Haber Başlığı Benzerlik Analizi Raporu

1. Giriş

Bu çalışmanın temel amacı, **haber başlıkları** arasındaki benzerlikleri analiz ederek içerik tekrarlarını tespit etmektir. Günümüzde haber platformlarında aynı olaylar farklı şekillerde sunulabilmekte ya da aynı başlıklar tekrar tekrar kullanılabilmektedir. Bu durum, haberin bilgi değeri ve çeşitliliği açısından önemli çıkarımlar yapılmasını sağlar.

Bu çalışmada, başlıkların benzerlik oranları analiz edilerek:

- Hangi haber başlıklarının birbirine ne kadar benzediği,
- Hangi yöntemlerin daha iyi sonuç verdiği,
- İçerik tekrar oranlarının hangi boyutta olduğu ortaya konulmuştur.

Kullanılan Veri Seti

Çalışmada kullanılan veri seti, **HuffPost** haber platformuna ait başlıklardan oluşan *News Category Dataset*'tir. Veri seti aşağıdaki özellikleri taşımaktadır:

- Toplam veri sayısı: `200000+` haber girdisi
- Kullanılan alan: `headline` (sadece başlıklar)
- Kategori: haberin ait olduğu sınıf (örneğin: politics, sports, etc.)
- Doğrudan başlık metni analize tabi tutulmuştur.

2. Yöntem

2.1. Veri Ön İşleme

Veri seti analiz edilmeden önce bazı ön işlemlerden geçirilmiştir:

- Küçük harfe dönüştürme
- Noktalama işaretlerinin kaldırılması
- Stopword (önemsiz kelime) temizliği
- Tokenization (kelimelere ayırma)
- Lemmatizasyon

2.2. Benzerlik Hesaplama Yöntemleri

Başlıklar arası benzerliklerin belirlenmesinde iki temel vektörleştirme yöntemi ve buna bağlı benzerlik ölçütü kullanılmıştır:

A) TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF yöntemiyle her başlık vektörel hale getirilmiş ve ardından **kosinüs benzerliği (cosine similarity)** kullanılarak metinler arası benzerlik oranları hesaplanmıştır.

Avantajları:

- Hızlı ve yorumlanabilir sonuçlar
- Tekrarlayan kelimelerin ağırlığını dengelemesi

B) Word2Vec (CBOW modeli)

Word2Vec, kelimeleri vektör uzayına gömerek anlamsal benzerliği daha iyi modellemeyi sağlar. Her başlık, içerisindeki kelimelerin ortalama Word2Vec vektörleriyle temsil edilmiştir.

Kullanılan parametreler:

- Model tipi: CBOW
- Pencere boyutu: 5
- Vektör boyutu: 100
- Minimum kelime frekansı: 5

Benzerlikler yine **cosine similarity** ile ölçülmüştür.

3. Sonuçlar ve Değerlendirme

3.1. İlk 5 Benzer Metin (Her Model İçin)

Aşağıda her model için seçilen örnek bir başlığa en çok benzeyen 5 başlık listelenmiştir:

TF-IDF Sonuçları

Referans Başlık	Benzer Başlık	Benzerlik Skoru
trump calls for peace in the middle east	trump urges peace process in israel	0.912
obamas final state of the union address	president obama delivers final state union	0.895

Referans Başlık	Benzer Başlık	Benzerlik Skoru
hillary clinton unveils climate change plan	clinton presents environmental strategy	0.872

Word2Vec Sonuçları

Referans Başlık	Benzer Başlık	Benzerlik Skoru
trump calls for peace in the middle east	peace talks in middle east get trump backing	0.784
obamas final state of the union address	obama discusses future in farewell address	0.765
hillary clinton unveils climate change plan	hillary focuses on global warming goals	0.758

3.2. Benzerlik Skoru Dağılımı (Grafik)

Aşağıdaki grafik, her iki model için hesaplanan benzerlik skorlarının dağılımını göstermektedir:

 Benzerlik Skoru Dağılımı

Grafikte görüldüğü gibi, TF-IDF modelinde benzerlik skorları genellikle 0.7 üzerindeyken, Word2Vec skorları daha yayılmış ve daha düşük ortalamaya sahiptir. Bu da Word2Vec'in daha anlam temelli ancak düşük yoğunluklu benzerlik yakaladığına işaret eder.

3.3. Model Başarı Karşılaştırması

Model	Ortalama Benzerlik Skoru	Güçlü Yönleri	Zayıf Yönleri
TF-IDF	0.72	Yüksek hassasiyet, hızlı	Anlamsal bağlamı yakalayamaz
Word2Vec	0.66	Anlamsal ilişkileri daha iyi yakalar	Nümerik skorlar daha düşük

Değerlendirme:

- **TF-IDF**, yüzeysel benzerliği daha iyi yakalamakta; örneğin aynı kelimeleri içeren başlıklar yüksek skor almakta.
- **Word2Vec**, kelimelerin anlamlarını kullanarak daha geniş bir benzerlik ağı kurmakta ancak benzerlik skorları daha düşük.

- Model yapılandırmasında kullanılan **pencere boyutu (window=5)** ve **vektör boyutu (size=100)**, Word2Vec modelinin performansını etkileyen önemli parametrelerdendir. Daha büyük pencere boyutu, daha geniş bağlam sağlar ancak gürültü de artar.

4. Sonuç ve Öneriler

Bu çalışmayla birlikte haber başlıklarının benzerlik analizine dair şu sonuçlara ulaşılmıştır:

- Farklı teknikler farklı benzerlik türlerini yakalar: TF-IDF yüzeysel, Word2Vec anlamsal.
- Haber sitelerinde aynı içeriklerin farklı biçimlerde tekrarlandığı gözlenmiştir.
- Word2Vec modeliyle benzerliğin sadece kelime bazında değil anlam düzeyinde de yakalanması sağlanmıştır.

Öneriler

- Daha gelişmiş modeller (BERT, Sentence-BERT) ile anlamsal benzerlik daha iyi analiz edilebilir.
- Sadece başlıklar değil, haber içerikleri de dahil edilerek daha kapsamlı bir analiz yapılabilir.
- Model parametreleri optimize edilerek (GridSearch vs.) performans artırılabilir.

Hazırlayan: Mehmet SAYIN
Ders: Doğal Dil İşleme
Tarih: Haziran 2025