به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۶

دی ماه ۱۴۰۲

*فهرست

بخش عملی – سوال ۱ (Categorization)

در این تمرین، وظیفه تمرین ۴ که با استفاده از BERT انجام شد را با استفاده از مدل‌های کلاسیک یادگیری ماشین انجام خواهید داد.

توجه داشته باشید که در این تمرین تعداد زیادی مدل ساده را آموزش خواهید داد و این موارد ممکن است زمان‌بر باشد. در مدیریت زمان، تحویل و انجام تکلیف، این مورد را در نظر بگیرید.

شرح دادگان

مجموعه داده این سوال، همان مجموعه داده نظرات پلتفرم گوگل پلی[1] بوده و وظیفه قابل انجام به صورت رده‌بندی سه رده‌ای (مثبت، منفی و خنثی) تعریف شده است. مجموعه دادگان این بخش به صورت فایل Zip در سامانه قابل دریافت است. مانند آنچه در تمرین ۴ گفته شد، احساسات نظرات را با استفاده از نمرات آن‌ها تقریب بزنید. نظرات با نمرات بالاتر از ۳، مثبت، نظرات با نمره ۳ به عنوان خنثی و نظرات با نمره پایین‌تر از ۳، نظرات منفی در نظر گرفته خواهند شد.

رده‌بندی با استفاده از Logistic Regression, SVM و Naïve Bayes

در ابتدا مطابق با آنچه در بخش شرح دادگان گفته شد، مجموعه داده را به فرم مورد نظر در بیاورید. سپس یک نمونه کوچک‌تر (حداقل ۵۰۰۰ مورد) را به گونه‌ای از مجموعه داده اصلی نمونه بگیرید که توزیع برچسب‌ها در مجموعه داده نمونه‌گیری شده، منطبق بر توزیع برچسب‌ها در مجموعه داده اصلی باشد.

الف) در مرحله بعد مجموعه اسناد را پیش‌پردازش کنید. برای این کار می‌توانید از تکنیک‌های مختلف پیش‌پردازش مانند حذف Stopwordsها، Stemming و دیگر روش‌های پیش‌پردازش استفاده کنید و دلیل خود را برای موارد مختلف با توجه به وظیفه مورد نظر که تشخیص احساسات است بیان کنید. یک نمونه از دادگان را پیش و پس از پیش‌پردازش در گزارش بیاورید و تغییرات را بیان کنید. پیشنهاد می‌شود برای این بخش از کتابخانه‌ی NLTK استفاده شود.

ب) یکی از مراحل مهم در وظایف رده‌بندی، انتخاب ویژگی‌هاست. به عنوان بخشی از این تمرین ویژگی‌های مختلفی را آزمایش خواهید کرد تا به بهترین ویژگی‌ها برای عملکرد بهتر مدل برسید. در ابتدا از تعداد وقوع برخی کلمات در متن برای رده‌بندی استفاده کنید و عملکرد آن را بر روی مدل‌ها گزارش

[1] Google Play

کنید. در مرحله‌ی بعد به جای تعداد وقوع کلمات از ویژگی tf-idf برای رده‌بندی استفاده کنید و تاثیر این دو ویژگی بر روی عملکرد مدل‌ها را با هم مقایسه کنید. در انتها سعی کنید حداقل یک ویژگی متنی دیگر که ممکن است به رده‌بندی بهتر مدل‌ها کمک کند را به عنوان روش سوم در نظر گرفته و عملکرد آن را بر روی مدل‌ها بررسی نمایید. در صورت زمان‌بر بودن آموزش مدل‌ها می‌توانید ابعاد ویژگی‌ها (max features) را محدود کنید. در نظر داشته باشید که این کار در کیفیت نهایی مدل شما تاثیر دارد و سعی کنید تا جای ممکن این کار را انجام ندهید.

رده‌بندهای Logistic Regression, SVM و Naïve Bayes را به عنوان رده‌بند خود در نظر گرفته، با روش 5-fold cross validation رده‌بندهای مختلف را بر روی داده‌های آموزشی آموزش دهید و نتایج را گزارش کنید. کدام رده‌بند و کدام روش بردارسازی عملکرد بهتری داشته است؟ با استفاده از نمودارهای مناسب، مقایسه و تحلیل کنید. برای ارزیابی مدل‌های خود از چهار معیار دقت[1]، صحت[2]، بازخوانی[3] و امتیاز F1[4] استفاده کنید و پیشنهاد کنید با توجه به ذات مسئله و توزیع برچسب‌ها در مجموعه داده، کدام معیار استفاده بهتری دارد؟

مقایسه عملکرد این مدل‌ها با نتایج تکلیف ۴ (BERT) و تحلیل نتایج نیز تاثیر مثبتی در نمره‌دهی و دید ذهنی شما خواهد داشت اما اجباری نیست.

