بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۱-۰۰

استاد: احسان‌الدین عسگری

تمرین پنجم | **برچسپ‌گذاری و تولید متن** | مهلت ارسال: ۳ مرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز 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 ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات کلی

در این تمرین شما با تولید متن و برچسپ‌گذاری دنباله آشنا می شوید. شما باید تا یک‌شنبه شب ۱۲ تیر ترک مورد نظر گروه خود را انتخاب کنید.

ترک استخراج موجودیت‌های نام‌دار فارسی دارای بخش برچسپ‌گذاری داده هست و بخشی از نمره این ترک را برچسپ‌گذاری فردی تشکیل می‌دهد.

استخراج آیات قرآن از متن

معرفی تسک و هدف: شناسایی آیات قرآن موجود در یک متن، یکی از مسائل مهمی است که در پژوهش‌های قرآنی و حدیثی کاربردهای زیادی دارد. ابزاری تحت عنوان Quranic Extractor با استفاده از regex برای اینکار توسعه داده شده است که شماره آیه و سوره عبارت یافت شده را نیز گزارش می‌کند؛ اما مشکل این است که صرف نظر از اینکه مقصود یک عبارت در نوشته ورودی اشاره به آیه قرآن است یا خیر، به علت تطابق عبارت یافت شده با قسمتی از آیه قرآن، این عبارت به عنوان یک (یا قسمتی از) آیه قرآن گزارش می‌شود؛ در صورتی که ممکن است این عبارت صرفا حاوی چند کلمه متوالی باشد که در قرآن نیز ذکر شده و مقصود از آن در نوشته ورودی، اشاره به آیه قرآن نبوده باشد. هدف از این تسکِ برچسب‌گذاری توالی [1]، پیدا کردن آیات موجود در متن ورودی با استفاده از مدل های برپایه توجه [2]است تا تلاش شود با درنظر گرفتن زمینه [3] متن، نتایجی که یافت می‌شوند واقعا اشاره به آیه‌ای از قرآن باشند و بتوان با استفاده از آن، خروجی های Quranic Extractor را اصلاح کرد.

یک نمونه: في ذبيحه الناصب و اليهودي و النصراني لا تاكل ذبيحته حتي تسمعه يذكر اسم اللّٰه عليه اما سمعت قول اللّٰه عز و جل و لا تاكلوا مما لم يذكر اسم الله عليه

این جمله به عنوان ورودی به ابزار Quranic Extractor داده شده و قسمت هایی که زیر آن‌ها خط کشیده شده است توسط ابزار به عنوان آیه قرآن تشخیص داده شده است. اما اگر به مفهوم جمله بالا توجه کنیم، روشن می‌شود که عبارت اول اشاره به آیه‌ای از قرآن ندارد؛ اما برخلاف آن، عبارت دوم اشاره‌ای است به یکی از آیات قرآن کریم (سوره انعام آیه ۱۲۱).

نحوه برچسب‌زنی مدل روی متن ورودی: اینکار با استفاده از سه برچسب B به معنای شروع آیه قرآن، I به معنای داخل آیه قرآن و O به معنای خارج از آیه قرآن انجام میشود.

مجموعه دادگان: مجموعه داده‌ای در اختیار شما قرار می‌گیرد که شامل جملات ورودی و آیات قرآن اشاره شده در آن است تا بتوانید برای آموزش مدل خود از آن استفاده کنید. همچنین ۲ روز مانده به پایان مهلت ارسال پاسخ تمرین، داده ارزیابی برای اندازه‌گیری کارایی مدل به شما داده می‌شود.

ارزیابی مدل: برای ارزیابی مدل خود چهار معیار Accuracy، Precision، Recall و F1-score و همچنین Confusion Matrix را روی داده‌های یادگیری و ارزیابی گزارش کنید. با توجه به کارایی مدل و ابتکار در حل مسئله، امکان اختصاص نمره اضافی نیز وجود دارد.

[1] Sequence Labeling
[2] Attention Based
[3] Context

استخراج موجودیت‌های نام‌دار فارسی بر روی دادگان خبری

معرفی تسک: شناسایی موجودیت‌های نام‌دار زیرمجموعه مسائل استخراج اطلاعات است که به دنبال مکان‌یابی و طبقه‌بندی موجودیت های نام‌دار در متن به دسته های از پیش تعریف شده نظیر نام افراد، سازمان ها، مکان ها، کدهای پزشکی، عبارات زمانی، مقادیر، ارزش های پولی، درصد و غیره است.
این تمرین دارای دو بخش انفرادی و گروهی است. بخش انفرادی در گام اول و بخش گروهی در گام دوم شرح داده شده است. تلاش‌های دوستان همراه با نام آن‌ها در خروجی‌های مربوط به این کار (مقاله پژوهشی) به عنوان نویسنده (در صورت مایل بودن) قرار خواهد گرفت.

