به نام خدا

نام : حسین سیم چی

تمرین : شماره دوم

درس : پردازش زبان های طبیعی

در ابتدا به بررسی پیکره ی ورودی که شامل ۳۰ سوال که هرکدام به ۴ شکل متفاوت نوشته شده اند میپردازیم ، سپس با استفاده از ابزار NLTK آن را TOKENIZE می کنیم و با ساخت مدل های ۱ تا ۴ گرام ، پیکره ی خود را میسازیم ، سپس معیار perplexity را محاسبه می کنیم و در انتها با بررسی توابع خروجی و ارزیابی ، کار خود را به پایان میرسانیم.

تمامی مراحل انجام کار به صورت فایل کد که در محیط JUPYTERLAB نوشته شده و قابل اجرا بر روی PYCHARM نیز میباشد ، به همراه تصاویر انجام کار به صورت ضمیمه در COURSEWARE بارگذاری شده است

متن ورودی ما به شرح زیر میباشد :

سلام کتابخانه کجاس؟

کتابخانه کجا میباشد ؟

میشه بگید کتابخانه کجاس؟

بخواهم بروم کتابخانه از کجا باید برم؟

اقای دکتر نوری امروز هست ؟

آقای دکتر نوری امروز کی میان؟

آقای دکتر نوری هستن؟ کار واجب دارم

با اقای دکتر نوری کار دارم

نماز خونه داره اینجا؟

نمازخونه کجا باید برم؟

وقت نمازه ، نمازخونه هست اینجا؟

نمازخونه کجاست؟

برگه ی من حاضره؟

برگه ی منو بدید

میشه برگه ی من رو بدید ؟

امکان دارد برگه ی من را بدهید؟

کارتون؟

با من کار داشتید؟

ببخشید امرتان؟

بله؟

این کتابو میخوام تمدید کنم

میشه برام تمدید کنید؟

کتاب رو میشود برای من تمدید نمایید ؟

چطوری میشود این کتاب را تمدید کرد؟

انصراف از تحصیل چجوریه؟

میشه انصراف از تحصیل داد؟

مراحل انصراف از تحصیل چگونه است؟

بی زحمت نامه انصراف از تحصیل من را بدهید

نمرات من چرا نمیاد ؟

میشه بفرمایید نمرات من کی میاد؟

باید از کی بپرسم نمرات من کی میاد؟

نمرات من تا کی وارد میشود؟

با رئیس دانشکده کار دارم

رئیس دانشکده کجاست؟

میشه بگید رئیس دانشکده کی میاد؟

رئیس دانشکده حضور دارند؟

جوایز تحصیلی را کی میدهید؟

جوایز تحصیلی را قرار است چه زمانی بدهید؟

زمان برگذاری جشن جوایز تحصیلی کی است؟

امسال جشن جوایز تحصیلی برگذار میشود؟

برگه فارغ التحصیلی من حاضر است؟لطفا بدهید

برگه فارغ التحصیلی منو بدید

لطفا برگه فارغ التحصلی من را بدهید

میشه اون برگه فار التحصیلی رو بدی

کلاس ۱۰۲ امروز تشکیل نمیشه؟

کلاس ۱۰۲ کنسل شد ؟

استاد کلاس ۱۰۲ امروز نمیان؟

میشه زودتر بگید کلاس ۱۰۲ تشکیل میشه یا نه؟

دانشگاه فردا باز است؟

دانشگاه فردا تعطیل است؟

دانشگاه از فردا به مدت چند روز تعطیل است؟

دانشگاه فردا تعطیله؟

ریزنمرات این ترم رو میخوام

میشه ریزنمرات این ترم را به من بدید

ریزنمرات رو بده

ریز نمرات را به من بدهید

مسئول این دانشگاه کیه؟

کی مسئوله اینجاست؟

مسئولش کیه؟

اینجا مسئول داره؟

خوابگاه کجاست؟

میشه راهنمایی کنید و بگید خوابگاه کجاست؟

خوابگاه از کدوم طرف باید رفت؟

از کجا باید خوابگاه برم؟

معرفی به استاد چجوریه؟

