به نام خدا

تمرین دوم درس پردازش زبان طبیعی

« آشنایی با Sequence Labeling به صورت POS tagging و NER و

بررسی تاثیر آن در بهبود دسته‌بندی متن »

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

نیم‌سال دوم ۰۲-۰۱

برای ارسال تمرین به نکات زیر توجه کنید.

۱- برای ارسال پاسخ تمارین این درس مجموعا ۷ روز زمان تاخیر مجاز وجود دارد و در صورت بیشتر شدن تایم تاخیر پاسخ ارسال شده بررسی نخواهد شد.

۲- هرگونه کپی‌برداری در انجام تمرین‌ها موجب کسر نمره خواهد شد.

۳- آخرین مهلت ارسال تمرین، **ساعت ۲۳:۵۵ روز ۱۶ اردیبهشت** می‌باشد.

۴- فایل های ارسالی خود شامل فایل‌های پیاده‌سازی و گزارش را فشرده کنید و با عنوان «**شماره دانشجویی_HW2**» مانند `HW2_97131912` ارسال کنید.

۵- زبان برنامه‌نویسی برای انجام تمرین‌ها، پایتون یا جاوا در نظر گرفته شده‌است.

۶- کدهای ارسالی خود را برای افزایش خوانایی و درک بهتر به صورت مناسب کامنت‌گذاری کنید.

۷- نحوه انجام پیش‌پردازش بر روی داده‌ها شامل کتابخانه مورد استفاده و مراحل انجام‌شده را در گزارش خود مکتوب کنید.

۸- برای انجام این تمرین می‌توانید از کتابخانه‌های آماده استفاده کنید(به جز مدل‌های آماری).

۹- اگر هرگونه سوال و ابهامی داشتید، از طریق ایمیل زیر می‌توانید در ارتباط باشید.

محمدهادی گلدانی m.h.goldani@gmail.com

تعریف مسئله و معرفی دادگان

در این تمرین هدف بررسی Sequence Labeling به صورت دو تکنیک پردازش زبان طبیعی POS tagging و NER می‌باشد و قرار است تاثیر بهره‌گیری از این دو تکنیک در دسته‌بندی متن مورد بررسی قرار بگیرد.
برای انجام این تمرین می‌توانید خودتان پیاده‌سازی انجام دهید و یا از کتابخانه‌هایی مانند NLTK استفاده نمایید. در نهایت لازم است از هرکدام از بخش‌ها در جهت ساخت بردار ویژگی در دسته بندی متن استفاده شود و در نهایت با استفاده از فایل‌های train.txt و Validatin.txt و ویژگی‌های برداری استخراج شده مدلی بهینه در جهت دسته‌بندی متن ساخته شود و بروی فایل test.txt ارزیابی انجام پذیرد و نتایج گزارش شود. لازم به ذکر است که به بهترین روش‌ها که به نتایج بهتری برسند نمره اضافی تعلق می‌گیرد.

بخش اول: POS tagging

هدف این قسمت از تمرین این هست که بهترین دنباله POS متناظر با جمله ورودی را به دست آورید.

الف) در این قسمت فایل آموزش و ارزیابی به همراه دنباله متناظر به شما داده شده است و باید مدلی ساخته شود و قادر باشد که یک فایل ورودی به نام in.txt را دریافت کند و متن برچسب زده شده را در فایل دیگری به نام out.txt تولید کند.

ب) فایل آزمون نیز به شما داده شده است، با یادگیری مدل توسط داده های آموزش و برچسب زنی داده های آزمون مقادیر Precisionو Recall را برای داده های آزمون به صورت Exact match به دست آورده و پس از به دست آوردن ماتریس Confusion بیشترین خطاهای مدل را به دست آوردید و نتایج را تحلیل کنید.

ج) در این قسمت برداری از ویژگی‌ها را به کمک POS tagging از جملات ورودی بسازید و روش خود را توضیح دهید.

بخش دوم: NER

هدف این قسمت از تمرین این هست که بهترین دنباله برای تشخیص موجودیت‌های نامدار متناظر با جمله ورودی را به دست آورید.

الف) دادگان مورد نیاز برای ساخت مدل به همراه صورت سوال با نامهای NERtr.in، NERtr.out به عنوان داده آموزش و NERvalid.in و NERvalid.out به عنوان داده ارزیابی داده شده است. مدلی بسازید که قادر باشد که یک فایل ورودی به نام in.txt را دریافت کند و متن برچسب زده شده را در فایل دیگری به نام out.txt تولید کند.

ب) با یادگیری مدل توسط داده های آموزش و برچسب زنی داده های آزمون مقادیر Precisionو Recallرا برای داده های

آزمون به صورت Exact match به دست آورده و مانند بخش اول پس از به دست آوردن ماتریس Confusion بیشترین خطاهای سیستم را به دست آوردید و تحلیل کنید.

ج) در این قسمت مشابه بخش اول برداری از ویژگی‌ها را به کمک NER از جملات ورودی بسازید و روش خود را توضیح دهید.

بخش سوم: ارزیابی دسته‌بندی متن

در این قسمت دادگانی شامل بخش‌های آموزش، ارزیابی و آزمون به شما داده شده است.

الف) با استفاده از متن دادگان و با استفاده از یک مدل Embedding ایستا مبتنی بر Word2vec مدلی جهت دسته‌بندی بسازید و و دقت و ماتریس confusion آن را به دست آورید.

ب) با استفاده از متن دادگان و با استفاده از مدل‌های مبدل[1] مدلی بسازید و و دقت و ماتریس confusion آن را به دست آورید.

ج) با بهره‌گیری از بردارهای استخراجگر ویژگی بخش اول و دوم علاوه بر ویژگی‌های دو قسمت قبل، تاثیر استفاده از این دو ویژگی را به تفکیک و باهم (POS و NER) در بهبود مدل را برای دو بخش الف و ب با گزارش دقت و ماتریس confusion به دست آورید و تحلیل خود را از نتایج به دست آمده بیان کنید.

موفق باشید

[1] Transformers