گزارش پروژه اول-نگین مشایخی-۹۸۲۴۳۰۵۴

در این پروژه از کتابخانه nltk فقط برای توکن بندی داده ها استفاده کردم.و بقیه توابع را پیاده سازی کردم.از جمله پیش پردازش داده های فارسی و لاتین.
این پروژه شامل چهار بخش اصلی معرفی توابع،اموزش،تست و اجرای اصلی برنامه است
پروژه از نوع دسته بندی است و با توجه به الگوریتم های پر استفاده در این حوزه من از الگوریتم naive bayes استفاده کردم
این الگوریتم با شمارش ویژگی ها در داده های اموزش مربوط به هر کلاس احتمال شرطی تخصیص داده ورودی به کلاس مورد نظر به شرط کلاس را برای همه کلاسها محاسبه میکند و بیشترین احتمال پاسخ است.
الگوریتم با استفاده از رابطه بیز برای احتمال های شرطی این احتمالات را محاسبه میکند که بیشتر توضیح خواهم داد.
زبان های انتخابی فارسی عربی انگلیسی المانی و فرانسوی هستند و به علت اینکه نتوانستم دیتاست مناسب برای زبانی دیگر با الفبای فارسی پیدا کنم زبان ششم در این پروژه ندارم.اما همه توابع نوشته شده قابل استفاده برای هر تعداد زبان هستند و مستقل از تعداد زبان ها کار میکنند.
در ابتدا جمع اوری دیتاست را با جستجو در ویکی پدیا به صورتی که اکثر جستجو در رابطه حوزه تخصصی این رشته بود و بخشی از ان را به جستجوی مطالب تصادفی و در رابطه با زبان های دیگر اختصاص دادم .به دلیل انکه که داده هایی داشته باشم که در ان کلمات به زبان های دیگر هم وجود داشته باشند و دیتاست یکدست به یک زبان نباشد و احتمال خطا در برنامه پایین بیاید.
طول داده ها نیز متفاوت بودند.و تعداد داده های همه زبان ها تقریبا یکسان است که احتمال خطا پایین بیاید.
سپس دیتاست را به اموزش و تست شکستم که حدود ۲۰ درصد رندوم به تست اختصاص داده و کنار گذاشتم تا با داده های اموزش اشتراک نداشته باشند.و پس از ان داده های اموزش را پیش پردازش کردم
برای انتخاب ویژگی ها(واژه های مهم):
ابتدا کلمات را با شیوه bag of word به کلمات بدون ترتیب ذخیره کردم
سپس کلمات دو حرفی و حروف و کاراکتر های خاص مثل @ و اعداد را از کلمات حذف کردم.این کلمات در تست های اولیه باعث خطا میشدند.چون در زبان ها مشترکند.
سپس مشکلی داشتم که اگر در یک جمله المانی کلمه ای مثل is میام داده را انگلیسی تشخیص میداد.چون is کلمه پرتکراریست و میتواند در احتمال وزن زیادی بگیرد. برای حل

ان کلمات stop word(ضمایر،افعال کمکی،حروف اضافه و ...) که پرتکرار و باعث این مشکلند را از ویژگی ها حذف کردم
سپس کلمات مانده(ویژگی ها) را در یک دیکشنری با تعداد تکرارشان گذاشتم
محاسبه احتمال:
برای محاسبه احتمال شرطی ذکر شده نیاز به محاسبه احتمال هر کلاس است
سپس نیازمند محسابه احتمال شرطی هر ویژگی به شرط کلاس است که باید برای همه ویژگی های داده نهایی حساب شود و در هم ضرب شوند.اینجا مشکلی که داشتم این بود که اگر یک کلمه جدید بود این احتمال را کلا صفر میکرد.راه اول در نظر نگرفتن ان ویژگی بود که به معنا یک بودن احتمال ان است که منطقی نبود.راه دوم تکنیک smooting (اضافه کردن عددی در صورت و مخرج) است.عدد صورت ضریبی از عدد مخرج(تعداد ویژگی های کل داده اموزشی) است و با ازمایش به عدد ۱ رسیدم.
چالش دیگر این مرحله این بود که ابتدا همه احتمالات را برای همه ویژگی ها حساب کرده بودم که موقع اجرا نیاز به محاسبه احتمال نباشد و به علت سربار زمانی و حافظه این روند را به محاسبه احتمال در زمان اجرا تغییر دادم.چون تعداد داده های تست و اجرا از اموزش کمتر است این راه محاسبات کمتری نیز دارد.
چالش دیگراین بود که اگر احتمال شرطی ویژگی به حد زیادی کوچک میشد این احتمال را برابر با صفر در نظر میگرفت و در انتها احتمال همه کلاس ها صفر میشد.برای حل این مساله تابع محاسبه را زمانی که احتمال همه کلاس ها به جز یکی صفر میشد متوقف میکردم.
در ادامه بخش اموزش بخش تست است که در طول پروژه از ان استفاده کردم به صورتی که داده های تست را اجرا کردم و خروجی را با لیبل داده مقایسه کردم و میزان دقت را برای داده ها سنجیدم و در طول پروژه با این تابع و رفع چالش ها به مرور این دقت را زیاد کردم.

بخش نهایی اجرای داده های بدون لیبل است که مراحل مانند بخش تست با داده های موجود در فایل خواهد بود