NLP Home work 5

مهدی فقهی

401722136

سوال اول)

Problem 1 Mention two common approaches to sequence modeling and discuss their advantages and disadvantages. (10 points)

Two common approaches to sequence modeling are:

a generative approach, HMM tagging,

and

a discriminative approach, MEMM tagging.

می‌دانیم Sequence modeling وظیفه آن اختصاص یک برچسب یا کلاس به هر واحد در یک دنباله است.

بنابراین دنباله ای از مشاهدات را به دنباله ای از برچسب ها نگاشت می کند.

HMM یک Sequence modeling احتمالی که با توجه به دنباله ای از واحدها (کلمات، حروف، تکواژها،جملات، هر چه باشد)،یک توزیع احتمال را بر روی توالی ممکن از برچسب ها محاسبه می کند و بهترین دنباله برچسب را انتخاب می کند.

یک زنجیره مارکوف زمانی مفید است که ما نیاز به محاسبه احتمال برای یک دنباله از رویدادهای قابل مشاهده داریم با این حال، در بسیاری از موارد، رویدادهایی که ما به آنها علاقه مند هستیم، پنهان هستند: ما آنها را مستقیماً مشاهده نمی کنیم.

یک مدل hidden Markov model (HMM) به ما امکان می دهد در مورد هر دو رویداد مشاهده شده صحبت کنیم.

در حالی که یک HMM می تواند به دقت بسیار بالایی دست یابد، به تعدادی نوآوری نیاز دارد تا مقابله کند با کلمات ناشناخته، پسوندها، پسوندها و غیره.

اگر بتوانیم ویژگی های دلخواه را مستقیماً و به روشی تمیز به مدل اضافه کنیم، بسیار ساده تر خواهد بود .

اما این برای مدل های Generative مانند HMM سخت است.

ما می توانیم رگرسیون لجستیک را به یک مدل توالی متمایز تبدیل کنیم

به سادگی با اجرای آن بر روی کلمات متوالی، با استفاده از کلاس اختصاص داده شده به کلمه قبلی به عنوان یک ویژگی در طبقه بندی کلمه بعدی ، وقتی رگرسیون لجستیک را به این شکل اعمال میکنیم به آن **MEMM** maximum entropy Markov model گفته می‌شود .

The HMM computes the **likelihood** of the observation given the hidden state, while the MEMM computes the **posterior** of each state

دلیل استفاده از یک مدل توالی Discriminative این است که ترکیب بسیاری از ویژگی ها آسان تر است.

یک مشکل مدل های MEMM و HMM این است که آنها به طور انحصاری از چپ به راست اجرا می‌شود.

Problem 2

In this problem, you should use the following sentences to implement POS tagging using HMM method on this sentence:

"Can Will hunt Mark?"

(a) First, apply POS tagging on the given sentences and create a table that shows the probability that each word is a noun, modal, or verb. You can use the first row of the following table as an example: (5 points)

Will mark hunt.

Mark will hunt.

Can Will mark?

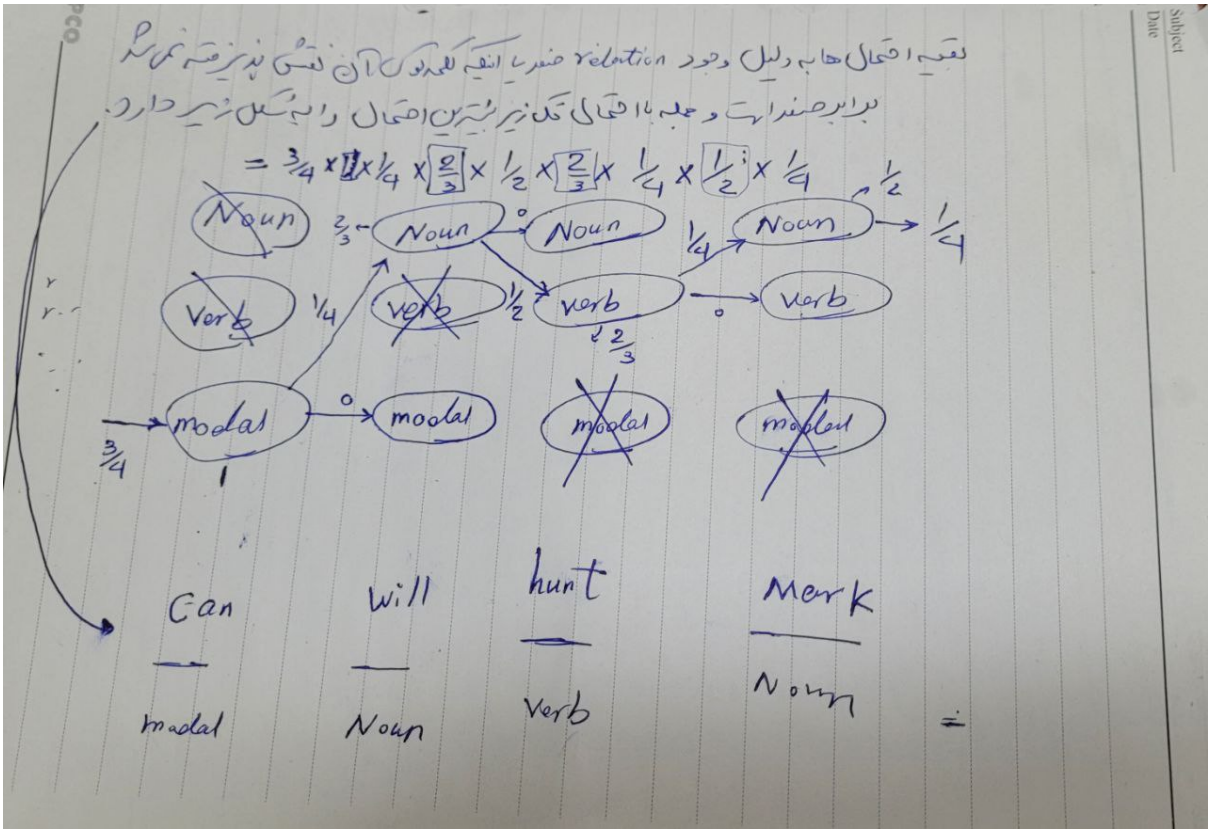
Mark can hunt.

