Doğal Dil İşleme Final Projesi Raporu

Proje Başlığı:

Spor Yorumlarına Dayalı Maç Özeti Öneri Sistemi (TF-IDF ve Word2Vec Karşılaştırması)

1. Amaç

Bu projede amacımız, Reddit'teki "r/soccer" başlığı altından elde edilen futbol yorumlarını kullanarak bir kullanıcı yorumuna benzer olan maç özetlerini otomatik şekilde önermekti. Modelleme için TF-IDF ve Word2Vec gibi kelime temsili yöntemleri kullanıldı, öneri kalitesinin anlamsal değerlendirmesi ve çeşitli metriklerle karşılaştırması yapıldı.

2. Veri Seti

- Kaynak: Reddit API (`r/soccer` başlığı)
- Yorum Sayısı: **18.061**
- Temizleme: Noktalama işaretleri kaldırıldı, küçük harfe çevirme, stopwords silme, lemmatization ve stemming uygulandı.

3. Kullanılan Yöntemler

Word Embedding:

Toplam **16 Word2Vec modeli** eğitildi:

- 2 Model tipi: **CBOW** ve **Skip-Gram**
- 2 Window size: **2**, **4**
- 2 Vektör boyutu: **100**, **300**
- 2 Ön işleme: **Lemmatized**, **Stemmed**

Ayrıca **2 TF-IDF modeli** eğitildi (lemmatized & stemmed versiyonlarla).

Benzerlik Yöntemi:

- Cosine Similarity: Hem TF-IDF hem Word2Vec modelleri için kullanıldı
- Jaccard Similarity: Modeller arısında öneri çıktılarının benzerliği için kullanıldı

4. Uygulama: Giriş Yorumu

feck bet penalti favor real madrid

Bu yorum modele input olarak verildi ve 18 modelin her biri tarafından bu yoruma en benzer 5 yorum çıkarıldı. Toplam **90 öneri yorum** elde edildi.

5. Cosine Similarity Ortalamaları

Her model için 5 yorumun cosine similarity ortalamaları hesaplandı.
En yüksek 5 model:

Mode	Avg. Cosine
lemmatized_cbow_window2_dim300	0.9999
lemmatized_cbow_window4_dim300	0.9999
lemmatized_cbow_window2_dim100	0.9998
lemmatized_cbow_window4_dim100	0.9998
stemmed_cbow_window2_dim300	0.9994

```
model_name,similarity
lemmatized_cbow_window2_dim300,0.9999184846878052
lemmatized_cbow_window4_dim300,0.9999096870422364
lemmatized_cbow_window2_dim100,0.9998514294624329
lemmatized_cbow_window4_dim100,0.9998461604118347
stemmed_cbow_window2_dim300,0.9994134902954102
stemmed_cbow_window4_dim300,0.999389386177063
stemmed_cbow_window2_dim100,0.9988985300064087
lemmatized_skipgram_window2_dim300,0.9988450884819031
stemmed_cbow_window4_dim100,0.998752748966217
lemmatized_skipgram_window2_dim100,0.9984229207038879
lemmatized_skipgram_window4_dim300,0.9971574187278748
lemmatized_skipgram_window4_dim100,0.9958354711532593
stemmed_skipgram_window2_dim300,0.9943438053131104
stemmed_skipgram_window2_dim100,0.9925349473953247
stemmed_skipgram_window4_dim300,0.9909379959106446
stemmed_skipgram_window4_dim100,0.9894066214561462
```

6. Anlamsal Değerlendirme (Elle Puanlama)

Tüm modellerin öneri yorumları 1–5 aralığında puanlandı:

