دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
بهار ۱۴۰۳
استاد: محمدحسین رهبان
گردآورندگان:

پروژه پردازش زبان‌های طبیعی | موضوع | مهلت ارسال: ۱ خرداد

۱ شرح مختصر پروژه

در این پروژه قرار است تا مقدمات پردازش زبان‌های طبیعی (NLP) آشنا شوید و به کمک یکی از مدل‌های زبانی (LM) مبتنی بر ترنسفورمر یک مسئله‌ی واقعی دسته‌بندی متن را حل کنید. مجموعه‌داده‌ای در این پروژه استفاده می‌شود یک مجموعه‌داده‌ی فیلم‌های فارسی است که با نام «persianmovies.csv» در مستندات پروژه قابل دسترس است. این مجموعه‌داده شامل توضیح مختصری از فیلم‌های فارسی و ویژگی‌های آن‌ها نظیر ژانر و سال تولیدشان به هر دو زبان فارسی و انگلیسی است.
در طی این پروژه با موارد زیر دست و پنجه نرم می‌کنید:

۱. کار با داده‌ها
۲. آموزش مدل دسته‌بندی متن

۲ کار با داده‌ها

یکی از مهم‌ترین کارها قبل از آموزش مدل و پیش‌بینی، جمع آوری داده و تمیز کردن داده است! قبل از هر چیزی نیاز است که داده‌هایمان را بشناسیم. در نوتبوک داده شده از شما خواسته شده که EDA و گام های مختلف تحلیل داده را برای دو زبان فارسی و انگلیسی به صورت جداگانه انجام دهید.
توصیه می‌شود که علاوه بر کارهای گفته شده، تحلیل‌ها و تنظیم‌های بیشتری را نیز روی مجموعه دادگان انجام دهید.
گام های خواسته شده از شما به طور کلی به صورت زیر هستند:

۱. پیش پردازش داده
۲. توصیف آماری
۳. تصویرسازی داده
۴. تحلیل ژانر
۵. تحلیل مبتنی بر زمان
۶. تحلیل رتبه بندی
۷. همبستگی و تحلیل چندمتغیره
۸. شناسایی مقادیر پرت
۹. متعادل سازی داده ها

دقت کنید که حتما اگر برای Balance کردن دیتاست، تصمیم به حذف دادگان یا اضافه کردن آنها می گیرید(منظور از حذف دادگان داده های null نیست و البته دقت کنید که داده های null نیز نباید در دیتاست باشند)، داده تست را از قبل جدا کنید که بعدا در اعتبارسنجی دچار خطای overfit نشوید! سپس با توجه به تمام تحلیل‌ها و نمودارها و تقسیم داده‌ها به دو بخش train و test مدل را آموزش دهید. (random_seed=42)
از هرگونه دستکاری عمدی و بی دلیل دادگان test برای بالاتر بردن دقت جواب نهایی بپرهیزید!

۳ آموزش مدل دسته‌بندی متن

در این بخش قرار است به کمک مدل زبانی ،BERT یک مدل دسته‌بندی برای ژانر فیلم‌ها آموزش دهید. به صورت خلاصه، ورودی مدل شما، خلاصه‌ای از فیلم است و خروجی آن یک دسته‌بندی بین ژانرهای موجود است.
روند کار به این صورت است که مدل زبانی، ورودی داده شده را به یک بازنمایی (Representation) می‌برد و سپس به کمک یک لایه‌ی Fully Connected دسته‌بندی مورد نظر انجام می‌شود.
روش‌های متنوعی برای انجام این کار وجود دارد که ما از Full Fine-tuning مدل به صورت end-to-end استفاده می‌کنیم. به این صورت که تمامی وزن‌های مدل زبانی BERT به همراه دسته‌بند موجود روی CLS head آموزش و به‌روز می‌شوند.
در این بخش ۴ مرتبه مدل BERT را روی داده‌هایی که در قسمت قبل آماده کرده‌اید، آموزش می‌دهید و نتایج را مشاهده می‌کنید:

۱. آموزش مدل ParsBERT روی داده‌های فارسی
   (آ) داده‌های فارسی پیش‌پردازش شده
   (ب) داده‌های فارسی پیش‌پردازش نشده
۲. آموزش مدل BERT روی داده‌های انگلیسی
   (آ) داده‌های انگلیسی پیش‌پردازش شده
   (ب) داده‌های انگلیسی پیش‌پردازش نشده

برای مدل BERT از نسخه‌ی bert-base-uncased و برای مدل ParsBERT از نسخه‌ی HooshvareLab/bert-fa-zwnj-base استفاده کنید؛ همچنین برای آموزش مدل می‌توانید از Trainer موجود در Huggingface استفاده کنید.
توجه: توصیه می‌شود برای آموزش مدل‌های حتما از Google Colab **استفاده کنید.**