به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش متن و زبان طبیعی

تمرین 2

فروردین ماه 1400

فهرست

مقدمه

در این تمرین قصد داریم با دو طبقه بند naïve Bayes و logistic regression شناسایی تنفر را بر روی داده‌های تمرین اول به دست آوریم. دادگان از شبکه اجتماعی توئیتر جمع‌آوری گردیده و کلاس‌های دادگان به صورت دوتایی می‌باشد (هر پیام، تنفر می‌باشد یا نه)

دادگان	داده‌های آموزشی		داده‌های تست	
	تنفر	غیر تنفر	تنفر	غیر تنفر
Hate Eval	10000		3000	
	4210	5790	1260	1740

1-پیش پردازش

همانند اکثر فعالیت‌های پردازش زبان طبیعی می‌بایست بر روی داده‌ها پیش پردازش‌هایی صورت بگیرد.

برای این تمرین می‌بایست پیش پردازش‌های زیر را انجام دهید:

- حذف ایموجی‌ها
- حذف لینک‌ها
- باز کردن هشتگ‌ها

دقت فرمائید داده‌های Hate Eval به صورت متوازن نیستند و همین امر می‌تواند موجب overfitting شود به همین دلیل لازم است قبل از انجام پیش پردازش داده‌ها را متوازن کنید. روش‌های متفاوتی را که برای متوازن کردن داده‌های متنی وجود دارد شرح دهید و حداقل یک مورد را پیاده‌سازی نمائید.

2-استخراج ویژگی‌ها

برای آن‌که بتوانیم یک رده بند را آموزش دهیم لازم است تعدادی ویژگی را استخراج نمائیم. تعداد و نوع ویژگی‌های انتخاب شده بر عهده خود شما می‌باشد می‌بایست به ازای هر ویژگی که تعریف می‌کنید دلیل انتخاب آن را نیز مطرح کنید و تأثیر هر ویژگی را به تنهایی و با در نظر گرفتن سایر ویژگی‌ها بررسی کنید و بیان کنید کدام ویژگی‌ها تأثیر بیشتری بر روی دقت طبقه بند شما دارد.

دقت نمائید ابتکار شما در طراحی ویژگی‌ها می‌تواند تأثیر به سزایی در نمره شما داشته باشد. همچنین می‌توانید از n-gram که در تمرین اول پیاده‌سازی نموده‌اید به عنوان یکی از ویژگی‌ها استفاده نمائید.

3-آموزش رده بند

در این مرحله با استفاده از ویژگی‌های به دست آمده در مرحله قبل می‌بایست دو رده بند **naïve Bayes** و **logistic regression** را بر روی داده‌های train آموزش دهید. برای این منظور می‌توانید از کتابخانه‌های scikit-learn یا NLTK استفاده نمائید.

برای انجام رده بندی از روش k-fold استفاده نمائید و مقدار k را برابر 5 در نظر بگیرید.

آیا با استفاده از k-fold می‌توان متوجه شد که overfitting رخ داده است یا خیر؟ شرح دهید.

4- ارزیابی مدل آموزش دیده

در این مرحله می‌بایست رده بندی را که آموزش داده‌اید، **بر روی داده‌های تست** ارزیابی کنید. معیارهای ارزیابی را **Accuracy**، **precision**،**Recall** و **F1** در نظر بگیرید و به ازای هر طبقه بند آن‌ها را گزارش دهید. نتایج به دست آمده را تحلیل کنید. کدام یک از رده‌بندها عملکرد بهتری داشته است؟ دلیل آن را شرح دهید.

لازم است confusion matrix **را به ازای دو طبقه بند آموزش دیده** رسم نمائید.

همچنین خروجی به دست آمده را با خروجی تمرین اول مقایسه کنید و دلایل تفاوت را ذکر نمائید.

ملاحظات (حتماً مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_HW2_StudendID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد. گزارش نهایی خود را حتماً به صورت PDFدر سایت درس بارگذاری نمائید.
- گزارش به صورت تایپ شده در قالبPDF شامل شرح آزمایش‌های انجام شده، پارامترهای آزمایش، نتایج و تحلیل‌ها باشد.
- مهلت تحویل تمرین به هیچ عنوان تمدید نخواهد شد. تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می‌شود که جریمه تأخیر تحویل تمرین تا یک هفته ۳۰ درصد است.
- توجه کنید این تمرین باید به صورت تک‌نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد . در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با این ایمیل در ارتباط باشید:

romina.oji@ut.ac.ir

مهلت تحویل بدون جریمه: سه شنبه 17 فروردین 1400

مهلت تحویل با تأخیر، با جریمه ۳۰ درصد: سه شنبه 24 فروردین 1400