بسم الله الرحمن الرحیم
فاطمه پرداختی 810197474

فاز اول: پیش پردازش داده ها
بتدا ایست واژه ها، کاراکترهای بی اهمیت و علایم نگارشی را از جملات حذف می کنیم سپس با کمک کتابخانه هضم جملات نرمال می شوند یعنی نیم فاصله ها و فاصله ها اصلاح می شوند. سپس جملات کلمه به کلمه شکسته می شوند تا بتوان هر کلمه را مستقل بررسی کرد. سپس کلمات ریشه یابی می شوند. ایست واژه ها کلمات پرتکراری هستند که در تمام انواع متن ها کاربرد دارند اما ارزش آموزش ندارند مانند انواع ضمیرها، افعال، کلمات ربط و ... . این عملیات را روی ترکیب دو ستون تایتل و دیسکریپشن انجام می دهم.

سوال یک: در متن دیتا ها ممکن است کلمات به اشکال مختلفی به کار رفته باشند اما ریشه آنها یکسان است. لذا با استفاده از این متد ها به ریشه کلمات میرسیم

فاز دوم:
در تابع classify برای هر سطر از فایل تست میزان احتمال تعلق آن داده به هر کلاس را حساب می‌کنیم. به این صورت که طبق قاعده بیزین احتمال هر کلمه از ستون‌ها را به شرط تعلق به کلاس مورد نظر به دست آورده و در هم ضرب می‌کنیم. داده متعلق به کلاسی است که احتمال بیشتری دارد. این حدس‌ها را در predicted_categories نگهداری می‌کنیم تا بعدا با categoryهای داده شده مقایسه شوند.

evidence =
همان عبارات موجود در ترکیب تایتل و دیسکریپشن است

likelihood =
احتمال آمدن جمله ایکس به شرط دانستن کلاس سی، که با استفاده از ضرب احتمال آمدن هر کلمه به شرط دانستن یک کلاس است. و این هم نسبت تعداد یک کلمه به تعداد تمام کلمات استفاده شده در یک کلاس است

prior =
احتمال آمدن یک کلاس، که نسبت آگهی های یک کلاس به تمام آگهی هاست

posterior =
احتمال عضو یک کلاس بودن به شرط دانستن عبارت که به دنبال محاسبه آن هستیم از طریق باقی احتمالات

سوال سوم:
شیر مایع سفیدی است

شیر حیوان قوی ای است.
در دو جمله بالا شیر با معنای متفاوتی است و با بررسی bigram میتوان این را تشخیص داد چون کلمه بعد شیر باعث تفکیک این دو جمله میشود.
اما اگر این اتفاق نیافتد به n-gram نیاز خواهیم داشت.
سوال چهار:
زیرا ممکن است یک کلمه برای اولین بار در یک کلاس به کار رود آنگاه برای محاسبه p(x|c مقدار صفر میگیرد و چون در باقی ضرب میشود کل این احتمال صفر خواهد شد.
سوال پنج:
هنگامی که این احتمال صفر شد از فرمول دیگری برای این محاسبه استفاده میکنیم.

$$p(x_i \mid c) = count(x_i, c) + 1 / count(c) + |v| + 1$$

سوال ششم:
نمودار ها رسم شدند

سوال هفت:

Precision

چه نسبتی از متن هایی که جز کلاس c تشخیص داده شده اند واقعا جز همین کلاس هستند. به تنهایی کافی نیست زیرا ممکن است متن هایی جز همین کلاس باشند اما تشخیض بدهیم عضو باقی کلاس ها هستند.

Recall:

چه نسبتی از متن های کلاس c را سیستم درست تشخیص داده است.
ممکن است مدل متن هایی از کلاس های دیگر را هم جز همین کلاس محسوب کند.

سوال هشت:
از میانگین گیری هندسی استفاده میکند. اینجا اهمیت دارد زیرا هم تشخیص هایی که به همین کلاس مرتبط هستند و اشتباه پیش بینی شده است و هم آنهایی که مرتبط با باقی کلاس ها هستند و اشتباه مرتبط با این کلاس پیش بینی شده اند.

سوال نه:
ماکرو میان کلاس ها میانگین گیری معمولی میکند
وزن دار با توجه به نمونه های مربوط به هر کلاس میانگین گیری وزن دار میکند

سوال ده:
الف وب انجام شد.

سوال یازده:
همانطور که قابل مشاهده است نتایج با addetiveSmoothing بسیار بهتر است.

سوال دوازده:
شامل کلمات تکراری که در همه کلاس ها هست مثل قیمت و نو است. میتوان تاثیر این کلمات را نادیده گرفت.