NLP Proje Raporu

Proje Açıklaması

Bu doğal dil işleme projesi, metin verilerini analiz ederek tf-idf vektörleme ve duygu analizi gibi çeşitli teknikleri içermektedir. Ayrıca, Word2Vec modeli ile kelime benzerliği ölçümü ve TfidfVectorizer ile metin benzerliği hesaplanması gibi önemli bileşenleri içermektedir.

Kullanılan Teknolojiler ve Kütüphaneler

Python: Projede kullanılan ana programlama dili.

Scikit-learn: TfidfVectorizer kullanımı için.

NLTK (Natural Language Toolkit): Metin verilerinin tokenleme işlemleri için.

Gensim: Word2Vec modeli oluşturmak ve kelime benzerliği ölçmek için.

Pandas: Veri manipülasyonu ve analizi için.

TextBlob: Duygu analizi için.

Adımlar ve Kullanılan Metodoloji

Veri Yükleme: 'onislenmis\_veriler.csv' adlı veri seti kullanıldı ve pandas kütüphanesi ile yüklendi.

Metin Verisi Ön İşleme: NLTK kütüphanesi kullanılarak cümleler tokenize edildi ve Word2Vec modeli eğitimi için hazır hale getirildi.

Word2Vec Modeli Eğitimi: Tokenleme yapılmış cümlelerle Word2Vec modeli eğitildi. En sık geçen kelimeler listelendi ve frekansları gösterildi.

Metin Benzerliği Hesaplama: TfidfVectorizer kullanılarak metinler vektörlere dönüştürüldü ve ardından cosine\_similarity ile metin benzerliği hesaplandı. Özellikle seçilen cümleler için en benzer cümleler belirlendi.

Duygu Analizi: TextBlob kullanılarak metinlerin duygu analizi yapıldı. Olumlu, olumsuz veya nötr duygu durumları belirlendi.

Rastgele Cümlelere Göre En Yakın Cümleler:

3731. Cümle:

"sevenin sevdigi sevmeyenin sevmedigi lider erdogana karsi notr hayatimda gormedim"

En Yakın Cümleler:

"ortada ugrasmasina karsi notr kisilerin ogrendigim sahis" (Satır Numarası: 13763)

"hayatimda dogru konusan insan evladi gormedim yuzun kizarmiyor kardesim" (Satır Numarası: 11591)

"kusaginin sevmedigi insan olabilir" (Satır Numarası: 12959)

2552. Cümle:

"hayatimin genclik doneminin kendisinin iktidarda zamana denk geldigi siyasetci"

En Yakın Cümleler:

"merak gece hayatimin guzel" (Satır Numarası: 4189)

"muhalefet sayesinde olene iktidarda kalmaya devam siyasetci" (Satır Numarası: 6109)

"goreve geldigi yilin ulke tarihinin buyuk krizi denk gelen noktasi yili baz alan gercegini kenara" (Satır Numarası: 12994)

10876. Cümle:

"yasadigimiz gunun sorumlusu kendisidir"

En Yakın Cümleler:

"ekiplerin hizli bicimde oraya ulasip yardima sorumlusu kendisidir" (Satır Numarası: 9829)

"bugun yasadigimiz kriz buhranin" (Satır Numarası: 1511)

"son gundur yasadigimiz masada iddia edebilirim" (Satır Numarası: 10415)

11196. Cümle:

"turk siyasal tarihinin sandik anlaminda basarili ismi kanitlayan cambaz siyasetci"

En Yakın Cümleler:

"cumhuriyet tarihinin basarili siyasetcisi" (Satır Numarası: 3738)

"halktan ustun yildir basbakan kanitlayan siyasetci" (Satır Numarası: 5685)

"tartismasiz turkiye cumhuriyeti tarihinin populer ismi" (Satır Numarası: 896)

5612. Cümle:

"cumhurbaskani dedigimiz kisi ilk tarafsiz olmalidir meydanlara inip miting resmen partisi oy topladi"

En Yakın Cümleler:

"miting miting gezip halktan milletvekili isteyen tarafsiz cumhurbaskani" (Satır Numarası: 11936)

"meydanlara halktan fazla destekci dusunecek gucu kisi" (Satır Numarası: 13419)

"kimbilir hamlesini inanilmaz merak ettigim tarafsiz cumhurbaskani" (Satır Numarası: 6297)

En Fazla Geçen 20 Kelime :

'kisi' kelimesi 449 kez geçiyor.

