Yusuf ERTAN -2102131028- Dizi/Film Senaryo Analizi Ödevi

1.GİRİŞ

Ödevin Amacı:

Ödevimin amacı IMSDB (Internet Movie Script Database) sitesi üzerinden filmlerin senaryolarını çekip, senaryolar üzerinden metin ön işleme adımları yapmak ve bu sayede benzer filmleri ortaya çıkarmaktır.

Veri Setim:

Kullandığım veri seti IMSDB (Internet Movie Script Database) sitesi üzerinden veri çok büyük olmasın diye 10 tane film türü seçtim ve her tür için 7 tane film senaryosu çektim.

2.YÖNTEM

Benzerlik Nasıl Hesaplandı:

Film senaryoları arasındaki benzerliği hesaplamak için önce metinleri temizledim, küçük harfe çevirdim, gereksiz kelimeleri çıkardım ve kök hallerini aldım. Yani metin ön işleme adımlarından geçirdim. Daha sonra her senaryo, TF-IDF yöntemiyle sayısal verilere dönüştürdüm. Hem lemmatize hem de stemmed yaptığım verileri ayrı olarak TF- IDF uyguladım. Daha sonra senaryolar arasındaki benzerliği, kosinüs benzerliği adlı yöntemle ölçtüm.

Hangi Modeller ve Teknikler Kullanıldı:

Tokenization, Lowercasing, Stop Word Removal, Lemmatization, Stemming, özel karakter ve noktalama işaretlerinin temizlenmesi, TF-IDF, CBOW (Continuous Bag of Words), Skip-gram, Cosine Similarity modellerini kullandım.

3.SONUÇLAR ve DEĞERLENDİRME

Her model için ilk 5 benzer metinler :

```
Lemmatized Word2Vec Modelleri - En benzer 5 metin:

Model: lemmatized_w2v_cbow_win2_vec100.model
1. Metin indeksi: 60, Benzerlik skoru: 0.8771
2. Metin indeksi: 28, Benzerlik skoru: 0.8744
3. Metin indeksi: 11, Benzerlik skoru: 0.8704
4. Metin indeksi: 0, Benzerlik skoru: 0.8700
5. Metin indeksi: 52, Benzerlik skoru: 0.8643

Model: lemmatized_w2v_skipgram_win2_vec100.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8791
2. Metin indeksi: 11, Benzerlik skoru: 0.8579
3. Metin indeksi: 67, Benzerlik skoru: 0.8555
```

4. Metin indeksi: 28, Benzerlik skoru: 0.8555
5. Metin indeksi: 4, Benzerlik skoru: 0.8529

Model: lemmatized_w2v_cbow_win4_vec100.model
1. Metin indeksi: 28, Benzerlik skoru: 0.8312
2. Metin indeksi: 11, Benzerlik skoru: 0.8193
3. Metin indeksi: 60, Benzerlik skoru: 0.8159
4. Metin indeksi: 0, Benzerlik skoru: 0.8066
5. Metin indeksi: 21, Benzerlik skoru: 0.7952

Model: lemmatized_w2v_skipgram_win4_vec100.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8593
2. Metin indeksi: 67, Benzerlik skoru: 0.8141
3. Metin indeksi: 54, Benzerlik skoru: 0.8140
4. Metin indeksi: 26, Benzerlik skoru: 0.8063
5. Metin indeksi: 27, Benzerlik skoru: 0.8051

Model: lemmatized_w2v_cbow_win2_vec300.model
1. Metin indeksi: 28, Benzerlik skoru: 0.8883
2. Metin indeksi: 60, Benzerlik skoru: 0.8854
3. Metin indeksi: 0, Benzerlik skoru: 0.8778
4. Metin indeksi: 64, Benzerlik skoru: 0.8704
5. Metin indeksi: 11, Benzerlik skoru: 0.8703

