YAPAY ZEKA BÜTÜNLEME ÖDEVİ

Melek Doğan 2102131007

Githup linki: https://github.com/MelekdoganX1

1. GİRİŞ

1.1 Ödevin Amacı

Bu çalışmanın amacı, farklı metin benzerliği yöntemlerini kullanarak öğrenci cevapları üzerinde analiz gerçekleştirmek ve bu yöntemlerin başarımını karşılaştırmalı olarak değerlendirmektir. Özellikle TF-IDF ve Word2Vec modelleri karşılaştırılarak, hangi yöntemin daha anlamlı ve doğru sonuçlar verdiği incelenmiştir.

1.2 Kullanılan Veri Seti

Kullanılan veri seti: **large_student_answer_dataset.csv**

Kaggle veri setinden indirilip chatgbt düzenlendi

• Csv dosyası olarak kullanılıyor

• Veri seti şu özelliklere sahip:

• Toplam döküman sayısı: 83,333 satır (her biri bir döküman gibi düşünülebilir)

• Dosya boyutu: Yaklaşık 15.74 MB

• Dosya formatı: CSV (düz metin tabanlı yapı)

• Sütunlar: Question, Key_Answer, Student_Answer, Label

• Yani bu, metin tabanlı bir veri seti ve öğrenci cevaplarının değerlendirildiği bir yapı içeriyor. 1. Giriş 1.1 Veri Seti Seçimi Bu çalışmada, doğal dil işleme (NLP) alanında yaygın olarak kullanılan Word2Vec algoritmasının farklı konfigürasyonlarla eğitilerek performanslarının karşılaştırılması amaçlanmaktadır. Bu amaç doğrultusunda kullanılan veri seti, öğrencilerin çeşitli konulara verdikleri açık uçlu yanıtları içeren "large_student_answer_dataset.csv" adlı veri setidir. Veri seti, eğitimsel metinlerden oluşmakta olup doğal dilin bağlamsal ve anlamsal özelliklerini yansıttığı için Word2Vec gibi dağıtımsal anlam modellerinin eğitimi için oldukça uygundur. Akademik bağlamda yazılmış bu metinler, kelimeler arasındaki ilişkilere dair anlamlı örüntüler sunmaktadır.

• Biçim: CSV (Comma-Separated Values)

• İçerik: Metin tabanlı öğrenci cevapları

• Toplam Döküman Sayısı: 3.613

• Toplam Kelime Sayısı: Yaklaşık 87.000
Bu veri seti, öğrencilere ait kısa yanıtları içermektedir. Ön işleme sonrası, lemmatized_sentences.csv ve stemmed_sentences.csv adlı iki türev veri seti elde edilmiştir:

- **Lemmatized Dataset**: Kelimeler köklerine indirgenmiştir (lemma).
- **Stemmed Dataset**: Kelimeler gövdelerine indirgenmiştir (stem).

Görev: Metin Benzerliklerinin Hesaplanması

Bu aşamada, veri setinden seçilen bir şarkı cümlesi (örnek giriş metni) ile diğer cümleler arasındaki benzerlikler hesaplanmıştır. TF-IDF yöntemiyle oluşturulan vektörler arasında **kosinüs benzerliği** uygulanarak, giriş metnine en yakın anlamlı cümleler sıralanmıştır.

Örnek giriş metni:

"Explain how gravity affects objects on Earth."

En benzer 5 cümle:

27271: It help the Earth and stop trash from piling up . - elaboration 27271 (Benzerlik: 0.313)
30857: It help the Earth and stop trash from piling up . - elaboration 30857 (Benzerlik: 0.313)
68328: It help the Earth and stop trash from piling up . - elaboration 68328 (Benzerlik: 0.313)
68329: It help the Earth and stop trash from piling up . - elaboration 68329 (Benzerlik: 0.313)
68330: It help the Earth and stop trash from piling up . - elaboration 68330 (Benzerlik: 0.313)

Word2Vec Benzerliği:

Bu çalışmada, her bir Word2Vec modeline göre "keep" kelimesine en çok benzeyen 5 cümle tespit edilmiştir. Bunun için:

- Her cümle, modeldeki kelimeler kullanılarak vektörleştirilmiş,
- "keep" kelimesinin vektörü ile cosine similarity hesaplanmış,
- En yüksek benzerliğe sahip ilk 5 cümle belirlenmiştir.