	Noun	Verb	Modal	Total Count
will	2	0	1	3
mark	2	2	0	4
hunt	1	2	0	3
can	0	0	2	2

(b) Second, calculate the probability of two labels occurring together. You can use following table structure: (5 points)

	<S>	Noun	Verb	Modal	<E>
<S>	0	3/4	0	1/4	0
Noun	0	0	1/2	1/2	1/4
Verb	0	1/4	0	0	3/4
Modal	0	1/4	3/4	0	0
<E>	0	0	0	0	0

Finally, use the results of previous sections to perform POS tagging on " Can Will hunt Mark?<E>" . You need to draw a graph of the sentence, then delete the edges that have zero probability to get the answer. (10 points)


Noun
Verb
modal
Noun
Verb
modal
Noun
verb
modal
Noun
verb
modal
Noun
verb
modal
Can
Will
hunt
Mark
modal
Noun
Verb
Noun

Problem 3:

نکته قسمت نشان دادن ۱۰ جمله شبیه و ۱۰ جمله دور از هم اول که در پایان قسمت B , C, D خواسته شده بود را به عنوان معیاری برای مقایسه در بخش آخر یعنی e آورده ام .

(a)

Set up your environment and visualize your dataset in a simple way. (10 points)

```
# Imports
from datasets import load_dataset
import pandas as pd
import numpy as np
from tqdm import tqdm
tqdm.pandas()



# Load the English STSB dataset
stsb_dataset = load_dataset('stsb_multi_mt', 'en')
stsb_train = pd.DataFrame(stsb_dataset['train'])
stsb_test = pd.DataFrame(stsb_dataset['test'])

# Check loaded data
print(stsb_train.shape, stsb_test.shape)
stsb_test.head()
```

به کمک کتابخانه load dataset دیتاست STSB که یک دیتاست مربوط به جفت جمله‌ای‌هایی هست که نظر مفهوم به همدیگر شبیه هستند را دانلود می کنیم.

دو قسمت train و test آن را از همدیگر جدا می کنیم .

مقدار shape هریک را بدست می‌آوریم .

و در نهایت برای دیدن داده‌ها ۵ تا اول آن را مشاهده می کنیم .

```
(5749, 3) (1379, 3)
```

	sentence1	sentence2	similarity_score
0	A girl is styling her hair.	A girl is brushing her hair.	2.5
1	A group of men play soccer on the beach.	A group of boys are playing soccer on the beach.	3.6
2	One woman is measuring another woman's ankle.	A woman measures another woman's ankle.	5.0
3	A man is cutting up a cucumber.	A man is slicing a cucumber.	4.2
4	A man is playing a harp.	A man is playing a keyboard.	1.5

در قدم بعد به کمک فانکشن text processing که نمونه‌های مشابه آن برای داده انگلیسی زیاد است ابتد غلط های املایی متن را به کمک textBlob درست می کنیم و سپس به کمک nlp جملات را توکن می کنیم و هر توکن را lemmatization و lowercase و در صورت stopword بودن از لیست کلمات حذف می کنیم و در لیست قرار می‌دهیم .

```python
from sklearn.metrics.pairwise import cosine_similarity
from textblob import TextBlob
import spacy
nlp = spacy.load("en_core_web_sm")

def text_processing(sentence):
    """
    Lemmatize, lowercase, remove numbers and stop words

    Args:
      sentence: The sentence we want to process.

    Returns:
      A list of processed words
    """
    sentence = ''.join(TextBlob(sentence).correct())
    # print(sentence)
    sentence = [token.lemma_.lower()
                for token in nlp(sentence)
                if token.is_alpha and not token.is_stop]


    return sentence
```

```
stsb_train = pd.read_csv('stsb_train.csv',index_col = [0])
stsb_train['sentence1'] = stsb_train['sentence1'].apply(lambda cw : text_processing(cw))
stsb_train['sentence2'] = stsb_train['sentence2'].apply(lambda cw : text_processing(cw))
stsb_train.head()
```

	sentence1	sentence2	similarity_score
0	[plane, take]	[air, plane, take]	5.00
1	[man, play, large, flute]	[man, play, flute]	3.80
2	[man, spread, screwed, cheese, penza]	[man, spread, shrouded, cheese, unhooked, penza]	3.80
3	[man, play, chess]	[man, play, chess]	2.60
4	[man, play, cell]	[man, seat, play, cell]	4.25

```
print(stsb_train.shape)
```

و داده‌های ما به شکل بالا خواهد شد .

