بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۲-۰۱

مدرس: احسان‌الدین عسگری

تمرین چهارم | **طبقه‌بندی سند - طبقه‌بندی کلمه** | مهلت ارسال: ۲ تیر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در تمرین‌هایی که چند چالش دارند، فقط یک نفر از هر گروه در سامانه CW باید چالش مورد نظر گروه را انتخاب کند. امکان تغییر چالش تا قبل از زمان ددلاین انتخاب چالش وجود دارد. البته ذکر این نکته ضروری است که هر چالش محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب چالش اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها مطابق زمان مشخص شده در تقویم، بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و مواردی از این دست) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- کد نهایی شما باید یک کلاس شامل تابع run داشته باشد که برای هر چالش طبق ورودی‌های مشخص شده، خروجی موردنظر را بدهد. شما می‌توانید به کمک روند معرفی شده در فایل CONTRIBUTION کدهای خود را به کتابخانه‌ی parsi.io اضافه کنید. در صورتی که pull-request شما پذیرفته شود نمره‌ی امتیازی به شما تعلق خواهد گرفت.
- دقت داشته باشید، موارد امتیازی که در این تمرین آمده است، صرفا بر روی امتیاز همین تمرین اثر دارد و بر روی نمرات تمارین و یا بخش‌های دیگر درس، تاثیر ندارد.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به تیم تدریس خودداری کنید.

ساز و کار تمرین (ابتدا این بخش را به صورت کامل مطالعه نمایید.)

در این تمرین هر گروه یکی از موضوع‌های پیشنهادی را انتخاب خواهد کرد. هر موضوع شامل دو بخش طبقه‌بندی سند[1] و طبقه‌بندی کلمه[2] است. درصورتی که در هرکدام از موضوعات تمایل دارید تا روی مجموعه‌دادگان دیگری کار کنید، توضیحات و آدرس مجموعه‌دادگان مدنظر را برای تیم تدریس در کوئرا ارسال کنید تا پس از بررسی و تایید تیم

[1] Document Classification
[2] Token Classification

تدریس، بتوانید بر روی آن تمرین خود را انجام دهید. همچنین در برخی از موضوعات که در قسمت توضیحات آن گفته شده است امکان برچسب زنی توسط شما برای دادگان اعلام شده وجود دارد و درصورتی که تمایل به این کار داشته باشید می‌توانید پس از هماهنگی با تیم تدریس (به منظور جلوگیری از همپوشانی تیم‌ها) اقدام به این کار نمایید. (برچسب زنی دادگان نمره امتیازی خواهد داشت و می‌توانید برای این منظور از روش‌هایی مانند prompt engineering به کمک مدل‌های زبانی بزرگ استفاده نمایید).

برای موضوعاتی که در ترم‌های قبل نیز توسط دانشجویان مورد بررسی قرار گرفته است، بهترین کد آن‌ها در اختیار شما قرار داده خواهد شد تا از دوباره‌کاری جلوگیری شود. به تبع انتظار می‌رود تلاش شما روی آن موضوع باعث بهبود عملکرد کدهای قبلی شود. یکی از ایده‌های اولیه برای ارتقای عملکرد مدل‌ها یادگیری ماشین، تحلیل خطا است. به این معنی که برخی از نمونه‌هایی که مدل آن‌ها را اشتباه پیش‌بینی می‌کند را بررسی کرده و درصورت مشاهده ساختارهای پرتکرار در آن‌ها که احتمالا دلیل بروز خطا در مدل هستند، تلاش کرد که مواردی که این ساختار را دارند اصلاح کرد یا مدل را نسبت به این موارد مقاوم نمود.

در بخش **طبقه‌بندی سند** نیاز است تا برای موضوع پیشنهادی، دو مدل زیر پیاده‌سازی شود:

۱. **مدل پایه:** اجرای مدل Logistic Regression یا Linear-SVM یا Nïve Bayes بر روی بردار ویژگی tf-idf

۲. **مدل اصلی:** استفاده از طبقه‌بندهای بر پایه ترنسفورمر (به عنوان مثال تنظیم کردن[3] پارامترهای مدل BERT)

در بخش **طبقه‌بندی کلمه** نیاز است تا برای موضوع پیشنهادی، دو مدل زیر پیاده‌سازی شود:

