امین عارف زاده
810195424

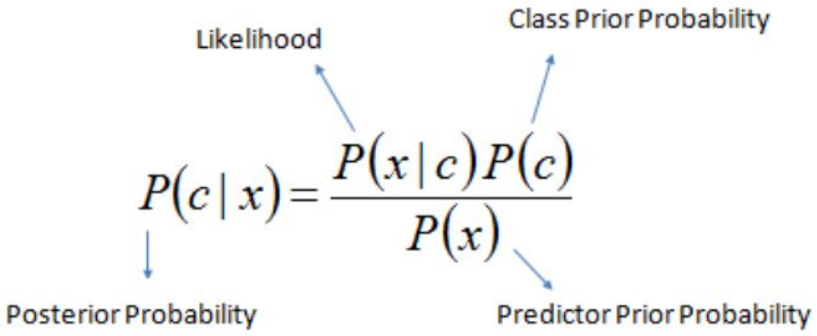
پروژه ۳

پیش پردازش داده:

کوچک کردن حروف: در صورت عدم انجام این کار کلمه مانند Roof و roof دو کلمه جدا از هم در نظر گرفته میشود در حالی که از لحاظ معنایی یکی هستند و که این عمر باعث میشود در تشخیص فرکانس تکرار کلمات دچار مشکل شویم و تشخیص های اشتباه بیشتر شود.

روش های stemming و lemmatization: در stemming تنها پیشوند و پسوند از کلمه حذف می‌شود در حالی که در lemmatization ریشه کلمه پیدا میشود و به مفهوم کلمه دقت میشود در نتیجه این روش نتیجه بهتری در کار های پردازش متن می گیرد.

قاعده بیزین:



در این رابطه c بیانگر کتگوری هر خبر و x در مدل bag of words احتمال wi است یعنی احتمال قرارگیری هر کلمه در پوزیشن i ام جمله است. البته در ادامه ما میتوانیم این احتمال ها را به شرط دانستن کتگوری از هم مستقل بگیریم تا محاسبات ساده تر شود.
در این رابطه posterior احتمال هر کدام از کتگوری ها به شرط دانستن جمله است. likelihood احتمال وجود آن کتگوری به شرط دانستن کتگوری است. class prior احتمال رخ داد هر کتگوری است که برای ساده سازی میشود احتمال همه‌ی کتگوری ها را یکسان در نظر گرفت و در مخرج هم احتمال وجود جمله بدون در نظر گرفتن کتگوری است که برای محاسبه‌ی آن میشود از ضرب احتمال شرطی likelihood و class prior و سامیشن روی کتگوری استفاده کرد که البته برای بدست آوردن جواب نهایی احتمال نیازی به محاسبه مستقیم این مخرج نیست.

ارزیابی:
بهتر است از هر کتگوری ۸۰ درصد دیتا را جدا کنیم تا از تمام دیتا ۸۰ درصد. اگه ۸۰ درصد دیتا رو برداریم برای train (بدون در نظر گرفتن کتگوری) و اگر این دیتا بزرگ باشد و کاملا بُر خورده باشد آنگاه احتمال از هر کتگوری همان ۸۰ درصد برداشته میشود ولی به صورت کلی معمولا به صورت رندوم به هر کدام از کتگوری ها درصدی بالا و پایین تخصیص داده میشود.
این رندومنس باعث خراب شدن precision و recall کتگوری های مختلف میشود زیرا ممکن است یه کتگوری oversample شود و کتگوری دیگر undersample

نکته‌ی بد دیگری که این مسئله دارد امکان oversampling و undersampling به صورت دلخواه را ندارد در نتیجه ممکن است مدل در زمان train شدن به سمت یه کتگوری خاص train شود و بالانس نباشد.

:confusion matrix

	Predicted BUTY	Predicted BUSINESS	Predicted TRAVEL
Actual BUTY	1765	67	166
Actual BUSINESS	65	654	136
Actual TRAVEL	138	144	2032

:evaluation

phase 1	TRAVEL	BUSINESS
Recall	92	81
Precision	93	81
Accuracy	89	

phase 2	TRAVEL	BUSINESS	BUTY
Recall	87	76	88
Precision	87	75	89
Accuracy	86		

:oversampling

با بالا بردن درصد نمونه برداری یه کتگوری (در مرحله‌ی نمونه برداری) recall مربوط به آن کتگوری بالا میرود اما precision پایین می‌آید. عملا با oversampling یک مسئله مدل ما به سمت آن کتگوری متمایل می‌شود و اگر اخباری را که واقعا آن کتگوری را دارد به آن بدهیم با درصد بالا ترین درست آن را تشخیص میدهد (ریکال بیشتر میشود) ولی اخباری که کتگوری های دیگه‌ائی دارد را هم بیشتر به صورت کتگوری که oversample شده تشخیص میدهد در نتیجه precision کمتر می شود که خوب نیست. باید بتوان درصدی مناسب پیدا کرد که مقدار precision و recall به هم نزدیک تر شوند. درصد های استفاده شده در این پروژه در زیر آمده است

```
categories = {
                'TRAVEL': 0.74,
                'BUSINESS': 0.84,
                'STYLE & BEAUTY': 0.77,
}
```

سوالات:

۱) قبلا پاسخ داده شده

۲) این شاخص مقدار frequency کلمات را در تشخیص متن در نظر میگیرد. در حالت bag of words معمولا ما تکرار یک خبر را در نظر نمیگیریم در واقع کلمه‌ائی که در یک خبر مربوط به Business چند بار آمده با کلمه ایي که در همان خبر تنها ۱ بار آمده تفاوتی ندارد. اگه بخوایم این شاخص را تاثیر دهیم میبایست تکرار کلمات در هر خبر را هم در نظر بگیریم مثلا برای اینکار میتوان به جای عدم حضور یا حضور کلمه در تعداد اخبار را در نظر بگیریم. فرکانس کلمه در کل اخبار در نظر بگیریم یعنی ببینیم چند درصد کلماتمان را این کلمه در اختیار دارد.

۳) ریکال هم اهمیت زیادی دارد. مثلا فرض کنید یه ماشین تشخیص باران داریم که بسیار حساس است و با کوچکترین باد یا ابری تشخیص باران میدهد. خوب precision این ماشین بالاست یعنی زمانی که واقعا باران می‌آید ماشین با احتمال بالایی آن را تشخیص میدهد اما ریکال آن پایین است یعنی خیلی وقت ها بارانی در کار نیست و اما ماشین به ما هشدار باران میدهد.

۴) خوب مطمئنا در این حالت که این کلمه در زمان train شدن تنها در یک دسته بندی آمده باعث میشه که احتمال آمدن این کلمه به شرط بقیه دسته بندی ها برای ما صفر باشد (چون ما آن را در باقی دسته بندی ها ندیدیم آن را صفر اندازه گیری کردیم) در نتیجه در این حالت خروجی ماشین ما همان دسته بندی است که کلمه در آن آمده و احتمال آن هم ۱۰۰ درصد است.