به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

درس هوش مصنوعی

پروژه سوم

پردازش متن و شبکه های بیزین

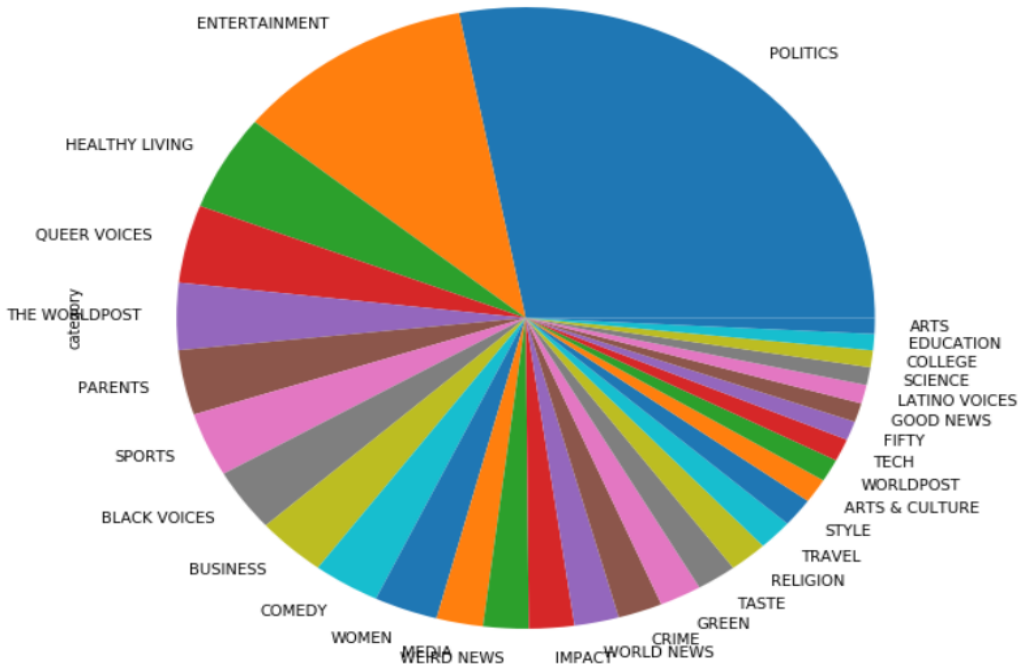
مهلت ارسال تا ۷ اردیبهشت

طراحان پروژه: تینا بهزاد ـ حمید تراشیون

بهار ۹۹-۹۸

مقدمه

در دنیای امروز که همه‌ی صنعت‌ها به سمت استفاده از هوش مصنوعی می روند، پردازش زبان‌های طبیعی و به عنوان یکی از مشتقات آن دسته‌بندی متن‌ها جایگاه مهمی پیدا کرده است. از جمله کاربردهای آن میشود به دسته‌بندی موضوعی کتاب‌های کتابخانه‌ها اشاره کرد که به کمک ابزارهای هوش مصنوعی در کمترین زمان و با کمترین نیروی انسانی انجام میشود. بحث استفاده از هوش مصنوعی برای دسته‌بندی متن مدت‌هاست که وجود دارد، از قدیمی ترین موارد استفاده‌ی آن میتوان به دسته‌بندی ایمیل‌ها و لیبل دادن به آن‌ها در Gmail اشاره کرد.
در عصر امروز که بازاریابی از موتورهای جست‌وجو به شبکه‌های اجتماعی رفته است که در آن بین برندها و مشتریان گفت‌وگو شکل میگیرد هم از دسته‌بندی متن‌ها برای شناخت بیشتر مشتریان با توجه به متن آنچه در شبکه‌های اجتماعی به اشتراک میگذارند و ارائه‌ی سرویس‌های شخصی‌سازی شده استفاده میشود. انتخاب کلیدواژه و تگ‌های مناسب برای متن بلاگ‌ها به گونه‌ای که به SEO کمک کند، سریع‌تر شدن کمک‌رسانی در شرایط اضطراری با تشخیص توییت‌هایی که متن آن‌ها نشان دهنده‌ی اضطرار و درخواست کمک است، دسته‌بندی پست‌ها در شبکه‌های اجتماعی نظیر اینستاگرام و بسیاری موارد دیگر استفاده‌ی آن نشان دهنده‌ی اهمیت این مسئله در دنیای هوش مصنوعی است.
کاربردی که در این پروژه می خواهیم آن را بررسی کنیم، دسته‌بندی موضوعی اخبار است. از سیستم‌های دسته‌بندی موضوعی اخبار میتوان برای دسته‌بندی اخبار در سایت خبرگزاری ها و دادن تگ مناسب به آن‌ها، جمع‌آوری خبرهای مربوط به یک دسته‌ی خاص برای کارهای تحقیقاتی، پیدا کردن ترندهای خبری در هر دسته از اخبار در شبکه‌های اجتماعی و... نام برد.
برای دسته‌بندی متن‌ها در کاربردهای صنعتی معمولا از روش‌های پیچیده‌تری استفاده میشود ولی در این پروژه خواهیم دید که با استفاده از مدل بیزین که در درس با آن آشنا شدید میتوان در دسته‌بندی متن اخبار به دقت خیلی خوبی رسید و در بسیاری از موارد اگر نیاز به سرعت بالای تشخیص در سیستم نباشد، استفاده از مدل بیزین به علت سادگی آن و همچنین دقت نسبتا خوب آن جز گزینه‌های روی میز است.



