تمرین Text Classification با Naive Bayes
آرش محرابی

لینک کد:
https://github.com/arash-mehrabi-z/data-mining/blob/main/naive_bayes/naive_bayes.ipynb

توضیحات درباره Crawl:
برای درخواست HTTP دادن از کتابخانه Requests پایتون و برای scrape کردن html از BeautifulSoup4 استفاده شده است.
به صورت اتوماتیک به تمام صفحات mstajbakhsh.ir درخواست می‌زند و پست های هر صفحه را برمی‌دارد و جداگانه به‌به آن پست ها هم درخواست می زند.
در پست ها نوشته هایی که در post-body هستند را به عنوان text بر میگردد و کتگوری ها را به عنوان class label .

توضیح مدل:
از تابع کتابخانه ای TfidfVectorizer برای درآوردن feature و vectorize کردن اطلاعات استفاده می‌کند و از تابع MultinomialNB برای مدل Naive Bayes.

تحلیل داده ها:
من حالت‌های مختلفی را امتحان کردم و نتایج مختلفی را گرفتم که به ترتیب توضیح می دهم.

۱. حالت یک:
در داده‌های سایت بعضی text ها چندین کلاس مثل (Tutorial, AndroJava و ...) به صورت همزمان دارند و بعضی‌ها تنها یک کلاس.
در این حالت برای تکست هایی که چند لیبل داشت من فقط لیبل اول را استفاده کردم چون با بررسی چشمی می‌توان نتیجه گرفت که تقریباً همیشه لیبل اول مرتبط ترین به متن پست ها بود.
در نتیجه تعداد تکست های ما دقیقاً برابر تعداد پست های سایت یعنی ۵۵ تا بود.
برای آموزش داده‌ها من از کل داده‌ها استفاده کردم و برای تست کردن ۱۰ پست اخیر وبسایت را انتخاب کردم. در نتیجه دیتای تست را مدل قبلاً دیده بود. نتایج به دست آمده بسیار خوب بود و از ۱۰ تا دیتای ترین ۹ تا را درست تشخیص داده بود.
دیتایی که اشتباه تشخیص داده یک متن مربوط به Home Assistant است که به جای کلاس IoT آن را Anonymity Networks تشخیص داده که احتمالاً به خاطر کم بودن دیتای ترین از کلاس IoT می باشد.

۲. حالت دو:
در این حالت همان سیاست استفاده از لیبل های حالت قبلی استفاده شده است.
اما در این حالت دیتای ترین را مدل قبلاً ندیده بود و فقط یکی از ۶ تا دیتای تست را درست تشخیص داده. تستی که درست تشخیص داده مربوط به Hidden Mail Service است که لیبل Anonymity Networks دارد و دلیل اینکه آن را تشخیص داده احتمالاً به خاطر استفاده از واژه‌هایی است که در دیگر پست های مربوط به این کاتگوری هم استفاده شده است مثل Tor, Hidden, server و ... .

۳. حالت سه:
در این حالت من همه لیبل ها را استفاده کردم و تکست هایی که چندین لیبل داشتند، به ازای هر لیبل یک بار دیگر text را وارد دیتاست کردم. پس مثلاً اگر تکست « Einstein was a great scientist. >> با دو کاتگوری academic و personal آمده باشد، انگار دو تکست داریم که متن هر کدامشان یکی است اما لیبل هایش متفاوت. طبیعتاً در این حالت چون از لیبل های زیادی استفاده می‌کنیم تعداد داده هایمان به جای ۵۵ تا ۱۳۲ تاست و لیبل های پرت زیادی وجود دارد به این معنی که ممکن است از یک لیبل خاص فقط یک تکست وجود داشته باشد که طبیعتاً عمل ترین کردن را خیلی سخت می کند. انتظار داریم که در این حالت درصد موفقیت به طور محسوسی پایین‌تر بیاید.
برای عمل ترین، دوباره مثل حالت یک در این حالت هم مدل قبلاً دیتای تست را دیده.

برای بررسی عمل کرد در این حالت من اینطور بررسی کرده‌ام که از پست هایی که چندین لیبل دارند حتی اگر یک لیبل را هم درست بگوید، آن را به عنوان عمل‌کرد درست می پذیرم. با این فرض می‌بینیم که در این حالت عمل‌کرد ۱۰۰ درصدی را شاهدیم. علت آن هم این است که طبیعتاً وقتی برای یک تکست چند جواب درست وجود داشته باشد، شانس اینکه مدل بتواند آن را درست تشخیص دهد بالاتر می رود.

۴. حالت چهار:
در این حالت همان سیاست استفاده از لیبل های حالت قبلی استفاده شده است.
اما در این حالت دوباره دیتای ترین را مدل قبلاً ندیده بود. می‌بینیم که در این حالت مدل موفق نشده که هیچ کدام از دیتای تست را به درستی تشخیص دهد. به طرز عجیبی می‌بینیم همه دیتای تست را که ۶ پست اخیر وبلاگ هستند را Tutorial پیش‌بینی کرده. اگر به نحوه distribution کاتگوری ها نگاه کنیم می‌بینیم که Tutorial با ۱۸ پست بعد از Personal با ۱۹ پست بیشترین تعداد پست ها را دارد، اما سایر کاتگوری ها مثل AndroJava و Anonymity Networks نیز تعداد خوبی پست دارند پس دیتا نسبتاً خوب توزیع شده و مدل می بایست خوب توزیع شده باشد اما عمل‌کرد مدل در این حالت عجیب است.

نتیجه گیری:
در حالت‌هایی که مدل قبلاً دیتای تست را به عنوان دیتای آموزش دیده بود، عمل‌کرد بسیار عالی (۹۰٪ و ۱۰۰٪) را داشتیم اما در حالت‌هایی که دیتای تست جدید بود عمل‌کرد بسیار بد (۱۶٪ و ۰٪) را دیدیم.
دلیل این را هم باید در کمی دیتای آموزش و پخش بودن زیادی کلاس لیبل ها جستجو کرد. برای بهبود عمل‌کرد مدل توصیه می‌شود که کاتگوری ها را از سایت برنداریم بلکه با دیتایی که داریم خودمان تعداد کمتری کلاس لیبل (مثلاً ۵ تا) در نظر بگیریم و به صورت supervised به هر تکست یک لیبل اختصاص بدهیم. پیش‌بینی می‌کنم که در این حالت عمل‌کرد مدل در حالت‌هایی که دیتای تست جدید هم باشد بسیار بهتر شود.