(b)

Use the number of common words between sentences as a measure of similarity. Compare your results with the dataset using the Spearman rank correlation coefficient. Visualize the top 10 most similar and the least similar pair sentences that your model found using a table. (20 points)

```
def find_number_of_common_word(sentence_one_list,sentence_two_list):

        return len(set(eval(sentence_one_list)) & set(eval(sentence_two_list)))
```

ابتدا لیست کلمات را درون set قرار می‌دهیم تا از هر کلمه فقط یکی را داشته باشیم و سپس دو مجموعه کلمه را با همدیگر and میکنیم تا فقط کلماتی که شبیه هم هستند باقی بماند و در نهایت اندازه سایز را برمی گردانیم به این شکل تعداد کلماتی که مشابه هم هستند در دو جمله بدست می‌آید .

```
stsb_train['common_word']  = stsb_train.apply(lambda row : find_number_of_common_word(row[0],row[1]),axis=1)
stsb_train.head(10)
```

	sentence1	sentence2	similarity_score	common_word
0	['plane', 'take']	['air', 'plane', 'take']	5.00	2
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	3
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	4
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	3
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	3
5	['man', 'fight']	['man', 'fight']	4.25	2
6	['man', 'smoke']	['man', 'state']	0.50	1
7	['man', 'play', 'piano']	['man', 'play', 'guitar']	1.60	2
8	['man', 'play', 'guitar', 'singing']	['woman', 'play', 'caustic', 'guitar', 'singing']	2.20	3
9	['person', 'throw', 'cat', 'ceiling']	['person', 'throw', 'cat', 'ceiling']	5.00	4

سپس براساس میزان شباهت برحسب تعداد کلمات یکسان برای محاسبه spearman rank که در آینده می‌خواهیم بدست آوریم rank می‌نماییم.

```
def ranking(ser_of_data):
    array = np.array(ser_of_data)
    temp = array.argsort()
    ranks = np.empty_like(temp)
    ranks[temp] = np.arange(len(array))
    ranks = 1 + ranks
    # print(sorted(ranks))
    return ranks

stsb_train['rank_by_common_word'] = ranking(stsb_train['common_word'])
stsb_train.head(10)
```

	sentence1	sentence2	similarity_score	common_word	rank_by_common_word
0	['plane', 'take']	['air', 'plane', 'take']	5.00	2	2042
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	3	3478
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	4	3891
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	3	3471
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	3	3469
5	['man', 'fight']	['man', 'fight']	4.25	2	2481
6	['man', 'smoke']	['man', 'state']	0.50	1	738
7	['man', 'play', 'piano']	['man', 'play', 'guitar']	1.60	2	2468
8	['man', 'play', 'guitar', 'singing']	['woman', 'play', 'caustic', 'guitar', 'singing']	2.20	3	3460
9	['person', 'throw', 'cat', 'ceiling']	['person', 'throw', 'cat', 'ceiling']	5.00	4	3905

همین طور کلمات را براساس میزان شباهت اصلی نیز rank می کنیم .

```
stsb_train['rank_by_similarity_score'] = ranking(stsb_train['similarity_score'])
stsb_train.head(10)
```

	sentence1	sentence2	similarity_score	common_word	rank_by_common_word	rank_by_similarity_score
0	['plane', 'take']	['air', 'plane', 'take']	5.00	2	2042	5749
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	3	3478	4143
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	4	3891	4145
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	3	3471	2508
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	3	3469	4917
5	['man', 'fight']	['man', 'fight']	4.25	2	2481	4916
6	['man', 'smoke']	['man', 'state']	0.50	1	738	604
7	['man', 'play', 'piano']	['man', 'play', 'guitar']	1.60	2	2468	1489
8	['man', 'play', 'guitar', 'singing']	['woman', 'play', 'caustic', 'guitar', 'singing']	2.20	3	3460	2019
9	['person', 'throw', 'cat', 'ceiling']	['person', 'throw', 'cat', 'ceiling']	5.00	4	3905	5504

سپس برای محاسبه spearman مقدار d که براساس حاصل تفاضل ranking اصلی از ranking که براساس شباهت جدیدی که فکر می کردیم درسته بدست می‌آوریم و یک ستون دیگر حاصل d به توان دو هست نیز اضافه می کنیم .

```
stsb_train['d'] = stsb_train.apply(lambda row : row[4] - row[5] ,axis=1)
stsb_train['d^2'] = stsb_train.apply(lambda row : row[6]**2,axis=1)
stsb_train.head(10)
```

	sentence1	sentence2	similarity_score	common_word	rank_by_common_word	rank_by_similarity_score	d	d^2
0	['plane', 'take']	['air', 'plane', 'take']	5.00	2	2042	5749	-3707	13741849
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	3	3478	4143	-665	442225
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	4	3891	4145	-254	64516
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	3	3471	2508	963	927369
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	3	3469	4917	-1448	2096704
5	['man', 'fight']	['man', 'fight']	4.25	2	2481	4916	-2435	5929225
6	['man', 'smoke']	['man', 'state']	0.50	1	738	604	134	17956
7	['man', 'play', 'piano']	['man', 'play', 'guitar']	1.60	2	2468	1489	979	958441
8	['man', 'play', 'guitar', 'singing']	['woman', 'play', 'caustic', 'guitar', 'singing']	2.20	3	3460	2019	1441	2076481
9	['person', 'throw', 'cat', 'ceiling']	['person', 'throw', 'cat', 'ceiling']	5.00	4	3905	5504	-1599	2556801

سپس r که همان spearman rank هست را از فرمول زیر بدست می‌آوریم .

