به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس پردازش متن و زبان طبیعی

تمرین شماره سه

نام و نام خانوادگی :

شماره دانشجویی :

فهرست سوالات

مقدمه

در این تمرین می‌خواهیم چند روش از حل مسائل sequential را بر روی دو مسئله‌ی Part-of-Speech Tagging و Named Entity Recognition تمرین کنیم و تفاوت‌ها و چالش‌های هر یک از آن‌ها را بررسی کنیم.

داده‌های این تمرین از دیتاست Penn Treebank است که بخشی از آن توسط کتابخانه‌ی nltk قابل دسترسی است.

۱- تعیین نقش کلمات

در این سوال قصد داریم تا با استفاده از دیتاست Penn Treebank مدل‌هایی برای پیش‌بینی نقش کلمات در جمله آموزش دهیم. برای استفاده از این دیتاست، می‌توانید از کتابخانه‌ی nltk استفاده کنید و دیتاست treebank را از بین corpusهای موجود در این کتابخانه import کنید. توجه کنید که در این تمرین باید از tagged scentences استفاده کنید.

الف) این جملات را می‌توانید هم به صورت عادی و هم با 'tag-set='universal لود کنید. تفاوت این دو در چیست؟ یک جمله را در هر یک از این حالات بررسی کنید و در ادامه‌ی این سوال، برای سادگی جملات را با مجموعه تگ‌های universal استفاده کنید.

ب) داده‌ها را به سه مجموعه‌ی train، validation و test افراز کنید. درصدی از داده که برای هر یک این دسته‌ها در نظر گرفتید را گزارش کنید.

پ) الگوریتم Viterbi را برای تعیین نقش کلمات در جمله پیاده‌سازی کنید. همچنین سودوکد آن را در گزارش نهایی خود ذکر کنید. دقت مدل را بر روی دادگان تست گزارش کنید.

ت) برخی از کلمات که نقش اشتباهی برای آن‌ها تشخیص داده‌اید را انتخاب کنید و حدس خود، در مورد علت این خطاها را توضیح دهید.

ث) از چه روشی برای برخورد با کلمات ناشناخته در داده‌ی تست استفاده کردید؟ به صورتی کلی برای تشخیص بهتر نقش این کلمات، چه راهکارهایی پیشنهاد می‌کنید؟

ج) این بار مسئله را با استفاده از مدل‌های بازگشتی یا RNNها حل کنید. با استفاده از داده‌ی validation، پارامترهای مدل خود را تنظیم کنید. ۳ مقدار برای اندازه‌ی hidden-layer گزارش کنید و تفاوت نتایج آن‌ها را توضیح دهید. اهمیت استفاده از داده‌ی validation به جای داده‌ی test برای تعیین این پارامترها را ذکر کنید.

چ) بعد از تعیین پارامترها، دقت را بر روی دادگان تست گزارش کنید.

ح) قسمت‌های (ج) و (چ) را برای LSTM و GRU تکرار کنید و نتایج را گزارش کنید. تفاوت عملکرد این سه مدل بازگشتی را چگونه توجیه می‌کنید؟

خ) گیت‌های مختلف LSTM را توضیح دهید و همچنین تفاوت آن با GRU را بیان کنید.

د) بهترین نتیجه از بین این سه مدل بازگشتی را با نتیجه‌ی قسمت (پ) مقایسه کنید و تحلیل خود را ارائه دهید.

۲- تشخیص گروه‌های اسمی

در این سوال هم از همان دیتاست Penn Treebank در سوال قبل استفاده خواهیم کرد.

الف) این بار tagها را در مود universal قرار ندهید و با استفاده از کتابخانه‌ی nltk، برای هر کلمه در سیستم BIO تگ مربوط به named entity را مشخص کنید. از این تگ‌ها به عنوان برچسب داده‌ها استفاده کنید.

ب) بعد از تقسیم داده به داده‌ی train و test، الگوریتم Viterbi پیاده‌سازی‌شده در بخش قبل را به صورتی تغییر دهید که با استفاده از آن بتوانید Named Entityها را تشخیص دهید. توجه کنید که باید از رخداد توالی‌های غیرممکن هم جلوگیری کنید. تغییرات خود را توضیح دهید.

پ) عملکرد مدل بر روی داده‌های تست، با معیارهای recall، precision و F1 را گزارش کنید.

ت) آیا می‌توانیم از مدل‌های بازگشتی برای حل مسئله‌ی named entity recognition استفاده کنیم؟ چه چالش‌هایی در این بین وجود دارد؟ برای حل آن معمولا از چه روش‌هایی استفاده می‌شود؟

ملاحظات (حتما مطالعه شود)

• تمامی کدها و نتایج شما باید در یک فایل فشرده با عنوان HW3LastNameStudentID.zip بارگذاری شود.

• خوانایی و دقت بررسی و تحلیل‌های شما در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. همچنین تمرین‌هایی که به صورت عکس در سایت بارگذاری شوند تصحیح نخواهند شد. گزارش نهایی را حتما به صورت pdf در کنار سایر فایل‌ها آپلود کنید.

• گزارش شما به صورت تایپ‌شده و pdf شامل شرح بررسی‌ها، پارامترها، نتایج و تحلیل‌ها خواهد بود.

• پاسخ‌های ارائه‌شده باید نتیجه‌ی فعالیت شخص شما باشد و در صورت مشاهده‌ی تقلب، نمره‌ی تمامی افراد درگیر، صفر خواهد بود.

• درصورت بروز هرگونه مشکل در فروم درس، و یا با ایمیل sahar.rajabi76@gmail.com در تماس باشید.