تشخیص نویسنده با استفاده از دیتاست Newsgroup20

محمدرضا تاجیک

400521198

هدف

آشنایی با نحوه استفاده از ماشین بردار پشتیبانی (SVM) برای دسته‌بندی متون و تشخیص نویسنده.

توضیحات پروژه

در این پروژه، هدف تشخیص نویسنده از طریق دسته‌بندی متون با استفاده از دیتاست 20 Newsgroupsاست. این پروژه شامل مراحل زیر است:

1. بارگذاری دیتاست 20 Newsgroups: دیتاست 20 Newsgroups شامل مقالات مختلفی است که به 20 دسته مختلف تقسیم‌بندی شده‌اند. این دیتاست از طریق تابع fetch_20newsgroups از کتابخانه sklearn بارگذاری می‌شود.
2. پیش‌پردازش داده‌ها و استخراج ویژگی‌ها با استفاده از TF-IDF: داده‌های متنی به ویژگی‌های عددی تبدیل می‌شوند. این تبدیل با استفاده از روش TF-IDF (Term Frequency-Inverse Document Frequency)انجام می‌شود که میزان اهمیت کلمات را در مستندات مختلف محاسبه می‌کند. همچنین، کلمات متوقف (stop words)حذف و تعداد ویژگی‌ها به 5000 محدود می‌شود.

3. تقسیم داده‌ها به مجموعه‌های آموزشی و تست :داده‌های پیش‌پردازش شده به دو بخش تقسیم می‌شوند: داده‌های آموزشی که برای آموزش مدل استفاده می‌شوند و داده‌های تست که برای ارزیابی عملکرد مدل به کار می‌روند. این تقسیم‌بندی به صورت تصادفی با نسبت 20/80 انجام می‌شود.
4. آموزش مدل SVM با کرنل خطی :مدل ماشین بردار پشتیبانی (SVM) با استفاده از کرنل خطی و مقدار پارامتر C برابر با 1 آموزش داده می‌شود. این مدل از داده‌های آموزشی برای یادگیری الگوهای دسته‌بندی متون استفاده می‌کند.
5. پیش‌بینی با استفاده از مدل آموزش دیده :مدل آموزش‌دیده برای پیش‌بینی دسته‌بندی مقالات در مجموعه داده تست استفاده می‌شود و دسته‌بندی‌های پیش‌بینی‌شده با دسته‌بندی‌های واقعی مقایسه می‌شوند.
6. محاسبه دقت مدل :دقت مدل با استفاده از معیار accuracy (نسبت درست‌پیش‌بینی‌ها به کل پیش‌بینی‌ها) محاسبه و گزارش می‌شود.
7. گزارش دسته‌بندی :گزارش دسته‌بندی شامل معیارهای مختلف مانندprecision ، recallو f1-score برای هر دسته محاسبه و نمایش داده می‌شود. این گزارش به تفصیل عملکرد مدل را در دسته‌بندی‌های مختلف نشان می‌دهد.
8. رسم ماتریس سردرگمی :برای مشاهده عملکرد مدل به صورت بصری، ماتریس سردرگمی (confusion matrix) رسم می‌شود. این ماتریس نشان می‌دهد که مدل در تشخیص دسته‌های مختلف چگونه عمل کرده است. در این نمودار، محور افقی نشان‌دهنده برچسب‌های پیش‌بینی‌شده و محور عمودی نشان‌دهنده برچسب‌های واقعی است. همچنین، این نمودار با استفاده از کتابخانه seaborn ایجاد می‌شود و هر سلول نشان‌دهنده تعداد مقالاتی است که به درستی یا نادرستی دسته‌بندی شده‌اند.