به نام خدا

درس پردازش زبان طبیعی

تکلیف برنامه نویسی: دسته بندی متون با استفاده از r

دانشگاه تهران
گروه علوم

در این تمرین، هدف دسته‌بندی متون بر مبنای بکارگیری مدل از پیش آموزش دیده BERT است. دادگان مورد نظر در این تمرین دادگان پرسیکا است. پرسیکا پیکره‌ای است حاوی متون خبری برگرفته از خبرگزاری ایسنا. متون این پیکره در یازده طبقه موضوعی شامل ورزشی، اقتصادی، فرهنگی، مذهبی، تاریخی، سیاسی، علمی، اجتماعی، آموزشی، حقوق قضایی و بهداشت طبقه‌بندی شده‌اند و پیش‌پردازش‌هایی به منظور قابل استفاده بودن در کاربردهای مختلف پردازش زبان طبیعی و داده‌کاوی بر روی آن‌ها انجام گرفته است. دادگان را به دو بخش آموزش (80%) و آزمون (20%) تفکیک کنید و از بخش آموزش برای پیش آموزش مدل استفاده نمایید. سامانه پیاده‌سازی شده را برای دسته‌بندی اسناد مجموعه آزمون اجرا کرده و دقت عملکرد آن را به ازای پارامترهای مختلف و مدل‌های مختلف پیش آموزش دیده (ParsBERT و مدل‌های چند زبانه) محاسبه نمایید و مشاهدات و نتایج حاصله را در قالب یک گزارش ارایه نمایید و با نتایج حاصله در تمرین قبل مقایسه نمایید.
برای دانلود دادگان پرسیکا و کسب اطلاعات بیشتر در خصوص آن به سایت https://www.peykaregan.ir/ مراجعه نمایید.

لطفا کد به همراه گزارش را در قالب یک فایل فشرده شده تا قبل از موعد اعلام شده به آدرس ایمیل ut.cs.exam@gmail.com ارسال نمایید.

Format : FirstName.LastName.HW5
EX: Bagher.BabaAli.HW5

با آرزوی سربلندی