

**دانشگاه تهران**

**گروه علوم کامپيوتر د**

به نام خدا

**درس پردازش زبان طبيعی**

**تکليف برنامه نويسی: دسته بندی متون با استفاده از Transformer**

|  |
| --- |
| در اين تمرين، هدف دسته­بندی متون بر مبنای بکارگيری مدل از پيش آموزش ديده BERT است. دادگان مورد نظر در اين تمرين دادگان پرسيکا است. پرسيکا پيکره‌اي است حاوي متون خبري برگرفته از خبرگزاري ايسنا. متون اين پيکره در يازده طبقه موضوعي شامل ورزشي، اقتصادي، فرهنگي، مذهبي، تاريخي، سياسي، علمي، اجتماعي، آموزشي، حقوق قضايي و بهداشت طبقه‌بندي شده‌اند و پيش‌پردازش‌هايي به منظور قابل استفاده بودن در کاربردهاي مختلف پردازش زبان طبيعي و داده‌کاوي بر روي آن‌ها انجام گرفته است. دادگان را به دو بخش آموزش (80%) و آزمون (20%) تفکيک کنيد و از بخش آموزش برای پيش آموزش مدل استفاده نماييد. سامانه پياده­سازی شده را برای دسته­بندی اسناد مجموعه آزمون اجرا کرده و دقت عملکرد آن را به ازای پارامترهای مختلف و مدل­های مختلف پيش آموزش ديده (ParsBERT و مدل­های چند زبانه) محاسبه نماييد و مشاهدات و نتايج حاصله را در قالب يک گزارش ارايه نماييد و با نتايج حاصله در تمرين قبل مقايسه نماييد.  برای دانلود دادگان پرسيکا و کسب اطلاعات بيشتر در خصوص آن به سايت <https://www.peykaregan.ir/> مراجعه نماييد.  لطفا کد به همراه گزارش را در قالب يک فايل فشرده شده تا قبل از موعد اعلام شده به آدرس ايميل  ut.cs.exam@gmail.com ارسال نماييد.  Format : FirstName.LastName.HW5  EX: Bagher.BabaAli.HW5 |

**با آرزوي سربلندي**