۱. **مدل پایه:** استفاده از مدل LSTM/CRF و یا HMM

۲. **مدل اصلی:** استفاده از مدل‌های بر پایه ترنسفورمر

به منظور استفاده از دادگان در هر بخش، می‌بایست دادگان خود را به سه بخش ۸۰، ۱۰ و ۱۰ درصد تقسیم کنید که به ترتیب دادگان آموزش، اعتبارسنجی و تست می‌باشد (بعضی از مجموعه دادگان به صورت پیش‌فرض برای این منظور تقسیم‌بندی شده‌اند، در این موارد نیازی به تقسیم بندی مجدد نیست و از همان تقسیم بندی پیش‌فرض خود مجموعه استفاده کنید تا مدل نهایی شما با دیگران قابل مقایسه باشد). در نهایت پس از بررسی کامل مدل و انتخاب تمامی هایپرپارامترها، عملکرد هرکدام از مدل‌ها را بر روی دادگان تست گزارش کنید. توجه داشته باشید که حتما ابتدا دادگان را تقسیم بندی کرده ذخیره کنید و سپس مدل‌های مختلف را بر روی آن‌ها تست نمایید، تا به این ترتیب مقایسه مدل‌های مختلف با هم عادلانه‌تر باشد. (توجه داشته باشید که در بخش استفاده از **مدل پایه ۱ بخش طبقه‌بندی سند** نیازی به ۱۰ درصد اعتبارسنجی نیست (دادگان اعتبارسنجی را با آموزش ترکیب کنید) و به این ترتیب دادگان به نسبت ۹۰ به ۱۰ تقسیم شده و به صورت Cross Validation با دادگان آموزش، معیارهای ارزیابی که در ادامه گفته شده است را محاسبه و میانگین و انحراف معیار آن‌ها را گزارش کنید).

برای هر دو بخش **طبقه‌بندی سند** و **طبقه‌بندی کلمه** نیاز است تا معیارهای F1(macro/micro)، Accuracy، Precision، Recall و ماتریس درهم‌ریختگی[4] محاسبه شود.

همچنین در بخش **طبقه‌بندی سند** نیاز است تا از روش‌های تفسیرپذیری مدل استفاده کنید و بر روی ۲۰ مورد از دادگان اعتبارسنجی که مدل شما اشتباه برچسب زده است و ۲۰ مورد از این دادگان که درست برچسب زده است بررسی کنید. برای تفسیرپذیری پیشنهاد می‌شود از کتاب‌خانه SHAP استفاده نمایید.

به منظور استفاده بهتر از مدل نهایی که شما توسعه داده اید و بررسی آن در ترم‌های آینده نیاز است تا بهترین مدل در هر بخش را در فضای Huggingface درس (این لینک) بارگذاری نمایید.

[3] Fine-tune
[4] Confusion matrix

تحلیل احساسات

تحلیل احساسات به معنای پردازش داده‌های متنی است که برای بررسی نظرات و احساسات افراد درباره یک موضوع خاص استفاده می‌شود. این احساسات به صورت معمول بین طیفی از احساسات منفی تا مثبت طبقه‌بندی می‌شوند.

طبقه‌بندی سند

برای این تمرین می‌توانید از دو مجموعه‌داده نظرات غذا که از سایت دیجی‌کالا جمع‌آوری شده، و یا نظرات فیلم که از سایت تیوال استخراج شده است، استفاده نمایید. ساختار کلی این دو مجموعه داده به صورت ذیل است:

Domain	Review	Sentiment	(Aspect, Sentiment)
Food & beverages	خیلی خیلی کادوی جذابیه هم بسته بندی شیک هم شکلات خوشمزه و قلبی شکل خصوصا که پاکت هم داره	Very positive	(بسته بندی، خیلی مثبت) (طعم، مثبت)
Food & beverages	در شگفت انگیز به قیمت خیلی پایین خریدم ولی به نظرم ارزش نداره و طعم خاصی جز شکر نداره	Negative	(ارزش خرید، منفی) (طعم، منفی)
Movie review	در جشنواره متاسفانه نتونستم ببینم ولی دیشب در اکران فیلم های جشنواره فجر در پردیس چارسو موفق به دیدن فیلم شدم. چه فیلم خوبی از فضای بصری زیبا و چشم نواز ، تا بازی فوق العاده حامد بهداد.....	Positive	(صحنه، مثبت) (بازی، خیلی مثبت)
Movie review	فیلمی بسیار ضعیف، علی الخصوص در زمینه ی تدوین و فیلم نامه پر از شعار زدگی، کلیشه و اغراق آمیز!!! واقعا خانم درخشنده توی این فیلم تنزل فاحشی پیدا کردن. بعد اصلا معلوم نیست اون زن دوم اون وسط چی میگه، از بس که شخصیت پردازی ضعیفه!	Very negative	(بازی، خیلی منفی) (داستان، خیلی منفی) (کارگردانی، خیلی منفی)
Movie review	فیلم از فضای نقد اجتماعی و سیاسی تهی است...یه قصه غیر قابل باور که هیجان خاصی نداشت...ریتم فیلم قابل قبول بود...الناز شاکردوست هم خیلی فراتر از انتظار بود...نمره 5 از 10	Mixed/borderline	(داستان، منفی) (بازی، خیلی مثبت)

این دو مجموعه داده شامل شش دسته زیر هستند. هدف از این بخش، پیاده‌سازی مدلی است که بتواند با دریافت نظر کاربر، تحلیل احساسات انجام دهد و یکی از شش دسته‌ی مذکور را به عنوان خروجی ارائه کند.

