به نام خدا

**درس پردازش زبان طبيعی (1400-1399)**

**تکليف برنامه نويسی: دسته بندی متون**

|  |
| --- |
| در اين تمرين هدف دسته­بندی متون بر مبنای مفاهيم ماتريس کلمه-سند؛ روش تجزيه ماتريسی SVD (LSA) است. ابتدا به کمک مفاهيم ماتريس کلمه-سند؛ تجزيه SVD هر سند به يک بردار تبديل کنيد و سپس به کمک سه دسته­بند مختلف کار دسته­بندی متون را انجام دهيد. 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. نيازی به پياده­سازی SVD و دسته­بندها نيست و مجاز به استفاده از ابزارها و کدهای آماده هستيد. سامانه پياده­سازی شده را برای دسته­بندی اسناد مجموعه آزمون اجرا کرده و دقت عملکرد آن را به ازای پارامترهای مختلف نظير طول بردار هر سند و غيره و روش­های مختلف دسته­بندی محاسبه نماييد.  برای دانلود دادگان پرسيکا و کسب اطلاعات بيشتر در خصوص آن به سايت <https://www.peykaregan.ir/> مراجعه نماييد. |

**با آرزوي سربلندي**