به نام خدا

تمرین سوم درس فهم زبان، پاییز ۱۴۰۱

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

در این تمرین به مسئله **پرسش-پاسخ استخراجی** (Extractive Question Answering) خواهیم پرداخت. در این مسئله با داشتن یک متن (Passage) و یک پرسش، تلاش می‌شود تا ناحیه‌ی پاسخ آن پرسش در متن پیدا شود. برای این تمرین از یک مجموعه داده فارسی استفاده خواهیم کرد. این مجموعه داده PQuAD نام دارد که در تمرین ضمیمه شده است. یک نمونه از این مجموعه داده را در جدول ۱ مشاهده می‌کنید. سه بخش آموزش (train)، اعتبارسنجی (validation) و آزمون (test) در این مجموعه داده به تفکیک آورده شده است که شما باید از آن‌ها برای دستیابی به بهترین مدل استفاده کنید.

متن	سوال	پاسخ	اندیس شروع پاسخ
والیبال یک ورزش گروهی و ششمین ورزش پر طرفدار و گسترده‌ترین ورزش در بعضی کشورهای جهان است که در آن بازیکنان در دو تیم شش نفره، در دو سوی توری قرار می‌گیرند و تلاش می‌کنند تا طبق قوانین بازی، توپ را از روی تور در زمین تیم مقابل فرود آورند. طول زمین والیبال ۱۸ متر و عرض آن ۹ است. هرتیم حداقل باید ۳ ست (یا دست) از ۵ پنج ست بازی را ببرد، تا بتواند پیروز مسابقه شود.	هرتیم والیبال چگونه می‌تواند پیروز مسابقه شود؟	هرتیم حداقل باید ۳ ست (یا دست) از ۵ پنج ست بازی را ببرد، تا بتواند پیروز مسابقه شود.	۲۸۱

جدول ۱ نمونه‌ای از مجموعه داده PQuAD

بخش ۱: برای انجام این تمرین در گام اول شما باید از یک مدل ترنسفومری که بر روی زبان فارسی Pretrain شده است استفاده کنید. توصیه می‌شود که از مدلی مانند ParsBERT استفاده کنید که از طریق لینک زیر در دسترس است.

https://huggingface.co/HooshvareLab/bert-fa-base-uncased

پس از انتخاب مدل مورد نظر شما باید فاز finetuning مدل را بر روی مجموعه داده PQuAD انجام دهید و مدل را ارزیابی کنید.

بخش ۲: در گام دوم، از مدل BERT آموزش ندیده که بر روی هیچ داده‌ای Pretrain نشده است استفاده کنید و با استفاده از مجموعه داده PQuAD فرایند آموزش را انجام دهید (از آنجایی که در این حالت داده آموزشی کم است، احتمالاً نیاز باشد اندازه مدل را کوچک کنید. همچنین به راحتی این مدل روی این حجم از داده آموزش خواهد دید. هدف گرفتن بهترین نتیجه است). تفاوت کارایی این مدل با مدل قبل را گزارش کنید. آیا Pretraining سبب بهبود نتایج شده است؟

بخش ۳ (۲۰ درصد امتیازی): برای این بخش شما باید مسئله پرسش پاسخ استخراجی را به صورت دو زبانه انجام دهید. منظور از دو زبانه، زبان فارسی و انگلیسی است. برای زبان فارسی از مجموعه داده PQuAD و برای زبان انگلیسی از مجموعه داده SQuAD 2.0 استفاده کنید. در این جا نیز باید از مدل BERT استفاده شود. لزومی ندارد که حتماً از مدل Pretrain استفاده

کنید و فقط گرفتن بهترین نتیجه مورد نظر است. توجه کنید که مدل باید **همزمان** بر روی این دو زبان آموزش ببیند و نباید مدل‌ها جدا جدا باشند.

ارزیابی: برای ارزیابی مدل شما باید از دو معیار F1 و Exact Match استفاده کنید و آنها را روی دادگان تست گزارش و تحلیل کنید.

مشابه تمرین اول و دوم برای اینکه ترقیب شوید به انجام هرچه بهتر تمرین و انتخاب بهترین استراتژی‌ها، به نفرات اول و دوم بر اساس F1-score به ترتیب ۱۰ و ۵ درصد تشویقی نمره اضافه داده خواهد شد (البته نتایج باید قابل قبول باشد)

در این تمرین امکان اینکه ایده‌های دیگری بجای موارد اینجا بزنید وجود دارد. مثلاً اینکه از مدل‌های جدیدتر استفاده کنید. نتایج روش پایه را در مقاله خانم دکتر ممتازی می‌توانید مشاهده کنید:

https://arxiv.org/abs/2202.06219