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2-1)}$$

```
r = 1 - (6*stsb_train['d^2'].sum())/(pow(len(stsb_train['d^2']),3) - len(stsb_train['d^2']))
print(r)
```

```
0.5773636963561097
```

در اینجا r برابر با 0.577 بدست آمده است .

(c)

Use the TF-IDF vectorizer to embed the dataset's sentences. Use cosine to measure the similarities of sentences' embeddings and compare your results with the dataset's score (Use Spearman rank correlation coefficient).
Visualize the top 10 most similar and the least similar pair sentences that your model found using a table. (20 points)

ابتدا یک لیست از تمامی جملاتی که داریم که حاصل concat کردن جملات ستون اول و جملات ستون دوم هست را می‌سازیم و سپس به کمک TfidfVectorizer از روی آن یک مدل بدست می‌آوریم.

```
from sklearn.feature_extraction.text import TfidfVectorizer
model = TfidfVectorizer(lowercase=True, stop_words='english')
# Train the model
X_train = pd.concat([new_stsb_train['sentence1'], new_stsb_train['sentence2']]).unique()
model.fit(X_train)
```

```
TfidfVectorizer(stop_words='english')
```

سپس به کمک ستون جمله‌های اول دوم را embedding می کنیم .

```
# Generate Embeddings on train
sentence1_emb = model.transform(new_stsb_train['sentence1'])
sentence2_emb = model.transform(new_stsb_train['sentence2'])
```

سپس از تابع cos_sim برای پیدا کردن میزان شباهت بین این vectorهای embed شده استفاده می کنیم و یک ستون به عنوان میزان شباهت براساس cos_sim براساس TfidfVectorizer به جدول اضافه می کنیم .

```
def cos_sim(sentence1_emb, sentence2_emb):
    """
    Cosine similarity between two columns of sentence embeddings

    Args:
      sentence1_emb: sentence1 embedding column
      sentence2_emb: sentence2 embedding column

    Returns:
      The row-wise cosine similarity between the two columns.
      For instance is sentence1_emb=[a,b,c] and sentence2_emb=[x,y,z]
      Then the result is [cosine_similarity(a,x), cosine_similarity(b,y), cosine_similarity(c,z)]
    """
    cos_sim = cosine_similarity(sentence1_emb, sentence2_emb)
    return np.diag(cos_sim)
```

```
# Cosine Similarity
new_stsb_train['TFIDF_cosine_score'] = cos_sim(sentence1_emb, sentence2_emb)

new_stsb_train.head()
```

	sentence1	sentence2	similarity_score	TFIDF_cosine_score
0	['plane', 'take']	['air', 'plane', 'take']	5.00	0.721472
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	0.869867
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	0.596340
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	1.000000
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	0.801722

سپس مانند قسمت قبل d , d به توان ۲ و در نهایت spearman rank را براساس آن بدست می‌آورم .

	sentence1	sentence2	similarity_score	TFIDF_cosine_score	rank_by_TFIDF_cosine_score	rank_by_similarity_score
0	['plane', 'take']	['air', 'plane', 'take']	5.00	0.721472	4305	5749
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	0.869867	5268	4143
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	0.596340	3384	4145
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	1.000000	5585	2508
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	0.801722	4876	4917
5	['man', 'fight']	['man', 'fight']	4.25	1.000000	5597	4916
6	['man', 'smoke']	['man', 'state']	0.50	0.236014	1108	604
7	['man', 'play', 'piano']	['man', 'play', 'guitar']	1.60	0.411897	2081	1489
8	['man', 'play', 'guitar', 'singing']	['woman', 'play', 'caustic', 'guitar', 'singing']	2.20	0.730670	4374	2019
9	['person', 'throw', 'cat', 'ceiling']	['person', 'throw', 'cat', 'ceiling']	5.00	1.000000	5604	5504

```
new_stsb_train['d'] = new_stsb_train.apply(lambda row : row[4] - row[5] ,axis=1)
new_stsb_train['d^2'] = new_stsb_train.apply(lambda row : row[6]**2,axis=1)
new_stsb_train.head(10)
```

