به نام خدا

دانشگاه تهران

دانشکده مهندسی

برق و کامپیوتر

درس داده کاوی

تمرین امتیازی

طراح	محمد صادق صادقی Mhmssadeghi74@gmail.com
تاریخ بارگذاری	۱۴۰۳/۴/۲
مهلت ارسال	۱۴۰۳/۴/۱۶

فهرست

مقدمه

Topic Modelingروشی در پردازش زبان طبیعی است که برای استخراج موضوعات مختلف از مجموعه‌ای از اسناد متنی به کار می‌رود. این روش به‌طور خودکار گروه‌هایی از کلمات را شناسایی می‌کند که معمولاً با هم ظاهر می‌شوند و موضوعات مختلف را نمایان می‌سازند. از سوی دیگر، Multi-label Classificationنوعی از یادگیری ماشین است که در آن هر نمونه می‌تواند به بیش از یک دسته تعلق داشته باشد. این روش در مواردی مفید است که هر سند یا شیء دارای ویژگی‌های چندگانه باشد و نیاز به طبقه‌بندی در چندین برچسب یا دسته‌بندی همزمان داشته باشد. در این تمرین از شما انتظار می‌رود مدلی را آموزش بدهید که توانایی تشخیص موضوعات اسناد را داشته باشد.

مجموعه داده

این مجموعه داده شامل مجموعه‌ای از چکیده و عنوان مقالات پژوهشی است که هدف اصلی آن، پیش‌بینی موضوعات مختلف برای هر مقاله پژوهشی خواهد بود. به عبارت دیگر، با استفاده از چکیده و عنوان هر مقاله، باید بتوانیم یک یا چند موضوع کلیدی که مرتبط با آن مقاله است را پیش‌بینی کنیم. این مجموعه داده به علت اهمیت برچسب‌گذاری در جستجو و توصیه مقالات علمی، بسیار ارزشمند است و برای مطالعات در حوزه‌های مختلفی مانند

- Computer Science
- Physics
- Mathematics
- Statistics
- Quantitative Biology
- Quantitative Finance

مناسب است.

بخش عملی

در این تمرین از شما انتظار می‌رود با به کارگیری شبکه عصبی مناسب، وظیفه‌ی دسته‌بندی اسناد موجود در مجموعه داده را انجام بدهید. همانطور که در توضیحات مجموعه داده، بیان شده است، هر سند در این مجموعه داده می‌تواند متعلق به چندین دسته‌ی مختلف باشد. بنابراین مدل آموزش داده شده باید توانایی داشته باشد چندین برچسب را به مستندات موجود در داده‌ی تست را اختصاص دهد.

1- در ابتدا لازم است که مجموعه داده مورد بررسی قرار گرفته و نحوه‌ی توزیع هر کدام از مستندات در هر کدام از دسته موضوعات را با نمودار مناسب نمایش دهید.
2- حداقل دو روش کلی را برای وظیفه‌ی multi label classification به صورت کامل بررسی کنید.
3- یکی از روش‌های اشاره شده در بخش دو را با کمک شبکه‌ی عصبی مناسب پیاده‌سازی نمایید. (گزارش کامل خود را از نحوه‌ی آموزش مدل و پارامترهای بکارگرفته شده بیان کنید)
4- مدل آموزش دیده را بر روی داده‌های تست، ارزیابی کرده و نتایج را بیان کنید.

توجه:

- **در این بخش شما میتوانید از کتابخانه‌هایی مانند Pytorch استفاده کنید.**
- **مجموعه داده را به نسبت ۸۰/۱۰/۱۰ به داده‌های آموزش، تست و اعتبارسنجی تقسیم کنید.**
- **به منظور استخراج ویژگی‌ها از متون می‌توانید با کمک کتابخانه‌های آماده و روش‌هایی مثل Tf-idf بهره ببرید.**
- **به تمرین‌هایی که صرفا پیاده‌سازی کد هستند نمره‌ای تعلق نخواهد گرفت و گزارش ارائه شده معیار اصلی نمره شما خواهد بود.**

ملاحظات

تمامی نتایج شما باید در یک فایل فشرده با عنوان DM_EXTRA_StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- بخش اصلی نمره به گزارش شما تعلق می‌گیرد و دستیاران الزامی برای اجرای تمام کدهای شما در صورتی که در گزارش به آن‌ها اشاره‌ای نکرده باشید ندارند. لطفا تمام موارد مورد نیاز را در گزارش ذکر کنید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- برای تحویل تمارین از چارچوب قرارداده شده در سامانه و کانال تلگرام استفاده کنید.
- در صورت قصد ارسال تمرین به صورت دیگر (انگلیسی، latex و ...)، لطفا پیش از ارسال با دستیارمسئول تمرین هماهنگ کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (هم‌فکری خارج از چارچوب و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب برای همه‌ی افراد مشارکت کننده، نمره تمرین، صفر در نظر گرفته خواهد شد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

  mailto:mhmssadeghi74@gmail.com

مهلت تحویل: ۱۴۰۳/۴/۱۶