گام اول (بخش انفرادی):
در این بخش شما باید به سامانه جمع‌سپاری مراجعه کنید و ۷۵ خبر را بر اساس ملاک‌هایی که در ادامه توضیح داده شده است، برچسب گذاری کنید. این کار حداکثر سه ساعت از شما وقت می‌گیرد. دقت شود که مهلت انجام این گام اول تمرین تا آخر **دوشنبه ۲۰ تیر** است و این مهلت نه تمدید می‌شود و نه امکان استفاده از تاخیر مجاز برای آن وجود دارد. برای ورود به سامانه جمع‌سپاری، ایمیل و رمز عبور هر نفر در CW به اشتراک گذاشته شده است. برای مشاهده اطلاعات ورود خود به صفحه درس در CW بخش نام کاربری و رمزعبور سامانه جمع‌سپاری مراجعه کنید. بعد از این که داده‌ها توسط خود دانشجویان برچسب‌زنی شد. داده‌ها در اختیارتان قرار می‌گیرد.

در این گام باید داده‌ها به صورت دستی برچسب‌زنی شوند. طرح حاشیه‌نویسی مجموعه دادگان به صورت BIO است و باید با هشت گروه اشخاص (PER) ، مکان (LOC) ، محل اصلی اتفاق خبر (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.

هر یک از موجودیت‌ها به صورت زیر تعریف می‌شوند.

- **اشخاص**: نام افراد.
- **مکان**: نشانه‌ها، سازه‌ها، ویژگی‌های جغرافیایی و موجودیت‌های ژئوپلیتیکی طبیعی و ساخت بشر.
- **محل اصلی اتفاق خبر**: این تگ زیر مجموعه تگ مکان (LOC) است و اولین کاربرد آن در متن تگ بخورد کافی است. چرا که در یک خبر می‌تواند به چندین مکان اشاره شود. در این صورت تنها با همین تگ محل اصلی اتفاق خبر را مشخص کنید.
- **سازمان**: شرکت ها، گروه های سیاسی، گروه های موسیقی، باشگاه های ورزشی، ارگان های دولتی و سازمان های عمومی. این موجودیت شامل ملیت ها نمی شود.
- **رویداد**: رویدادهای تاریخی، اجتماعی و طبیعی.
- **ملیت**: ملیت و قوم‌ها
- **عبارت زمانی TIM**: تمامی عبارات زمانی با طول کمتر از یک روز
- **عبارت زمانی DAT**: تمامی عبارات زمانی با طول بیشتر از یک روز

در برچسب‌زنی داده‌ها باید قوانین زیر رعایت شوند.:

۱. صفات که داخل کلمات هستند برچسب موجودیت می‌گیرند مانند «خلیج همیشگی فارس» که کلمه همیشگی برچسب موجودیت می‌گیرد.

۲. شاخص های اول اسم‌ها برچسب موجودیت نمی‌گیرند مثلا در «دکتر ظریف» نباید «دکتر» برچسب بگیرد. تنها در صورت شاخص اول کلمات برچسب می‌گیرد که حذف آن شاخص موجب شود کلمات باقی‌مانده معنی اسامی خاص ندهند. مثلا در کلمه «امام زمان» باید کلمه «امام» هم برچسب PER بگیرد.

۳. برای عبارت‌های زمانی پیچیده که هم شامل زمان هم و هم تاریخ است مانند «ساعت ۸ صبح روز دوشنبه» باید دقت داشته باشید که بخش زمان و بخش تاریخ باید جدا برچسب زنی شوند. بدین ترتیب «ساعت ۸ صبح» تگ‌های BIO زمان و عبارت «روز دوشنبه» تگ‌های BIO مربوط به تاریخ را می‌گیرد.

۴. مواردی مانند «مدرسه ۱۵ تیر» که نام مدرسه یک عبارت زمانی است باید کل این عبارت به عنوان سازمان در نظر گرفته شود.

در صورت وجود ابهام در مورد نحوه برچسب‌زنی مواردی که قانون آن به صراحت ذکر نشده است، سوالات خود را با دستیاران آموزشی در کوئرا مطرح کنید.

