ملیکا بهروزرازق

93522018

فاز دوم پروژه(پیاده سازی naïve bayes )

برای قسمت اول پروژه ابتدا با استفاده از فایل normalization-and-sentence tokenization دو فایل ورودی book1.txt و book2.txt که رد فاز قبلی پروژه نیز استفاده شده بود را با استفاده از کتابخانه هضم نرمال کرده و سپس sentence tokenizing را انجام داده و داده ها را به صورت مرتب به طوری که هر جمله در هر خط است در فایل های

Train1.txt و train2.txt نوشته می شود.

سپس در فایل probability of classes نیز دو تابع برای محاسبه احتمال کلاس اول و احتمال کلاس دوم نوشته شده است.

سپس در فایل split data and naive bayes ابتدا یک عدد رندوم تولید کرده و با استفاده از آن داده های خود را به صورت رندوم تقسیم کرده و در فایل های randomtrain1.txt و randomtrain2.txt و randomtest.txt می ریزیم.

هنگام نوشتن در فایل تست جمله ها را نیز لیبل گذاری می کنیم به صورتی که برای جملات مربوط به کتاب اول در ابتدای جمله عدد 1 و برای جملات مربوط به کتاب دوم در ابتدای جمله عدد 2 را اضافه می کنیم.

سپس احتمالات کلمات مربوط به randomtrain1.txt را در dictionary به اسم first می نویسیم و احتمالات کلمات مربوط به randomtrain2.txt را در dictionary به اسم second می نویسیم.

سپس کلمات هر جمله مربوط به randomtest.txt را با استفاده از dictionary های first و second محاسبه می کنیم .

و در آخر مجموع احتمالات کلمات هر جمله را در معکوس احتمال کلاس ضرب می کنیم.(زیرا مجموع لگاریتم احتمالات کلمات عدد منفی است و باید در معکوس احتمال کلاس که عددی مثبت است ضرب شود تا نتیجه درست و بهتری گرفته شود.)

اگر احتمال محاسبه شده با کلاس اول بزرگتر بود مربوط به کلاس اول است و در غیر اینصورت مربوط به کلاس دوم است.

لازم به ذکر است پیاده سازی smoothing naïve bayes نیز صورت گرفته است که به صورت کامنت در کد مشاهده می شود و برای تست آن باید خط بالای کامنت که مربوط به non smoothing naïve bayes است را کامنت کنیم.

برای قسمت دوم پروژه نیز در فایل vowpalwabbitpreparation.txt ابتدا دو فایل ورودی را نرمال کرده و سپس جمله جمله جدا کرده و لیبل 1 و -1 در ابتدای جملات زده و کدی نوشته شده که تعداد جملات فایل اول و دوم را برابر در نظر گرفته است.

سپس با استفاده از command های زیر نتایج را می بینیم.

Vw –d vw-result.txt –loss\_function logistic –ngram 2 –f model.vw

Vw –I model.vw –t –d vw-randomtest.txt –p n2.txt

به جای 2 می توان ngram های دیگر را نیز محاسبه کرد.