	sentence1	sentence2	similarity_score	TFIDF_cosine_score	rank_by_TFIDF_cosine_score	rank_by_similarity_score	d	d^2
0	['plane', 'take']	['air', 'plane', 'take']	5.00	0.721472	4305	5749	-1444	2085136
1	['man', 'play', 'large', 'flute']	['man', 'play', 'flute']	3.80	0.869867	5268	4143	1125	1265625
2	['man', 'spread', 'screwed', 'cheese', 'penza']	['man', 'spread', 'shrouded', 'cheese', 'unhoo...	3.80	0.596340	3384	4145	-761	579121
3	['man', 'play', 'chess']	['man', 'play', 'chess']	2.60	1.000000	5585	2508	3077	9467929
4	['man', 'play', 'cell']	['man', 'seat', 'play', 'cell']	4.25	0.801722	4876	4917	-41	1681
5	['man', 'fight']	['man', 'fight']	4.25	1.000000	5597	4916	681	463761
6	['man', 'smoke']	['man', 'state']	0.50	0.236014	1108	604	504	254016
7	['man', 'play', 'piano']	['man', 'play', 'guitar']	1.60	0.411897	2081	1489	592	350464
8	['man', 'play', 'guitar', 'singing']	['woman', 'play', 'caustic', 'guitar', 'singing']	2.20	0.730670	4374	2019	2355	5546025
9	['person', 'throw', 'cat', 'ceiling']	['person', 'throw', 'cat', 'ceiling']	5.00	1.000000	5604	5504	100	10000

```
r = 1 - (6*new_stsb_train['d^2'].sum())/(pow(len(new_stsb_train['d^2']),3) - len(new_stsb_train['d^2']))
print(r)
```

```
0.6521460459919839
```

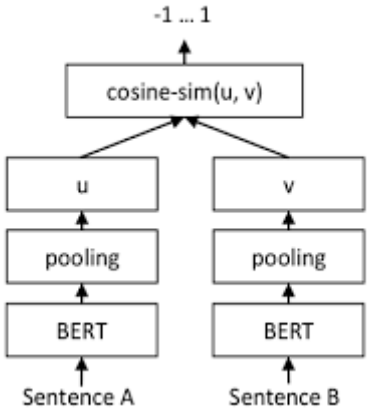
و همانطور که می‌بیند در اینجا میزان spear rank به میزان 0.65 رسیده است .

(d)

(Bonus section)

Use a deep learning model to embed the sentences and compare your results with the dataset (Use Spearman rank correlation coefficient).

Visualize the top 10 most similar and the least similar pair sentences that your model found using a table. (10 bonus points)



برای این قسمت از مدل Bert که بروی STSB یادگیری شده است استفاده می کنیم .

ماژول text2vec-transformers به شما امکان می دهد از یک مدل ترانسفورماتور زبان از پیش آموزش دیده به عنوان یک ماژول برداری Weaviate استفاده کنید. مدل‌های ترانسفورماتور با Contextionary متفاوت هستند، زیرا به شما امکان می‌دهند یک ماژول NLP از پیش آموزش‌دیده مخصوص مورد استفاده خود را وصل کنید. این بدان معناست

که مدل هایی مانند DilstilROBERTa، RoBERTa، DilstBERT، BERT، و غیره را می توان خارج از جعبه با Weaviate استفاده کرد.

از tokenizer خود این مدل استفاده نکردم برای مقایسه صرفا بر روی نوع embedding که انجام می‌دهد پس مدل pre train شده صرفا جهت embedding مانند Tf idF استفاده کردم و سپس مانند قبل cos similarity را بین دو ستون embed شده را بدست آوردم به ازای هر جفت .

```
from sentence_transformers import SentenceTransformer

# Load the pre-trained model
model = SentenceTransformer('sentence-transformers/stsb-mpnet-base-v2')

# Generate Embeddings
sentence1_emb = model.encode(new_stsb_train['sentence1'], show_progress_bar=True)
sentence2_emb = model.encode(new_stsb_train['sentence2'], show_progress_bar=True)

# Cosine Similarity
new_stsb_train['SBERT_BiEncoder_cosine_score'] = cos_sim(sentence1_emb, sentence2_emb)
```

و مانند گذشته رنک کردم و d , d ^2 را بدست آوردم به ازای معیار شباهت جدید .

	sentence1	sentence2	similarity_score	TFIDF_cosine_score	rank_by_TFIDF_cosine_score	rank_by_similarity_score	d	d^2
0	plane take	air plane take	5.00	0.721472	4305	5749	-1444	2085136
1	man play large flute	man play flute	3.80	0.869867	5268	4143	1125	1265625
2	man spread screwed cheese penza	man spread shrouded cheese unhooked penza	3.80	0.596340	3384	4145	-761	579121
3	man play chess	man play chess	2.60	1.000000	5585	2508	3077	9467929
4	man play cell	man seat play cell	4.25	0.801722	4876	4917	-41	1681
5	man fight	man fight	4.25	1.000000	5597	4916	681	463761
6	man smoke	man state	0.50	0.236014	1108	604	504	254016
7	man play piano	man play guitar	1.60	0.411897	2081	1489	592	350464
8	man play guitar singing	woman play caustic guitar singing	2.20	0.730670	4374	2019	2355	5546025
9	person throw cat ceiling	person throw cat ceiling	5.00	1.000000	5604	5504	100	10000

```
[27] r = 1 - (6*new_stsb_train_brt['d^2'].sum())/(pow(len(new_stsb_train_brt['d^2']),3) - len(new_stsb_train_brt['d^2']))
     print(r)

     0.8005006328318197
```

و درنهایت میزان r را بدست آوردم که برابر با 0.80 در سوال بود .

