به نام پروردگار

پردازش زبان طبیعی تمرین اول

موعد تحویل :

زینب منتظری (zeinab.montazeri@gmail.com)

آنالیز احساسات

در این تمرین یک طبقه‌بند ساده Naïve Bayes را برای طبقه‌بندی احساسات پیاده‌سازی می‌کنید. برای این منظور از کرپوس نقد فیلم‌ها استفاده می‌کنیم. NLTK نسخه‌ای از این مجموعه داده را فراهم کرده است. این مجموعه داده هر نقد را به دو دسته مثبت یا منفی دسته‌بندی کرده است. برای استفاده از این مجموعه داده می‌توانید از دستور زیر استفاده کنید:

```
import nltk
nltk.download('movie_reviews')
```

مراحل انجام تمرین

1- پیش‌پردازش

مانند اکثر فعالیت‌های پردازش زبان طبیعی، پیش‌پردازش‌های لازم را بر روی داده‌ها انجام دهید. این پیش‌پردازش‌ها می‌تواند شامل حذف Stop Wordها، Tokenization، حذف puncuationها، Normalization و... باشد.

برای انجام این عملیات می‌توانید از ابزار NLTK استفاده کنید.

http://nltk.org/

2- استخراج ویژگی

برای اینکه بتوانیم یک رده‌بند را آموزش دهیم، لازم است تعدادی ویژگی را از متن استخراج کنیم. تعداد و نوع ویژگی‌هایی که استخراج می‌کنید کاملا اختیاری و نمره این قسمت با توجه به ابتکارتان برای انتخاب ویژگی‌ها در نظر گرفته می‌شود.

3- آموزش رده‌بند

در این مرحله با استفاده از ویژگی‌های استخراج شده در مرحله قبل باید یک رده‌بند را به روش Naïve Bayes آموزش دهید. برای پیاده‌سازی این رده‌بند می‌توانید از کتابخانه Scikit Learn یا NLTK استفاده کنید.

http://scikit-learn.org/

http://nltk.org/

برای انجام رده‌بندی از روش k-fold validation بر روی داده‌های آموزش استفاده کنید و مقدار k را 5 در نظر بگیرید.

4- ارزیابی مدل آموزش دیده

در این مرحله باید رده‌بندی را که آموزش داده‌اید، ارزیابی کنید. برای ارزیابی باید معیارهای Accuracy، Recall، Precission و امتیاز F1 را گزارش دهید.

5- گزارش نهایی

قسمت اصلی ارزیابی پروژه شما با توجه به گزارش پروژه انجام می‌شود. این گزارش باید شامل موارد زیر باشد:

- تحلیل تاثیر پیش‌پردازش‌ها بر نتایج رده‌بندی
- توضیح ویژگی‌های استخراج شده و دلیل انتخاب آن‌ها
- تحلیل نتایج رده‌بندی

✓ موارد تحویل

- گزارش
- کدهای پیاده‌سازی شده

✓ نکته: برای تمامی مراحل، آزاد به استفاده از کتابخانه‌های موجود هستید.

لطفا به قواعد حل تمرین که در CECM قرار داده شده است توجه کنید.