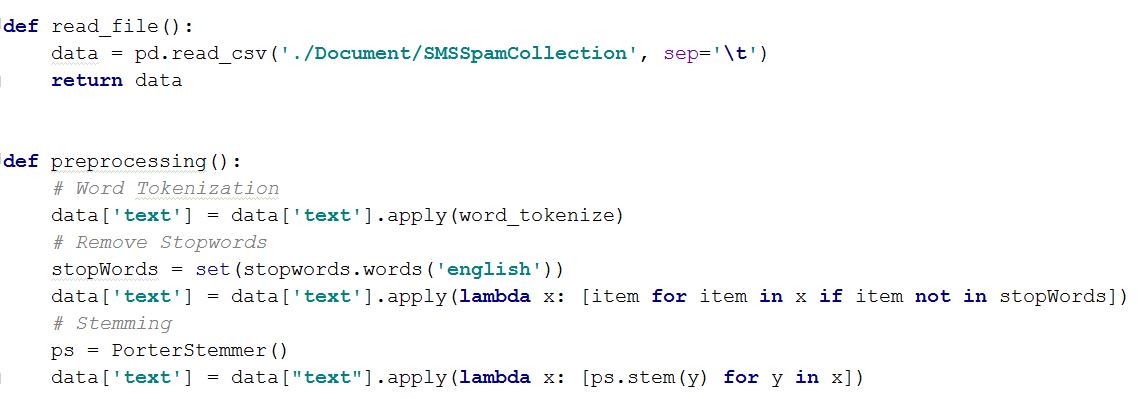
به نام حق

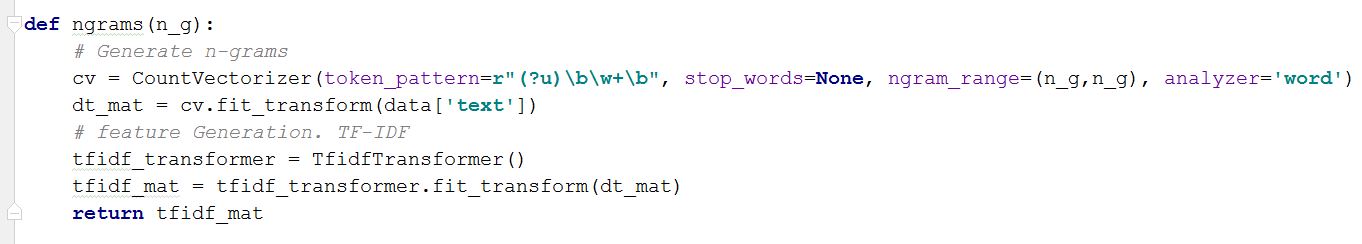
گزارش پروژه پایانی در بازیابی اطلاعات و جستجو وب علیرضا پرچمی 943618113004

Text Classification جهت پیش‌بینی label سندها در Data mining و Information Retrieval استفاده می‌شود. در این پروژه از 3 کتابخانه Scikit-Learn (sklearn) جهت آموزش و دسته‌بند کردن داکیومنت‌ها، Pandas جهت کار کردن بهتر و راحت‌تر با داده‌ها و nltk جهت پیش‌پردازش داده‌ها که شامل حذف StopWordها و Stemming استفاده شده‌است.

در شکل زیر، توابع read\_file و preprocessing جهت خواندن و پیش‌پردازش سندها استفاده شده‌است:



در تابع ngrams که سه بار با ورودی های {1و2و3} خوانده می‌شود، ابتدا n-gramها تولید می‌شود و سپس با استفاده از TF-IDF به تولید ویژگی می‌پردازیم:



در پایان نیز داده‌ها را به براساس Test و Train و براساس Label و Text به 4 دسته تقسیم کرده‌ام و پس از آن عمل Train را انجام می‌دهیم. گزارش کاملی نیز در پایان هر مرحله از Accuracy، Precision، Recall برای هر دسته داده شده و در پایان Confusion Matrix نیز نمایش داده می‌شود:

