1.GİRİŞ

Bu çalışmanın amacı, birinci ödevde ön işlemden geçirilmiş CV verisiyle eğitilen TF-IDF ve Word2Vec modellerini kullanarak metinler arası benzerlik hesaplamaları yapmak ve bu modellerin performanslarını karşılaştırmalı olarak değerlendirmektir.

Veri seti, Türkçe özgeçmiş (CV) cümlelerinden oluşmaktadır. Ön işleme süreçlerinde cümleler lemmatizasyon ve stemleme yöntemleriyle temizlenmiş, ardından 2 adet TF-IDF ve 16 adet Word2Vec modeli eğitilmiştir.

2.Yöntem

2.1 TF-IDF Benzerliği

TF-IDF yöntemiyle her cümle, kelime sıklıklarına dayalı vektörler ile temsil edilmiştir. Giriş cümlesinin TF-IDF vektörü, tüm veri setindeki cümlelerle cosine similarity yöntemiyle karşılaştırılarak en benzer 5 cümle bulunmuştur.

Kullanılan araçlar: TfidfVectorizer, cosine_similarity

📌 Görsel Ekle:

Stemmed için çıktı ve skorlar:

```
cv_data = pd.read_csv("data/cv_dataset.csv")
print("=== TF-IDF (Stemmed) Benzer Sonuçlar ===")
for idx in top_5_indices:
    print(f"\nCV #{idx} - Skor: {similarities[idx]:.4f}")
    print(cv_data.iloc[idx]["on_yazi"])
```

```
=== TF-IDF (Stemmed) Benzer Sonuçlar ===

CV #2245 - Skor: 1.0000
İlgili alandaki tecrübelerim ve becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum.

CV #591 - Skor: 1.0000
İlgili alandaki tecrübelerim ve becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum.

CV #3122 - Skor: 1.0000
İlgili alandaki tecrübelerim ve becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum.

CV #3662 - Skor: 1.0000
İlgili alandaki tecrübelerim ve becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum.

CV #596 - Skor: 1.0000
İlgili alandaki tecrübelerim ve becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum.
```

Lemmatized için çıktı ve skorlar:

```
print("=== TF-IDF (Lemmatized) Benzer Sonuçlar ===")
for idx in top_5_indices:
    print(f"\nCV #{idx} - Skor: {similarities[idx]:.4f}")
    print(lemmatized_df.iloc[idx]["on_yazi"])
```

```
=== TF-IDF (Lemmatized) Benzer Sonuçlar ===

CV #1853 - Skor: 1.0000
ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum

CV #3654 - Skor: 1.0000
ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum

CV #3655 - Skor: 1.0000
ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum

CV #1529 - Skor: 1.0000
ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum

CV #4255 - Skor: 1.0000
ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum
```

2.2 Word2Vec Benzerliği

16 farklı Word2Vec modeli (CBOW ve Skip-gram, farklı pencere boyutu ve vektör boyutlarıyla) kullanılarak, giriş metni ve tüm veri setindeki metinler için ortalama kelime vektörleri alınmış ve yine cosine similarity hesaplanmıştır. Her model için en benzer 5 cümle elde edilmiştir.

Kullanılan araçlar: gensim.models.Word2Vec, cosine_similarity, NumPy

3.Sonuçlar ve Değerlendirme

3.1 Her Modelin İlk 5 Benzer Metni

Her model için, giriş metniyle en çok benzerlik gösteren ilk 5 cümle ve benzerlik skorları tablo halinde aşağıda sunulmuştur.

Model Bazlı İlk 5 Sonuç ve Anlamsal Puanlar:

```
pd.set_option('display.max_colwidth', None)
pd.set_option('display.max_rows', 100)

# Tüm tabloyu stil ile göster
display(HTML("<h3>📄 Model Bazlı İlk 5 Sonuç ve Anlamsal Puanlar</h3>"))
display(evaluation_df.style.set_caption("Her Model İçin İlk 5 Benzer CV ve Puanları").background_gradient(cmap='YlGnBu'))
```