Model: lemmatized_w2v_skipgram_win2_vec300.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8814
2. Metin indeksi: 67, Benzerlik skoru: 0.8607
3. Metin indeksi: 28, Benzerlik skoru: 0.8564
4. Metin indeksi: 1, Benzerlik skoru: 0.8531
5. Metin indeksi: 11, Benzerlik skoru: 0.8490

Model: lemmatized_w2v_cbow_win4_vec300.model
1. Metin indeksi: 28, Benzerlik skoru: 0.8636
2. Metin indeksi: 0, Benzerlik skoru: 0.8269
3. Metin indeksi: 60, Benzerlik skoru: 0.8231
4. Metin indeksi: 51, Benzerlik skoru: 0.8155
5. Metin indeksi: 11, Benzerlik skoru: 0.8086

Model: lemmatized_w2v_skipgram_win4_vec300.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8509
2. Metin indeksi: 67, Benzerlik skoru: 0.7977
3. Metin indeksi: 54, Benzerlik skoru: 0.7953
4. Metin indeksi: 26, Benzerlik skoru: 0.7885
5. Metin indeksi: 61, Benzerlik skoru: 0.7785

Stemmed Word2Vec Modelleri - En benzer 5 metin:

Model: stemmed_w2v_cbow_win2_vec100.model
1. Metin indeksi: 28, Benzerlik skoru: 0.8565
2. Metin indeksi: 0, Benzerlik skoru: 0.8448
3. Metin indeksi: 11, Benzerlik skoru: 0.8424
4. Metin indeksi: 51, Benzerlik skoru: 0.8382
5. Metin indeksi: 60, Benzerlik skoru: 0.8353

Model: stemmed_w2v_skipgram_win2_vec100.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8825
2. Metin indeksi: 54, Benzerlik skoru: 0.8555
3. Metin indeksi: 67, Benzerlik skoru: 0.8514
4. Metin indeksi: 27, Benzerlik skoru: 0.8456

```
5. Metin indeksi: 26, Benzerlik skoru: 0.8455

Model: stemmed_w2v_cbow_win4_vec100.model
1. Metin indeksi: 0, Benzerlik skoru: 0.7603
2. Metin indeksi: 28, Benzerlik skoru: 0.7360
3. Metin indeksi: 11, Benzerlik skoru: 0.7270
4. Metin indeksi: 54, Benzerlik skoru: 0.7264
5. Metin indeksi: 51, Benzerlik skoru: 0.7264

Model: stemmed_w2v_skipgram_win4_vec100.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8605
2. Metin indeksi: 54, Benzerlik skoru: 0.8165
3. Metin indeksi: 26, Benzerlik skoru: 0.8046
4. Metin indeksi: 67, Benzerlik skoru: 0.8042
5. Metin indeksi: 61, Benzerlik skoru: 0.7940

Model: stemmed_w2v_cbow_win2_vec300.model
1. Metin indeksi: 11, Benzerlik skoru: 0.8559
2. Metin indeksi: 62, Benzerlik skoru: 0.8505
3. Metin indeksi: 28, Benzerlik skoru: 0.8487
4. Metin indeksi: 0, Benzerlik skoru: 0.8480
5. Metin indeksi: 60, Benzerlik skoru: 0.8374

Model: stemmed_w2v_skipgram_win2_vec300.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8795
2. Metin indeksi: 54, Benzerlik skoru: 0.8493
3. Metin indeksi: 67, Benzerlik skoru: 0.8476
4. Metin indeksi: 27, Benzerlik skoru: 0.8412
5. Metin indeksi: 23, Benzerlik skoru: 0.8392

Model: stemmed_w2v_cbow_win4_vec300.model
1. Metin indeksi: 0, Benzerlik skoru: 0.7822
2. Metin indeksi: 28, Benzerlik skoru: 0.7586
3. Metin indeksi: 54, Benzerlik skoru: 0.7507
4. Metin indeksi: 51, Benzerlik skoru: 0.7488
5. Metin indeksi: 11, Benzerlik skoru: 0.7474

Model: stemmed_w2v_skipgram_win4_vec300.model
1. Metin indeksi: 0, Benzerlik skoru: 0.8600
2. Metin indeksi: 54, Benzerlik skoru: 0.8087
3. Metin indeksi: 67, Benzerlik skoru: 0.7975
4. Metin indeksi: 26, Benzerlik skoru: 0.7968
5. Metin indeksi: 27, Benzerlik skoru: 0.7873
```

Benzerlik skorları da matrisler şeklinde gösteriyorum.

