مباحثی در علوم کامپیوتر **سارا بابایی خانه سر**

تمرین دوم: تشخیص هرزنامه‌ها **۹۷۱۳۰۰۵**

در این پروژه، ایمیل‌های spam، با استفاده از الگوریتم kNN شناسایی می‌شوند. به این صورت که ابتدا نمونه‌های آموزشی خوانده و پردازش می‌شوند. سپس نمونه‌های آزمایشی خوانده و spam بودن یا نبودن آن‌ها در دو نوبت مجزا، یک بار با معیار تشابه کسینوسی و بار دیگر با معیار تشابه tf-idf، تعیین می‌گردد. در انتها عملکرد الگوریتم با هر یک از معیارهای تشابه، با معیارهای precision، recall و F1-measure سنجیده و اعلام می‌شود.

پیش‌پردازش

تمام داده‌ها (آمورشی و آزمایشی) پس از خوانده شدن از فایل به صورت لیستی از تمام کلمات موجود در متن، ذخیره می‌شوند. سپس کلمات شامل کاراکترهایی غیر از حروف فارسی از این لیست حذف می‌شوند. بعد از آن، ایست‌واژه‌ها با استفاده از لیستی که از مخزن hazm به‌دست آمده مشخص و حذف می‌شوند و واژه‌های باقی‌مانده، با استفاده از ماژول parsivar، با ریشه‌هایشان جایگزین می‌گردند.

در نهایت لیست پردازش‌شده‌ی کلماتِ هر داده، به یک دیکشنری از کلمات و تعداد وقوع آن‌ها در آن داده تبدیل می‌شود.

آموزش

بعد از پیش‌پردازش روی تمام داده‌های آموزشی، دو دیکشنری با مجموعه کلید تمام واژه‌های موجود در داده‌ها و ارزش‌های ـ یکی به صورت تعداد داده‌های آموزشی عادی شامل آن کلمه و دیگری به صورت تعداد داده‌های آموزشی هرزنامه‌ی شامل آن کلمه ـ ساخته می‌شود.
(spam_words_documents_frequency و ham_words_documents_frequency)

سپس با استفاده از معیار chi-square، ۵۰۰ واژه‌ی مهم‌تر به عنوان text featureها شناسایی و ذخیره می‌شوند.

در نهایت تمام داده‌های آموزشی عادی و هرزنامه، به صورت شی‌ای از document که شامل نام فایل (name)، بردار featureهای متناظر با فایل (words_vector) و مشخصه‌ی هرزنامه بودن یا نبودن (is_spam) است، در ماژول ذخیره می‌شوند.

آزمایش

پس از آموزش مدل، تمام داده‌های آزمایشی خوانده، پیش‌پردازش شده و مانند نمونه‌های آموزشی به صورت document ذخیره می‌شوند. در ادامه، با دو معیار شباهت، k نزدیک‌ترین همسایه‌ی هر یک از نمونه‌های آزمایشی پیدا شده و کلاس مربوط به آن تعیین می‌شود.

ابتدا با استفاده از معیار تشابه کسینوسی، الگوریتم kNN روی تمام داده‌های آزمایشی انجام شده و مطابق تصویر زیر، ۱۱ هرزنامه، به اشتباه ایمیل عادی و ۱۰ ایمیل عادی، به اشتباه هرزنامه تشخیص داده شده و الگوریتم با ٪۹۴ = F1 اجرا می‌شود.

```
Confusion Matrix:
                  Actual
                  Spam  ham
Predicated  Spam:  189   10
            ham:   11    190

Precision:  94.9748743718593
Recall:  94.5
F1-measure:  94.73684210526316
```

سپس با استفاده از معیار tf-idf، الگوریتم kNN مجددا روی تمام داده‌های آزمایشی انجام شده و مطابق تصویر زیر، تمام هرزنامه‌ها به درستی شناسایی شده ولی ۱۰۶ ایمیل عادی نیز به اشتباه، هرزنامه تشخیص داده می‌شوند و درنتیجه این بار ٪۷۹ = F1 است.

```
Confusion Matrix:
                  Actual
                  Spam  ham
Predicated  Spam:  200   106
            ham:   0    94

Precision:  65.359477124183
Recall:  100.0
F1-measure:  79.05138339920948
```

لازم به ذکر است که در هر دو حالت، مقدار k با آزمون و خطا در حدود ۴۵ تعیین شده است.

نتیجه‌گیری

با وجود این که با درنظر گرفتن معیار tf-idf، تمام هرزنامه‌ها به درستی تشخیص داده شده‌اند، ولی تعداد زیادی از ایمیل‌های عادی نیز به اشتباه هرزنامه تشخیص داده شده‌اند که باعث کم‌تر شدن F1-measure در این معیار نسبت به درنظر گرفتن تشابه کسینوسی شده است. بنابراین در این مسئله، تشابه کسینوسی معیار مناسب‌تری به شمار می‌آید.

همچنین آزمایش با هر دو اندازه‌ی واژگان (تعداد text featureها) ۲۰۰ و ۵۰۰ تکرار شد ولی این تغییر، در نتیجه‌ی حاصل (F1) هیچ‌یک از دو معیار، تاثیر چشم‌گیری نداشت.