YZM509 Ara Sınav Projesi

Son teslim tarihi: 26.04.2023- 23.59

Aşağıda verilen 3 sorudan dilediğiniz bir tanesini cevaplayınız.

1. Türkçe dilinde doğal dil işleme üzerine son yıllarda yapılan çalışmaları araştırınız. Hangi araştırma grupları ve firmalar aktif olarak çalışmaktadır? Türkçe NLP çalışmalarında karşılaşılan zorluklar nelerdir? Açık kaynaklı olarak yayınlanan ve toplumla paylaşılan Türkçe veri ve kod kaynaklarını araştırarak raporlaştırınız. İlginizi çeken bir kaynaktan elde edeceğiniz kodları çalıştırarak sonuçları yorumlayınız.

2. İlgi alanınıza göre edineceğiniz bir corpus kullanarak kelime vektörlerini hesaplatınız. İlgilendiğiniz anahtar kelimelere benzer olan kelimeleri elde ederek bu kelimeler arasındaki ilişkiyi yorumlayınız. Skip-gram, CBOW, gensim ve tsne gibi yöntem ve fonksiyonları kullanabilirsiniz. Python kodlarınızı açıklayarak, sonuçları yorumlayınız.

3. Twint veya API kullanarak twitter üzerinde veri çekiniz. Hashtag, cashtag, kullanıcı adı, tarih aralığı, dil, konum gibi filtreler uygulayınız. Python kodlarınızı açıklayarak, sonuçları yorumlayınız.

NOT: Çözüm dosyanızı Numara-ad soyad.docx olarak gönderiniz. Kodları da gönderiniz.

2)Bu sorunun çözümünde gutenberg.org'dan elde edilen Charles Dickens'ın yazdığı "**A Tale of Two Cities**" ekitabı kullanıldı.Külliyattaki noktalama işaretleri silindi.Kodlar Google Collab üzerinde yazıldı.

```
[55] import numpy as np
     from gensim.models import Word2Vec
     from sklearn.manifold import TSNE
     import matplotlib.pyplot as plt
```

Dizilerle çalışabilmek için numpy kullanıldı.Daha sonra gensim modellerinden word2vec'i import edildi.TSNE ve plt görselleştirme yapmak istenirse kullanılır.

```
[119] from google.colab import drive
      drive.mount('/content/gdrive')
```

Külliyatı kullanabilmek için drive mount edildi.

```
[120] %cd /content/gdrive/MyDrive/Colab Notebooks/

/content/gdrive/MyDrive/Colab Notebooks
```

Külliyatın bulunduğu klasöre erişim sağlandı.

```
[123] !ls

'A TALE OF TWO CITIES.txt'                                  word2vec.ipynb
 NlpVize                                                    word2vec.model
'Topic models with Gensim.ipynb adlı not defterinin kopyası'
```

Klasörde bulunan dosyaların listesi kontrol edildi.

```
[126] f = open('A TALE OF TWO CITIES.txt', 'r', encoding='utf8')
      text = f.read()
      t_list = text.split('\n')

      corpus = []

      for cumle in t_list:
          corpus.append(cumle.split())
```

"open" komutunda,külliyat text'imizi 'r' komutu ile utf8 unicode ile okuduk.f değişkeni artık bu dosyayı tutacak.Bu dosya daha sonra text değişkenine yönlendirildi.Dosyadaki her cümlenin bir satırda olduğu bilindiğinden split ile her cümle t_list içerisine alındı.Bir corpus listesi oluşturuldu.t_list içerisindeki

her bir cümle boşluktan sonra split yapılıp kelimeler halinde append ile corpus listesine eklendi.

```
[204] print(corpus[:10])

[['chapter', 'i'], ['the', 'period'], ['it', 'was', 'the', 'best', 'of', 'times', 'it', 'was', 'the', 'worst', 'of', 'times', 'it', 'was', 'the',
```

İlk 10 kelimeyi kontrol etmek için print komutu kullanıldı.

```
[205] model = Word2Vec(corpus, vector_size=100, window=5, min_count=5, sg=1)
```