- very positive
- positive
- neutral
- negative
- very negative
- mixed/borderline

بخش امتیازی:

یکی از تسک‌های پردازش زبان طبیعی، تحلیل احساسات از جنبه‌های مختلف است. در مجموعه دادگان مذکور دیدیم که نظرات کاربران بر اساس جنبه‌های مختلف نظردهی، دسته‌بندی شده و سپس احساسات برای آن مورد مشخص شده است. شما می‌توانید مدلی آموزش دهید که با دریافت نظر کاربر، و جنبه‌ی مدنظر، تحلیل احساسات را انجام دهد. به عنوان نمونه می‌توانید به این دمو رجوع نمایید.

طبقه‌بندی کلمه

در این بخش، هدف پیاده‌سازی مدلی برای پرسش و پاسخ استخراجی [۵] است. در مدل پرسش و پاسخ استخراجی، یک سند یا متن به عنوان ورودی به مدل داده می‌شود و سپس با مطرح کردن سوالاتی از متن، مدل بازه‌ای از متن که شامل پاسخ است را به عنوان خروجی بازمی‌گرداند.
برای این بخش می توانید از مجموعه داده‌ی subjqa در دو بخش نظرات غذا و یا نظرات فیلم استفاده نمایید.
به این منظور شما باید یک مدل پرسش و پاسخ را بر روی مجموعه‌دادگان آموزش دهید و امتیاز f۱ و EM [۶] را بر روی داده آزمایش گزارش دهید. امتیاز EM برای مسئله پرسش و پاسخ میزان هم‌پوشانی پاسخ مرجع با پاسخ تولید شده است.

بخش امتیازی:

در صورت تمایل شما می‌توانید از مجموعه‌دادگانی که در بخش طبقه‌بندی سند استفاده کرده‌اید، در این بخش نیز استفاده نمایید. به این منظور نیاز است که کلمه‌ها را استخراج نمایید و بر اساس احساسات و همان شش دسته، برچسب‌گذاری را انجام دهید. سپس مدلی پیاده‌سازی نمایید که طبقه‌بندی و تحلیل احساسات را در سطح کلمه انجام دهد.
توجه نمایید که درصورت انجام این بخش، لازم به پیاده‌سازی مدل پرسش و پاسخ نیست و شما نمره‌ی امتیازی نیز دریافت می‌نمایید.

[5] Extractive Question Answering
[6] Exact Matching

گونه‌ی محاوره‌ای در مقابل گونه‌ی رسمی

طبقه‌بندی سند

هدف این بخش تشخیص گونه‌ی محاوره‌ای از رسمی در زبان فارسی است. شما باید مدلی پیاده‌سازی کنید که با دریافت یک جمله از ورودی، محاوره‌ای یا رسمی بودن آن را تشخیص دهد. همچنین کلمات کلیدی هر کلاس را که بیشترین تاثیر را داشته‌اند مشخص کنید.

طبقه‌بندی کلمه

هدف این بخش طبقه‌بندی کلمه‌های موجود بر اساس محاوره، رسمی و خنثی بودن کلمات در متون است. به این معنا که باید مدلی آماده کنید تا بتواند موجودیت‌های محاوره‌ای، رسمی و خنثی را تشخیص دهد. خنثی به این معنی است که کلمه قابل شکستن نیست و تفاوتی در صورت محاوره‌ای و رسمی آن وجود ندارد. برای برچسب‌گذاری این بخش از سه حرف F[۷]، C[۸] و N[۹] استفاده شده که به ترتیب نشان‌دهنده‌ی رسمی، محاوره‌ای و خنثی بودن کلمه است یک نمونه از این کار در جدول زیر برای جمله‌ی «دارم میرم خانه» آمده است:

کلمه	برچسب
دارم	N
میرم	C
خانه	F

مجموعه دادگانی از جملات محاوره‌ای و معادل رسمی آنها تولید شده و در اختیار شما قرار می‌گیرد. توجه نمایید که این ترک در تمرین ترم گذشته پیاده‌سازی شده و برای طبقه‌بندی کلمه، برچسب‌گذاری نیز توسط دانشجویان انجام شده است. هدف شما در این تمرین باید ارتقای زحمات دوستانتان در ترم گذشته باشد. به عنوان یک مورد، می‌توانید به مجموعه دادگان داده اضافه کرده و برچسب‌گذاری انجام دهید. برای داشتن اطلاعات بیشتر، یک نمونه گزارش به همراه کد از دانشجویان ترم پیش در دسترس شما قرار خواهد گرفت. می‌توانید از گزارش و کد مورد نظر، برای بهبود کار خودتان نسبت به کار پیشین انجام شده استفاده کنید. صرفا کپی کردن کدها بدون داشتن ایده‌ی جدید باعث از دست دادن نمره می‌شود.