شرایط معرفی به استاد چجوریه؟

آیا میتوان درسی را معرفی به استاد گرفت؟

برای معرفی به استاد باید چکار کنم؟

چجوری واحدهامو انتقال بدم؟

ایا میشه یسری واحدهارو انتقال داد ؟

شرایط انتقال واحد چیه؟

چطوری میشه یسری درسها رو انتقال داد ؟

امکان تغییر ساعت کلاس برای درس هوش وجود دارد ؟

میشه ساعت کلاس هوش و عوض کنید؟

چجوری میشه ساعت کلاس هوش رو عوض کرد؟

شرایط تغییر ساعت کلاس هوش چیست؟

تاریخ حذف و اضافه کیه؟

میشه تاریخ حف و اضافه رو بگید؟

لطفا تاریخ حذف و اضافه را اعلام نمایید

حذف و اضافه رو چه زمانی اعلام میکنید ؟

شرایط مرخصی تحصیلی چیه؟

میشه بیش از ۱ تزم مرخضی تصیلی گرفت ؟

لطفا شرایط مرخصی تحصیلی را اعلام نمایید

نکات مربوط به مرخصی تحصیلی چیست؟

اشتغال به تحصیل برای بیمه میخوام

میشه برگه استغال به تحصیل منو برای بیمه بدهید ؟

برگه اشتغال به تحصیل من اماده است؟

امده ام تا برگه ی اشتغال به تحصیل را بگیرم

شرایط مجوز خروج از کشور چیه؟

لطفا شرایط لازم برای خروج از کشور را بفرمایید

ایا من میتونم مجوز خروج از کشور بگیرم؟

چجوری میشه مجوز خزوج از کشور گرفت؟

کارت دانشجویی من رابدهید

میشه کارت دانشجویی من را بدهید؟

کارت دانشجویی من اماده است؟

کارت دانشجویی من رو بده

ارائه پایان نامه تا کی وقت داره؟

ایا هنوز میتوان پایان نامه را ارائه داد؟

تا چه زمانی میتوان پایان نامه را ارائه داد؟

مهلت ارائه پایان نامه را بفرمایید

شرایط مهمان شدن به دانشگاه شما چیست ؟

شرایط مهمان شدن چیه؟

ایا میشه به عنوان مهمان اینجا درس خوند؟

چجوری میشه اینجا مهمان شد ؟

شرایط نخبگی چیست؟

چه شرایطی برای نخبه شدن وجود دارد؟

میشه شرایط نخبگی را بفرمایید؟

شرایط لازم برای پذیرش نخبه چیست؟

چند جلسه میشود غیبت کرد؟

در یک ترم چندجلسه میتوان غیبت کرد؟

چندجلسه میشه غیبت کرد؟

تعداد جلسات مجاز برای غیبت در هر درس

شرایط وام دانشجویی چیست ؟

لطفا شرایط وام دانشجویی را بفرمایید ؟

چه مدارکی برای دریافت وام دانشجویی باید بدهم؟

به چه کسانی وام دانشجویی میدهید؟

شرایط خوابگاه برای ورودی امسال چیست؟

میشه شرایط گرفتن خاوبگاه رو بفرمایید ؟

به منم خواگاه تعلق میگیرد؟

چگونه میتوان خوابگاه دانشجویی گرفت؟

شرایط وام دانشجویی را بفرمایید ؟

ایا میتوان وام دانشجویی دریافت کرد؟

به چه کسانی وام دانشجویی تعلق میگیرد؟

شرایط وام دانشجویی چیه؟

در ابتدا با استفاده از دستور زیر متن ورودی را می خوانیم :

```
file = open("C:\\Users\\Lenovo\\Desktop\\question.txt" , encoding = "UTF-8")
text = file.read()
print(text)
```

سپس کتابخانه های مورد نیاز خود را import می کنیم و متن خود را TOKENIZE می کنیم :

```
from nltk.tokenize import sent_tokenize , word_tokenize
from nltk.collocations import BigramCollocationFinder:
```

