DOĞAL DİL İŞLEME PROJESİ – HABER BAŞLIĞI ÇOĞALTMA TESPİTİ

1. GİRİŞ

Bu proje, doğal dil işleme (NLP) teknikleri kullanılarak haber başlıklarının benzerliğini tespit etmeyi amaçlamaktadır. Günümüzde haber içerikleri farklı kaynaklar tarafından yeniden yazılarak sunulmaktadır. Bu durum, haber çoğaltımı (duplicate detection) problemini ortaya çıkarır. Projenin temel hedefi, NewsAPI'den elde edilen başlıklar üzerinden benzerlik analizi yaparak, aynı içeriğe sahip olabilecek farklı başlıkları tespit etmektir.

2. VERİ SETİ VE ÖN İŞLEME

Veri seti NewsAPI aracılığıyla toplanan Türkçe haber başlıklarından oluşmaktadır. Başlıklar üzerinde iki ayrı ön işleme uygulanmıştır:

- **Lemmatizasyon** (Zemberek ile)
- **Stemleme** (kök alma)

Oluşturulan `lemmatized.csv` ve `stemmed.csv` dosyalarına kaydedilmiştir.

3. MODELLEME

3.1 TF-IDF

- Modeller: `tfidf_lemmatized`, `tfidf_stemmed`
- Cosine similarity ile benzerlik ölçümü

3.2 Word2Vec

- Toplam 16 model (8 lemmatized, 8 stemmed; CBOW/SkipGram, farklı window ve vektör boyutları)
- Cümle vektörleri kelime vektörlerinin ortalamasıyla elde edildi

4. BENZERLİK ANALİZİ

Giriş Metni:
Ataşehir'de "Çedes ve Geleneksel Çocuk Oyunları Şenliği" düzenlendi

- **TF-IDF Ort. Semantik Skor:** 2.60
- **Word2Vec Ort. Semantik Skor:** 5.00

En başarılı Word2Vec modelleri:

- `w2v_lemmatized_1`
- `w2v_stemmed_1`
- `w2v_lemmatized_2`

5. SIRALAMA TUTARLILIĞI (Jaccard)

18×18 Jaccard matrisleri ile model tutarlılıkları analiz edildi; benzer yapılandırmalar en yüksek tutarlılığı sağladı.

6. DEĞERLENDİRME

TF-IDF, anlamsal bağlantıyı tam yansıtamadı; Word2Vec ise kelime gömlemesiyle anlamı yakalayarak başarılı oldu.

7. SONUÇ VE ÖNERİLER

Word2Vec önerilir: CBOW düşük veri koşullarında bile iyi performans gösterdi. Haber benzerliği tespitlerinde anlam odaklı yöntemler tercih edilmelidir.