[7] Formal
[8] Colloquial
[9] Neutral

دسته‌بندی کالا

طبقه‌بندی سند

هدف از این بخش، تعیین دسته‌بندی کالا بر اساس توضیحات آن است. به این منظور مجموعه داده‌ای که از سایت دیوار جمع‌آوری شده است، در اختیار شما قرار می‌گیرد و لازم است بر اساس توضیحاتی که کاربر برای کالای خود ارائه نموده، دسته‌بندی کلی آن را تشخیص دهید. چالشی که در این بخش وجود دارد، نیاز به پیش‌پردازش و مرتب‌سازی داده است. شما باید دسته‌بندی را بر اساس کلی‌ترین دسته‌ی محصولات، یعنی چهار دسته‌ی home-kitchen, vehicles, electronic-devices, real-state انجام دهید.

طبقه‌بندی کلمه

در این بخش، هدف پیاده‌سازی مدلی برای پرسش و پاسخ استخراجی [10] است. در مدل پرسش و پاسخ استخراجی، یک سند یا متن به عنوان ورودی به مدل داده می‌شود و سپس با مطرح کردن سوالاتی از متن، مدل بازه‌ای از متن که شامل پاسخ است را به عنوان خروجی بازمی‌گرداند.
برای این بخش می توانید از بخشی از مجموعه داده‌ی قراردادهای حقوقی و تجاری cuad که شامل متن قراردادها و سوالاتی از متن است، و یا مجموعه داده‌ی subjqa در بخش نظرات کالاهای الکترونیکی استفاده نمایید.
به این منظور شما باید یک مدل پرسش و پاسخ را بر روی مجموعه‌دادگان آموزش دهید و امتیاز f۱ و EM [11] را بر روی داده آزمایش گزارش دهید. امتیاز EM برای مسئله پرسش و پاسخ میزان هم‌پوشانی پاسخ مرجع با پاسخ تولید شده است.

بخش امتیازی:

در صورت تمایل شما می‌توانید از مجموعه‌دادگانی که در بخش طبقه‌بندی سند استفاده کرده‌اید، در این بخش نیز استفاده نمایید. به این منظور نیاز است که برای یک دسته مانند real-state کلمه‌ها را استخراج نمایید و بر اساس موجودیت‌ها برچسب‌گذاری را انجام دهید. موجودیت‌هایی که می توانید بر اساس آن‌ها برچسب‌گذاری را انجام دهید می‌تواند به این صورت باشد:

- Locality (L)
- Total Price (P)
- Land Area (LA)
- Cost per land area (C)
- Contact name (N)
- Contact telephone (T)
- Attributes of the property (A)
- Other (O)

سپس مدلی پیاده‌سازی نمایید که طبقه‌بندی را در سطح کلمه انجام دهد.

توجه نمایید که درصورت انجام این بخش، لازم به پیاده‌سازی مدل پرسش و پاسخ نیست و شما نمره‌ی امتیازی نیز دریافت می‌نمایید.

[10] Extractive Question Answering
[11] Exact Matching

دسته‌بندی اخبار و مقالات

طبقه‌بندی سند

هدف از این بخش، طبقه‌بندی موضوعی اخبار یا مقالات است و شما باید مدلی پیاده‌سازی نمایید که با دریافت یک سند، عنوان دسته‌ی موضوعی آن را خروجی دهد. برای این تمرین می‌توانید از مجموعه داده‌ی مجله‌ی دیجی‌کالا (دیجی‌کالامگ) در این لینک استفاده نمایید، این مجموعه داده شامل هفت دسته‌ی موضوعی است.
درصورت تمایل می‌توانید از مجموعه داده‌ی اخبار آزمایشگاه استفاده کنید. این مجموعه داده در اختیار شما قرار خواهد گرفت. در ترم گذشته تلاش‌هایی برای دسته‌بندی موضوعی این مجموعه داده انجام شده است و در صورت انتخاب این مجموعه داده، نیاز به بهبود نتایج دانشجویان ترم‌های گذشته است.