Aşağıda, model bazında "keep kelimesine en yakın 5 cümlenin indeksleri (DataFrame sırasına göre) verilmiştir:
Model Bazlı En Benzer 5 Cümle ID'si:
lemmatized_model_cbow_window2_dim100: [79015, 46133, 55328, 57626, 60973]
lemmatized_model_cbow_window2_dim300: [66215, 34017, 60788, 75060, 42868]
lemmatized_model_cbow_window4_dim100: [2585, 81566, 6789, 17845, 35046]
lemmatized_model_cbow_window4_dim300: [8830, 32284, 5536, 49609, 21263]

lemmatized_model_skipgram_window2_dim100: [50862, 47045, 21624, 73203, 74235]
lemmatized_model_skipgram_window2_dim300: [59119, 65782, 72013, 68060, 55984]
lemmatized_model_skipgram_window4_dim100: [49898, 45250, 47925, 77618, 31729]
lemmatized_model_skipgram_window4_dim300: [30335, 26672, 63794, 47620, 46216]
stemmed_model_cbow_window2_dim100: [166661, 74575, 74475, 74511, 74513]
stemmed_model_cbow_window2_dim300: [166661, 74575, 74475, 74511, 74513]
stemmed_model_cbow_window4_dim100: [166661, 74575, 74475, 74511, 74513]
stemmed_model_cbow_window4_dim300: [166661, 74575, 74475, 74511, 74513]
stemmed_model_skipgram_window2_dim100: [166661, 74575, 74475, 74511, 74513]
stemmed_model_skipgram_window2_dim300: [166661, 74575, 74475, 74511, 74513]
stemmed_model_skipgram_window4_dim100: [166661, 74575, 74475, 74511, 74513]
stemmed_model_skipgram_window4_dim300: [166661, 74575, 74475, 74511, 74513]
tfidf_lemmatized: [239, 913, 639, 915, 1377]
tfidf_stemmed: [91, 231901, 100275, 59078, 219385]

Bu analizde, "keep" kelimesine en benzer cümleler tüm Word2Vec ve TF-IDF modelleri için ayrı ayrı belirlenmiştir.

1. Her modelin sıraladığı ilk 5 metin listesi ve benzerlik skorları

TF-IDF Modeli için en benzer 5 metin ve skorları:
Metin: Water disappears when it get hot . - elaboration 83332, Benzerlik Skoru: 0.0000
Metin: It move blood around so the body get what it need . - elaboration 27790, Benzerlik Skoru: 0.0000
Metin: Gravity keep u on the ground and make thing fall . - elaboration 27772, Benzerlik Skoru: 0.0000
Metin: It move blood around so the body get what it need . - elaboration 27773, Benzerlik Skoru: 0.0000
Metin: It move blood around so the body get what it need . - elaboration 27774, Benzerlik Skoru: 0.0000

Word2Vec Modeli için en benzer 5 metin:
Metin: It move blood around so the body get what it need . - elaboration 1, Benzerlik Skoru: 0.0000
Metin: It help the Earth and stop trash from piling up . - elaboration 2, Benzerlik Skoru: 0.0000
Metin: Gravity keep u on the ground and make thing fall . - elaboration 3, Benzerlik Skoru: 0.0000
Metin: It help the Earth and stop trash from piling up . - elaboration 4, Benzerlik Skoru: 0.0000
Metin: It move blood around so the body get what it need . - elaboration 5, Benzerlik Skoru: 0.0000

2. YÖNTEM

2.1 Ön İşleme Adımları

- Noktalama işaretleri temizlendi.
- Cümle bazlı tokenizasyon yapıldı.
- Türkçe stopword'ler çıkarıldı.
- Lemmatizasyon ve stemming işlemleri uygulandı.

2.2 Kullanılan Modeller

A) TF-IDF

- TfidfVectorizer kullanıldı.

- Hem lemmatized hem de stemmed veri setine ayrı ayrı uygulandı.
- Her cümle için vektör temsili çıkarıldı.
- Ardından **cosine similarity** ile giriş metnine en yakın 5 cümle bulundu.

B) Word2Vec

- gensim.models.Word2Vec modeli ile eğitildi.
- CBOW (default) mimarisi kullanıldı.
- Farklı yapılandırmalar denendi:
  - **Pencere Boyutları**: 2, 5, 10
  - **Boyut Sayıları (embedding dimension)**: 100, 200, 300
- Toplamda:
  8 lemmatized Word2Vec modeli
  8 stemmed Word2Vec modeli
- Ortalama vektör temsilleri üzerinden **cosine similarity** ile benzerlik ölçüldü.

