به نام خدا

مبانی داده کاوی

خرداد ۱۴۰۰

تاریخ تحویل: ۱۴۰۰/۰۳/۲۵

پروژه پایانی

کد برنامه نوشته شده به همراه فایل گزارش آن)شامل شرح کار انجام شده، توضیح کد و خروجی‌های دریافتی شما(را در یک فایل با نام STnumber_DM_Final زیپ کرده و ارسال کنید.

---

۱. **[پیاده‌سازی- دسته‌بندی متون فارسی با بیز ساده]** یک برنامه رایانه‌ای بنویسید که با استفاده از دادگان مجموعه *زبرا*، متون فارسی را با روش بیز ساده دسته‌بندی کند. برای این کار، هر کدام از ۷ پوشه دادگان زبرا را به عنوان ۷ موضوع (دسته) در نظر بگیرید. در این دادگان مجموعه داده تست و آموزش از همدیگر جدا شده‌اند.

برای انتخاب کلمات (واژگان) ابتدا ۵۰۰ کلمه پرکاربرد کل متن‌ها (آموزش و آزمون) را استخراج کرده و سپس برای آنها مدل‌ها (احتمال‌ها) را در یک (یا چند فایل) از روی داده آموزشی تولید کند. در صورت لزوم برای جلوگیری از صفر شدن احتمال‌ها، از روش هموارسازی لاپلاس با $\alpha = 1$ استفاده کنید.

پس از ساخت مدل‌ها، بر روی داده تست، مقدار Accuracy و ماتریس درهم‌ریختگی (Confusion Matrix) آن را محاسبه کنید. همچنین مقدار سه معیار Precision، Recall و F-Measure را بدست آورید.

---

۲. **[پیاده‌سازی- دسته‌بندی متون فارسی با نزدیک‌ترین همسایه]** یک برنامه بنویسید که با استفاده از دادگان مجموعه *زبرا*، متون فارسی را با روش نزدیک‌ترین همسایه دسته‌بندی کند. برای این کار، از هر سند (فایل) یک بردار ویژگی استخراج کنید و بردارها را با روش فاصله کسینوسی با همدیگر مقایسه کنید. برای استخراج ویژگی از هر سند، یک بردار ۵۰۰ بعدی در نظر بگیرید که هر کدام یک کلمه است و تعداد رخداد هر کلمه در هر سند را به عنوان بردار ویژگی آن سند در نظر بگیرید. به این روش در فراوانی عبارت (TF: Term Frequency) گفته می‌شود.

برای این روش نیز مقدار Accuracy را برای سه مقدار K=1، K=5 و K=15 و ماتریس درهم‌ریختگی آنها را بدست آورید. حال برای هر سند بردار TF-IDF آن را بدست آورید و مراحل بالا را تکرار کنید. (از فاصله کسینوسی برای سنجش شباهت اسناد استفاده کنید). در مرحله آخر بررسی کنید که پس از حذف Stop Words های فارسی چه تغییری در دقت مدل ساخته ایجاد میشو د؟

فایل stop word های فارسی در لینک زیر آورده شده است :
https://github.com/kharazi/persian-stopwords