طبقه‌بندی کلمه

در این بخش، هدف پیاده‌سازی مدلی برای پرسش و پاسخ استخراجی [۱۲] است. در مدل پرسش و پاسخ استخراجی، یک سند یا متن به عنوان ورودی به مدل داده می‌شود و سپس با مطرح کردن سوالاتی از متن، مدل بازه‌ای از متن که شامل پاسخ است را به عنوان خروجی بازمی‌گرداند.
برای این بخش می توانید از مجموعه داده‌ی پرس و پاسخ PersianQA استفاده نمایید.
به این منظور شما باید یک مدل پرسش و پاسخ را بر روی مجموعه‌دادگان آموزش دهید و امتیاز f۱ و EM[۱۳] را بر روی داده آزمایش گزارش دهید. امتیاز EM برای مسئله پرسش و پاسخ میزان هم‌پوشانی پاسخ مرجع با پاسخ تولید شده است.

بخش امتیازی:

در صورت تمایل شما می‌توانید از مجموعه‌دادگانی که در بخش طبقه‌بندی سند استفاده کرده‌اید، در این بخش نیز استفاده نمایید. به این منظور نیاز است که کلمه‌ها را استخراج نمایید و بر اساس موجودیت‌های نامدار برچسب‌گذاری را انجام دهید. موجودیت‌های نامدار به صورت BIO است و باید با هشت گروه اشخاص (PER) ، مکان (LOC) ، محل اصلی اتفاق (mainLoc) ، سازمان (ORG) ، رویداد (EVE) ، ملیت و اقوام (NAT) ، عبارت زمانی کمتر از یک روز (TIM) و عبارت زمانی بیش از یک روز (DAT) برچسب‌زنی شوند. سپس مدلی پیاده‌سازی نمایید که طبقه‌بندی را در سطح کلمه انجام دهد.
توجه نمایید که درصورت انجام این بخش، لازم به پیاده‌سازی مدل پرسش و پاسخ نیست و شما نمره‌ی امتیازی نیز دریافت می‌نمایید.

[12] Extractive Question Answering
[13] Exact Matching

داده‌های کلینیکال

مجموعه دادگان n2c2

یکی از مجموعه دادگانی که می‌توانید در این ترک از آن‌ها استفاده کنید دادگان ترک‌های ۱ و ۲ چالشی n2c2 سال ۲۰۱۸ هستند که از این لینک قابل دسترسی‌اند. به منظور تسهیل فرآیند توسعه، این دادگان از طریق این لینک (رمز: nlp1401) در اختیار شما می‌باشد. با توجه به این که دسترسی به این دیتاست نیازمند تایید تامین کننده داده است، لطفا فقط برای این تمرین استفاده شود. دادگان ترک ۱ برای طبقه‌بندی و دادگان ترک ۲ برای طبقه‌بندی کلمه قابل استفاده هستند.

مجموعه دادگان NCBI

یکی دیگر از مجموعه دادگانی که در این تمرین می‌توانید در بخش دسته‌بندی کلمه از آن استفاده کنید، مجموعه دادگان مربوط به NCBI می‌باشد که از طریق این لینک قابل دسترسی است. هر سر از این دادگان گزارش پزشک برای یک بیمار می‌باشد که به کلمه‌های سازنده آن شکسته شده است و کلمه‌ها شامل ۳ برچسب متفاوت هستند به این صورت که برچسب **0** به معنی «عدم اشاره به بیماری»، برچسب **1** به معنی «اولین کلمه مربوط به بیماری» و برچسب **2** به معنی «زیرمجموعه‌ای از کلمه‌های مربوط به بیماری» می‌باشد.

طبقه‌بندی سند

در این بخش می‌بایست از دادگان n2c2 استفاده شود. هدف پیاده‌سازی یک مدل برای استخراج اطلاعاتی راجع به بیمار است. این دادگان براساس شرح حال ۲۰۲ بیمار بوده و شامل ۱۳ برچسب هستند که می‌توانند دومقدار met و not met داشته باشند. مدل شما باید بتواند با توجه به شرح حال بیمار، مواردی که در زیر فهرست شده‌اند را تشخیص دهد.

- ABDOMINAL
- CREATININE
- MAJOR-DIABETES

طبقه‌بندی کلمه

در این بخش می‌توانید از هر دو دادگان استفاده کنید. درصورتی که از دادگان n2c2 استفاده کنید؛ باید مدلی پیاده‌سازی کنید که با دریافت مشخصات بیمار و نسخه‌ی او در ورودی، موجودیت های زیر را تشخیص دهد.

