به نام خدا

Language Detection

در پیاده‌سازی این پروژه از دیتاست WiLI-2018(Wikipedia Language Identification Dataset) استفاده شده‌است. این دیتاست شامل جملاتی از 256 زبان مختلف می‌باشد که از هر زبان 500 دیتا برای train و 500 دیتا برای test موجود است. برای تولید دیتای آموزش این پروژه، دیتای زبان‌های فارسی، آلمانی، انگلیسی، عربی، فرانسوی و ترکی از مجموعه‌های آموزش و تست این دیتاست استخراج شده‌اند؛ یعنی در مجموع از هر زبان 1000 داده برای آموزش مدل استفاده شده‌است.

برای انجام پیش پردازش داده‌های موجود، ابتدا اعداد و علائم نگارشی و نشانه‌گذاری (مانند: ~!@#$%^&*()_+|{}:"?><`/) از هر داده حذف و سپس حروف بزرگ به حروف کوچک تبدیل می‌شوند.

برای تبدیل دیتای داده شده به ویژگی های عددی(استخراج ویژگی)، از تابع CountVctorizer از کتابخانه‌ی sklearn استفاده شده‌است که بیشترین تعداد ویژگی را برابر با 100 قرار داده‌ایم. شئ ساخته‌شده توسط این تابع با داده‌های آموزشی ذکر شده آموزش می‌بیند و سپس بر روی دیتای تست نیز fit می‌شود.

برای انجام عمل طبقه‌بندی از طبقه‌بند Multinomial Naïve Bayes استفاده شده‌است که یکی از معروف‌ترین طبقه‌بندها برای طبقه‌بندی داده‌های متنی است. برای پیاده‌سازی این طبقه‌بند کلاس Multinomial_NB را ساخته‌ایم که دوتابع fit و predict دارد. تابع fit برای آموزش مدل و محاسبه‌ی prior و likelihood کلاس هاست و تابع predict با استفاده از مدل آموزش‌دیده، کلاس مربوط به هر یک از داده‌های تست را پیش‌بینی کرده و برمیگرداند.

برای محاسبه‌ی دقت این طبقه‌بند از رابطه‌ی زیر استفاده شده‌است:

$$\text{accuracy} = \frac{\#\text{truly classified data}}{\#\text{all test data}}$$

برای پیداکردن تعداد داده‌هایی که به‌درستی طبقه‌بندی شده‌اند، کلاس واقعی هر داده را به‌صورت دستی مشخص کرده و با کلاس پیش‌بینی شده توسط مدل مقایسه کرده‌ایم و در نهایت به خروجی زیر رسیده‌ایم که نشان می‌دهد همه‌ی داده‌های تست به‌درستی طبقه‌بندی شده‌اند:

Accuracy is: 1.0

برای ذخیره‌ی نتایج به‌دست آمده، از کتابخانه‌ی*csv* استفاده شده‌است. Header های این فایل به صورت Id,Category هستند. ستون Id در این فایل نشان دهنده‌ی دیتاهای تست و ستون Category نشان دهنده‌ی کلاس پیش‌بینی شده برای هر داده است. این فایل در همان محل فایل پروژه و با نام *test_data* ذخیره می‌شود.

لینک دیتاست :

841984.zenodo/10.5281/org.doi

ریحانه خوارزمی

۹۷۲۴۳۰۳۲