik dereceleri hesaplanabilir. Bu iki yöntem birlikte kullanıldığında çok güçlü bir metin analiz aracı sunar.