- Drug
- Strength
- Form
- Dosage
- Duration
- Frequency
- Route
- ADE
- Reason

درصورتی که از دادگان NCBI استفاده کنید؛ باید مدلی پیاده سازی کنید که با دریافت یک عبارت که گزارش پزشک راجع به یک بیمار می‌باشد، به ازای کلمه‌های مختلف گزارش دهد که چه بخشی مرتبط با بیماری است و در نهایت آن بخش‌هایی که مرتبط با بیماری است را از متن استخراج کند و به عنوان خروجی اعلام کند. به عنوان مثال:

- **Input:** Identification of APC2, a homologue of the adenomatous polyposis coli tumour suppressor.
- **Output:** adenomatous polyposis coli tumour

زبان توهین‌آمیز

طبقه‌بندی سند

در این بخش، شما می‌بایست با استفاده از مجموعه دادگان OLID که شامل مجموعه‌ای از متون و برچسب‌های متناظر با آن‌هاست، مدل دسته‌بندی آموزش دهید که بتواند جملات را به دو دسته «توهین آمیز» و «غیرتوهین آمیز» طبقه‌بندی کند (همان سطح A [۱۴] در دیتاست OLID) :

- OFF
- ON

در ترم‌های گذشته بر روی این دادگان تلاش‌هایی صورت گرفته است، در این بخش نیاز است تا ابتدا شما بهترین کد ترم گذشته را بررسی کرده و سپس تلاش کنید تا ارتقا مناسبی در عملکرد نسبت به آن‌ها داشته باشید.

طبقه‌بندی کلمه

در این بخش، شما می‌بایست در ابتدا با استفاده از مجموعه داده Toxic Spans (توضیح بیشتر مجموعه داده در این لینک)، یک دیتاست برچسب‌گذاری شده در سطح کلمه بسازید. برای اینکار، در ترم‌های گذشته کلمه‌هایی که در ستون probability ذکر شده‌اند و امتیازشان بیشتر از ۰.۵ است را برچسبی هم‌نام با گونه‌ای در ستون type زده‌اند که بیشترین امتیاز را دارد. سایر کلمه‌ها را برچسب none زده‌اند. نیاز است تا ابتدا شما بهترین کد ترم گذشته را بررسی کرده و سپس تلاش کنید تا در ابتدا اگر ایرادی به این روش برچسب‌گذاری وارد است آن را حل کرده و سپس با مجموعه دادگان تولید شده مدلی آموزش دهید تا بتواند به ازای هر کلمه ورودی، یکی از لیبل‌های type دیتاست Toxic Spans یا none را برچسب بزند:

- insult
- threat
- identity-based attack
- profane/obscene
- other toxicity
- none

[14] Level

قرآن و حدیث

طبقه‌بندی سند

یکی از کاربردهای رایج طبقه‌بندی متون در علوم اسلامی، طبقه‌بندی موضوعی احادیث است. به طور کلی، اغلب تالیفات در حوزه‌ی علم حدیث، به طور موضوعی دسته‌بندی شده‌اند. لذا با داشتن یک مجموعه از احادیث به همراه موضوع هر یک، می‌توان یک مدل آموزش داد که بتواند با گرفتن متن حدیث، موضوع آن را پیش‌بینی کند. این مدل می‌تواند به پیدا کردن احادیثی که در یک کتاب حدیث، احتمالا اشتباه دسته‌بندی شده‌اند هم کمک کند. برای این منظور، با استفاده از مجموعه‌داده‌ی اصول کافی شامل احادیث و دسته‌بندی هر یک، مدلی آموزش دهید که بتواند با گرفتن متن حدیث، آن را دسته‌بندی کند و سپس مدل خود را ارزیابی کنید.
بخش امتیازی: مساله‌ی طبقه‌بندی موضوعی، یک مثال از مسائل دسته‌بندی بین چند کلاس با یک برچسب بود. یک نوع دیگر از مسائل دسته‌بندی در این حوزه، تخمین راویان یک حدیث با داشتن متن حدیث است که یک مساله‌ی دسته‌بندی با چند کلاس و چند برچسب است. به عنوان یک تلاش اختیاری، می‌توانید برای توسعه‌ی چنین دسته‌بندی سعی کنید.

طبقه‌بندی کلمه

در این بخش، تنها کافی است یکی از دو قسمت «بازشناسی موجودیت‌های نام‌دار» یا «پرسش و پاسخ» را انجام دهید:

