به نام خدا

محمدحسین ملک‌پور

9613425

مباحثی در علوم کامپیوتر

تمرین دوم: تشخیص هرزنامه ها

ابتدا کتابخانه ها و ماژول های مورد نیاز را import میکنیم. سپس دیتای هر فولدر را خواندیم و در قالب یک دیتافریم همرا با ستون lable ذخیره کردیم. در مرحله پیش پردازش ابتدا داده ها را نرمالیز کردم که خروجی آن بعد از نرمالیز کردن چاپ شده. سپس بوسیله ی تابع split هر ایمیل را در قالب یک لیست توکنایز کردم و بعد از حذف stop wordها، به کمک کتابخانه hazm عملیات حذف پیشوند، پسوند و یافتن ریشه کلمات شبیه بهم را انجام دادم. در قسمت وکتورایز کردن به روش tfidf ابتدا کلمات یکتا در داده ی اموزشی و تست را پیدا کردم و به کمک تابع tfidf_vectorize عملیات وکتورایز را انجام میدهیم. به وسیله ی تابع chi2 از کتابخانه sklearn هم 500 فیچر که تاثیرگذاری بیشتری دارند را پیدا میکنیم. دو تابع knn پیاده سازی شد که شباهت را با cos similarty و tfidf score محاسبه میکنند. دو تابع knnرا یک بار روی وکتورهای اصلی و یک بار روی وکتورهایی که توسط chi2 بدست اوردیم از k=2 تا 50 امتحان میکنیم تا بهترین k و بیشترین دقت را بدست اوریم. دقت تابع tfidf_knn روی وکتوری که با chi2 بدست اوردیم 10 درصد بیشتر از حالت عادی شد. ولی دقت تابع cos_sim_knn روی هر دوتا وکتور تقریبا 95درصد بود و در مجموع از tfidf_knn بهتر عمل کرد. پس مدل نهایی ما روی وکتور منتخبchi2 و cos_sim_knn با k = 3 اموزش می‌بیند.