5: Aynı anlam

4: Çok benzer

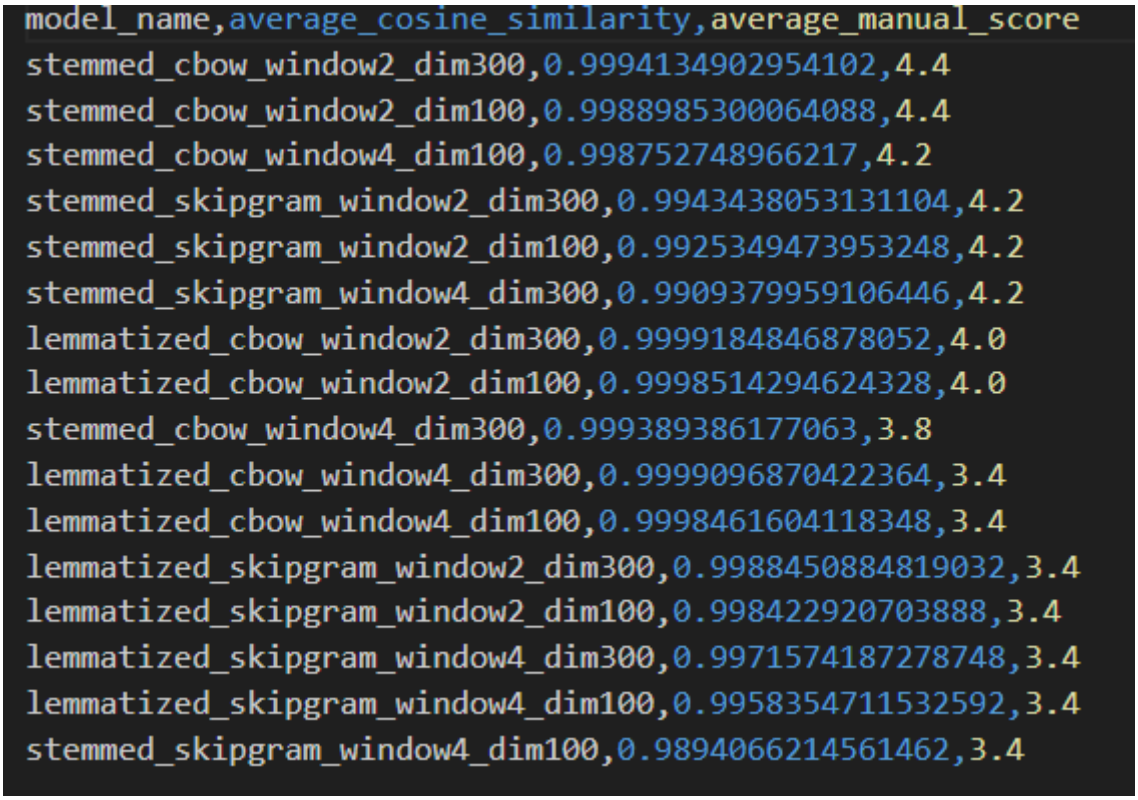
3: Bağlamsal benzerlik var

2: Kısmi alaka

1: Alakasız

En yüksek puan ortalaması alan modeller:

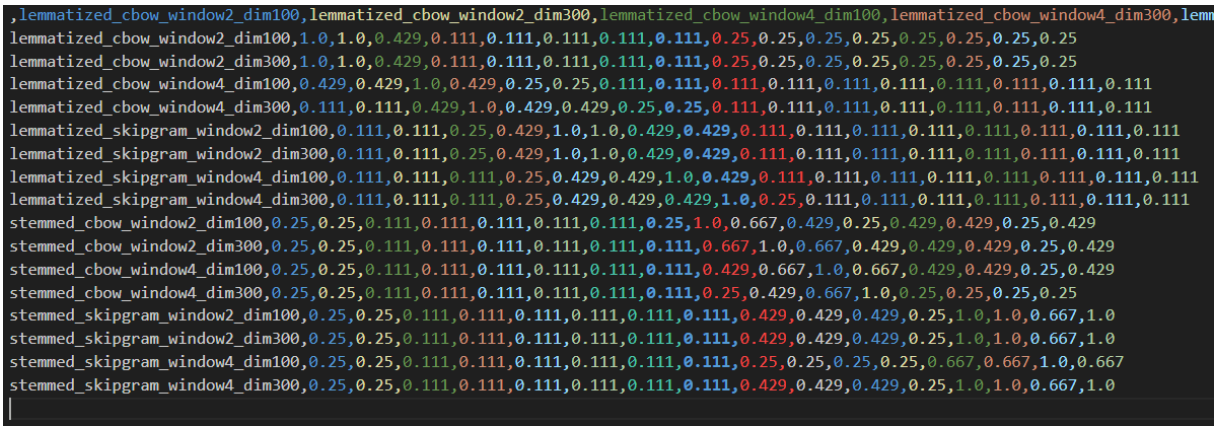
Model	Avg. Anlamsal Skor
stemmed_cbow_window2_dim300	4.4
stemmed_cbow_window2_dim100	4.4
stemmed_cbow_window4_dim100	4.2



```
model_name,average_cosine_similarity,average_manual_score
stemmed_cbow_window2_dim300,0.9994134902954102,4.4
stemmed_cbow_window2_dim100,0.9988985300064088,4.4
stemmed_cbow_window4_dim100,0.998752748966217,4.2
stemmed_skipgram_window2_dim300,0.9943438053131104,4.2
stemmed_skipgram_window2_dim100,0.9925349473953248,4.2
stemmed_skipgram_window4_dim300,0.9909379959106446,4.2
lemmatized_cbow_window2_dim300,0.9999184846878052,4.0
lemmatized_cbow_window2_dim100,0.9998514294624328,4.0
stemmed_cbow_window4_dim300,0.999389386177063,3.8
lemmatized_cbow_window4_dim300,0.9999096870422364,3.4
lemmatized_cbow_window4_dim100,0.9998461604118348,3.4
lemmatized_skipgram_window2_dim300,0.9988450884819032,3.4
lemmatized_skipgram_window2_dim100,0.998422920703888,3.4
lemmatized_skipgram_window4_dim300,0.9971574187278748,3.4
lemmatized_skipgram_window4_dim100,0.9958354711532592,3.4
stemmed_skipgram_window4_dim100,0.9894066214561462,3.4
```

