به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

# هوش مصنوعی

## پروژه سوم - Naive Bayes Classifier

مهلت تحویل: ۲۸ آبان ۱۳۹۹

طراحان: مهدیه بیاری‌نژاد، ژیوار صورتی، محسن فیاض

بهار ۹۹

## مقدمه:

در این پروژه قصد داریم با استفاده از Naive Bayes Classifier به تجزیه و تحلیل نگرش[1] کامنت‌ها بپردازیم و مثبت یا منفی بودن نظرات را طبق متن آنها تشخیص دهیم.

## معرفی مجموعه داده:

مجموعه داده دیجیکالا در فرمت csv در اختیار شما قرار گرفته است. در این داده، تیتر و متن نظرات و اینکه نویسنده آن نظر، کالا را پیشنهاد می‌کند یا نه مشخص شده است.

| | title | comment | recommend |
|---|---|---|---|
| 0 | زیبا اما کم دوام | با وجود سابقه خوبی که از برند ایرانی تهرین سرا... | not_recommended |
| 1 | بسیار عالی | بسیار عالی | recommended |
| 2 | سلام | برای کسایی...\r\nمن الان ۳ هفته هست استفاده میکنم | not_recommended |
| 3 | به درد نمیخورههههه | عمرش کمه تا یه هفته بیشتر نمیشه استفاده کرد یا... | not_recommended |
| 4 | کلمن آب | فکر کنین کلمن بخرین با ذوق. کلی پولشو بدین. به... | not_recommended |

دو فایل در اختیار شما قرار گرفته است که یکی برای آموزش و دیگری برای ارزیابی مدل شما است. فایل مربوط به آموزش مدل به عنوان comment_train.csv و همینطور فایلی که مربوط به ارزیابی مدل شما است با نام comment_test.csv در اختیار شما قرار گرفته‌است.

[1] Sentiment Analysis

## فاز اول: پیش‌پردازش داده

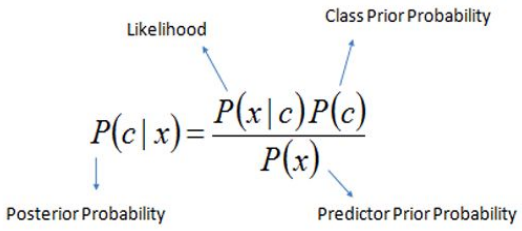
در فاز اول باید اطلاعات متنی داخل مجموعه داده را برای تحلیل های بعدی پیش‌پردازش کنیم. برای این کار می‌توانید از کتابخانه‌ی هضم[2] استفاده کنید یا خودتان موارد مورد نیازتان را پیاده سازی کنید. شما باید متن و تیتر هایی که موجود است را تا حد ممکن Normalize کنید. (روش های ممکن، شامل حذف کلمات پرتکرار یا همان stop words، تبدیل کلمات به ریشه آنها و ... است.)

دقت کنید که این کار هم روی داده های train و هم روی داده های test باید انجام شود و لزوما اجرای هر نوع پیش‌پردازشی باعث بالا رفتن دقت مدل شما نخواهد شد. روش های متفاوت را با استفاده از کتابخانه یا بدون آن امتحان کنید و ترکیب هر کدام از آنها که به مدل شما بیشتر کمک می‌کرد را اجرا کنید.

1) در گزارش کار خود، جایگزین کردن کلمات با روش stemming یا lemmatization را توضیح دهید و تاثیر آن را تحلیل کنید.

## فاز دوم: فرآیند مسئله

در این مسئله می‌خواهیم با استفاده از Naive Bayes براساس متن و تیتر کامنت‌ها، تشخیص دهیم که نویسنده آن، کالا را پیشنهاد کرده یا پیشنهاد نکرده است. در این مسئله از مدل bag of words استفاده می‌کنیم. به این صورت که هر کلمه را مستقل از جایگاهش در جمله در نظر می‌گیریم. feature های این مسئله را تعداد هر کلمه در کلاس مربوطه در نظر بگیرید. یعنی هر چه تعداد یک کلمه در یک کلاس بیشتر باشد، احتمال اینکه آن کلمه متعلق به آن کلاس باشد بیشتر است. برای حل این مسئله به صورت کلی از naive bayes استفاده می‌کنید که مفهوم پشت آن با توجه به مفاهیم احتمالی زیر قابل بحث است.




$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

[2] https://github.com/sobhe/hazm

2) در گزارش کار خود، توضیح دهید که هر کدام از (evidence, likelihood, prior, posterior) بیانگر چه مفهومی در این مسئله هستند و چگونه محاسبه می‌شوند.

دقت کنید که نیازی نیست عبارت Evidence در مخرج کسر به صورت مستقیم محاسبه شود.

