«گزارش آمار و احتمال»

پروژه صفر

برنا فروهری –810101480

در ابتدا یک لیست از علائم نگارشی و یک مجموعه از اعداد تشکیل میدهیم.و سپس شش category را در لیستی از شش dictionary قرار دادیم.و با توابع کتابخانه hazm متن در دو فایل را پردازش کرده و در یک لیست آن را ذخیره میکنیم .سپس دو فایل csv را خوانده و در دو dataframe جداگانه ذخیره میکنیم.

Normalize:

ابتدا با هدف حذف نیم فاصله ها و فاصله و نیو لاین ها و تشدید و سکون و همچنین اعراب موجود در متن و به طور کلی یکپارچه سازی و تبدیل فرم غیررسمی به رسمی برای متن صورت میگیرد .

Tokenize:

جملات را شکسته و به لیستی از کلمات جدا از هم تبدیل میکند مثلاً «رفته است» به‌عنوان یک توکن («رفته\_است») در نظر گرفته می‌شود.

Lemmatizer:

در کتابخانه هضم وظیفه ریشه یابی را بر عهده دارد مثلا ریشه ی کلمات: کتاب ها و کتابی هر دو کتاب می باشد و در متن فرقی ندارند

Stopwords:

که در جملات ما وجود دارد و باعث کاهش دقت ما می شود و اهمیت

معنایی خاصی ندارند که از روی آن ها به دسته بندی خاصی رسید

Remove\_puncs\_and\_nums:

در مرحله بعد علائم نگارشی که در ابتدا در یک لیست تشکیل داده بودیم و تاثیری در محاسبه احتمال ما ندارد را از متن حدف می کنیم.

و همچنین اعداد موجود در متن را نیز حذف میکنیم .

در مرحله بعدی کلمات تکراری را حذف کرده تا به لیستی از کلمات یونیک شده از فایل ترین برسیم

مرحله بعد تشکیل bow است

در مرحله بعدی برای تشکیل bow لیستی با 6 دیکشنری که هر کدام یک نماینده از 6 category ما خواهد بود می سازیم.

و سپس کلمات uniquedرا در هریک از دیکشنری ها ریخته و تعداد هر یک از کلمات را درکل train محاسبه میکنیم با استفاده از تابع create\_bow و در مقادیر دیکشنری قرار می دهیم.

در تابع sum\_of\_words\_of\_catsتعداد کلمات در هر category را محاسبه کردیم.

برای محاسبه احتمال با توجه به اینکه احتمال هریک از کتگوری ها p(c)) برابر و مساوی 6/1 است را از محاسبات حذف میکنیم همچنین احتمال کلماتی که در متن وجود دارند(p(X)) در هر کتگوری برابر است و آن را نیز حذف می کنیم.

حال ما احتمال x به شرط c را با احتمال c به شرط ایکس برابر در نطر گرفتیم و آن را مقایسه میکنیم.

در این مرحله بیشترین احتمال تعلق هر کتاب به هر کتگوری را محاسبه و آن کتگوری ذخیره میکنیم.probabilty\_book\_category

در نهایت کتگوری اصلی محاسبه شده را بررسی میکنیم و دقت کل کتاب های تست را نشان میدهیم.project\_percent\_truth

پاسخ پرسش اول) اگر به این مورد بر بخوریم احتمال صفر حساب میشود با توجه به اینکه جمله صفر در باقی احتمال ها ضرب میشود کل احتمال یک کتاب برای یک دسته بندی را صفر میکند و در نهایت دقت محاسبات مارا به شدت پایین می آورد برای حل این مشکل می توانیم به جای مقدار اولیه صفر ، 0.1 )الفا) قرار می دهیم

با توجه به فرمول additive smoothing در صورت و مخرج به اندازه الفا \* تعداد کلمات مربوط کم شده است

پاسخ پرسش دوم) چون احتمال های هر کلمه در هم ضرب می شوند زمان زیادی می برد چون محاسبه ضرب تعداد زیادی عدد کوچک است

پس ما از log این اعداد استفاده و با یکدیگر مقایسه میکنیم و جای ضرب از جمع میان logها استفاده می کنیم

دقت درستی حدس کتاب ها با در نظر گرفتن lemmatize و حذف stopwords برابر 80.444 و با لحاظ نکردن آن ها 80.666 می شوند با استفاده از روش additive smoothing

دقت درستی حدس کتاب ها 16.666 می شوند بدون استفاده از روش additive smoothing