7. Jaccard Benzerlik Matrisi

- Her modelin 5 önerisi alındı
- Toplam 18x18 Jaccard matrisi oluşturuldu
- Aynı önerileri sunan modellerde **1.0**, farklılarda **0–0.3** skorlar gözlendi



```
,lemmatized_cbow_window2_dim100,lemmatized_cbow_window2_dim300,lemmatized_cbow_window4_dim100,lemmatized_cbow_window4_dim300,lemm
lemmatized_cbow_window2_dim100,1.0,1.0,0.429,0.111,0.111,0.111,0.111,0.111,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25
lemmatized_cbow_window2_dim300,1.0,1.0,0.429,0.111,0.111,0.111,0.111,0.111,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25
lemmatized_cbow_window4_dim100,0.429,0.429,1.0,0.429,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111
lemmatized_cbow_window4_dim300,0.111,0.111,0.429,1.0,0.429,0.429,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111
lemmatized_skipgram_window2_dim100,0.111,0.111,0.25,0.429,1.0,1.0,0.429,0.429,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111
lemmatized_skipgram_window2_dim300,0.111,0.111,0.25,0.429,1.0,1.0,0.429,0.429,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111
lemmatized_skipgram_window4_dim100,0.111,0.111,0.111,0.25,0.429,0.429,1.0,0.429,0.111,0.111,0.111,0.111,0.111,0.111,0.111,0.111
lemmatized_skipgram_window4_dim300,0.111,0.111,0.111,0.25,0.429,0.429,0.429,1.0,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.111
stemmed_cbow_window2_dim100,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.25,1.0,0.667,0.429,0.25,0.429,0.429,0.25,0.429
stemmed_cbow_window2_dim300,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.667,1.0,0.667,0.429,0.429,0.429,0.25,0.429
stemmed_cbow_window4_dim100,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.429,0.667,1.0,0.667,0.429,0.429,0.25,0.429
stemmed_cbow_window4_dim300,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.25,0.429,0.667,1.0,0.25,0.25,0.25,0.25
stemmed_skipgram_window2_dim100,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.429,0.429,0.429,0.25,1.0,1.0,0.667,1.0
stemmed_skipgram_window2_dim300,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.429,0.429,0.429,0.25,1.0,1.0,0.667,1.0
stemmed_skipgram_window4_dim100,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.25,0.25,0.25,0.25,0.667,0.667,1.0,0.667
stemmed_skipgram_window4_dim300,0.25,0.25,0.111,0.111,0.111,0.111,0.111,0.111,0.429,0.429,0.429,0.25,1.0,1.0,0.667,1.0
```

8. Gözlem ve Yorum

- **Stemmed + CBOW + window=2** kombinasyonu hem cosine similarity hem de anlamsal değerlendirmede öne çıktı.
- CBOW yöntemi SkipGram'e kıyasla daha tutarlı oldu.
- TF-IDF modelleri, Word2Vec'e göre daha yüzeysel kaldı.
- Jaccard benzerliği ile aynı çıktılar veren modeller tespit edilebildi.

9. Sonuç

Bu çalışma, kelime temsili yöntemlerinin spor yorumları üzerindeki etkisini ve öneri sistemlerinde kullanım potansiyelini ortaya koymuştur. En başarılı model kombinasyonları belirlenerek hem istatistiksel hem anlamsal değerlendirme başarıyla uygulanmıştır.

Elde edilen sonuçlara göre, stemmed + CBOW kombinasyonları, kısa cümleli ve hızlı eşleşme gerektiren öneri sistemleri için oldukça uygundur. Öte yandan, SkipGram modelleri daha derin semantik ilişkiler kurabilir ve özellikle az geçen (low frequency) kelimeler için daha iyi performans sunar. TF-IDF modelleri ise yüzeysel ama hızlı bir sıralama aracı olarak kullanılabilir. 300 boyutlu vektörlerle eğitilen Word2Vec modelleri anlam açısından daha güçlü, 100 boyutlu olanlar ise daha hızlı ve düşük maliyetlidir.

10. Ek Dosyalar (GitHub):

- `reddit_soccer_clean_comments.csv`
- `model_outputs.csv`
- `manual_scoring_filled.csv`

- `model_average_similarity_scores.csv`
- `model_comparison_scores.csv`
- `jaccard_matrix.csv`
- Kodlar: `veri_cek.py`, `preprocess.py`, `train_16_models.py`, `recommend.py`, `compare_models_scores.py`, `jaccard_matrix_generator.py`