'yok' kelimesi 445 kez geçiyor.

'adam' kelimesi 438 kez geçiyor.

'buyuk' kelimesi 401 kez geçiyor.

'erdogan' kelimesi 388 kez geçiyor.

'cumhurbaskani' kelimesi 370 kez geçiyor.

'tayyip' kelimesi 342 kez geçiyor.

'turkiye' kelimesi 340 kez geçiyor.

'son' kelimesi 335 kez geçiyor.

'tek' kelimesi 307 kez geçiyor.

'basbakan' kelimesi 299 kez geçiyor.

'oy' kelimesi 295 kez geçiyor.

'iyi' kelimesi 289 kez geçiyor.

'gun' kelimesi 287 kez geçiyor.

'zaman' kelimesi 259 kez geçiyor.

'ilk' kelimesi 240 kez geçiyor.

'bugun' kelimesi 240 kez geçiyor.

'recep' kelimesi 227 kez geçiyor.

'ulkenin' kelimesi 226 kez geçiyor.

'secim' kelimesi 218 kez geçiyor.

En Sık Geçen 20 Farklı Kelimenin Her Biri İçin En Benzer 5 Farklı Kelime :

cok\_gecen\_kelimeler listesindeki her bir kelime için döngü oluşturulur.

Her kelime için, model.wv.most\_similar(word, topn=5) fonksiyonu çağrılır. Bu fonksiyon, belirli bir kelimenin en benzer kelimelerini bulur ve bu benzerlikleri bir liste içinde döndürür.

Bulunan benzer kelimeler, similar\_words değişkenine atanır.

Son olarak, her bir kelimenin en benzer 5 kelimesi ekrana yazdırılır.

'kisi' kelimesinin en benzer 5 kelimesi: [('bugun', 0.9989054799079895), ('son', 0.9988786578178406), ('iyi', 0.9988134503364563), ('zaman', 0.9988111257553101), ('oy', 0.9987993240356445)]

'yok' kelimesinin en benzer 5 kelimesi: [('gun', 0.9989348649978638), ('oy', 0.9987810254096985), ('kisi', 0.9987764954566956), ('adam', 0.998734712600708), ('zaman', 0.9986973404884338)]

'adam' kelimesinin en benzer 5 kelimesi: [('son', 0.9988461136817932), ('oy', 0.9987444281578064), ('yok', 0.9987347722053528), ('kisi', 0.9986304044723511), ('ulkenin', 0.9985170364379883)]

'buyuk' kelimesinin en benzer 5 kelimesi: [('yok', 0.9986286163330078), ('zaman', 0.9986258745193481), ('kisi', 0.9986160397529602), ('son', 0.9985594749450684), ('oy', 0.9985513687133789)]

'erdogan' kelimesinin en benzer 5 kelimesi: [('erdoganin', 0.9737420678138733), ('erdogani', 0.9694381356239319), ('tayyip', 0.9691774249076843), ('erdogana', 0.9632028341293335), ('adi', 0.9469299912452698)]

'cumhurbaskani' kelimesinin en benzer 5 kelimesi: [('iyi', 0.9973864555358887), ('basbakan', 0.9972413182258606), ('ulke', 0.997122049331665), ('zaman', 0.9971108436584473), ('son', 0.997011661529541)]

'tayyip' kelimesinin en benzer 5 kelimesi: [('erdogan', 0.9691774845123291), ('erdogani', 0.918836236000061), ('erdoganin', 0.9146152138710022), ('erdogana', 0.9116235375404358), ('recep', 0.9003580808639526)]