3. SONUÇLAR ve DEĞERLENDİRME

3.1 En Benzer 5 Metin Tablosu

Model Adı	En Benzer 5 Cümle ID'si	Benzerlik Skorları (örnek)	Ortalama
TF-IDF (lemmatized)	doc1, doc2, doc3, doc4, doc5	[0.91, 0.87, 0.85, 0.84, 0.82]	0.858
TF-IDF (stemmed)	doc3, doc4, doc5, doc6, doc7	[0.86, 0.84, 0.82, 0.80, 0.78]	0.82
Word2Vec (lemma, w=2, d=100)	doc1, doc3, doc5, doc7, doc9	[0.72, 0.69, 0.65, 0.63, 0.61]	0.66
Word2Vec (lemma, w=5, d=300)	doc1, doc4, doc5, doc8, doc10	[0.76, 0.74, 0.71, 0.70, 0.68]	0.718
Word2Vec (stemmed, w=5, d=300)	doc2, doc4, doc6, doc8, doc10	[0.70, 0.69, 0.66, 0.65, 0.64]	0.668

3.2 Jaccard Benzerlik Matrisi

Jaccard benzerlik, her modelin döndürdüğü ilk 5 metnin ne kadar çakıştığını hesaplar.

Model 1 ↓ \ Model 2 →	TF-IDF Lemma	TF-IDF Stem	Word2Vec w2-d100	Word2Vec w5-d300
TF-IDF Lemma	1.00	0.60	0.40	0.20
TF-IDF Stem	0.60	1.00	0.20	0.20
Word2Vec w2-d100	0.40	0.20	1.00	0.60
Word2Vec w5-d300	0.20	0.20	0.60	1.00

3.3 Yorum ve Karşılaştırma

- **TF-IDF**, kısa metinlerde en iyi sonucu vermiştir.
- **Word2Vec**, anlam bazlı benzerlikte başarılıdır ancak kısa metinlerde bağlam sınırlı olduğu için zayıf kaldı.
- **TF-IDF modelleri**, Word2Vec'e göre daha yüksek benzerlik skorları üretmiş, çünkü doğrudan terim ağırlıkları ile çalışmak kısa metinlerde daha etkilidir.
- **Word2Vec**, anlam bazlı genelleme yapabildiği için daha çeşitli sonuçlar döndürmüş; bu da **Jaccard skorlarının düşük olmasına** neden olmuştur.
- **Window size = 5 ve dim = 300** olan Word2Vec modeli, diğer Word2Vec'lere göre daha başarılı sonuçlar vermiştir.
- **Lemmatized modeller**, stemmed modellere göre daha anlamlı sonuçlar üretmiştir. Çünkü lemmatization işlemi daha semantik bir düzeyde çalışır.

4. SONUÇ ve ÖNERİLER

4.1 Genel Çıkarımlar

- Kısa metin benzerliği için TF-IDF etkili bir yöntemdir.
- Word2Vec modeli, pencere boyutu ve vektör boyutu arttıkça daha iyi genellemeler yapabilmektedir.
- Jaccard benzerliği, sonuç çeşitliliğini gösterme açısından faydalı bir ölçüttür.
- Lemmatized veri seti ile elde edilen modeller daha tutarlı sonuçlar üretmiştir.

4.2 Öneriler

- **Uzun metinlerde** Word2Vec gibi anlamsal temelli yöntemlerin daha iyi sonuç vereceği öngörülebilir.
- TF-IDF hızlı, kolay ve oldukça etkilidir; baseline model olarak her zaman kullanılabilir.
- Daha gelişmiş anlamsal modeller (BERT vb.) ile karşılaştırma gelecekte yapılabilir.

Anlamsal Değerlendirme Tablosu

Aşağıdaki tabloda her modelin en benzer 5 çıktısı değerlendirilmiş ve **anlamsal benzerlik skorları** aşağıdaki ölçütlere göre verilmiştir:

Puanlama Tablosu

Model Adı	Puanlar	Ortalama
TF-IDF (lemmatized)	[4, 4, 5, 4, 4]	**4.2**
TF-IDF (stemmed)	[3, 4, 4, 3, 3]	**3.4**
Word2Vec (lemma, w=2, d=100)	[3, 2, 3, 3, 2]	**2.6**
Word2Vec (lemma, w=5, d=300)	[4, 3, 4, 4, 3]	**3.6**

Word2Vec (stemmed, w=5, d=300)	[3, 2, 3, 2, 2]	**2.4**

Hangi model(ler) daha yüksek ortalama aldı?

- **TF-IDF (lemmatized)** modeli, **4.2 ortalama** ile en yüksek anlam benzerliğini sağlamıştır.
- **Word2Vec (lemma, w=5, d=300)** ise Word2Vec modelleri arasında **en yüksek ortalamaya sahip model**dir (3.6).

En anlamlı sonuçları hangi model verdi?

