هدف تمرین:

- آشنایی با نحوه آماده‌سازی داده‌های متنی برای عملیات دسته‌بندی
- آشنایی با دسته‌بندی متن

مراحل:

۱. در این تمرین هدف ایجاد یک دسته‌بند متن است که خبرهای درست و واقعی (True) را از خبرهای جعلی و غیر واقعی (Fake) تشخیص دهد. در این تمرین شما با دو فایل Fake.csv و True.csv کار می‌کنید که فایل اول حاوی خبرهای جعلی و فایل دوم حاوی خبرهای واقعی است. این مجموعه داده‌ها به صورت csv هستند و شامل ۴ ستون می‌باشند (ستون‌های title، text، subject و date). در این تمرین ما با ستون‌های متنی یعنی text و title کار می‌کنیم.

۲. پیشنهاد می‌شود برای راحتی کار با فایل csv از کتابخانه pandas استفاده کنید. همچنین برای ایجاد دسته‌بندها باید از کتابخانه scikit-learn استفاده کنید.

۳. برای اینکه هم اطلاعات عنوان و هم اطلاعات متن خبر را در دسته بندی در نظر بگیریم، ابتدا ستون جدیدی ایجاد کنید که از به هم چسباندن متن خبر و عنوان خبر ایجاد می‌شود.

۴. مراحل پیش‌پردازش را روی متن ایجاد شده برای هر خبر اعمال کنید.
   a. تمام علائم نگارشی را حذف کنید به صورتی که فقط اعداد و حروف باقی بمانند.
   b. کلمات پرتکرار (stopwords) را حذف کنید و عمل lemmatize را برای هر کلمه به وسیله WordNetLemmatizer انجام دهید.

۵. توجه کنید که داده‌هایی که در اختیار شما قرار داده شده است، ستون برچسب (label) را ندارند اما هر فایل به صورت جداگانه در بر گیرنده خبرهای با برچسب خاص (fake یا true) است که می‌توانید بر این اساس ستون برچسب داده‌ها را ایجاد کنید.

۶. همانطور که می‌دانید ورودی الگوریتم‌های دسته‌بندی به صورت برداری است. برای اینکه متون و کلمات را به بردار تبدیل کنیم، روش‌های مختلفی وجود دارد. Bag of words و Tf*Idf از جمله روش‌های ساده‌ای است که می‌توان به کمک آنها متون را به بردار تبدیل کرد تا آماده استفاده در دسته‌بندها شوند.

۷. بررسی کنید نحوه اعمال Bag of words و Tf*Idf در scikit-learn به چه صورت است؟ سپس این دو روش را به صورت جداگانه بر مجموعه داده‌ها اعمال کنید. میخواهیم نتیجه دسته‌بندی به وسیله الگوریتم‌های مختلف را با استفاده از این روش‌ها به صورت جداگانه بررسی کنیم.

۸. در ادامه مجموعه داده‌ها را به دو بخش آموزش و تست تقسیم کنید. بخش تست باید ۲۰ درصد از کل مجموعه داده باشد.

۹. در نهایت میخواهیم با استفاده از دو الگوریتم Naive Bayes و SVM به دسته‌بندی متون اخبار بپردازیم.

۱۰. بررسی کنید چگونه می‌توان در scikit-learn الگوریتم‌های Naive Bayes و SVM را آموزش دهیم؟

۱۱. به ازای هر الگوریتم، یک بار با استفاده از ویژگی‌های Bag of Words و یکبار با استفاده از ویژگی‌های Tf*Idf آن را آموزش دهید.

۱۲. در این مرحله باید ۴ دسته‌بند آموزش داده شده داشته شده باشد(دو نوع الگوریتم و به ازای هرکدام دو نوع ویژگی جهت آموزش دادن). در ادامه مقادیر صحت (accuracy)، دقت (precision)، بازیابی (recall) و f1-score را برای هر الگوریتم بدست آورده ،گزارش کنید و نتایج را تحلیل کنید.

۱۳. ماتریس در هم ریختگی (confusion matrix) هر الگوریتم را بدست آورید و مقادیر آنرا تحلیل کنید. و با مقادیر بدست آمده در بخش ۱۲ مقایسه کنید.

- نتیجه نهایی را در قالب یک فایل Jupyter notebook و یک گزارش PDF در یک فایل فشرده zip که نام آن با فرمت NLP_Name_Family_Ex6 (به جای Name اسم و به جای Family نام‌خانوادگی خودتان را قرار دهید) است قراردهید و به ایمیل abedi.a1997@gmail.com ارسال کنید.

موفق باشید