(e)

Compare your results from different methods and explain the advantages and disadvantages of each method. (20 points)

ابتدا بیاییم ببینم معیار **Spearman rank correlation coefficient** نشان دهنده چیست :

می‌توان نشان داد که ضریب همبستگی رتبه‌ای اسپیرمن نیز مانند ضریب همبستگی خطی پیرسون، مقداری بین ۱ و ۱- دارد. به این ترتیب مقدار ۱ نشان دهنده انطباق کامل بین رتبه‌ها و مقدار ۱- نیز انطباق معکوس بین رتبه‌ها را نشان می‌دهد.

یعنی هرچقدر به یک مقدار بدست آمده نزدیک‌تر باشد نشان می‌دهد همانندی بیشتری بین مدل ما و expert برای نزدیکی جملات از لحاظ مفهوم وجود دارد .

با این حساب پس مدل با pre training بهترین شباهت سپس مدل با TfidfVectorizer و در انتها مدل که تعداد کلمات بکار رفته شبیه بهم در دو جمله را معیار شباهت قرار می‌دهد در انتها قرار می گیرد .

اما بگذاریم قسمت ranking هر یک را نیز مشاهده کنیم قسمت ranking از ۱۰ تا بالا و ۱۰ تا پایین .. ابتدا ۱۰ تای اول مدل اول را نگاه می کنیم .

index	sentence1	sentence2	similarity_score
3117	Still, the "somewhat ambiguous ruling" might be a setback for Static Control depending on how it developed its competing product, Merrill Lynch analyst Steven Milunovich said.	But Merrill Lynch analyst Steven Milunovich said the "somewhat ambiguous ruling" by regulators might be a setback for Static Control depending on how it developed its competing product.	5.0
2473	Tom Kraynak, manager of operations and resources for the Canton, Ohio-based East Central Area Reliability Council, said that scenario is one among many that investigators are considering.	Tom Kraynak, manager of operations and resources for the Canton, Ohio-based East Central Area Reliability Council, said investigators are considering the scenario.	4.25
3599	hong kong universities collaborate with universities, businesses and government sectors of mainland china to coordinate training programs and research centers to promote high-tech research, commercialization, and technology transfer.	hong kong universities have collaborated with tertiary education, business and government sectors of mainland china to direct training programs and research centers to promote research commercialization and technology transfer.	5.0
3586	saferworld team leader on transfer controls and small arms roy isbister stated-- the eu embargo prohibits direct or indirect supply of military equipment for use in myanmar.	saferworld team leader on transfer controls and small arms roy isbister stated that -- the eu embargo explicitly states that no military equipment should be supplied either directly or indirectly for use in myanmar.	5.0
3737	the resolution requires all 192 united nations member states to adopt laws to prevent terrorists, black marketeers and other non-state actors from manufacturing, acquiring or trafficking in nuclear, biological or chemical weapons or the materials to make them.	resolution 1540 requires all countries to adopt laws to prevent non-state actors from manufacturing, acquiring or trafficking in nuclear, biological or chemical weapons, the materials to make them, and the missiles and other systems to deliver them.	3.5999999046325684
3641	brazil's strategic affairs minister roberto mangabeira unger stated brazil's new national defense plan calls for establishing partnerships with countries including russia and france to build a state-of-the-art weapons industry.	unger stated brazil's new national defense plan calls for establishing partnerships with countries including russia and france to build a state-of-the-art weapons industry.	3.799999952316284
3467	switzerland's trade and diplomatic relations with iran have been criticized in recent months after foreign minister micheline calmy-rey traveled to tehran in march 2008 to sign a deal with iran's state gas firm.	switzerland's trade and diplomatic relations with the islamic republic have been criticized in recent months after foreign minister micheline calmy-rey traveled to tehran in march 2008 to sign a gas deal.	3.799999952316284
3703	iraq has been lobbying for the security council to stop using the country's oil revenue to pay compensation to victims of the 1991 gulf war and the salaries of the united nations monitoring, verification and inspection commission inspectors and to have all money remaining in the united nation's oil-for-food accounts transferred to the government's development fund.	iraq's new leaders have been lobbying for the united nations security council to stop using the iraq's oil revenue to pay the salaries of the inspectors and to have all money remaining in the united nation's oil-for-food account transferred to the iraqi government.	4.0
3643	variants of the advanced light helicopter (alh) contain rocket launchers from belgium; rockets, guns and engines from france; brake systems from italy; fuel tanks and gearboxes from britain; self-protection equipment from a swedish company and have received crucial design development and engine control manufacturing from german companies.	the alh's arsenal includes-- rocket launchers from belgium rockets, guns and engines from france brake systems from italy fuel tanks and gearboxes from britain self-protection equipment from a swedish company.	3.5999999046325684
3605	russian president vladimir putin states that the decision to cease implementation of the conventional forces in europe treaty is a response to u.s. plans to establish missile defense sites in eastern europe and to nato's failure to ratify an amended version of the treaty.	russian president vladimir putin states that the decision is a response to u.s. plans to establish missile defense sites in eastern europe and to nato's failure to ratify an amended version of the treaty.	4.199999809265137

خوب همانطور که انتظار هم داشتیم ۱۰ جمله اول از جملاتی که شبیه بهم تشخص داده از جملاتی بوده‌اند که بیشترین تعداد کلمات را داشته‌اند یعنی در جملات طولانی ، بسیار خوب عملکرده است و توانسته است به خوبی این جملات را نزدیک به شباهتی که expert مدنظر داشته است پیش بینی کند .