بازشناسی موجودیت‌های نام‌دار

با توسعه‌ی روش‌ها و فنون جدید در هوش مصنوعی و به‌ویژه پردازش زبان‌های طبیعی، امروزه می‌توان برخی از فرایندهای استنباط فقهی را به کمک کامپیوتر خودکارسازی کرد. یکی از این فرایندها، علم رجال[۱۵] است که به بررسی صحت و ضعف روایات حدیثی و شناسایی روایات معتبر و روایات غیرمعتبر می‌پردازد. از اولین اقداماتی که برای بررسی احادیث در این علم به آن نیاز است، جدا کردن اسامی خاص در متن حدیث است.
برای این منظور نیاز است تا با استفاده از روش‌های بازشناسی موجودیت‌های نام‌دار[۱۶]، دسته‌های مختلف از اسامی را در متن برچسب‌گذاری کنید. مجموعه‌داده‌ای که برای این منظور موردنظر است، مجموعه‌داده‌ی CANER است و نیازمندی این است که مدلی طراحی کنید که بتواند برچسب‌های مربوط به موجودیت‌های این مجموعه‌داده را تشخیص دهد و سپس ارزیابی مدل را با متریک‌های گفته شده در بخش سازوکار تمرین گزارش کنید.
بخش امتیازی: متنی که مجموعه‌داده‌ی CANER از آن ساخته شده است متن کتاب صحیح بخاری[۱۷] است. با استفاده از مدل آموزش داده شده روی این مجموعه‌داده، موجودیت‌های نامدار در احادیث کتاب اصول کافی را پیدا کنید و برای روایان احادیث در این کتاب، PageRank را محاسبه کنید و افراد با PageRank بالا را گزارش کنید.

پرسش و پاسخ

یکی از اجزای اصلی یک سامانه‌ی هوشمند فقهی، سامانه‌ای است که بتواند با گرفتن مجموعه‌ی وسیعی از منابع فقهی، به سوالات مرتبط پاسخ بدهد. مطابق با یکی از مهم‌ترین اصول علوم و مهندسی، خوب است کار با یک مساله‌ی ساده شروع شود و یکی از محبوب‌ترین مسائل ساده‌ای که در حوزه‌ی طراحی سامانه‌های پرسش و پاسخ در علوم اسلامی به آن پرداخته شده، پرسش و پاسخ روی متن قرآن کریم است. مجموعه‌داده‌ی QuranQA مجموعه‌ای از بخش‌هایی از متن قرآن کریم به همراه سوال و جواب از هر بخش آن است و نیازمندی این است که سامانه‌ای طراحی کنید که بتواند به طور خودکار، پاسخ هر سوال را از روی متن داده شده، به‌دست بیاورد.

[۱۵] علم الرجال المتعلق بالحدیث
[۱۶] Named Entity Recognition
[۱۷] از کتب صحاح اهل سنت

تشخیص قصد

تشخیص قصد از روی متن به صورت کلی باعث بهبود تعاملات بین انسان و ماشین می‌شود. یکی از مهم‌ترین کاربردهای آن در سیستم‌های دستیار صوتی و متنی است که با گرفتن یک دستور می‌بایست قصد را استخراج کرده و عملیات مدنظر را انجام دهند. برای این منظور یک مجموعه داده شامل زبان‌های مختلف از جمله فارسی، که متعلق به الکسا آمازون می‌باشد، در اختیار شما قرار می‌گیرد. (مجموعه دادگان)
در ستون utt (string) متن دستور قرار دارد ، scenario (class label) یک برچسب کلی از قصد دستور گفته شده، در ستون intent (class label) قصد دستور به صورت جزئی‌تر و در نهایت در ستون annot_utt (string) کلمه یا کلماتی که عنصر اصلی دستور بوده اند دسته‌بندی شده‌اند. برای جزئیات بیشتر می‌توانید به صفحه توضیحات دادگان مراجعه کنید.

طبقه‌بندی سند

در این بخش انتظار داریم تا بر روی دادگان زبان فارسی این مجموعه داده مدل خود را به نحوی آموزش دهید که بتواند کلاس‌های سطح scenario (class label) را پیش‌بینی کند.
مثالی از ورودی و خروجی:

- **ورودی:** مرا جمعه ساعت نه صبح بیدار کن.
- **خروجی:** Alarm

طبقه‌بندی کلمه

در این بخش انتظار داریم تا بر روی دادگان زبان فارسی این مجموعه مدلی آموزش دهید که بتواند قصد هرکدام از لغات را تشخیص دهد. البته در محدوده کلاس‌هایی که این دادگان در اختیار شما قرار داده است که در ستون annot_utt (string) می‌توانید مشاهده کنید.
مثالی از ورودی و خروجی:

- **ورودی:** مرا جمعه ساعت نه صبح بیدار کن. (فرمت کلمه کلمه شده: {م، را، جمعه، ساعت، نه، صبح، بیدار، کن})
- **خروجی:** [none, none, date, none, time, time, none, none]

بخش امتیازی:

با توجه به این که مدل‌های چندزبانی مبتنی بر ترنسفورمر عملکرد مناسبی تا به حال از خود نشان داده اند؛ می‌توانید یک مدل ترنسفورمر چند زبانی را روی یک یا چند زبان از این مجموعه داده آموزش داده و سپس روی یک یا چند زبان دیگر تست کنید و عملکرد مدل را تحلیل کنید. این فرآیند را در هر دو بخش بررسی کنید.