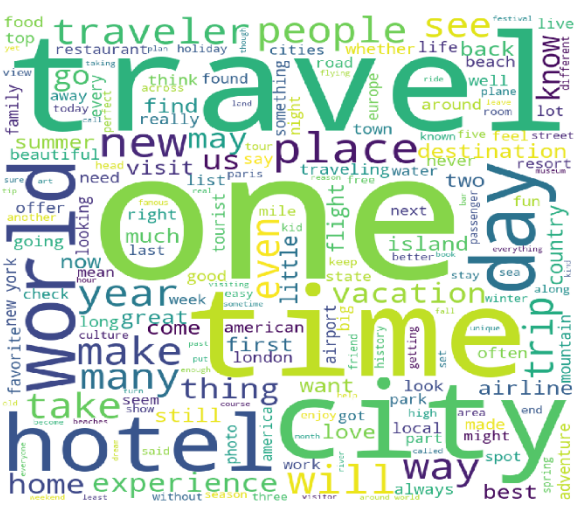
نمودار دایره‌ای سهم هر دسته از اخبار بر اساس داده های مربوط به یک خبرگزاری

تعریف مسأله

در این پروژه قرار است با استفاده از قاعده بیزین، با داشتن شرح کوتاهی از خبر بتوانیم دسته بندی موضوعی آن خبر را پیدا کنیم.

مسیر راه حل

راه حل در نظر گرفته شده برای حل این مسئله روش bag of words می باشد. به این صورت که هر کلمه از متن اخبار را به عنوان یک فیچر در نظر می گیریم و تعداد بار تکرار آن کلمه در متن اخبار را بدست می آوریم. در نهایت برای هر کلمه بدست می آوریم که به چه احتمالی در هر یک از دسته بندی های اخبار حضور دارند.
حال برای پیش بینی دسته بندی یک خبر جدید می توانیم به سراغ کلمات و احتمال وجود آنها در دسته های مختلف اخبار برویم و با استفاده از قاعده بیزین احتمال ها را محاسبه کرده و با هم مقایسه کنیم.
نمودارهای word cloud دید خوبی از کلیت کلمات موجود در هر دسته به شما نشان میدهند که میتواند در مسائل به شما کمک کند که آیا روش bag of words برای داده‌ای که در اختیار دارید مناسب است یا نه. به کمک آن ها میشود فهمید آیا واقعا کلمات استفاده شده در دسته‌های مختلف به قدری متفاوت هستند که به کمک احتمال حضور آنها در یک خبر بتوان دسته‌ی آن خبر را تشخیص داد یا خیر. در زیر نمودار مربوط به دو دسته از دسته‌های مورد نظر ما در این پروژه رسم شده است. نمودار سمت راست مربوط به دسته‌ی Business و نمودار سمت چپ مربوط به Travel. در این نمودارها کلمات بسته به فرکانس تکرارشان اندازه‌ای متناسب با دیگر کلمات میگیرند.



