به نام خدا

درس پردازش زبان طبیعی

تکلیف برنامه نویسی: دسته بندی متون

دانشگاه تهران
گروه علوم کامپیوتر

در این تمرین هدف دسته‌بندی متون بر مبنای مفاهیم ماتریس کلمه-سند؛ روش تجزیه ماتریسی SVD (LSA) است. ابتدا به کمک مفاهیم ماتریس کلمه-سند؛ تجزیه SVD هر سند به یک بردار تبدیل کنید و سپس به کمک سه دسته‌بند مختلف کار دسته‌بندی متون را انجام دهید. 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 (۸۰%) و آزمون (۲۰%) تفکیک کنید. نیازی به پیاده‌سازی SVD و دسته‌بندها نیست و مجاز به استفاده از ابزارها و کدهای آماده هستید. سامانه پیاده‌سازی شده را برای دسته‌بندی اسناد مجموعه آزمون اجرا کرده و دقت عملکرد آن را به ازای پارامترهای مختلف نظیر طول بردار هر سند و غیره و روش‌های مختلف دسته‌بندی محاسبه 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.HW4
EX: Bagher.BabaAli.HW4

با آرزوی سربلندی