با استفاده از این دستور و استفاده از آن میتوان Bigram های موجود متن را شناسایی کرد

```
from nltk.collocations import TrigramCollocationFinder:
```

با استفاده از این دستور و استفاده از آن میتوان Trigram های موجود متن را شناسایی کرد

```
from nltk.collocations import TrigramAssocMeasures:
```

با استفاده از این دستور در ادامه میتوان معیار ارزیابی خاصی برای انتخاب Trigram های خود انتخاب کرد

```
from nltk.collocations import BigramAssocMeasures:
```

با استفاده از این دستور در ادامه میتوان معیار ارزیابی خاصی برای انتخاب Bigram های خود انتخاب کرد

```
text_tokenize = word_tokenize(text)
print(text_tokenize)
```

پس از انجام مراحل بالا با استفاده از دستورات زیر Unigram و Bigram های موجود در متن را با استفاده از معیار LIKELIHOOD_RATIO انتخاب کرده و درون یک LIST آن را قرار میدهیم ، معیار ارزیابی ما اینگونه عمل میکند که بر اساس بیشترین تعداد دفعات وقوع Bigram ها نسبت به یکدیگر آن ها را به LIST اضافه می کند به عنوان مثال ما در این کد ، n را مساوی عدد ۵۰۰ در نظر گرفته ایم که به معنای این میباشد که ۵۰۰ بایگرام اول که بیشترین تعداد دفعات تکرار را نسبت به بقیه Bigram ها در متن را داشته اند را به ما برمیگرداند درضمن میتوان از توابع دیگری به جای nbest برای تولید خروجی نیز استفاده کرد ، به عنوان مثال دیگر میتوان از تابع Score_ngrams نیز استفاده کرد که این تابع به هر یک از Bigram های موجود یه امتیازی را نسبت میدهد :

```
def bag_of_words(words):
   return dict([(word,True) for word in words])
def bag_of_bigrams(words ,
score_fn=BigramAssocMeasures.likelihood_ratio,n=500):
      bcf = BigramCollocationFinder.from_words(words)
      bigrams=bcf.nbest(score_fn,n)
      return bag_of_words(words+bigrams)
Bigram = bag_of_bigrams(text_tokenize)
Bigram_list= list(Bigram)
print(Bigram_list)
```

سپس با استفاده از دستور زیر Trigram های درون متن را شناسایی میکنیم و آنها را به Bigram_list خود اضافه می کنیم .

```
tcf = TrigramCollocationFinder.from_words(text_tokenize)
```

```
Bigram_list.extend(tcf.nbest(TrigramAssocMeasures.likelihood_ratio,500))
print(Bigram_list)
```

با استفاده از دستور زیر طول list خود را بدست اورده سپس از آن را به دو بخش Train و Test تقسیم میکنیم :

```
print(len(Bigram_list))
train_set = []
test_set = []
cutoff = int(len(Bigram_list)*0.9)
train_set.extend(Bigram_list[:cutoff])
test_set.extend(Bigram_list[cutoff:])
print(len(train_set))
print(len(test_set))
```

در ادامه با استفاده از دستورات زیر که از منابع کد در سایت NLTK نیز استفاده شده است Perplexity را ازمایش می کنیم . کد به این صورت عمل میکند که با توجه به ورودی ما و محاسبه احتمال ان در بین Train_set خروجی را بدست می اوریم . هرچه میزان perplexity کاهش یابد یعنی مدل ما بهتر کار میکند :

```
import collections, nltk
def unigram(train_set):
    model = collections.defaultdict(lambda: 0.01)
    for f in train_set:
        try:
            model[f] += 1
        except KeyError:
```

```
        model [f] = 1
        continue
    N = float(sum(model.values()))
    for word in model:
        model[word] = model[word]/N
    return model

def perplexity(testset, model):
    testset = testset.split()
    perplexity = 1
    N = 0
    for word in testset:
        N += 1
        perplexity = perplexity * (1/model[word])
    perplexity = pow(perplexity, 1/float(N))
    return perplexity
testset1 = "را به من"

model = unigram(train_set)
print (perplexity(testset1, model))
```