Hangi modeller daha başarılı:

En başarılı olarak `Model: lemmatized_w2v_cbow_win2_vec300.model`

Bu modeli görüyorum sayıları en yüksek bu model. Daha sonra

```
Model: stemmed_w2v_skipgram_win2_vec100.model
Model: stemmed_w2v_cbow_win2_vec300.model
Model: stemmed_w2v_skipgram_win2_vec300.model

Bu modelleri daha başarılı olarak yorumluyorum.
```

Model yapılandırmalarının başarıya etkileri:

Ödevimde farklı pencere boyutları denenmiştir, genellikle geniş pencere boyutları modelin anl am ilişkilerini yakalama yeteneğini artırmıştır. Ancak, çok büyük pencere boyutları modelin g enel performansını düşürebilir çünkü anlam karmaşıklaşabilir.

Projede 100 ve 300 boyutlu vektörler kullanılmış ve 300 boyutlu modeller genellikle daha yük sek benzerlik skorları elde etmiştir.

Bu çalışmada her iki model türü de (CBOW ve Skip-gram) kullanılarak karşılaştırılmış, mode l türüne bağlı olarak performans değişiklikleri gözlemlenmiştir.

4.SONUÇ ve ÖNERİLER

Genel Çıkarımlar:

Ödevimde film senaryoları metinleri üzerinde metin benzerliği tespiti amacıyla çeşitli metin ön işleme teknikleri (lemmatizasyon, stemming) ve farklı Word2Vec yapılandırmaları (CBOW, Skip-gram, pencere boyutu, vektör boyutu) uygulanmıştır. Ayrıca TF-IDF tabanlı benzerlik analizi ile Word2Vec temelli yöntemler karşılaştırılmıştır.Elde edilen bulgulara göre:

Lemmatizasyon ile elde edilen metinler, stemming yöntemine kıyasla daha anlamlı temsiller üretmiş ve benzerlik analizinde daha başarılı sonuçlar vermiştir.

Skip-gram modeli, özellikle daha az geçen kelimeler üzerinde daha başarılı olup, anlamsal benzerliği daha iyi yansıtmıştır.

300 boyutlu vektörlerle eğitilen modeller, 100 boyutlu modellere göre daha yüksek benzerlik skorları üretmiş, bu da daha zengin anlamsal temsiller sağladığını göstermektedir.

TF-IDF yöntemi, temel düzeyde etkili sonuçlar verse de, kelime sırasını ve anlamsal ilişkileri göz önünde bulundurmadığı için Word2Vec tabanlı modellere göre daha yüzeysel kalmıştır.

Hangi model, hangi tür görevler için uygun olabilir:

CBOW modeli, daha hızlı eğitim süresi sayesinde büyük veri setlerinde ve genel amaçlı görevlerde tercih edilebilir.

Skip-gram modeli, özellikle anlam derinliği gerektiren metin benzerliği gibi görevlerde daha isabetli sonuçlar vermektedir.

TF-IDF yöntemi, ön analiz veya içerik sınıflandırma gibi dilin semantiğinden ziyade frekansa dayalı uygulamalarda yeterli olabilir.

Word2Vec (özellikle Skip-gram, 300 boyut, geniş pencere ile) modelleri, içerik tabanlı öneri sistemleri, benzer senaryo bulma, tema analizi gibi anlamsal derinlik gerektiren görevlerde daha uygundur.