به نام خدا

تمرین دوم درس پردازش زبان طبیعی

«آشنایی با برچسب‌گذاری دنباله‌ای[1] به صورت POS Tagging[2] و NER[3]»

استاد درس: دکتر ممتازی

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

بهار ۱۴۰۳

برای ارسال تمرین به نکات زیر توجه کنید.

۱- برای ارسال پاسخ تمرین‌های این درس، **مجموعا ۱۰ روز** زمان تاخیر مجاز در نظر گرفته شده‌است و در صورتی که مجموع زمان تاخیرها از این مقدار بیشتر باشد، پاسخ ارسال‌شده مورد بررسی قرار نخواهد گرفت.

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

۳- آخرین مهلت ارسال تمرین، **ساعت ۲۳:۵۵ روز شنبه ۲۲ اردیبهشت** می‌باشد.

۴- فایل‌های ارسالی جهت نمره‌دهی باید شامل پیاده‌سازی و گزارش تمرین باشد، فایل‌های خود را فشرده نمایید و به‌صورت **«شماره دانشجویی_HW2»** مانند HW2_400131022 نام‌گذاری کنید.

۵- زبان برنامه‌نویسی برای انجام این تمرین، تنها می‌تواند **پایتون** باشد.

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

۷- در این تمرین شما باید **موارد خواسته‌شده را پیاده‌سازی نمایید** و استفاده از کتابخانه‌های آماده مجاز نمی‌باشد.

۸- در صورت هرگونه سوال یا مشکل می‌توانید با تدریس‌یار درس از طریق ایمیل زیر در ارتباط باشید.

محمد نعیمی mohammad.naeimi+nlp@aut.ac.ir

1 Sequence Labeling
2 Part-of-Speech Tagging
3 Named-Entity Recognition

بخش اول: تعریف مسئله و معرفی مجموعه‌داده

در این تمرین هدف بررسی مدل برچسب‌گذاری دنباله‌ای در دو وظیفه پردازش زبان طبیعی POS Tagging و NER می‌باشد. لازم است با استفاده از توضیحات این تمرین و مجموعه‌داده‌های ارائه‌شده، این دو وظیفه پیاده‌سازی شوند.
دو مجموعه‌داده متفاوت در این تمرین، هرکدام شامل سه قسمت Train، Val و Test که به ترتیب برای آموزش[۴]، اعتبارسنجی[۵] و آزمون[۶] می‌باشند، در اختیار شما قرار گرفته‌است. این دو مجموعه‌داده مربوط به وظایف POS Tagging و NER می‌باشند که مجموعه‌ای از جملات فارسی به همراه دنباله برچسب‌های متناظر می‌باشند.

الف) هر قسمت از مجموعه‌داده ارائه‌شده برای POS Tagging شامل متن داده و برچسب‌های POS آن می‌باشد و تعداد ورودی‌های هریک از قسمت‌ها در مجموعه‌داده POS Tagging به‌شرح زیر است:

نام فایل	تعداد ورودی‌ها
Train	10,000
Val	2,000
Test	2,000

ب) در مجموعه‌داده NER نیز هر قسمت از مجموعه‌داده ارائه‌شده شامل متن و برچسب NER آن می‌باشد و تعداد ورودی‌های هریک از قسمت‌ها در مجموعه‌داده NER به‌شرح زیر است:

نام فایل	تعداد ورودی‌ها
Train	10,000
Val	2,000
Test	2,000

* راهنمایی: می‌توانید برای خواندن فایل‌های مجموعه‌داده‌ها از ابزار pandas.read_json() استفاده نمایید.

4 Train
5 Validation
6 Test

بخش دوم: علامت‌گذاری اجزای سخن "Part-of-Speech Tagging" (۴۰ امتیاز)

هدف ما در این قسمت از تمرین این هست که بهترین دنباله POS Tagging متناظر با جمله ورودی را به دست آوریم.

الف) در این قسمت مجموعه‌داده‌ای از متن و دنباله POS Tagging متناظر به شما داده شده‌است و باید با استفاده از یکی از روش‌های شبکه عصبی بازگشتی (RNN)، یک مدل POS Tagging ساخته‌شود که قادر باشد یک جمله ورودی را دریافت کند و دنباله برچسب‌های متناظر با آن را تولید کند.

ب) با آموزش و تنظیم مدل توسط داده‌های آموزش و ارزیابی و سپس علامت‌گذاری داده‌های آزمون، مقادیر Accuracy و همچنین Recall ،Precision و F1-score را برای داده‌های آزمون به دست آورید و گزارش کنید.

ج) به مدل پیاده‌سازی شده در بند «الف» یک لایه CRF اضافه کنید و مجدداً نتایج را براساس بند «ب» ارزیابی کنید.

د) برای بهترین مدل پیاده‌سازی‌شده خود بین بند «الف» و «ج» تجزیه و تحلیل خطا[۷] انجام دهید. با استفاده از Confusion Matrix بیشترین خطاهای مدل را به دست آورده، گزارش نمایید و نتایج را تحلیل کنید.

بخش سوم: شناسایی موجودیت نام‌گذاری‌شده "Named-Entity Recognition" (۴۰ امتیاز)

هدف ما در این قسمت از تمرین این هست که بهترین دنباله NER متناظر با جمله ورودی را به دست آوریم.

الف) در این قسمت مجموعه‌داده‌ای از متن و دنباله NER متناظر به شما داده شده‌است و باید مشابه بخش دوم با استفاده از یکی از روش‌های شبکه عصبی بازگشتی (RNN)، یک مدل NER ساخته شود که قادر باشد یک جمله ورودی را دریافت کند و دنباله برچسب‌های متناظر با آن را تولید کند.

ب) مدل را توسط داده‌های آموزش و ارزیابی، آموزش‌داده و تنظیم نمایید، داده‌های آزمون را با استفاده از مدل آموزش‌دیده برچسب‌زنی نمایید و مقادیر Recall ،Precision و F1-score را برای داده‌های آزمون به دو صورت token level و entity level به دست‌آورده و گزارش نمایید.

ج) در این وظیفه نیز با اضافه کردن یک لایه CRF به مدل نتایج را مجدداً براساس بند «ب» بررسی و تحلیل کنید.

د) برای بهترین مدل پیاده‌سازی‌شده خود بین بند «الف» و «ج» تجزیه و تحلیل خطا انجام دهید. با استفاده از Confusion Matrix بیشترین خطاهای تخصیص نوع موجودیت[۸] مدل را به دست آورده، گزارش نمایید و نتایج را تحلیل کنید.

---

7 Error Analysis
8 Entity Type

بخش چهارم: استفاده از روش‌های ارزیابی و تحلیل نتایج (۲۰ امتیاز)

الف) همانطور که توضیح داده‌شد، برای بررسی عملکرد مدل‌های برچسب‌زنی دنباله‌ای، معیارهای ارزیابی Accuracy، Precision، Recall و F1-score را بر روی مجموعه‌داده آزمون محاسبه می‌کنیم. این معیارهای ارزیابی در برچسب‌زنی دنباله‌ای، به خصوص در وظیفه NER، اندکی متفاوت با وظایف دیگر، مانند دسته‌بندی[۹]، پیاده‌سازی و استفاده می‌شوند. این معیارهای ارزیابی را متناسب با وظیفه برچسب‌زنی دنباله‌ای پیاده‌سازی نمایید و توضیح دهید.

ب) دو روش token level و entity level در معیارهای ارزیابی مدل‌های برچسب‌زنی دنباله‌ای را توضیح دهید و تفاوت آن‌ها را بیان نمایید.

موفق باشید

محمد نعیمی

9 Classification