پروژه هوش مصنوعی

ترانه بهار-98243074

در این بخش از پروژه (بخش اول) از ما خواسته شده به کمک الگوریتم naïve bays سوالات و جملاتی را که به آن ورودی میدهیم را پیش بینی و دسته بندی کند، و به ما بگوید این ورودی متعلق به کدامیک از 5 دسته خواسته شده است.

ابتدا فایل train حاوی حدود 3000 نمونه سوال ها و جملات مختلف، از جمله ادبی یا محاوره ای، را از فایل csv داده شده میخوانیم و سوالات را به 5 دسته خواسته شده تقسیم بندی میکنیم. در اینجا برای هر 5 دسته این اعمال را تکرار میکنیم:

1) ابتدا تمامی جملاتی که در فایل train ما دارای عدد label مورد نظر هستند، جدا میکنیم.
2) سپس یک جدول متشکل از تمامی لغات استفاده شده در جملات میسازیم و تعداد هر لغت در هر جمله را مشخص میکنیم.
3) در مرحله بعد، تعداد تمامی لغت های موجود در هر بخش (متعلق به هر label) را میشماریم.
4) تعداد کل لغات در هر بخش را نیز با هم جمع میکنیم.
5) از اینجا، بخش جملات مورد پیش بینی (مربوط به فایل test)، که از پیش آن ها را خوانده و توکنایز کرده ایم(یعنی تمامی بخش ها و لغت های جمله را از هم جدا کرده ایم)، به این بخش از کد میدهیم.

6) سپس با استفاده از فرمول احتمال شرطی بیز، احتمال وجود هر لغت در هر یک از 5 بخش را حساب میکنیم.
7) در بخش شماره 6 این احتمالات برای هر یک از کلمات جمله توکنایز شده به صورت مجزا حساب شده اند. حالا برای اینکه تنها یک عدد برای احتمال هر جمله در هر بخش داشته باشیم، لازم است تمامی اعداد احتمال بدست آمده مربوط به هر کلمه جمله را در یکدیگر ضرب کنیم.
8) پس از انجام عمل ضرب، خروجی احتمال برای هریک از 5 بخش، تنها یک عدد اعشاری خواهد بود.
9) در آخر، پس از انجام تمامی این محاسبات،کافی ست 5 عدد بدست از 5 بخش را با هم مقایسه کنیم. جمله متعلق به بخشی ست که بیشترین عدد را دارا باشد.

مراحل 1 تا 8 را برای تک تک بخش های 1الی5 تکرار میکنیم و سپس بخش 9 را به صورت کلی در پایان برنامه اجرا کرده و خروجی را دریافت میکنیم و آن را در یک فایل csv به نام result، به همراه شماره خط آن مینویسیم.

این عمل را میتوانیم با هر فایل test.csv که حاوی حاوی جملات باشد، انجام دهیم.