طبقه‌بندی محتوایی کتاب

طبقه‌بندی سند

امروزه با توجه به حجم بالای تولیدات متنی و همچنین متونی که تا به امروز وجود دارند، نگهداری و طبقه‌بندی و همچنین جست و جو میان حجم بالایی از این کتب، به یک چالش تبدیل شده است. برای این منظور در این بخش تصمیم داریم تا با استفاده از دادگان مسابقه GermEval 2019 که شامل متن ابتدایی کتاب‌ها آلمانی به همراه برچسب سلسله مراتبی از این کتاب‌هاست، تجربه ای در جهت توسعه سیستم هوشمند دسته بندی داشته باشیم. دسته‌بندی ای که این مجموعه داده ارائه کرده است، شامل چند سطح است؛ اما در این تمرین انتظار داریم تا فقط سطح اول دسته‌بندی گفته شده را به عنوان هدف قرار دهید. بنابراین مدل شما به ازای هر سند باید یکی از ۸ دسته زیر را اعلام کند:

- Lit- eratur Unterhaltung (Literature Entertain- ment)
- Ratgeber (Counsel)
- Kinderbuch Jugend- buch (Books for Children and Young Adult Read- ers)
- Sachbuch (Nonfiction)
- Ganzheitliches Be- wusstsein (Holistic Awareness)
- Glaube Ethik (Belief Ethics)
- Ku¨nste (Arts)
- Architektur Garten (Architecture Gardening)

طبقه‌بندی کلمه

در این بخش، هدف پیاده‌سازی مدلی برای پرسش و پاسخ استخراجی [۱۸] است. در مدل پرسش و پاسخ استخراجی، یک سند یا متن به عنوان ورودی به مدل داده می‌شود و سپس با مطرح کردن سوالاتی از متن، مدل بازه‌ای از متن که شامل پاسخ است را به عنوان خروجی بازمی‌گرداند. در این بخش انتظار داریم تا با کمک دادگان subjqa در بخش کتاب به حل چالش گفته شده (یافتن محل پاسخ از قسمتی از متن) بپردازید.

به این منظور شما باید یک مدل پرسش و پاسخ را بر روی مجموعه‌دادگان آموزش دهید و امتیاز f۱ و EM [۱۹] را بر روی داده آزمایش گزارش دهید. امتیاز EM برای مسئله پرسش و پاسخ میزان هم‌پوشانی پاسخ مرجع با پاسخ تولید شده است.

[18] Extractive Question Answering
[19] Exact Matching

استنتاج زبان طبیعی

طبقه‌بندی سند

در این بخش شما می‌بایست یک دسته‌بند از نوع استنتاج زبان طبیعی[۲۰] آموزش دهید. مسائل استنتاج زبان طبیعی به یادگیری ارتباط دو متن با یکدیگر می‌پردازند و ۳حالت ارتباط را بین آنها تشخیص می‌دهند. مجموعه دادگانی که در این ترک می‌توانید از آن استفاده کنید، مجموعه دادگان FarsTail است که شامل سه دسته می‌باشد. هر سطر این داده شامل دو متن است که این دو متن سه حالت با یکدیگر دارند که به صورت زیر است:

- یک متن را بتوان از دیگری استنباط کرد. (اثبات)
- اگر دو متن با بکدیگر در تناقض باشند. (تضاد)
- هیچ یک از موارد فوق نباشد و کاملا بی‌ربط باشند (خنثی)

این مجموعه دادگان شامل دو ستون premise و hypothesis که هر کدام به صورت string به مدل ورودی داده می‌شوند و یک ستون label که در واقع برچسب این دادگان است می‌باشد و می‌بایست در نهایت توسط مدل تشخیص داده شود.

طبقه‌بندی کلمه

در این بخش، هدف پیاده‌سازی مدلی برای پرسش و پاسخ استخراجی [۲۱] است. در مدل پرسش و پاسخ استخراجی، یک سند یا متن به عنوان ورودی به مدل داده می‌شود و سپس با مطرح کردن سوالاتی از متن، مدل بازه‌ای از متن که شامل پاسخ است را به عنوان خروجی بازمی‌گرداند.
برای این بخش می توانید از مجموعه داده‌ی پرس و پاسخ PersianQA استفاده نمایید.
به این منظور شما باید یک مدل پرسش و پاسخ را بر روی مجموعه‌دادگان آموزش دهید و امتیاز f۱ و EM[۲۲] را بر روی داده آزمایش گزارش دهید. امتیاز EM برای مسئله پرسش و پاسخ میزان هم‌پوشانی پاسخ مرجع با پاسخ تولید شده است.

[20] Natural Language Inference (NLI)
[21] Extractive Question Answering
[22] Exact Matching