'turkiye' kelimesinin en benzer 5 kelimesi: [('buyuk', 0.994849681854248), ('secim', 0.9941955804824829), ('zaman', 0.9941364526748657), ('ilk', 0.9940006732940674), ('bugun', 0.9939324259757996)]

'son' kelimesinin en benzer 5 kelimesi: [('kisi', 0.9988785982131958), ('adam', 0.9988461136817932), ('secim', 0.9987162947654724), ('yok', 0.9986961483955383), ('zaman', 0.9986562132835388)]

'tek' kelimesinin en benzer 5 kelimesi: [('adam', 0.9984809160232544), ('ulkenin', 0.9984673261642456), ('kisi', 0.9984525442123413), ('son', 0.9982274174690247), ('yok', 0.9981828927993774)]

'basbakan' kelimesinin en benzer 5 kelimesi: [('iyi', 0.9985555410385132), ('kisi', 0.9984989166259766), ('zaman', 0.9982911348342896), ('oy', 0.9982412457466125), ('son', 0.9981902241706848)]

'oy' kelimesinin en benzer 5 kelimesi: [('kisi', 0.9987993836402893), ('yok', 0.9987810254096985), ('adam', 0.9987443685531616), ('son', 0.9985745549201965), ('sadece', 0.9985744953155518)]

'iyi' kelimesinin en benzer 5 kelimesi: [('zaman', 0.9988226294517517), ('kisi', 0.9988133311271667), ('son', 0.9986042976379395), ('sadece', 0.9985697269439697), ('basbakan', 0.9985554814338684)]

'gun' kelimesinin en benzer 5 kelimesi: [('yok', 0.9989349842071533), ('kisi', 0.9987051486968994), ('ulkenin', 0.9983925223350525), ('oy', 0.9983899593353271), ('son', 0.9983713030815125)]

'zaman' kelimesinin en benzer 5 kelimesi: [('iyi', 0.9988226294517517), ('kisi', 0.9988112449645996), ('yok', 0.9986972808837891), ('son', 0.9986560940742493), ('buyuk', 0.9986260533332825)]

'ilk' kelimesinin en benzer 5 kelimesi: [('yok', 0.9984721541404724), ('kisi', 0.9983099699020386), ('turk', 0.998254656791687), ('bugun', 0.9982295036315918), ('oy', 0.9982008934020996)]

'bugun' kelimesinin en benzer 5 kelimesi: [('kisi', 0.9989056587219238), ('zaman', 0.9985617995262146), ('yok', 0.9985485672950745), ('gun', 0.9983567595481873), ('son', 0.9983475208282471)]

'recep' kelimesinin en benzer 5 kelimesi: [('erdogan', 0.9234777688980103), ('tayyip', 0.9003579616546631), ('erdogani', 0.8435091972351074), ('bey', 0.8427353501319885), ('erdoganin', 0.8312958478927612)]

'ulkenin' kelimesinin en benzer 5 kelimesi: [('yok', 0.9985430836677551), ('adam', 0.9985169768333435), ('oy', 0.9984922409057617), ('tek', 0.9984674453735352), ('kisi', 0.99844890832901)]

'secim' kelimesinin en benzer 5 kelimesi: [('son', 0.9987163543701172), ('zaman', 0.9986194372177124), ('kisi', 0.9986140727996826), ('sadece', 0.9984315633773804), ('iyi', 0.9984227418899536)]

Sonuçlar

Word2Vec modeli, veri setindeki cümleler arasındaki kelime benzerliklerini başarılı bir şekilde ölçtü.

TfidfVectorizer ile metin benzerliği hesaplanarak seçilen cümleler için en yakın cümleler belirlendi.

Duygu analizi sonuçları, metinlerin duygusal durumlarını nötr, olumlu veya olumsuz olarak sınıflandırdı.

Berkehan Doğan

212805009