


| | |
|---|---|
|  | <p>تمرین دوم – درس پردازش زبان طبیعی آماری</p> <p>دکتر ممتازی</p> <p>ترم زمستان ۹۷-۱۳۹۶ – دانشکده کامپیوتر، دانشگاه صنعتی امیرکبیر</p> <p>زمان تحویل: ۸ اردیبهشت ۹۷</p> |
|---|---|

۱. در جدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین‌ها آورده شده است:

| میزان نمره منفی | تاخیر (روز) |
|-----------------|-------------|
| هر روز ۰.۵٪ | از ۱ الی ۲ |
| هر روز ۱.۰٪ | از ۳ الی ۶ |

توجه داشته باشید در صورت تاخیر بین ۷ تا ۱۴ روز، نمره تمرین از ۵۰٪ محاسبه شده و پس از نمره‌ای تعلق نمی‌گیرد.

۲. هدف از انجام تمرین‌ها، یادگیری عمیق‌تر مطالب درسی است. در نتیجه هرگونه کپی‌برداری موجب کسر نمره خواهد شد.

۳. تا ساعت ۲۳:۵۵ روز ۸ اردیبهشت فرصت دارید تا تمرین را در مدل بارگذاری کنید. تمام فایل‌های پیاده‌سازی را به همراه فایل، pdf مربوط به گزارش تمرین، در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود قرار دهید. (برای مثال HW2_95131105)

۴. زبان برنامه‌نویسی برای انجام تمرین‌ها پایتون، جاوا و یا متلب در نظر گرفته شده است.

۵. برنامه‌های نوشته شده خوانا باشد و کامنت‌گذاری مناسب باشد (طوری‌که روند کار کاملاً مشخص باشد).

۶. در فایل گزارش درباره کد توضیح ندهید! فقط کافی است نتیجه به دست آمده را در گزارش قرار داده و مختصراً آن را تحلیل نمایید.

۷. در صورت وجود هرگونه سوال می‌توانید از طریق ایمیل با تدریس‌یاران درس در ارتباط باشید:

rahbararman@aut.ac.ir a.heidarnasab@aut.ac.ir

در این تمرین هدف بررسی دو تکنیک پردازش زبان طبیعی POS tagging و NER می‌باشد. در این تمرین استفاده از تمامی ابزارها مجاز است. به عنوان مثال می‌توانید از ابزارهایی مانند Stanford pos tagger و Stanford NER استفاده کنید.

POS tagging

در این قسمت از تمرین هدف این است که با استفاده از مجموعه داده بی‌جن‌خان محدود شده و ابزارهای موجود بهترین دنباله POS متناظر با جمله ورودی را پیدا کنید. کد ارسال شده قادر باشد که یک فایل ورودی به نام in.txt را دریافت کند و متن برچسب‌زده شده را در فایل دیگری به نام out.txt تولید نماید.

الف) همراه با صورت تمرین، دو فایل آموزش POSStr.txt و آزمون POSTe.txt موجود است. مدل مخفی مارکوف را با استفاده از مجموعه داده آموزشی، آموزش داده و سپس توسط مجموعه داده آزمون Accuracy مدل را بدست آورید.

ب) برای داده‌های آزمون Confusion Matrix را بدست آورید.

ج) Confusion Matrix را نرمال کرده و تحلیل نمایید که بیشترین خطا ناشی از چه بوده است.

مثلاً:

$$confusion = \begin{bmatrix} 240 & 560 \\ 120 & 180 \end{bmatrix}$$
$$confusion_normal = \begin{bmatrix} 240/800 & 560/800 \\ 120/300 & 180/300 \end{bmatrix} = \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$$

NER

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

با یادگیری مدل مناسب توسط داده‌های آموزش و برچسب‌زنی داده‌های آزمون مقادیر Precision و Recall را برای داده‌های آزمون بصورت:

۱- Type match

۲- Boundary match

۳- Exact match

و برای حالت Exact match ماتریس Confusion را نیز بدست آورید.

در پایان شرح مختصری از ابزارهای استفاده شده نیز در گزارش ارائه نمایید.

توجه: در صورت تمایل می‌توانید روش‌های POS Tagging و یا NER را خودتان پیاده‌سازی نمایید. به عنوان مثال مدل مخفی مارکف را طراحی کنید و سپس توسط الگوریتم ویتربی بهترین دنباله POS متناظر با جمله ورودی را پیدا کنید.