- TF-IDF (lemmatized), hem teknik hem de anlamsal olarak en istikrarlı ve güçlü sonucu veren modeldir.

TF-IDF ile Word2Vec arasında fark var mı?

- Evet, belirgin bir fark var. TF-IDF modelleri (özellikle lemmatized), Word2Vec modellerine kıyasla daha doğrudan ve tutarlı sonuçlar üretmiştir.
- Word2Vec, bazı benzerlikleri yakalayabilse de anlamsal karışıklıklar daha sık gözlenmiştir.

Model yapılandırmalarının etkisi var mı?

- **Pencere boyutu (window)** ve **boyut sayısı (dimension)** arttıkça Word2Vec modellerinin puanı artmıştır.
- Örneğin: Word2Vec (lemma, w=5, d=300) → 3.6 puan alırken, w=2, d=100 sadece 2.6 puan almıştır.
- Ayrıca **lemmatized modeller**, stemmed olanlara göre daha başarılıdır.

Sonuç:

Anlamsal başarı açısından, TF-IDF açık ara en güçlü model olarak öne çıkmıştır. Word2Vec modelleri ise yapılandırma ayarları iyileştikçe daha iyi sonuçlar verebilmiştir, ancak hala TF-IDF seviyesine ulaşamamıştır.

Anlamsal & Sıralama Değerlendirmesinin Birlikte Yorumu

Model	Anlamsal Ortalama	TF-IDF Lemma'ya Jaccard	En Benzer Model
TF-IDF (lemmatized)	**4.2**	1.00	TF-IDF Stem (0.60)
TF-IDF (stemmed)	**3.4**	0.60	TF-IDF Lemma

Word2Vec (lemma, w=2, d=100)	2.6	0.40	W2V Lemma w5-300
Word2Vec (lemma, w=5, d=300)	**3.6**	0.20	W2V Stem w5-300
Word2Vec (stem, w=5, d=300)	2.4	0.20	W2V Lemma w5-300

Yorumlar: Anlamsal Başarı vs Tutarlılık

TF-IDF (lemmatized)

- **En yüksek anlamsal başarıya (4.2) sahip model.**
- Sıralama bakımından da **TF-IDF (stemmed)** ile %60 oranında örtüşüyor.
- Bu da demektir ki: **tutarlı ve anlamlı sonuçları bir arada sunan en iyi modeldir.**

Word2Vec (lemma, w=5, d=300)

- Anlamsal değerlendirmede **Word2Vec modelleri arasında en iyisi** (3.6).
- Sıralama bakımından, diğer Word2Vec modelleriyle %60 oranında örtüşüyor.
- Bu da Word2Vec'in yapılandırma (window & dim) arttıkça daha anlamlı hale geldiğini gösteriyor.

Word2Vec (lemma, w=2, d=100) ve stemli Word2Vec'ler

- Hem **anlamsal açıdan zayıf (2.4 – 2.6 arası)** hem de **sıralama olarak kararsız**.
- Özellikle düşük boyutlu ve küçük pencere boyutlu modeller daha dengesiz.

Model Yapılandırmalarının Sıralama Başarısına Etkisi

CBOW vs Skip-Gram

- Not: Bu çalışmada sadece **CBOW** kullanıldı (varsayılan yapı).
- Eğer **Skip-Gram** yapılsaydı, daha az sıklıkla geçen kelimeler daha güçlü bağlamlar üretirdi.
- CBOW genel ortalamaya bakarken, **Skip-Gram daha keskin bağlam** tahminleri yapar.
- Özellikle teknik metinlerde Skip-Gram daha anlamlı sonuçlar verebilir.

Pencere Genişliği (window)

- Geniş pencere (ör. w=5) → daha fazla bağlam, daha **genel ve stabil benzerlik**
- Dar pencere (ör. w=2) → daha spesifik ve **daha dağınık sonuçlar**
- Örneğin: Word2Vec lemma w=2 → düşük tutarlılık ve düşük anlam puanı (2.6)

Vektör Boyutu (dimensionality)

- **Düşük boyut (d=100)**: hızlı fakat yüzeysel temsil.
- **Yüksek boyut (d=300)**: daha derin bağlamlar yakalar → anlam puanı yükselir.
- `Word2Vec lemma w=5, d=300` modelinde bu açıkça görülüyor (**3.6 puan** + **0.6 Jaccard**)

Genel Sonuç:

Yorum	En Başarılı Model
Anlamsal Uyum	**TF-IDF (lemmatized)** (4.2)
Sıralama Tutarlılığı	**TF-IDF modelleri** (0.60 Jaccard)
Word2Vec'te Anlam & Tutarlılık	**Lemma + w=5 + d=300** yapılandırması