دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

گزارش کار درس مباحثی در علوم‌کامپیوتر

نام دانشجو:

پرنیان تقی پور

شماره دانشجویی:

۹۷۱۳۰۱۰

تقسیم داده آموزش و توسعه
از اون جایی که KNNبه نحوی پیاده سازی شده است کهاز داده ای که بعد از اموزش مسیبیند ترین نشود پس اسن جدا سازی را انجام نمیدهم. و از همان تست استفاده میکنم.

پیش پردازش داده‌ها
در این بخش از کتابخانه ی حضم برای استریم کردنو لمینایزرو ... استفاده شده است.

remove_non_farsi: کلیه کاراکترهای غیر فارسی اعم از انگلیسی، علائم نگارشی را حإف و حروف فارسی را نگه میدارد.

Fianl_clean کلمات ایست حذف شده و کلمات باقیمانده با ریشه خود جایگزین می‌شوندو کتب خانه ی حضم در این تابع استفاده میشود.

وکتورایز کردن

از تابع های سایکیت لرن برای وکتورایز کردن داده ها استفاده میکنیم.

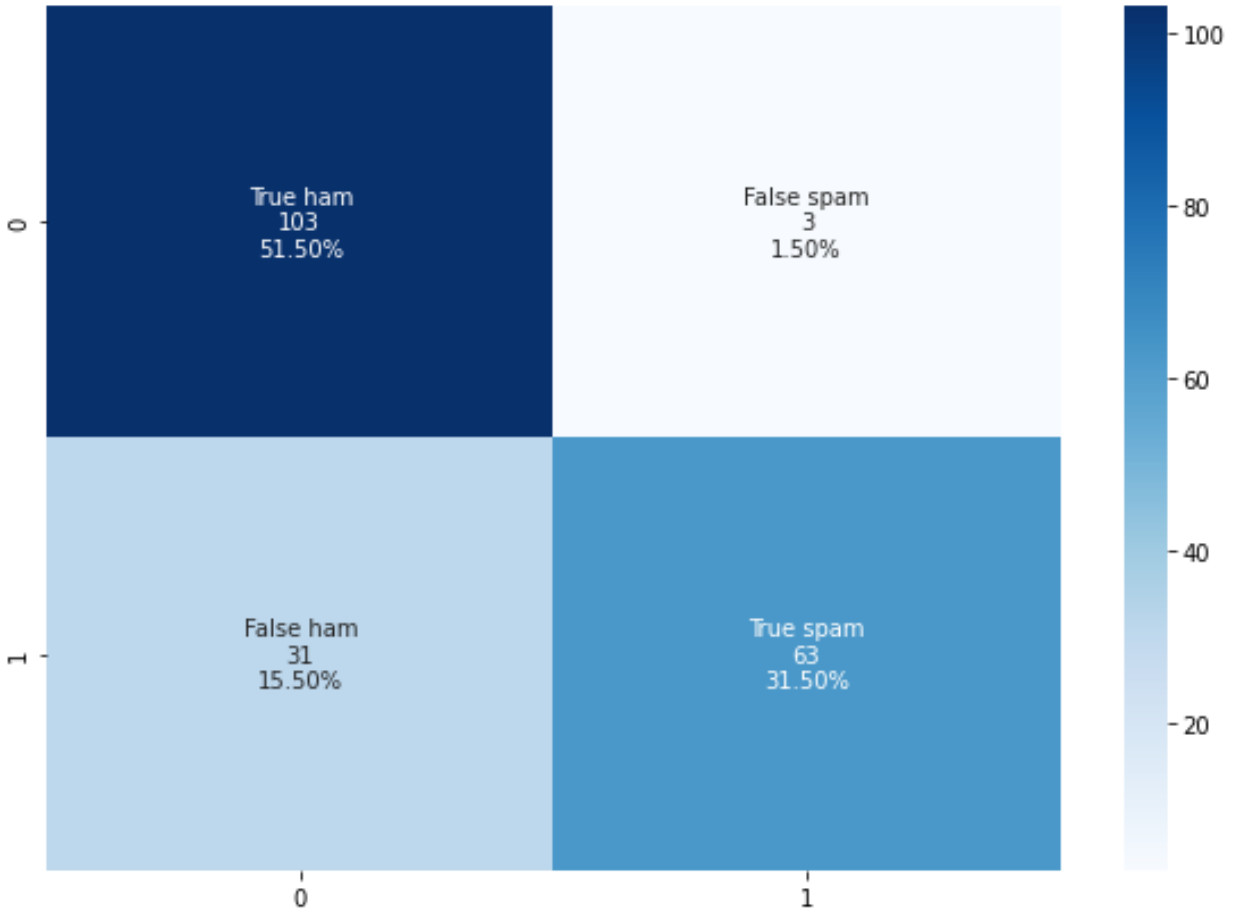
تعیین مهم‌ترین کلمات
حالا با توجه به این که از tfidfبرای وکتورایز کردن استفاده کرده بودیم. یک بار دیگر عملیات را با حال با استفاده از معیار χ^2و با معیار های ۵۰۰و ۲۰۰ استفاده کرده و کامات مهم را هم در دیتای ترین و هم دیتای تست استفاده میکنیم.(خروجی به صورت نمودار در کد نشان داده شده است)