پیش‌پردازش داده

داده ای که برای این پروژه در فایل data.csv در اختیار شما قرار گرفته، شامل سطرهایی است که در هر یک از آن‌ها اطلاعات مربوط به یک خبر قرار گرفته‌است. اطلاعات شامل دسته‌ای که خبر در آن قرار میگیرد، تیتر ، نویسندگان ، لینک، توضیحی کوتاه و تاریخ انتشار خبر است. آنچه در این پروژه انتظار میرود تشخیص دسته‌ی اخبار تنها با ستون مربوط به توضیح کوتاه(short description) در مورد خبر است و با استفاده از اطلاعات موجود در همین ستون میتوان به دقت لازم رسید و نیازی نیست در روابط بیزین مورد استفاده احتمالات مربوط به اطلاعات ستون های دیگر را درنظر بگیرید ولی استفاده از آن‌هابلامانع است و می توانید برای بالا بردن دقت مدل خود، احتمال های مربوط به آن‌ها را هم در محاسبات خود وارد کنید.
دسته‌هایی که برای این پروژه انتخاب شدند و قرار است در ادامه سیستم شما آن‌ها را تشخیص دهد سه دسته‌ی BUSINESS، STYLE & BEAUTY و TRAVEL است و داده‌ای که در اختیار شما قرار گرفته تنها شامل همین دسته‌هاست.

مرحله‌ی اول هر پروژه‌ی پردازش متن آماده کردن داده برای یادگیری سیستم است. در این مرحله لازم است توضیح خبرهای مختلف را تا جایی که میتوانید Normalize کنید.این فرایند باید حداقل شامل موارد زیر باشد:

- حذف کلمات پرتکرار از متن‌ها مانند this، that، the ـ که به آنها stop words میگویند ـ.
- تبدیل همه‌ی حروف بزرگ به کوچک
- حذف علائم نگارشی
- جایگزین کردن کلمات با ریشه‌ی آن ها با روش‌هایی نظیر stemming یا lemmatization

هر چه بیشتر بتوانید متن‌ها را یکپارچه کنید، در انتها خروجی دقیق‌تری کسب خواهید کرد . برای این کار میتوانید از کتابخانه‌ی nltk استفاده کنید.
در گزارش خود در مورد تاثیر داشتن یا نداشتن:

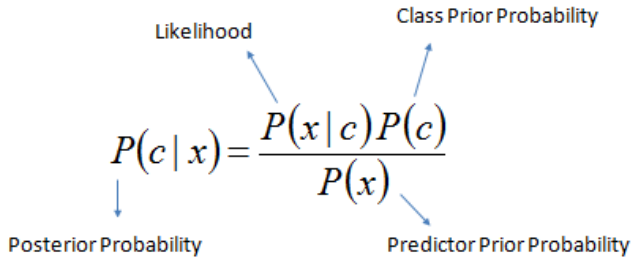
۱.کوچک کردن حروف بزرگ

۲. جایگزین کردن کلمات با ریشه‌ی آن‌ها در به کمک هر یک از روش‌های stemming یا lemmatization

توضیح دهید و این تاثیرات را تحلیل کنید.

قاعده ی بیزین

پس از پیش‌پردازش متن اخبار، داده آماده است تا به مدل داده شود. نیاز است پیش از هر کار، مشخص کنید مقادیر مورد نیاز برای استفاده از قاعده‌ی بیزین شامل (posterior، prior، likelihood، evidence) در برنامه ی شما بیانگر چه هستند و چگونه محاسبه میشوند. حتما در گزارش خود آن‌ها را بنویسید.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

توجه کنید که نیازی نیست عبارت evidence در مخرج کسر را به صورت مستقیم حساب نکنید.

فرآیند مسئله

در فاز اول می خواهیم میان دسته های TRAVEL و BUSINESS پردازش و پیش بینی کنیم و بعد از بدست آوردن نتیجه ی دلخواه میان این دو دسته، در فاز دوم دسته ی سوم "BEAUTY & STYLE" را هم به دو دسته ی قبلی اضافه می کنیم و دوباره پیش بینی را انجام می دهیم.