میتوان با استفاده از تابع خروجی Score_ngrams که در ابتدا نیز به آن اشاره شد ، امتیازی بر اساس تعداد دفعات وقوع هر Bigram یا Trigram به آن اختصاص داد :

```
text_tokenize = word_tokenize(text)
y = []
```

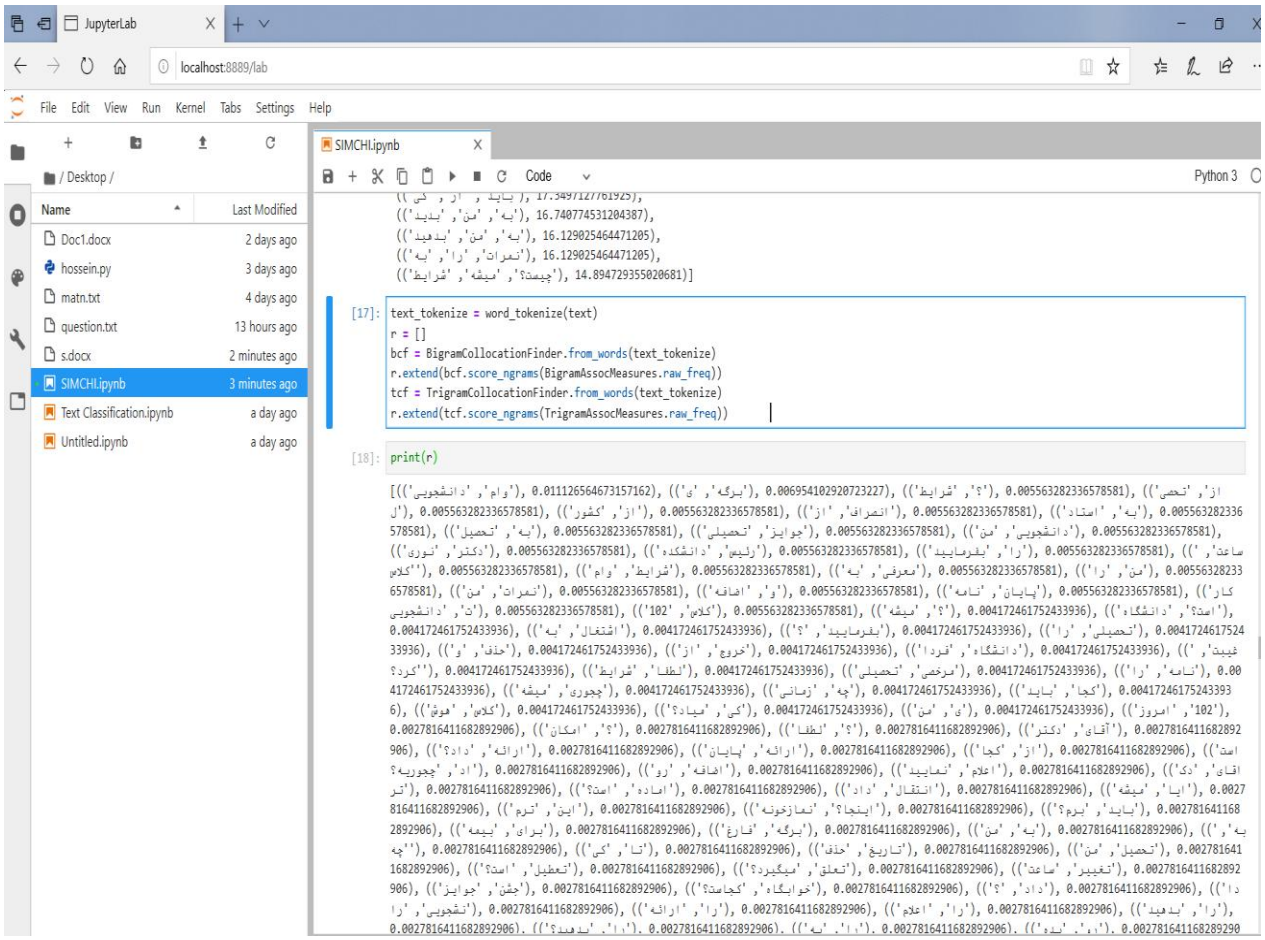
```
y.extend(text_tokenize)
y.extend(bcf.score_ngrams(BigramAssocMeasures.likelihood_ratio))
tcf = TrigramCollocationFinder.from_words(text_tokenize)
y.extend(tcf.score_ngrams(TrigramAssocMeasures.likelihood_ratio))
print(len(y))
train_set = []
test_set = []
cutoff = int(len(y)*0.9)
train_set.extend(y[:cutoff])
test_set.extend(y[cutoff:])
```

