تمرین سوم – درس پردازش زبان طبیعی آماری دکتر ممتازی بهار ۱۳۹۸ – دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر	

۱. درجدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین ها آورده شده است:

تاخیر (روز)	میزان نمره منفی
از ۱ الی ۲	هر روز ۵٪
از ۳ الی ۶	هر روز ۱۰٪

توجه داشته باشید در صورت تاخیر بین ۷ تا ۱۴ روز، نمره تمرین از ۵۰٪ محاسبه شده و پس از آن نمره ای تعلق نمی گیرد.

۲. هدف از انجام تمرین ها یادگیری عمیق تر مطالب درسی است. در نتیجه هرگونه کپی برداری موجب کسر نمره خواهد شد.

۳. تا ساعت ۲۳:۱۵ روز ۱۰ خرداد فرصت دارید تمرین را در مودل بارگذاری کنید. تمام فایل های پیاده سازی را به همراه فایل pdf مربوط به گزارش تمرین را در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود (برای مثال HW1_97131024) قرار دهید.

۴. زبان برنامه نویسی برای انجام تمرین: پایتون، جاوا و یا متلب در نظر گرفته شده است.

۵. برنامه های نوشته شده خوانا باشد و کامنت گذاری مناسب باشد (طوری که روند کار کاملا مشخص باشد).

۷. در صورت وجود هر گونه سوال می توانید از طریق ایمیل با تدریس یاران درس در ارتباط باشید:

javadforough@gmail.com و amin.ghsm@aut.ac.ir

تمرین

در این تمرین هدف بررسی دو تکنیک پردازش زبان طبیعی POS tagging و NER می باشد.

برای انجام این تمرین استفاده از تمامی ابزارها مجاز هست. برای مثال می توانید از کتابخانه‌هایی مانند nltk ، Stanford POS tagger و یا Stanford NER استفاده نمایید. در صورت تمایل می‌توانید بدون استفاده از ابزارهای فوق بخش اول را با استفاده از مدل مخفی مارکوف و بخش دوم را با استفاده از ماکزیمم آنتروپی پیاده‌سازی نمایید.

در بخش POS tagging هدف کار با داده فارسی و در بخش NER هدف کار با داده انگلیسی است.

بخش اول – POS tagging

هدف این قسمت از تمرین این هست که با استفاده از مجموعه داده بی جن خان محدود شده و ابزار های موجود بهترین دنباله POS متناظر با جمله ورودی را به دست آوردی. کد ارسال شده 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.

الف) همراه با صورت تمرین، دو فایل آموزش POStr.txt و آزمون POSte.txt موجود می باشد. ابزارهای مذکور را با استفاده از مجموعه داده آموزشی، آموزش داده و سپس توسط مجموعه داده آزمون Accuracy مدل را بدست آوردید.

ب) برای داده های آزمون Confusion Matrix را بدست آورید.

ج) Confusion Matrix را نرمال کرده و تحلیل نمایید که بیشترین خطا ناشی از چه بوده است (راهنمایی: با استفاده از پیدا کردن اندیس بزرگترین مولفه های غیر قطری).

برای مثال:

$$confusion = \begin{bmatrix} 360 & 240 \\ 190 & 210 \end{bmatrix}$$

$$confusion\_normal = \begin{bmatrix} 360/600 & 240/600 \\ 190/400 & 210/400 \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 \\ 0.475 & 0.525 \end{bmatrix}$$

در این مثال عدد مربوط به سطر ۲ و ستون ۱ بزرگترین عنصر غیر قطری است که نشان می دهد بیشترین خطا ناشی از شناسایی برچسب ۱ به جای برچسب ۲ می باشد

بخش دوم – NER

دادگان مورد نیاز به همراه صورت سوال با نامهای NERtr.txt و NERte.txt داده شده است.

با یادگیری مدل توسط داده های آموزش و برچسب زنی داده های آزمون مقادیر Precision و Recall را برای داده های آزمون به صورت Exact match به دست آورده و مانند بخش اول پس از به‌دست آوردن ماتریس Confusion ، بیشترین خطاهای سیستم را به‌دست آوردید.

گزارش

در گزارش تمرین علاوه بر ارائه نتایج بخش‌های اول و دوم شرح مختصری از ابزار های استفاده شده در دو بخش قبل را نیز ارائه نمایید.