دانشگاه علم و صنعت ایران

دانشکده ریاضی و علوم کامپیوتر

پروژه پایانترم برنامه سازی پیشرفته

Sentiment Analysis

استاد: ابراهیم اردشیر لاریجانی

توضیحات اولیه

در این پروژه که بسیار واقعی‌ست(!) می‌خواهیم با استفاده از دانشی که تا الان در پایتون کسب کرده‌اید، و مقداری دانش جدید که کسب خواهید کرد به تحلیل احساسات بپردازید.

تحلیل احساسات(Sentiment analysis) یکی از زیرشاخه‌هایNLP می‌باشد، که در آن متن نوشته شده توسط کاربر را پردازش کرده و به مثبت یا منفی بودن این نظر پی می‌بریم.

شما قرار است با دیتایی که از نظرات کاربران در سایت Amazon دارید، و ساخت یک مدل ماشین لرنینگ این کار را انجام بدهید. این دیتاست شامل 20000 نظر از سایت Amazon در مورد کالاها و برنامه‌های مختلف است. هرکدام از این دیتاها دارای لیبل positive به معنای نظر مثبت و یا negative به معنای نظر منفی می‌باشد.

این پروژه را باید با استفاده از N-gram انجام بدهید.

در شکل زیر چند سطر از دیتاست را مشاهده می‌کنید.

	reviewText	Sentiment
383	Enjoyable game. I love those little birds and ...	Positive
13934	This is really a good APP. I use it on my Kin...	Positive
4643	Love this game when I just want to relax. It'...	Positive
17797	Don't waste your money and get csipsimple FREE...	Negative
15529	I Hate This App. Just Hate. >____< Its M...	Negative

مراحل پروژه

---

0. import کتابخانه‌های مورد نیاز:

کتابخانه‌های pandas، numpy، matplotlib، sklearn و... را import کنید. (ممکن است از همه این کتابخانه‌ها استفاده نکنید، و یا به کتابخانه‌های دیگری نیز نیاز پیدا کنید)

1. خواندن دادگان:

- ابتدا فایل csv پیوست داده شده را با استفاده از تابع read_csv در pandas بخوانید، و آن را در یک dataframe بریزید.

2. پیش پردازش داده:

- ابتدا مقادیر Positive را به عدد 1 و مقادیر Negative را به عدد 0 تغییر بدهید.
- داده‌ها را به چهار بخش X_train, X_test, y_train, y_test تقسیم کنید. 20 درصد از داده‌ها را برای تست بردارید.
- از روش‌های Tokenization و Stopword removal و Stemming و یا روش‌های دیگر برای حذف داده‌های بیهوده و آماده کردن کلمات برای تبدیل شدن به بردار استفاده کنید.
- سپس از روش CountVectorizer و تنظیم پارامتر ngram_range برای تبدیل کلمات به بردار استفاده کنید.

3. ساخت مدل:

- می‌توانید از مدل‌های Logistic Regression، Naive Bayes، SVM و یا هر مدل ساده دیگری که می‌خواهید استفاده کنید.
- مدل ساخته شده را با استفاده از دادگان train آموزش دهید.

4. تست مدل:

- با استفاده از دادگان X_test روی مدل خود، لیبل‌ها را تخمین بزنید.
- با معیار accuracy_score دقت مدل را بسنجید.

نکات کلیدی

1. برای انجام این پروژه، بهتر است که از Jupyter Notebook استفاده کنید.
2. پروژه باید در قالب تیم‌های دو نفره انجام شود.
3. کامنت گذاری و نوشتن کد تمیز(clean code) الزامی ست.
4. Visualization کل پروسه با استفاده از کتابخانه matplotlib نمره امتیازی دارد.
5. در هنگام ارائه، هر دو نفر باید توانایی توضیح و تغییر کد را داشته باشند.
6. معیار سنجش پروژه، دقت مدل شما در انتها و نحوه پیاده سازی آن است. این که از چه توابعی و یا چه مدلی استفاده می‌کنید اهمیتی ندارد.

موفق باشید

تیم حل تمرین برنامه سازی پیشرفته