به نام خدا

نام استاد : دکتر شمس فرد

نام اعضا فعال گروه : حسین سیم چی ، علی یزدانی

درس : پردازش زبان های طبیعی

عنوان : پروژه ی پایانی (یافتن سوال مشابه درFAQ سها و سامانه پرسش و پاسخ)

پاییز و زمستان ۹۸

فهرست

مقدمه :

پروژه ی پایانی درس پردازش زبان های طبیعی شامل دو بخش میباشد که در بخش اول به بررسی سوال مشابه در سامانه ی FAQ سها میپردازیم ، که در آن با دادن سوال مشابه به سامانه سعی در یافتن سوال مشابه آن می کنیم که در بخش های بعد مفصلا درباره آن بحث خواهیم نمود .

در بخش دوم به بررسی سامانه ی پرسش و پاسخ می پردازیم به این گونه که با توجه به سوالات موجود در دیتاست و جوابهای پیش بینی شده برای هر کدام ، سوالی از سیستم پرسیده میشود (طبق دیتاست) و سپس سیستم پاسخ پیش بینی شده که قبلا به آن گفته ایم را به ما برمیگرداند که این بخش نیز مفصلا درباره آن بحث خواهیم نمود .

توجه : گروه فعال در این پروژه شامل آقایان حسین سیم چی و علی یزدانی میباشد .

تمامی تصاویر و فایل کد نیز جداگانه ارسال خواهد شد

در پایان این بخش از زحمات سرکار خانم اعتضادی که در طی انجام این پروژه با فراهم نمودن دیتاست و کمک های بسیار عالی ما را یاری نمودند بسیار متشکریم .

یافتن سوال مشابه در سامانه ی FAQ سها :

در این بخش دیتاستی که شامل سوالاتی مشابه ولی با بیانهای متفاوت میباشد به سیستم داده میشود و سپس سوالی از سیستم پرسیده می شود و سیستم بر اساس معیاری شباهت را با تک تک جملات دیتاست میسنجد و سپس بیشترین میزان شباهت و بهترین جمله که از این لحاظ شبیه به سوال پرسیده شده میباشد را به ما برمیگرداند .

معیار استفاده شده جهت سنجش میزان شباهت جملات در هر دو پروژه :

در ابتدای کار بایستی با درنظر گرفتن حلقه ی For تک تک جملات ورودی را با جمله ی وارد شده مقایسه کنیم . برای اینکار به عنوان مثال اولین جمله ی موجود در دیتاست را وارد لیست دیگری کرده و همچنین سوال درنظر گرفته شده را نیز با همین رویکرد وارد لیست می کنیم .

سپس باید هر جمله را به برداری تبدیل کنیم که این بردار به تعداد کلمات موجود در جمله ی ما مولفه دارد ، در انتها نیز باید بررسی کنیم که ترکیب دوجمله از چند مولفه تشکیل شده است که در واقع برابر است با تعداد کلمات هر دو جمله .

به عنوان مثال درنظر بگیرید :

اسم من یزدانی هست >>>>> [۱ , ۱ , ۱ , ۱]

من علی یزدانی هستم >>>>> [۱ , ۱ , ۱ , ۱]

بردار مجموع دو جمله برابر است با :

[۱ , ۱ , ۱ , ۱ , ۱ , ۱ ,۱ , ۱]

حال فرض کنید که بخواهیم میزان شباهت جمله اول با مجموع دو جمله را بدست اوریم در نتیجه باید ببینیم که جمله ی اول چند مولفه یا کلمه ی مشترک با مجموع دو جمله دارد که در واقع برابر است با : { من ، یزدانی }

اگر در ابتدا ریشه ی هر کلمه را در نظر بگیریم میتوان " هست " را نیز به عنوان کلمه ی مشترک درنظر گرفت .

که در انتها جاهایی که کلمات جمله دارای اشتراک با مجموع دو جمله است عدد یک و جاهایی که اشتراک ندارند عدد صفر را به خود اختصاص میدهند و بدین گونه بردار ما برای محاسبه ی میزان شباهت بین جملات ساخته می شود .

برای میزان شباهت از معیار Cosine_similarity استفاده می کنیم که کتابخانه ی Sklearn در پایتون اینکار را به راحتی برای ما انجام می دهد .

در واقع بردار محاسبه شده را که به کمک توابع اماده موجود در پایتون در مرحله ی قبل انجام دادیم به عنوان ورودی به آن میدهیم و به صورت درصد میزان شباهت را به ما برمیگرداند .

در قسمت اول پروژه ما سوالی را از سامانه می پرسیم و سپس سوال ما با تک تک جملات موجود در دیتاست مقایسه میشود و میزان شباهت هرکدام با توجه به توضیحات داده شده محاسبه می گردد و این مقادیر در هربار به لیستی اضافه می شود و در انتها نیز سوالی که بیشترین شباهت را به سوال ما دارد برگردانده میشود . با دستوری که در انتها نوشته شده است بار دیگر از کاربر سوال می شود که آیا مایل است سوالی را بپرسد ؟ اگر جواب آن بله باشد باید عدد یک و اگر جواب آن خیر باشد باید عدد دو تایپ نماید بدین شکل اگر سوال دیگری داشته باشد بار دیگر تمامی مراحل فوق تکرار میشود در غیر این صورت پیغام " خداحافظ " را چاپ می کند و برنامه خاتمه می یابد ، بدین شکل قسمت اول پروژه به پایان میرسد .