البته در بین جملات طولانی جملاتی را هم به اشتباه نزدیک تشخص داده است هرچند که میزان شباهت آنان در وسط هست اما اینکه بتوانند در بالاترین سطح قرار بگیرند درست نیست که علت آن هم می‌تواند این باشد که این مدل صرف نظر از اینکه جملات چگونه کنار هم بکار رفته اند و صرف تعداد قرار گرفتن آنان نسبت به همدیگر می‌سنجد به خاطر همین جملاتی که طول کوتاهی دارند ولی بسیار شبیه هم هستند در بالا قرار نمی‌گیرند چون تعداد کلمات مشابه آنان نسبت بهم بسیار کمتر از جملات با شباهت کمتر ولی طول بیشتر و احتمال رخداد کلمات شبیه هم بیشتر در آن هست می‌شود . **برای حل این مشکل می‌شود تعداد بدست آمده را تقسیم بر تعداد کلمات موجود هر دو مجموعه جمله کرد .** که یک نوع میانگین است و شاید بتواند ranking را بهتر کند .اما مشکل استقلال کلمات هنوز وجود دارد .

1 to 10 of 10 entries Filter

index	sentence1	sentence2	similarity_score
5748	Putin spokesman: Doping charges appear unfounded	The Latest on Severe Weather: 1 Dead in Texas After Tornado	0.0
800	A gun is being fired.	A potato is being peeled.	0.5
4791	Senate to vote on moving ahead on Hagel nod	Benedict comes home to new house and new Pope	0.0
813	The two dogs were in the pen.	The cowboy rode his horse in the desert.	0.0
4784	Egypt protesters 'to be dispersed'	Abduction teacher to be sentenced	0.0
817	A man is kicking oots of water.	A woman is slicing a red pepper.	0.25
818	A man is eating.	A woman is rock climbing.	0.0
821	The woman is peeling lemon.	The man is balancing on the wire.	0.4000000059604645
826	Someone is cutting a circle out of a pink sheet of paper.	Someone is pouring tomato sauce into a pot of meat.	0.20000000298023224
829	A girl is eating a cupcake.	A man is playing a trumpet.	0.0

از بررسی ۱۰ جمله آخر که قاعدتا شبیه بهم نیستن اطلاعات زیادی بدست نمی‌آوریم زیرا مدل توی این قبیل موارد خوب عمل کرده است .

سراغ مدل بعدی که با TF_IDF embed کردیم و با cos_sim میزان شباهت را بررسی کردیم می‌رویم .

1 to 10 of 10 entries Filter

index	sentence1	sentence2	similarity_score
195	Two dogs swim in a pool.	Dogs are swimming in a pool.	4.199999809265137
77	A monkey pushes another monkey.	The monkey pushed the other monkey.	4.800000190734863
1517	boy doing tricks on a skateboard	A boy does a trick on a skateboard.	4.800000190734863
2011	There's no chance of a fair trial.	there is no chance at a fair trial.	5.0
4582	Exclusive-UPDATE 2-Egypt pro-Mursi alliance signals flexibility in talks	EXCLUSIVE-UPDATE 1-Egypt pro-Mursi alliance signals flexibility in talks	4.599999904632568
2176	the problem isn't who has money.	the problem is who doesn't have money.	2.4666666984558105
2473	Tom Kraynak, manager of operations and resources for the Canton, Ohio-based East Central Area Reliability Council, said that scenario is one among many that investigators are considering.	Tom Kraynak, manager of operations and resources for the Canton, Ohio-based East Central Area Reliability Council, said investigators are considering the scenario.	4.25
502	A toy train is striking a toy car.	A toy train strikes a toy car.	5.0
4776	7 killed in attacks in Iraq	27 killed in attacks across Iraq	2.0
5341	Snowden sees 'no chance' for US fair trial	Snowden sees "no chance" to get fair trial in U.S.	5.0

Show 25 per page

توی این مدل میزان حضور جملات با شباهت بیشتر در ۱۰ حالت شبیه ترینش بیشتر از مدل قبل است با این تفاوت که جملات با دقت کامل ۵ نسبت به مدل قبل کمتر است و جملات با طول طولانی کمتر از مدل قبل در ۱۰ تا اول حضور دارند ولی به صورت میانگین شاید بهتر از مدل قبل عمل کرده است .
مهمترین مشکل این مدل در حالت فعلی که می‌شود با چشم دید در حضور دو مدل با ۲ cos sim در بین بالاترین‌هاست .

7 killed in attacks in Iraq

27 killed in attacks across Iraq

و

The problem isn’t who has money.

The problem is who doesn’t have money .

تشخیص هر دو جمله بیشتر از هرچیز دیگر نیازمند به تحلیل منطقی از جملات دارد که نشان داده این هست که مدل توانایی درک عمیق از جملات نظارت و شباهت ظاهری جملات عامل فریب این مدل است .