برای انجام دو فاز بالا لازم است داده‌هایی که در اختیار دارید را به دو دسته‌ی یادگیری[1] و ارزیابی[2] تقسیم کنید. این کار از آن جهت است که مدل را با داده‌ی دسته‌ی یادگیری بسازیم و به کمک داده‌ی دسته‌ی ارزیابی آن را بسنجیم چرا که داده‌هایی که مدل با آن‌ها آموزش دیده داده‌های مناسبی برای سنجیدن مدل نیستند و برای اینکه ببینیم آیا واقعا مدل توانایی تشخیص را دارد یا نه باید داده‌هایی به آن بدهیم که پیش از آن ندیده است. معمولا بخش عمده‌ی داده (حدود 80 درصد) برای آموزش و باقی آن برای ارزیابی استفاده میشود. توجه کنید که بهتر است از هر دسته ۸۰ درصد را جدا کنید و در کنار هم قرار بدهید نه اینکه از کل داده مستقل از دسته بندی ها ۸۰ درصد را جدا کنید. (چرا؟) برای ارزیابی مدل توسط داده های دسته‌ی ارزیابی از روابط بخش بعد استفاده می کنیم.

معیار ارزیابی مدل

$$\text{Recall} = \frac{\textit{Correct Detected Category}}{\textit{All Category}}$$

$$\text{Precision} = \frac{\textit{Correct Detected Category}}{\textit{Detected Category (This includes wrong detections)}}$$

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

Correct Detected Category: تعداد خبر هایی که مدل شما به درستی آن را" Category" تشخیص داده است.
Detected Category : تعداد خبر هایی که مدل شما آن را جزو دسته "Category" تشخیص داده است.
Correct Detected : تعداد خبرهایی که مدل شما کلن دسته بندی خود را درست تشخیص داده است.

Total : تعداد کل اخبار

- برای فاز اول که پیش بینی میان دو دسته ی "TRAVEL" و "BUSINESS" می باشد، مقادیر مطلوب برای precision، recall و accuracy بالای 77 درصد مطلوب می باشد.
- برای فاز دوم که پیش بینی میان سه دسته ی "TRAVEL" ، "BUSINESS" و "STYLE & BEAUTY" می باشد، مقادیر مطلوب برای precision، recall و accuracy بالای 70 درصد مطلوب می باشد.
- دقت کنید که مدل رندوم دقت حدود 50 درصد دارد و ممکن است به طور اتفاقی 60 درصد هم بشود. پس رسیدن به درصد 55 به عنوان مثال، ارزشی ندارد.
- برای مدلی که با سه دسته آموزش دیده confusion matrix را در گزارش خود بیاورید. در مورد اینکه این ماتریس چیست و هر خانه‌اش بیانگر چه چیزیست جست‌وجو کرده و در گزارش خود ذکر کنید.
- در گزارش خود دو جدول مانند زیر برای دو فاز مختلف باید داشته باشید.

---

[1] Train
[2] Evaluation

phase1	Travel	Business
Recall		
Precision		
Accuracy		

phase2	Travel	Business	Style & Beauty
Recall			
Precision			
Accuracy			

نمونه برداری

در مرحله‌ی ارزیابی اگر recall و precision دسته‌های مختلف را در خروجی چاپ کنید ممکن است متوجه تفاوت زیادی بین این مقادیر در دسته‌های مختلف شوید، برای حل این مساله در مورد oversampling جست‌وجو کنید و داده‌های مربوط به دسته‌ها را مقایسه کنید. سپس سعی کنید این مشکل را حل کنید به گونه‌ای که تفاوت بین recall و precision در دسته‌های مختلف زیاد نباشد. در گزارش خود در مورد نحوه‌ی حل این مساله و نتیجه‌ی آن توضیح دهید.

ارزیابی نهایی

یک فایل به نام test.csv در کنار فایل‌های مربوط به پروژه آپلود شده اند. این فایل حاوی سطرهایی از اخبار مانند داده‌ی اولیه‌ای که در اختیارتان گذاشتیم است با این تفاوت که دسته‌ی این خبرها مشخص نیست. دسته‌ی خبرهای مربوط به فایل test را با مدلی که برای سه دسته آموزش دیده تشخیص دهید و در فایلی با نام output.csv با ستون‌های index، category در کنار کدها و گزارش خود آپلود کنید.
توجه کنید که مدلی که برای تشخیص دسته‌ها از آن استفاده می کنید همان مدلی باشد که با داده‌های دسته‌ی train اموزش دیده و داده‌های ارزیابی را با آن تشخیص دادید و برای تشخیص دسته‌ی این خبرها مدل جدیدی با کل داده‌های داده شده آموزش ندهید!