در زیر تصاویری از خروجی و نحوه ی کد نوشته را ملاحظه می نمایید :

```
#Hossein_Simchi & Ali_Yazdani
Document1 = input('؟ سوالتون چیه !!! بهتون بگم من میتونم سوال مشابه رو ...')
sentencess=[]
grade = []
win = []
def question():
    global Document1,grade,win
    for i in range(len(sentences)):
        text = sentences[i]
        Document2 = text
        #print(Document2)
        corpus = [Document1,Document2]
        count_vect = CountVectorizer()
        X_train_counts = count_vect.fit_transform(corpus)
        from sklearn.feature_extraction.text import TfidfVectorizer
        vectorizer = TfidfVectorizer()
        trsfm=vectorizer.fit_transform(corpus)
        from sklearn.metrics.pairwise import cosine_similarity
        sim = cosine_similarity(trsfm[0:1],trsfm)
        sim1 = float(min(sim[0]))
        #print(sim1)
        grade.append(sim1)
    win = max(grade)
    index = grade.index(max(grade))
    print(sentences[index])


question()
while(1000):
    ask_me_q = int(input('[(خیر)2 یا (بله)1] سوالي هست که بخواین بپرسین؟'))
    if ask_me_q == 1 :
        Document1 = input("؟ سوالتون چیه !!! بهتون بگم من میتونم سوال مشابه رو ...")
        count = 0
        win = 0
        grade = []
        question()
    else:
        print('خداحافظ')
        break
```

```
'''
========== RESTART: C:\Users\Lenovo\Desktop\NLP-PROJECT\nlp_projrct.py =========
تعداد دانشجویان ارشد ازمایشگاه طراحي خودکار چندتا است؟... من میتونم سوال مشابه رو بهتون بگم !!! سوالتون چیه ؟
تعداد دانشجویان کارشناسي ارشد ازمایشگاه طراحي خودکار مدارهاي مجتمع چند نفر است؟
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
لیست پروژه هاي ازمایشگاه نرم افزار چیه؟... من میتونم سوال مشابه رو بهتون بگم !!! سوالتون چیه ؟
لیست پروژه هاي در حال انجام ازمایشگاه اتوماسیون مهندسي نرم افزار چیست؟
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
اسم دانشجویان ارشد ازمایشگاه بیسیم چیه؟... من میتونم سوال مشابه رو بهتون بگم !!! سوالتون چیه ؟
نام دانشجویان کارشناسي ارشد ازمایشگاه امنیت و شبکه هاي بیسیم چیست؟
2سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
خداحافظ
>>>
```

سامانه ی پرسش و پاسخ دانشگاه :

در این قسمت نیز همانند قسمت قبلی ، ابتدا دیتاست خود را که شامل سوالات درنظر گرفته شده و جوابهای پیش بینی شده برای هرکدام را در برنامه ی خود باید وارد نماییم .

در گام بعدی با توجه به معیار شباهت سنجی که در بخش قبلی به طور کامل به آن اشاره کردیم میزان شباهت سوال پرسیده شده و تک تک سوالات موجود در دیتاست را میسنجیم . همانطور که مشخص است هر سوال به ۴ صورت بیان شده و برای هر سوال تنها یک جواب را درنظر گرفته ایم که هر سوالی که بیشترین شباهت را با سوال پرسیده شده داشته باشد جواب مربوط به آن در خروجی برای اطلاع کاربر چاپ میگردد . در این قسمت نیز همانند قسمت قبلی در پایان از کاربر سوال میشود که آیا سوالی دارد یا خیر ، اگر بازهم سوال داشت همین مراحل تکرار میشود در غیر این صورت برنامه ی ما پایان می یابد

در زیر نمونه ایی از خروجی را مشاهده می نمایید :

```
================== RESTART: C:\Users\Lenovo\Desktop\part2_nlp.py ================
کجا میتونم دکتر طباطبایي رو پیدا کنم؟سوالتون چیست؟
اتاق دکتر طباطبایی، طبقه 4 انتهای راهرو
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
دفتر دکتر شمس فرد کجاست؟سوالتون چیست؟
طبقه سوم
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
کي نمرات رو میذارن رو سایت؟سوالتون چیست؟
!از استادتان بپرسید
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
فرم موضوع پایان نامه را کجا بدهم؟سوالتون چیست؟
به خانم زندی طبقه سوم
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
دبیرخانه کجاست؟سوالتون چیست؟
.دبیرخانه دانشکده، طبقه سوم، روبه روی راهروی کتابخانه
1سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
محل برگذاري کلاس هوش؟سوالتون چیست؟
اتاق 105
2سوالي هست که بخواین بپرسین؟ [1(بله) یا 2(خیر)]
خداحافظ
>>>
```

در پایان از زحمات تمامی افراد در طول ترم گذشته سپاسگزاریم .

پایان