برای محاسبات، می‌توانید دو ستون تیتر و کامنت را یکی کنید که در آن صورت یک ستون برای داده خواهید داشت و یک ستون recommended. می‌توانید یکی از دو ستون تیتر یا کامنت را کلا نادیده بگیرید، یا می‌توانید با وزن دادن به کلمات هر کدام از آنها، یکی را بر دیگری اولویت دهید. فقط توجه کنید که در بخش ارزیابی، باید دقت شما روی داده‌ی تست از حداقل گفته شده بیشتر باشد.

## Additive Smoothing

ممکن است در کامنت‌هایی که مربوط به دسته recommended هستند، کلمه‌ای وجود داشته باشد که در کامنتهای not_recommended نباشد و بالعکس، یا حتی کلمه‌ای در کامنت جدیدی که می‌خواهیم بررسی کنیم باشد که در هیچکدام از دو کلاسی که در داده train دیدیم نباشد. اگر مثلا کلمه‌ی "دیجی‌کالا" در دسته recommended باشد ولی در دسته not_recommended نباشد، مدل ما با قطعیت تشخیص می‌دهد که هر کامنتی که در متن آن کلمه "دیجی‌کالا" وجود دارد متعلق به کلاس recommended است در حالیکه می‌دانیم اینطور نیست.

3) در گزارش خود با در نظر داشتن naive bayes توضیح دهید چرا این اتفاق رخ می‌دهد.

4) درباره روش Additive Smoothing تحقیق کنید و با پیاده‌سازی آن در پروژه، این مشکل را برطرف کنید. در گزارش خود این روش را توضیح دهید و بگویید چطور به حل این مشکل کمک می‌کند.
(در بخش ارزیابی، تفاوتی که استفاده از این روش بر دقت می‌گذارد را باید گزارش کنید)

## فاز سوم: ارزیابی

برای ارزیابی مدل خود باید از 4 معیار زیر استفاده کنید.

$$Accuracy = \frac{Correct\ Detected}{Total}$$

$$Precision = \frac{Correct\ Detected\ Recommended}{All\ Detected\ Recommended\ (Including\ Wrong\ Ones)}$$

$$Recall = \frac{Correct\ Detected\ Recommended}{Total\ Recommended}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Correct Detected Recommended: تعداد کامنت هایی که به درستی recommended تشخیص داده شده.

All Detected Recommended: تعداد تمام کامنت هایی که recommended تشخیص داده شده.

Total Recommended: تعداد تمام کامنت هایی که در مجموعه داده تست recommended بودند.

5) در گزارش کار خود توضیح دهید که چرا مقدار Precision و Recall هر کدام به تنهایی برای ارزیابی مدل کافی نیست؟ برای هر کدام مدلی را مثال بزنید که در آن، این معیار مقدار بالایی دارد ولی مدل خوب کار نمی‌کند.

6) در گزارش کار خود توضیح دهید معیار F1 از چه نوع میانگین‌گیری بین precision و recall استفاده می‌کند؟ علت آن به نظر شما چیست؟

مدل خود را که با استفاده از naive bayes و براساس داده‌ی train ساخته‌اید، روی داده‌ی تست که در اختیارتان قرار دارد اجرا کنید و برای هر کدام از سطرهای آن تشخیص مدلتان را بدست آورید. سپس براساس آن، معیارهای بالا را محاسبه کنید.

7) در گزارش کار خود، در یک جدول، مقدار 4 معیار گفته شده را برای 4 حالت زیر به دست بیاورید.

a) استفاده از پیش‌پردازش داده ها و Additive Smoothing

b) استفاده تنها از Additive Smoothing

c) استفاده تنها از پیش پردازش داده ها

d) استفاده نکردن از هیچ‌کدام از پیش‌پردازش داده ها و Additive Smoothing

مقدار accuracy و F1 در حالت a باید به ترتیب بیش از 86 و 86 باشند.

8) در گزارش خود، مقادیر بدست آمده در بخش قبل را تحلیل کنید.

9) در گزارش خود 5 مورد از کامنت‌هایی که در داده‌ی تست هستند و مدل شما آنها را به اشتباه تشخیص داده است بیاورید. (در حالتی که از پیش‌پردازش و smoothing استفاده کردید)

به نظر شما چه بخش یا بخش‌هایی از راه حلی که پیش گرفتیم باعث شده این موارد اشتباه تشخیص داده شوند؟

## نکات پایانی:

- توجه داشته باشید که چون مجموعه داده متوازن است، مدل رندوم می‌تواند به دقت حدود 50 درصد برسد.
- دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده بطور مثال نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید.
- نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA3_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی های خواسته شده بخشی از نمره این تمرین را تشکیل می‌دهد. از نمایش درست خروجی های مورد نیاز در فایل html مطمئن شوید.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت توسط ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.