1 to 10 of 10 entries Filter

index	sentence1	sentence2	similarity_score
5748	Putin spokesman: Doping charges appear unfounded	The Latest on Severe Weather: 1 Dead in Texas After Tornado	0.0
1752	A snowboarder rides a snowboard down a railing beside a flight of steps.	A boy is grinding a skateboard down a concrete wall.	0.800000011920929
4735	US releases initial report on fracking impacts on water	Men in China detained after taking girls to hotel	0.0
1745	A black dog retrieves in the snow.	A skateboarder jumping in the street.	0.0
1739	An old man in a black trench coat standing in a marketplace.	The baby is wearing a red shirt and walking in a plaza.	1.2000000476837158
5437	Many dead as asylum boat sinks off Australia	Mandela spends third day in hospital	0.0
5435	China ‚Äòwill protect interests‚Äô of foreign business	Obama in ‚Äòdirect‚Äô confrontation with Putin on Ukraine	0.0
1712	Man rowing a boat.	a man doing a trick on skateboard	0.0
1708	Two dogs are playing around in the dirt.	A dancer posing for the camera in a red and white dress.	0.0
1706	A climber is standing on the rock face next to the pink rope.	The children are running in the snow with fences in the background.	0.0

Show 25 per page
Like what you see? Visit the data table notebook to learn more about interactive tables.

ولی این مدل در تشخیص ۱۰ تا آخر بسیار بهتر از مدل قبل عمل کرده است .

و در انتها سراغ مدل نهایی می‌رویم که نزدیکترین پیش بینی را به پیش بینی‌های expert داشته است .

index	sentence1	sentence2	similarity_score
575	A girl jumps on a car.	A girl is jumping onto a car.	4.800000190734863
2317	Religion that does that I have a problem with.	Irreligion that doesn't do that I don't have a problem with.	1.399999976158142
527	The dog is barking at the toy.	A dog is barking at a toy.	5.0
16	The polar bear is sliding on the snow.	A polar bear is sliding across the snow.	5.0
511	A woman is brushing some shrimp.	A woman brushes some shrimp.	5.0
4236	Soldier 'can't recall' massacre	US soldier 'doesn't recall massacre'	4.400000095367432
3931	Assange appeals extradition to UK's top court	Assange to appeal extradition to UK's top court	4.0
5341	Snowden sees 'no chance' for US fair trial	Snowden sees "no chance" to get fair trial in U.S.	5.0
4252	Avalanche buries at least 100 Pakistani soldiers	Avalanche buries over 100 Pakistani soldiers	4.599999904632568
247	The boy is playing the piano.	A boy plays a piano.	5.0

مدل pretrain شده ما نسبت به مدل‌های قبلی بسیار بهتر عملکرد اما خوب این مدل هم با این همه در تله یک خطای logic افتاد که توسط انسان به صورت منطقی قابل تشخیص هست اما برای ماشین نه ولی در این مدل هم سهم جملات با طول بالاتر کمتر از مدل اول است .

وجود جمله با شباهت ۱ که این اتفاق حتی در مدل‌های یک و دو هم رخ نداده بود در بالاترین قسمت نشان دهنده این است که در مواجه با جملات پیچیده مدل ما شکست می‌خورد که نیاز به preprocess بهتر وعمیقتر قبل از مرحله learn دارد که ما در این پروژه preprocess را به صورت خیلی ساده انجام دادیم که امکان از دست رفتن اطلاعات حیاتی برای مدل نیز وجود دارد .

جمله :

Religion that does that I have a problem with .

irreligion that doesn’t that I don’t have a problem with.

1 to 10 of 10 entries Filter

index	sentence1	sentence2	similarity_score
262	A man is riding a horse.	A woman is washing a freezer.	0.0
404	A man is jumping rope outside.	A woman is slicing a cucumber.	0.0
1514	The two boys play in the field.	The two people have their heads covered with scarves.	0.0
1422	Two women are sitting in a cafe.	Two men catching a fish in a swamp.	0.0
395	A dog is jumping on a trampoline.	A boy is playing a guitar.	0.0
335	A man is breaking water jugs.	A woman is peeling garlic.	0.20000000298023224
1578	Children jumping on a trampoline.	A man is sledding on an orange slide.	0.0
1185	Two brown horses standing in grassy field.	Two men sailing in a small sailboat.	0.0
1857	Four kids holding hands jump into a swimming pool.	Four dogs playing in the snow.	0.0
3718	this is particularly true in the poorest countries	this was particularly after ethiopia withdrows its forces	0.4000000059604645

در پایین ترین سطح البته فوق العاده است .

پس فهمیدم که برای متن‌های طولانی مدل یک برای بالاآوردن آنها در بالاترین جایگاه عملکرد بهتری از دو تای قبل دارد ولی در متن‌های کوتاه تر و درک متن از دو مدل قبل جایگاه بسیار پایین تری دارد.

مدل دوم نسبت به مدل سوم از قدرت درک متن کمتری برخورد دار است اما با این حال هر دو در درک متن‌های پیچیده و با پسوندهای منفی ضعیف عمل می کنن که اگر درک جملات با پسوند منفی را بهتر درک کنند احتمالا بسیار قوی‌تر ظاهر می‌شوند و برای این کار نیاز به preprocess بهتر در اول کار داریم .