گام دوم (بخش گروهی):

در این گام باید از یک مدل از پیش آموزش دیده شده با مکانیزم توجه (مانند پارس‌برت آزمایشگاه هوشواره) استفاده کنید و مدلی آموزش دهید که موجودیت‌های نام دار را استخراج کند. در نهایت بر روی بخش دادگان تست نیز دقت مدل خود را گزارش کنید. از کتابخانه پایتون seqeval به منظور راحتی بیشتر برای ارائه دقت‌ها می‌توانید استفاده کنید.

ارزیابی: دقت کنید که بخشی از نمره شما متناسب با عملکرد مدل شما روی مجموعه دادگان ارزیابی خواهد بود پس در هنگام برچسب زنی و آموزش مدل به این نکته دقت کنید. بقیه بخش‌های ارزیابی نیز مربوط به کد و گزارش و دقت‌های بدست آمده است. لطفا حتما فراموش نکنید که ماتریس درهم ریختگی را نیز برای خروجی مدل خود نیز ارائه دهید. مجموعه دادگان ارزیابی نیز حداقل دو روز قبل از پایان یافتن مهلت تمرین به شما داده خواهد شد.

تبدیل الگوی نوشتاری به الگوی واجی تلفظ برای زبان فارسی

معرفی تسک و اهمیت آن: تبدیل الگوی نوشتاری به الگوی واجی تلفظ[۴] یکی از تسک‌های مورد توجه در پردازش زبان طبیعی است. هدف از این تسک پیدا کردن تلفظ کلمه یا کلمات ورودی داده شده است. از جمله مهمترین کاربردهای این تسک می‌توان به تبدیل متن به صوت[۵] اشاره کرد.

هدف از تمرین: هدف از این تمرین آن است که شما با استفاده از یکی از مدل‌های توالی به توالی [۶] و با استفاده از دادگان داده شده در مساله مدلی طراحی کنید که بتواند تلفظ جملات ورودی را خروجی دهد.

مجموعه دادگان و کد: در اینجا یک مجموعه‌داده ابتدایی می‌توانید پیدا کنید که این مجموعه داده شامل تعدادی از کلمات فارسی و معادل تلفظی آن‌هاست. یک نمونه کد برای درک از این پروژه را نیز در اینجا می‌توانید پیدا کنید.

نمونه ورودی و خروجی: نمونه‌های ورودی و خروجی مورد انتظار را می‌توانید با توجه به اطلاعات درون این لینک اعتبار سنجی کنید. همچنین می‌توانید از هر کدام از فرمت‌های موجود خروجی معتبر استفاده کنید.

بخش اضافه: استفاده از دادگان اضافه،ابتکار در حل مساله همراه با بهبود و خروجی با استفاده از فرمت‌های مختلف از جمله بخش‌هایی است که می‌تواند نمره اضافه بگیرد.

ارزیابی: کد، تمیزی کد و منظم بودن بخش‌ها ، گزارش (می‌تواند در نوت بوک بیشتر نوشته شود)، عملکرد و دقت‌های دست یافته از جمله مواردی است که در ارزیابی لحاظ می‌شود. به‌عنوان ارزیابی دقت مدل نیز باید معیارهای BLEU-۱ و BLEU-۲ و BLEU-۳ و BLEU-۴ و ROUGE-L را روی داده اعتبارسنجی[۷] گزارش نمایید.

[4] Grapheme to Phoneme
[5] Text to Speech
[6] Seq2Seq
[7] Validation

ترجمه زبان‌ها و گویش‌های بومی

هدف از این سوال پیاده‌سازی یک سیستم برای ترجمه‌ی یک زبان به زبان دیگر می‌باشد.
در ابتدا باید یک زبان مبدا و یک زبان مقصد انتخاب کنید توصیه می‌شود که از زبان‌های بومی یا زبان‌هایی که سیستمی برای ترجمه‌ی آن‌ها وجود ندارد استفاده کنید.
سپس باید مجموعه دادگان موازی از هر دو زبان پیدا کرد برای این کار می‌توانید از مجموعه دادگان زیر استفاده کنید:
۱ـ زیر نویس فیلم‌ها که به چند زبان موجود است، برای این دادگان می‌توانید به این لینک مراجعه فرمایید.
۲ـ دادگان کتاب‌های مقدس که به چند زبان موجود هستند به عنوان مثال می‌توانید به این لینک مراجعه فرمایید.
۳ـ دادگان صفحات ویکیپدیا که به چند زبان ترجمه شده‌اند.
در آخر باید یک مدل بر روی دادگان پیدا شده برای ترجمه، آموزش دهید.
برای بررسی عملکرد مدل خود از معیار BLEU استفاده کنید.