الگوریتم KNN با استفاده از تشابه کسینوسی و TI_IDF
برای این بخش ابتدا کلاس مربوطه را تعریف کیکنیم این کلاس ۳ تابع دارد.
تابع ترین برای گرفتن داده های اموزشی و ذخیره ان استفاده میشود.تابع پریدیک عادی برای استفاده ازفاصله ی کسینوسی استفاده میشود. و تابع دیگر برای محاسبه ی tf-idf استفاده میشود. اگوریتم ان ساده است با توجه به معیارنزدیکی ایمیل جدید به ایمیل های سابق را محاسبه کرده و به تعداد مرود نظر بیشترین تعداد تکرار ایگرگ را انخاب میکنیم برای نتیجه.

تابع cosبا استفاده از لایبری پیاده سازی شده است. تی اف ای دی اف را ولی با دئ تابع IDFکه خروجی ان کلمات است و عدد مربوط به هرکدام. و تابع اصلی برای محاسبات نهایی استفاده میشود.(همان گونه که در توضیح پروژه امده است)

تنظیم هایپرپارامترها
در این بخش الگوریتم KNNرا تا K=10ران کردم و با استفاده از سایکیت لرن اکیورسی مدل ها را مقایسه کردم ماکزیمم در ۷ اتفاق افتاد.

ماتریس سردرگمی



تخلیل کانفیوژن ماتیرس یکی از اجرا ها برا ینمونه(برای KNN)

تعداد اسپم هایی که مدل گفته اپم هستند ولی نبودن ۳ تا هست که چون ۴۰۰ داده ذاریم میشود ۱/۵ درصد. همزمان ۶۳ تا از چیزایی که گفته اسپم هستند اسپم بودند یعنی ۳۱/۵ درصد . در واقع مدل به اسپم تشخیص ندادن بایاس است که بهتر است زیرا هز چقدز کمتر اسپم های غلط داشنه باشیم در این مسیله بهتر است.

تعداد هم های اشتباه برابر ۳۱ است و هم های درست ۱۰۳ که به تناسب بنا به ۴۰۰ ایمیل درصد دارند.

الگوریتم Naïve Bayes

تابع نیو بیز ما ۴ تا تابع در خود دارد.

قضیه نیو بیزین (naive bayesian) روش محاسبه احتمال posterior خلفی(، (P (c | x)، (P (c، P (x)و P (x | c) را فراهم می کند. دسته بندی naive bayesian فرض می کند که اثر ارزش یک پیش بینی (x) بر یک کلاس داده (c) مستقل از مقادیر پیش بینی کننده های دیگر است. این فرض استقلال شرطی طبقه است.

با توجه با این اگوریتم از دو تابع اول برای ترین و از دئ تابع دوم برای پریدیکشن و تست استفاده میکنیم.

تابع تست از تابع getprobبرای محاسبه ی احتمال ها استفاده میکند.

مقایسه NBوKNN

بنا به نتایج به دست امده اکیورسی NBاز TF-IDF بیشتر و از KNN cos similarity کمتر است.
و لحاظ زمانی TF-IDF بیشتر زمان میبرد. چون محاسبات ان بیشتر است.

KNNبا تعداد فیچر های متفاوت

مدل را با تعداد فیچر های ۲۰۰ و ۵۰۰ هم حساب کردم و خروجی گرفتم.

به صورت ترمال پیاده سازی ها اکیورسی های نزدیکی نشان داده اند. نتایج به صورت کامل در کد وجود دارد