به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

پردازش زبان های طبیعی

تمرین دوم

مدرسین: دکتر فیلی، دکتر یعقوب زاده

TA: امیر حسین فدایی

afadaei2@gmail.com

زمستان 1400

مقدمه

در انجام این تمرین کامپیوتری، با استفاده از Naïve Bayes تلاش خواهیم کرد تا Classification را بر روی دو دیتاست انجام بدهیم. در ابتدا بر روی یک دیتاست SMS های Spam و غیر Spam را از هم تفکیک می کنیم و در ادامه در یک دیتاست دیگر یک ابزار طبقه بندی برای تشخیص دروغ و حقیقت آموزش می دهیم. هدف از انجام این تمرین آشنایی بیشتر با این نوع طبقه بند، تمرین استخراج ویژگی های مناسب و آشنایی با ابزارهای موجود است.

دیتاست ها

دیتاست Spam collection SMS زیر را که در سایت Kaggle (لینک زیر) موجود است، دانلود کنید.

https://www.kaggle.com/uciml/sms-spam-collection-dataset

این دیتاست حاوی 5572 ردیف SMS است که Spam بودن یا نبودن آن ها مشخص شده است.

دیتاست دوم، دیتاست Sentimental LIAR است که می توانید آن را از لینک زیر دانلود کنید.

https://github.com/UNHSAILLab/SentimentalLIAR

پیش پردازش ها

در ابتدا پیش پردازش های لازم برای تولید یک مدل زبانی از این SMS ها را مطابق آنچه در بخش های قبل آموخته اید، انجام دهید. این پیش پردازش های می تواند حاوی حذف Stop Word ها، Tokenization، حذف Punctuation ها، Normalization و ... باشد. توجه داشته باشید که وجود یا عدم وجود برخی Punctuation ها می تواند در طبقه بندی موثر باشد (مثلا وجود ! می تواند یک نشانگر مناسب برای Spam باشد)، پس اگر قصد استفاده از این ویژگی ها برای طبقه بندی را دارید، این اطلاعات را برای مراحل بعدی نگه دارید.

برای انجام اکثر این پیش پردازش ها می توانید از ابزارهای موجود مانند NLTK نیز استفاده کنید:

https://www.nltk.org/

استخراج ویژگی ها

برای هر یک از دو طبقه بند زیر نیاز هست تا تعدادی ویژگی از متن استخراج کنیم. در هر مورد، تعداد و نوع ویژگی های استخراج شده کاملا اختیاری است و نمره این قسمت با توجه به ابتکار شما در انتخاب ویژگی های مناسب با توجه به هدف طبقه بند در نظر گرفته می شود.

1) طبقه بند SMS های Spam از SMS های معمولی. برای اینکار از دیتاست UCIML (دیتاست اول) استفاده خواهیم کرد.
2) طبقه بند دروغ و حقیقت. برای اینکار از دیتاست Sentimental LIAR (دیتاست) دوم استفاده خواهیم کرد.

متن ها در ستون statement و label آن ها بر اساس رای مردم در ستون label موجود است. برای سادگی label های barely true، pants fire و FALSE را دروغ و label های half-true ، mostly-true و TRUE را به عنوان حقیقت در نظر بگیرید. برای انجام این طبقه بندی اجازه دارید که ستون های sentiment (مثبت یا منفی بودن)، و ستون های anger، fear، joy، disgust و sad که نرخ احساسات را نشان می دهند نیز به عنوان ویژگی در آموزش طبقه بند (به ابتکار خود) استفاده کنید. از ستون های دیگر این دیتاست استفاده نکنید.

آموزش طبقه بندها

برای آموزش طبقه بندها می توانید از ابزار های آماده زیر استفاده کنید:

https://scikit-learn.org/

https://www.nltk.org/

در آموزش دیتاست UCIML، 80% داده را به عنوان داده آموزش و مابقی را به عنوان داده تست در نظر بگیرید. در آموزش دیتاست Sentimental LIAR از فایل train برای آموزش و از فایل test به عنوان داده تست استفاده کنید.

ارزیابی

در این مرحله برای هر طبقه بند آموزش داده شده، معیار های Accuracy، Recall، Precision و امتیاز F1 را گزارش دهید.

گزارش نهایی

ارزیابی اصلی پروژه شما بر اساس گزارش پروژه انجام خواهد شد. در این گزارش نیاز است تا موارد زیر توضیح داده شود:

1) تحلیل تاثیر پیش پردازش ها بر روی طبقه بندی

2) توضیح ویژگی های انتخاب شده برای هر طبقه بند و دلیل انتخاب آن

3) گزارش نتایج طبقه بندی و تحلیل آن

ملاحظات (حتما مطالعه شود):

- تمامی کد ها و گزارش مربوطه بایستی در یک فایل فشرده با عنوان NLP_CA2_StudentID تحویل داده شود.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه کد های پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرای مجدد آنها نیاز به تنظیمات خاصی می باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخهای ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش میگردد.
- در صورت وجود هر گونه سوال، می توانید با دستیار آموزشی مربوطه در تماس باشید.