به عنوان مثال بخشی از خروجی Test_set مانند زیر است :

```
[(('۲۸,۹۶۶۵۹۰۸۴۶۵۱۲۸۴ ,('کجاست؟', 'میشه', 'بگید'),
 (۲۸,۸۷۳۵۹۷۸۱۱۲۴۰۴ ,('اینجا', 'مهمان', 'شد')),
 (۲۸,۸۷۳۵۹۷۸۱۱۲۴۰۴ ,('درس', 'هوش', 'وجود')),
 (۲۸,۸۷۳۵۹۷۸۱۱۲۴۰۴ ,('میشود؟', 'با', 'رئیس')),
 (۲۸,۸۵۱۲۳۵۷۵۳۷۳۳۴۲ ,('را', 'به', 'من')),
 (۲۸,۸۱۰۳۶۴۷۲۶۱۵۶۱۲۴ ,('تشکیل', 'میشه', 'یا')),
 (۲۸,۸۱۰۳۶۴۷۲۶۱۵۶۱۲۴ ,('وجود', 'دارد؟', 'میشه')),
 (۲۸,۷۸۸۴۷۶۳۹۴۶۱۰۸ ,('درسها', 'رو', 'انتقال')),
 (۲۸,۷۸۸۴۷۶۳۹۴۶۱۰۸ ,('مجاز', 'برای', 'غیبت')),
 (۲۸,۷۸۸۴۷۶۳۹۴۶۱۰۷۹۷ ,('ام', 'تا', 'برگه')),
 (۲۸,۶۸۰۸۶۵۶۵۴۰۳۶۵۱۳ ,('بدید', 'لطفا', 'برگه')),.....
```

به عنوان پایان کار میخواهیم از متد ارزیابی raw_freq استفاده کنیم ، این متد اینگونه عمل میکند که فرکاس رخداد هر Bigram یا Trigram را نیز به ما برمیگرداند :

```
text_tokenize = word_tokenize(text)
r = []
bcf = BigramCollocationFinder.from_words(text_tokenize)
r.extend(bcf.score_ngrams(BigramAssocMeasures.raw_freq))
tcf = TrigramCollocationFinder.from_words(text_tokenize)
r.extend(tcf.score_ngrams(TrigramAssocMeasures.raw_freq))
```



از فرکانس رخداد هر یک از Bigram ها و Trigram ها میتوان دقت و احتمال وقوع جملات پیش بینی نشده را محاسبه کرد .

نکته آخر :

در این تمرین سعی شده است با استفاده از متن ورودی و قطعه بندی آن ، ابتدا مدلهای گفته شده از آن استخراج شود سپس با تقسیم list به دو بخش Train_set و Test_set داده های خود را تفکیک کرده و در ادامه با بررسی و نوشتن یک ساختار احتمالی برای Perplexity و محاسبه ی آن بپردازیم .

در انتها نیز با استفاده از توابع دیگر برای خروجی توانستیم امتیاز هر Bigram یا Trigram را بر اساس تعداد دفعات وقوع آن در متن برگردانیم همچنین فرکانس رخداد هر کدام را محاسبه کردیم

تمامی فایلها شامل فایل کد برنامه ، فایل متن وروری و عکسهای لازم جهت مشخص کردن مراحل انجام کار در COURSEWARE بارگذاری شده است .

باتشکر از زحمات شما ، حسین سیم چی

پایان