گزارش آمار و احتمال

پروژه 0

کسری کاشانی نژاد

810101490

در ابتدا اطلاعات دو فایل csv را خوانده و ذخیره کردیم. سپس این اطلاعات را توسط توابع کتابخانه hazm مرتب کردیم:

- متد normalize برای بومی سازی متن و حذف whitespace ها و حروفی که چند بار تکرار شده اند.

- متد tokenize برای جدا کردن متن و کلمه کلمه کردن آن داخل یک لیست.

- متد lemmatize برای تبدیل هر کلمه به ریشه آن و حذف ضمایر و ریشه یابی برای احتمال دقیق تر. (***امتیازی اول***)

- حذف stopword ها که در اکثر جملات وجود دارند و باعث کاهش دقت ما میشوند. (***امتیازی دوم***)

- حذف اعداد و علایم نگارشی فارسی از لیست کلمات.

کلمات تکراری را از لیست کلمات train حذف کردیم تا به لیست کلمات unique برسیم.

برای تشکیل bag of words لیستی از 6 دیکشنری درست کردیم که هرکدام نماینده یک category باشد و در هرکدام تمام کلمات unique را ریختیم.

سپس در هر category تعداد هرکدام از کلمات را در کل train محاسبه میکنیم و در مقادیر دیکشنری قرار میدهیم.

در تابع number\_of\_words\_of\_a\_category تعداد کل کلمات را در هر category محاسبه کردیم تا مخرج احتمال ها محاسبه شوند.

برای محاسبه احتمال چون احتمال انتخاب هر category یعنی p(c)یکسان بدست میاید (برابر 6/1) از محاسبات و مقایسه حذفش میکنیم. همچنین احتمال کلماتی که در متن وجود دارند یعنی p(X) برای هر کلاس و category یکسان است پس از محاسبه حذف میکنیم و در مقایسه فرقی نمیکند.

حالا ما احتمالc به شرط X را با احتمال X به شرط c مدل کردیم و همان ها را مقایسه میکنیم.

در مرحله محاسبه احتمال یعنی تابع compute\_probability\_of\_a\_category و تکه کد آخر بیشترین احتمال تعلق هر کتاب به هر category را پیدا کرده و آن category را ذخیره میکنیم.

در انتها category اصلی و محاسبه شده را بررسی و مقایسه میکنیم و دقت کل کتاب های test را نشان میدهیم. (تعداد کتاب های درست براورد شده با category درست به روی تعداد کل کتاب ها یعنی 450 و برای درصدگیری ضرب در 100)

***پرسش اول)*** اگر در توضیحات موجود درباره هر کتاب، کلمه ای در BoW وجود نداشته باشد و تعداد آن 0 باشد چون احتمال هر کلمه در یکدیگر ضرب میشود (به علت مستقل بودن) با ضرب 0 کل احتمال برابر 0 میشود. پس به جای مقدار اولیه ی 0 برای تعداد هر کلمه آن را با یک ضریبی مقدار دهی میکنیم که ما آن را 0.1 در نظر گرفتیم.

با توجه به فرمول additive smoothing نیز ما به تعداد کلمات ضرب در این ضریب 0.1 صورت احتمال را زیاد کردیم و لذا در مخرج نیز تعداد کل کلمات category ضرب در 0.1 را زیاد کردیم و چون ما احتمال ها را مقایسه میکنیم مشکلی پیش نمی آید به علت انتخاب درست و کم این ضریب یعنی 0.1.

***پرسش دوم)*** اگر کلمات و متن یک کتاب طولانی باشد زمان زیادی برای محاسبه و ضرب این احتمال ها در یکدیگر صرف میشود و عدد خیلی کوچکی با اعشار زیاد بدست می آید. برای اینکار میتوان از هرکدام از احتمال کلمات لگاریتم گرفت و آن ها را با یکدیگر جمع کرد (طبق این خاصیت که در لگاریتم ها ضرب به جمع تبدیل میشود.) بنابراین از احتمال ها لگاریتم میگیریم و جمع آن ها را محاسبه میکنیم و چون لگاریتم تابعی صعودی است مقایسه احتمال ها با مقایسه لگاریتم احتمال ها تفاوتی ندارد.

بدون متد lemmatize و حذف stopword ها دقت ما 80.6 است و بدون روشadditive smoothing نیز دقت ما 16.6 است ولی با انجام تمام این کار ها دقت ما به 80.4 دست یافت. (***امتیازی سوم***)