سوالات

۱. برای پیدا کردن ریشه‌ی کلمات روش‌های متفاوتی وجود دارد که دو تا از آن‌ها stemming و lemmatization هستند که هر دو در کتابخانه‌ی nltk پیاده‌سازی شدند. در مورد تفاوت‌های آن‌ها جست‌وجو کنید و ببینید استفاده از کدام روی داده‌های شما نتیجه‌ی بهتری میدهد.
۲. یکی از شاخص‌هایی که در پردازش متن بسیار مورد استفاده قرار میگیرد، شاخص tf-idf است. در مورد نحوه‌ی محاسبه‌ی این شاخص توضیح دهید و بگویید اگر قرار بود تشخیص سیستمتان را با استفاده از این شاخص بهتر کنید، چگونه در روابط بیزین از آن استفاده می کردید.
۳. چنانچه برای ارزیابی یک مدل ماشین لرنینگ فقط به مقدار precision توجه شود، چه مشکلی پیش می آید؟ برای مثال یک مدل ماشین لرنینگ معرفی کنید که precision بالایی دارد ولی خوب کار نمی کند.

۴. اگر در متن یکی از خبرهای فایل تست واژه‌ی Tabriz بیاید و این کلمه در داده‌های مربوط به یادگیری تنها یکبار و فقط در یکی از دسته‌ها آمده باشد ، چه اتفاقی می افتد و سیستم شما چه تشخیصی خواهد داد؟ برای پاسخ به این سوال به نحوه‌ی محاسبه‌ی احتمال بودن خبر در هر یک از دسته‌ها توجه کنید.

ملاحضات

- موعد تحويل غيرحضوري تا پايان روز يكشنبه ۷ ارديبهشت ماه مي‌باشد.
- تمامي نتايج بايد در يك فايل فشرده با عنوان AI_CA3_<#STID>.zip تحويل داده شود. اين فايل بايد شامل موارد زير باشد:
  - يك پوشه به نام Code شامل كدهاي تمام قسمت‌هايي از تمرين كه پياده‌سازي نموده‌ايد.
  - فايل output.csv بدست آمده از ارزيابي نهايي
  - گزارش پروژه با فرمت PDF و شامل شرح تمامي كارهاي انجام شده، نتايج به دست آمده و تحليل‌ها و بررسي‌هاي خواسته شده در صورت پروژه.
  - درصورتی که از Jupyter Notebook استفاده می‌کنید نیازی به ارسال جداگانه کدها و گزارش نیست و هردو را می‌توانید در یک فایل Notebook ارائه دهید. حتما خروجی html فایل Notebook خود را نیز همراه فایل Notebook ارسال کنید.
- توجه داشته باشيد كه علاوه بر ارسال فايل‌هاي پروژه، اين پروژه به صورت حضوري نيز تحويل گرفته خواهد شد. بنابراين تمام بخش‌هاي پروژه بايد قابليت اجراي مجدد در زمان تحويل حضوري را داشته باشند. همچنين درصورت عدم حضور در تحویل حضوری نمره‌ای دریافت نخواهید کرد.
- هيچگونه شباهتي در انجام اين پروژه بين افراد مختلف پذيرفته نمي‌شود. در صورت كشف هرگونه تقلب براي همه افراد متقلب نمره ۱۰۰- در نظر گرفته مي‌شود.
- استفاده از مراجع با ارجاع به آنها بلامانع است. اما در صورتي كه گزارش شما ترجمه عيني از آن‌ها باشد، يا از گزارش افراد ديگر استفاده كرده باشيد كار شما تقلب محسوب می‌شود.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کننده، در غیر این صورت به طراحان پروژه ایمیل بزنید یا حضوری از یکی از آن‌ها بپرسید.

htarashion@gmail.com
tina.behzad@gmail.com

موفق باشید!