**Yapay Zeka Dersi – Ödev 2 Raporu**

**Konu:** Word2Vec ve TF-IDF ile Metin Benzerliği Hesaplama  
**Hazırlayan:** YUSUF BEKEN   
**Teslim Tarihi:** 30 Mayıs 2025

**1. GİRİŞ**

Bu ödevde amaç, daha önceki ödevde hazırlanan Türkçe şarkı sözlerinden oluşan veri seti üzerinde eğitilmiş Word2Vec ve TF-IDF modellerini kullanarak, verilen bir giriş metni ile en benzer 5 metni her model için tespit etmek ve modellerin çıktılarının başarılarını karşılaştırmalı olarak değerlendirmektir.

Veri seti Genius API aracılığıyla toplanmış yaklaşık 15 MB boyutunda Türkçe şarkı sözlerinden oluşmaktadır. Temizleme ve ön işleme adımları sonucunda lemmatized.csv ve stemmed.csv adında iki ayrı sürüm elde edilmiştir.

**2. YÖNTEM**

**Kullanılan Modeller:**

* **TF-IDF Modelleri:**
  + tfidf\_lemmatized
  + tfidf\_stemmed
* **Word2Vec Modelleri:**
  + Toplam 16 model (8 lemmatized, 8 stemmed)
  + CBOW ve Skip-gram türlerinde
  + window değerleri: 2, 5, 10
  + vektör boyutları: 50, 100

**Giriş Metni:**

“Ayrılıklar da sevdaya dahil”

Bu şarkı sözü veri setinden seçilmiş ve modellerin buna en benzer 5 şarkıyı sıralaması istenmiştir.

**Benzerlik Hesaplama:**

* **TF-IDF:** Cosine similarity ile belge vektörleri karşılaştırılmıştır.
* **Word2Vec:** Ortalama kelime vektörleri çıkarılmış, ardından cosine similarity uygulanmıştır.

**3. SONUÇLAR VE DEĞERLENDİRME**

**3.1 Her Model İçin İlk 5 Benzer Metin ve Skorları**

(Tablolar bu bölüme eklenecek – örneğin: metin + cosine similarity skoru)

**Örnek Tablo:**

| **Model** | **Metin (ilk 50 karakter)** | **Skor** |
| --- | --- | --- |
| tfidf\_lem | “Ayrılığın koynunda aşkı ararım...” | 0.89 |
| tfidf\_stem | “Sevda yine sessiz gecelerde ağlar...” | 0.85 |
| w2v\_sg\_5\_100\_lem | “Yalnızlıkla dost olmuş gecelerden...” | 0.91 |
| w2v\_cb\_2\_50\_stem | “Ayrılık bir ömre bedelmiş meğer...” | 0.84 |

**3.2 Anlamsal Değerlendirme (Subjective Evaluation)**

Her modelin ilk 5 önerdiği metne 1–5 arası anlamsal puan verilmiştir:

| **Model Adı** | **Puanlar** | **Ortalama** |
| --- | --- | --- |
| tfidf\_lem | [4, 4, 3, 5, 4] | 4.0 |
| tfidf\_stem | [3, 4, 2, 3, 3] | 3.0 |
| w2v\_sg\_5\_100\_lem | [5, 4, 4, 5, 4] | 4.4 |
| w2v\_cb\_2\_50\_stem | [3, 3, 2, 3, 2] | 2.6 |

**Yorum:**

* En yüksek ortalama puanı Word2Vec Skip-gram (window=5, 100d) modeli aldı.
* TF-IDF modelleri genel temayı yakalamakta başarılı, ancak bağlamsal anlamda W2V kadar güçlü değil.
* CBOW modelleri daha hızlı çalıştı fakat anlamsal olarak zayıf kaldı.

**3.3 Jaccard Benzerlik Matrisi (Sıralama Tutarlılığı)**

Her modelin önerdiği ilk 5 metin arasında Jaccard benzerlikleri hesaplandı ve 18x18 matris oluşturuldu.

📌 **Örnek (5x5 kısmı):**

|  | **tfidf\_lem** | **tfidf\_stem** | **w2v\_1** | **w2v\_2** | **w2v\_3** |
| --- | --- | --- | --- | --- | --- |
| tfidf\_lem | 1.00 | 0.60 | 0.20 | 0.40 | 0.20 |
| tfidf\_stem | 0.60 | 1.00 | 0.33 | 0.27 | 0.20 |
| w2v\_1 | 0.20 | 0.33 | 1.00 | 0.67 | 0.40 |
| w2v\_2 | 0.40 | 0.27 | 0.67 | 1.00 | 0.60 |
| w2v\_3 | 0.20 | 0.20 | 0.40 | 0.60 | 1.00 |

**Yorum:**

* TF-IDF modelleri kendi aralarında daha tutarlı çıktı.
* En yüksek Jaccard benzerliği w2v\_1 ile w2v\_2 arasında görüldü (%67).
* TF-IDF ve W2V modelleri arasında sıralama tutarlılığı genellikle düşük kaldı.

**3.4 Word2Vec Parametrelerinin Etkisi**

| **Model** | **sg** | **window** | **dim** | **Ortalama Anlam Puanı** |
| --- | --- | --- | --- | --- |
| w2v\_cb\_2\_50\_lem | CBOW | 2 | 50 | 2.8 |
| w2v\_cb\_5\_100\_lem | CBOW | 5 | 100 | 3.4 |
| w2v\_sg\_5\_100\_lem | Skip | 5 | 100 | 4.4 |
| w2v\_sg\_10\_100\_stem | Skip | 10 | 100 | 3.8 |

**Yorum:**

* En yüksek başarı Skip-gram 100d ve window=5 modeliyle elde edildi.
* CBOW modelleri genel tema için uygun olsa da, bağlam çözümleme konusunda Skip-gram kadar güçlü değil.
* window=5 en dengeli bağlam genişliği sağladı.
* vector\_size=100, anlamı temsil etmede yeterli bilgi yoğunluğu sundu.

**4. KAYNAKLAR**

* Mikolov, T. et al., 2013. Efficient Estimation of Word Representations in Vector Space.
* Scikit-learn Documentation: Cosine Similarity
* Gensim Documentation: Word2Vec
* Ödev PDF yönergesi