راهنمایی: تقریبا تمامی موارد مورد استفاده در این سوال در scikit learn موجود هستند و در صورت استفاده از پردازنده‌های شرکت intel ، می‌توانید برای آموزش و اجرای سریع‌تر مدل‌ها از پچ ارائه شده توسط intel برای scikit learn استفاده کنید.

لازم به ذکر است فهم و درک مسئله و همین طور گزارش کامل در بخش‌های مختلف، بخش اصلی نمرات این بخش را تشکیل می‌دهند و به تمرین‌هایی که صرفا پیاده‌سازی کد است و یا نتایج بدون توضیح تشریحی هستند، هیچ نمره‌ای تعلق نخواهد گرفت.

سوال امتیازی (۲۰ نمره)

در این بخش قصد داریم وظیفه رده‌بندی بخش قبل را با استفاده از Few Shot Learning به کمک مدل‌های زبانی بزرگ انجام داده و نتایج را مقایسه کنیم. در نوع مدل زبانی و ابزار رده‌بندی محدودیتی وجود ندارد. با توجه به بزرگی این مدل‌ها، لازم نیست تمام مجموعه داده را برچسب بزنید و حداقل ۸۰

[1] Precision
[2] Accuracy
[3] Recall
[4] F1-Score

مورد کافی است. برای این وظیفه می‌توانید از API های OpenAI برای نسخه رایگان ChatGPT ، API های گوگل برای نسخه رایگان Gemini و یا Inference API های Huggingface برای مدل‌های زبانی متن‌باز استفاده کنید. توجه داشته باشید که در صورت استفاده از ChatGPT ، OpenAI محدودیت ۳ درخواست بر دقیقه را برای حساب‌های کاربری رایگان دارد و این مورد را می‌توانید با روش‌های مختلفی مثل timeout به چالش بکشید. پیشنهاد ما استفاده از مدل‌های متن باز مثل Zephyr و استفاده از Inference API در Huggingface است. این روش در عین آسان بودن پیاده سازی، محدودیت‌های کمتری هم دارد. می‌توانید از مدل‌های کوچک‌تر مثل phi-2 نیز استفاده کرده و آن‌ها را روی Google Colab اجرا کرده و بدون استفاده از API نتایج را به دست آورید.

الف) به کمک یکی از روش‌های گفته شده در بالا، رده‌بندی را انجام دهید. داده‌ها را به صورت خام استفاده می‌کنید یا پیش‌پردازش شده؟ دلایل خود را بیان کنید.

ب) نتایج رده‌بندی به دست آمده را با نتایج بخش قبل مقایسه کنید. احتمالا عملکرد رده‌بندی با معیارهای ما ضعیف‌تر از مدل‌های بخش قبل باشد. اما مدل‌های زبانی بزرگ، درک متن به مراتب بالاتری نسبت به مدل‌های کلاسیک و حتی BERT دارند. سعی کنید این تناقض را توجیه کرده و دلایل خود را با مثال برای نمونه‌هایی از مجموعه داده که خروجی مدل زبانی بزرگ با برچسب تعیین شده توسط ما تفاوت دارد ارائه دهید.

راهنمایی ۱: توجه داشته باشید که خروجی مدل‌های زبانی بزرگ همیشه قابل پیش‌بینی نیستند. برای آن‌که بتوانید رده پیش‌بینی شده را در خروجی مدل بیابید، باید خروجی مدل خود را بررسی کرده و با توجه به مدل مورد استفاده، به کمک روش‌هایی مثل محدود کردن شکل خروجی در Prompt ورودی (تاکید بر خروجی یک کلمه‌ای و یا فرمت json) می‌توانید این چالش را برطرف کنید. این سوال امتیازی، با هدف آشنایی شما با مبحث داغ و پرتقاضای مدل‌های زبانی بزرگ در دنیای پژوهش و صنعت طرح شده است. در صورت دیدن تلاش شما، مواردی مثل تازه بودن موضوع و فناوری در نمره‌دهی لحاظ خواهد شد 😊

راهنمایی ۲: به منظور علمی بودن تحلیل و مقایسه معنادار، بهتر است عملکرد مدل‌های کلاسیک را دقیقا روی داده‌هایی که با LLM رده‌بندی کرده‌اید به دست آورید و مقایسه را با استفاده از آن انجام دهید.