📄 Model Bazlı İlk 5 Sonuç ve Anlamsal Puanlar

Her Model İçin İlk 5 Benzer CV ve Puanları

	Model	Sıra	CV_Index	Benzerlik_Skoru	On_Yazı	Anlamsal_Skor
0	cv_lemmatized_model_cbow_window2_dim100	1	3393	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	4
1	cv_lemmatized_model_cbow_window2_dim100	2	2560	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	4
2	cv_lemmatized_model_cbow_window2_dim100	3	2561	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	3
3	cv_lemmatized_model_cbow_window2_dim100	4	3882	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	5
4	cv_lemmatized_model_cbow_window2_dim100	5	2566	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	4
5	cv_lemmatized_model_cbow_window2_dim300	1	2814	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	2
6	cv_lemmatized_model_cbow_window2_dim300	2	4230	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	3
7	cv_lemmatized_model_cbow_window2_dim300	3	4229	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	2
8	cv_lemmatized_model_cbow_window2_dim300	4	2779	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	3
9	cv_lemmatized_model_cbow_window2_dim300	5	312	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	1
10	cv_lemmatized_model_cbow_window4_dim100	1	4389	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	3
11	cv_lemmatized_model_cbow_window4_dim100	2	2172	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	3
12	cv_lemmatized_model_cbow_window4_dim100	3	3239	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	4
13	cv_lemmatized_model_cbow_window4_dim100	4	134	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	4
14	cv_lemmatized_model_cbow_window4_dim100	5	132	1.000000	ilgili alandaki tecrübelerim becerilerim doğrultusunda katkı sağlayabileceğime inanıyorum	3

3.2 Anlamsal Değerlendirme

Her modelin önerdiği 5 cümle, giriş cümlesiyle anlamsal benzerliği açısından 1-5 arası puanlanmıştır.

1: Çok alakasız

5: Çok güçlü benzerlik

Ortalama Anlamsal Skorlar:

```
# Anlamsal değerlendirme tablosunu görsel olarak göster
from IPython.display import display, HTML

# Ortalama skorları göster
ortalama_df = (
    evaluation_df
    .groupby("Model")["Anlamsal_Skor"]
    .mean()
    .reset_index()
    .rename(columns={"Anlamsal_Skor": "Ortalama_Anlamsal_Skor"})
    .sort_values(by="Ortalama_Anlamsal_Skor", ascending=False)
)

# Tabloyu görselleştir
display(HTML("<h3>🔍 Anlamsal Değerlendirme Ortalamaları</h3>"))
display(ortalama_df.style.set_caption("Model Başına Ortalama Anlamsal Skor").background_gradient(cmap='Blues'))
```

🔍 Anlamsal Değerlendirme Ortalamaları

Model Başına Ortalama Anlamsal Skor

	Model	Ortalama_Anlamsal_Skor
0	cv_lemmatized_model_cbow_window2_dim100	4.000000
2	cv_lemmatized_model_cbow_window4_dim100	3.400000
1	cv_lemmatized_model_cbow_window2_dim300	2.200000

🎯 Yorum:

En yüksek ortalama: cv_lemmatized_model_cbow_window2_dim100 (4.0)

En düşük ortalama: cv_lemmatized_model_cbow_window4_dim300 (2.2)

TF-IDF modelleri genel olarak Word2Vec'e kıyasla daha düşük ortalama skorlar aldı.

Word2Vec içinde yapılandırmalar belirleyici: düşük pencere boyutu ve düşük vektör boyutu daha iyi sonuç verdi.

3.3 Jaccard Benzerliği

Modellerin sıraladığı ilk 5 sonucu karşılaştırarak Jaccard benzerliği matrisi oluşturulmuştur. Böylece modeller arası benzerlik tutarlılığı analiz edilmiştir.

🎯 Yorum:

Aynı yapılandırmaya sahip modellerin (ör. CBOW + aynı pencere boyutu) Jaccard skorları genelde daha yüksek.

cv_lemmatized_model_cbow_window2_dim100 ve cv_lemmatized_model_cbow_window2_dim300 Jaccard ≈ 0.8 → oldukça tutarlı sıralama üretmiş.

Yüksek Jaccard + yüksek Anlamsal Skor bir araya gelince model daha tutarlı ve anlamlı sonuçlar veriyor.

4.Sonuç ve Öneriler

Word2Vec modelleri, TF-IDF modellerine göre daha anlamlı ve güçlü benzerlikler sunmuştur.

En başarılı model: cv_lemmatized_model_cbow_window2_dim100

CBOW mimarisi, küçük pencere boyutu (window=2) ve küçük vektör boyutu (dim=100) ile daha iyi sonuçlar alınmıştır.

Jaccard analizi, bazı modellerin benzer sıralamalar yaptığını göstermiştir, bu da model yapılandırmasının sıralama tutarlılığına doğrudan etkisini gösteriyor.

Öneriler:

Daha geniş veri setleriyle Skip-gram mimarisi test edilebilir.

Anlamsal puanlar başka kullanıcılar tarafından ortalamalı olarak alınarak değerlendirmenin nesnelliği artırılabilir.

Sadece cosine benzerliği değil, başka metriklerle de karşılaştırma yapılabilir (örn. Euclidean distance).

5.Ekler

Github Reposu: https://github.com/00GunneR00?tab=repositories

Çalıştırma yönergeleri: README.md içerisinde yer alacaktır