

دانشکده ریاضی و علوم کامپیوتر

پروژه پایانترم برنامه سازی پیشرفته

# **Sentiment Analysis**

استاد: ابراهیم اردشیر لاریجانی

# توضيحات اوليه

در این پروژه که بسیار واقعی ست(!) می خواهیم با استفاده از دانشی که تا الان در پایتون کسب کرده اید، و مقداری دانش جدید که کسب خواهید کرد به تحلیل احساسات بیردازید.

تحلیل احساسات(Sentiment analysis) یکی از زیرشاخههای NLP 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 می کنید.

	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

## نیاز: import کتابخانههای مورد نیاز:

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

### 1. خواندن دادگان:

• ابتدا فایل csv پیوست داده شده را با استفاده از تابع csv بیوست داده شده را با pandas بریزید.

### 2. پیش پردازش داده:

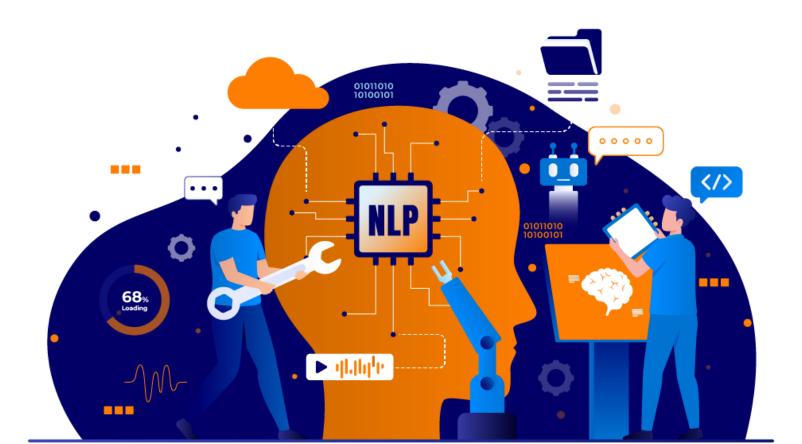
- ابتدا مقادیر Positive را به عدد 1 و مقادیر Negative را به عدد 0 تغییر دهند.
- دادهها را به چهار بخش X\_train, X\_test, y\_train, y\_test تقسیم کنید. 20 درصد از دادهها را برای تست بردارید.
- از روشهای Tokenization و Tokenization و Stopword removal و Tokenization و یا روشهای دیگر برای حذف دادههای بیهوده و آماده کردن کلمات برای تبدیل شدن به بردار استفاده کنید.
- سپس از روش CountVectorizer و تنظیم پارامتر mgram\_range برای تبدیل کلمات به بردار استفاده کنید.

#### 3. ساخت مدل:

- می توانید از مدلهای Logistic Regression، Naive Bayes، Logistic Regression، و یا هر مدل ساده دیگری که می خواهید 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.



# نكات كليدي

- 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 می کنید اهمیتی ندارد.

موفق باشيد

تیم حل تمرین برنامه سازی پیشرفته