انجام وظیفه باید به صورت استفاده از API و یا استفاده از مدل زبانی لوکال با استفاده از کد باشد و به پاسخ‌هایی که به صورت دستی به مدل‌های زبانی بزرگ ورودی داده و خروجی گرفته‌اند نمره‌ای تعلق نخواهد گرفت.

بخش عملی – سوال ۲ (Topic Modeling)

در این تمرین سعی داریم با استفاده از روش LDA و کتابخانه genism ، به مدل‌سازی موضوعی مقالات دیجیکالامگ[1] پرداخته و دسته‌های موضوعی را پیدا کرده و با موضوعات اصلی مجموعه داده مقایسه کنیم.

شرح دادگان

دادگان مورد استفاده در این بخش، اطلاعات استخراج شده از سایت دیجیکالامگ است که توسط آزمایشگاه عزیز هوشواره تهیه و در دسترس عموم قرار داده شده است. این مجموعه داده شامل ۸۵۱۵ مقاله استخراج شده از دیجیکالامگ به همراه دسته‌بندی موضوعی آن‌هاست. مجموعه داده از اینجا قابل دریافت است.

مدل‌سازی موضوعی به کمک LDA

الف) از هر نوع برچسب موجود در فایل train.csv مجموعه دادگان ارائه شده، ۵۰ مورد را نمونه‌گیری کرده و با مجموعه داده حاصله تمرین را ادامه دهید. پیش‌پردازش های لازم برای مدل‌سازی موضوعی را با دلیل بررسی کرده و انجام دهید. می‌توانید برای این منظور از ابزارهای فارسی موجود برای پیش‌پردازش متن فارسی مثل کتابخانه hazm استفاده کنید.

ب) مدل‌سازی موضوعی را برای تعداد موضوع بین ۳ تا ۱۵ انجام داده، برای هر مرحله c_v coherence را محاسبه کرده و با رسم نمودار مناسب، منسجم‌ترین مدل‌سازی موضوعی را انتخاب کنید.

پ) سعی کنید برای هر دسته نامی بیابید. (استفاده از نمودار بخش ت می‌تواند در این تحلیل به شما کمک کند). تعداد و نگاشت موضوعات را با تعداد و عنوان رده‌های اصلی در مجموعه داده مقایسه کنید و در صورت کم‌تر یا بیشتر بودن و یا موجود نبودن یک یا چند رده مجموعه داده اصلی در موضوعات پیدا شده توسط LDA، تحلیل‌های لازم را برای دلایل احتمالی این پدیده شرح دهید.

[1] https://www.digikala.com/mag/

ت) با استفاده از کتابخانه pyLDAvis یک مصورسازی از مدل‌سازی نهایی خود داشته باشید و آن‌را تحلیل کنید. خوب است که دسته‌های نشان داده شده در یک جای نمودار باشند یا در دو بعد پخش شده باشند؟ چرا؟ در مورد میزان هم‌پوشانی دسته‌های موضوعی نیز تحلیل‌های خود را بیان کنید.

راهنمایی ۱: پیشنهاد می‌شود در ابتدا مدل‌سازی موضوعی را برای تعداد موضوع منطبق با تعداد دسته‌بندی‌ها انجام دهید و با بررسی گروه کلمات خروجی، در روش پیش‌پردازش خود تغییرات لازم را اعمال کنید.

راهنمایی ۲: در صورت استفاده از Google Colab ، برای نمایش موضوعات به کمک pyLDAvis ممکن است به مشکل تداخل نسخه کتابخانه‌ها برخورد کنید. در این صورت pandas را به نسخه ۱.۵.۳ و pyLDAvis را به نسخه ۲.۱.۲ کاهش دهید.

بخش تشریحی – سوال ۱ (Naïve Bayes Classification)

فرض کنید سند زیر داده شده است. در صورت استفاده از طبقه بند Naïve Bayesبه همراه هموار سازی add one ، سند زیر چه برچسبی خواهد گرفت؟

"I loved the poor play"

"I hated the play movie"

Document	Text	Class
1	I loved the movie	+
2	I hated the movie	-
3	a great movie. good movie	+
4	poor acting	-
5	great acting. a good movie	+

آیا نتایج به دست آمده از نظر شما، منطقی است؟ چرا؟

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA6_StudentID 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.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- امکان ارسال با تاخیر برای این تمرین وجود ندارد.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه 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 بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:vahyd@live.com

مهلت تحویل: ۲۷ دی ماه ۱۴۰۲