Corpus içerisinde uzunluğu 100 olan vektörler oluştururken kelimenin 5 adet sağındaki 5 adet solundaki kelimeleri dikkate alacağımızı söyledik.Frekansı en az 5 olan yani ender görülen kelimeleri dışarıda bırakacağımızı söyledik.Ve skip-gram kullandığımız için sg=1 ifadesini kullandık.

```
model.wv['french']
```

```
array([-0.17210887,  0.25673553, -0.11236692,  0.03830841,  0.03294627,
       -0.21152633,  0.1450864 ,  0.4057691 , -0.12727253, -0.09938605,
       -0.05605832, -0.09149766, -0.05144686, -0.01636974,  0.07454301,
       -0.19031188, -0.02905647, -0.02610502, -0.05732597, -0.23680063,
        0.03036786,  0.02805902,  0.09094775, -0.13360554, -0.12147025,
       -0.1190185 , -0.09016347,  0.02096185, -0.28899953,  0.17981817,
       -0.00145444, -0.08856203,  0.1398791 , -0.22624955, -0.06042317,
        0.11413305,  0.04867197,  0.02637405,  0.00821785, -0.16619436,
        0.04722928, -0.03680338, -0.33368796,  0.01073698,  0.20749518,
       -0.0740619 ,  0.02934433,  0.03434779,  0.14767382,  0.11811719,
        0.07422854, -0.13015783,  0.02464882,  0.00974805, -0.09404378,
        0.07601232,  0.01712469, -0.07403509, -0.24966373,  0.0524031 ,
       -0.06839131,  0.0516002 , -0.0302113 , -0.1184094 , -0.11253074,
        0.24045135,  0.06132016,  0.18395059, -0.30596942,  0.29882205,
       -0.00265056,  0.12071525,  0.19510385, -0.22537173,  0.153164  ,
        0.23686068, -0.0486274 ,  0.04192673, -0.05082124, -0.01330517,
       -0.02080114,  0.04902384, -0.1523593 ,  0.07607736, -0.02168647,
       -0.05982541,  0.03641941,  0.01248172,  0.12820244,  0.22546479,
        0.26744705,  0.07280552,  0.0331792 , -0.0541895 ,  0.26252276,
       -0.04684075,  0.10328938, -0.05211307,  0.1858421 , -0.15960328],
      dtype=float32)
```

Vektörlerin oluştuğunu görmek için corpusumuzda bulunan “french” tokeninin vektörleri yazdırıldı.

Benzer kelimeleri bulmak için anahtar kelime olarak “evil”,”queen”,”night”,”chair” ve “power” kelimeleri kullanıldı.

```
[207] model.wv.most_similar('evil')
```

```
[('law', 0.995233416557312),
 ('crime', 0.9950929880142212),
 ('hazard', 0.9948596954345703),
 ('patients', 0.9946296811103821),
 ('creatures', 0.9945647120475769),
 ('emigrants', 0.9944233894348145),
 ('uncertain', 0.9944191575050354),
 ('gives', 0.994235634803772),
 ('difference', 0.994171679019928),
 ('outward', 0.9939135909080505)]
```

Law,crime,hazard,creatures gibi alakalı kelimeler olsa da uncertain,gives,difference gibi alakasız kelimelerde var.

```
[208] model.wv.most_similar('queen')
```

```
[('flies', 0.9951318502426147),
 ('spectacle', 0.9949846863746643),
 ('leaf', 0.9948805570602417),
 ('ghosts', 0.9943143725395203),
 ('growing', 0.9939749836921692),
 ('visible', 0.993800163269043),
 ('surrounded', 0.9937908053398132),
 ('rapidly', 0.9936684966087341),
 ('formed', 0.9936357140541077),
 ('feeble', 0.9935444593429565)]
```

Benzer kelime yok.En yakın alakalı kelime muhtemelen devrimciler tarafından kraliçenin etrafının sarılmasından gelen surrounded kelimesidir.

```
model.wv.most_similar('night')
```

```
[('last', 0.9363060593605042),
 ('shadows', 0.9220950603485107),
 ('soldiers', 0.9181480407714844),
 ('house', 0.918071985244751),
 ('prisoners', 0.9095385670661926),
 ('days', 0.9086292386054993),
 ('day', 0.9085057973861694),
 ('summer', 0.9059508442878723),
 ('trees', 0.9032012224197388),
 ('footsteps', 0.9024361371994019)]
```

Shadows,days,day,summer gibi benzer ve alakalı kelimeler mevcut.

```
model.wv.most_similar('chair')
```

```
[('shoulder', 0.9752479195594788),
 ('softly', 0.973059892654419),
 ('shook', 0.9722753167152405),
 ('seat', 0.9686557650566101),
 ('beside', 0.9680477976799011),
 ('table', 0.9650373458862305),
 ('walked', 0.9640761613845825),
 ('leaned', 0.9622665047645569),
 ('arms', 0.9577630162239075),
 ('towards', 0.957377016544342)]
```

Seat,table gibi benzer kelimeler mevcut.

```
model.wv.most_similar('power')
```

```
[('reasons', 0.9914357662200928),
 ('remembrance', 0.9888221025466919),
 ('chance', 0.9879240989685059),
 ('means', 0.9867265820503235),
 ('innocent', 0.9851906299591064),
 ('probably', 0.9846175312995911),
 ('secret', 0.9830580353736877),
 ('course', 0.9829707741737366),
 ('grown', 0.9827238321304321),
 ('feelings', 0.9826700687408447)]
```

Benzer kelime yok ama külliyattaki hikayeye göre alakalı kelimeler mevcut.

Bazı kelimelerde diğer kelimelere göre daha çok başarı var.Ama istenen keskinlikte bir öğrenme mevcut değil.Belki de böyle soyutluk ve mecaz içerebilen külliyatlarda bu öğrenme şekli yetersizdir.