به نام خدا

مباحثی در علوم کامپیوتر
تمرین دوم: تشخیص هرزنامه‌ها
۴ آذر ۱۳۹۹

استاد درس: دکتر اکبری
دانشگاه صنعتی امیرکبیر
دانشکده ریاضی و علوم کامپیوتر

امروزه هرزنامه‌ها [۱] به یک مشکل بزرگ تبدیل شده‌اند. با رشد سریع کاربران اینترنت، ایمیل‌های ناخواسته نیز در حال افزایش است. مردم از آن‌ها برای کارهای غیرقانونی و غیراخلاقی، تقلب و کلاه‌برداری استفاده می‌کنند؛ ارسال لینک‌های مخرب از طریق ایمیل‌های هرزه که می‌تواند به سیستم شما آسیب برساند و هم‌چنین می‌تواند در سیستم شما به‌راحتی جستجو کند. ایجاد یک حساب ایمیل جعلی برای اسپمرها بسیار آسان است، آن‌ها در ایمیل‌های خود نقش یک فرد واقعی را بازی می‌کنند. این اسپمرها افرادی را مورد هدف قرار می‌دهند که از این کلاه‌برداری ها آگاهی ندارند. بنابراین، لازم است هرزنامه‌ها را شناسایی کنید که به جلوگیری و یا حتی کاهش تبعات این ایمیل‌ها کمک کنید.

۱ شرح پروژه

در این پروژه با استفاده از تکنیک‌های یادگیری ماشین، اسپم‌ها را شناسایی می‌کنید. شما باید ماژولی را بنویسید که طبقه‌بندی را با الگوریتم KNN اجرا کند. برای این، هر ایمیل را به عنوان برداری از تعداد کلمات نـمایش بدهید، که طول بردار هم‌اندازه واژگان [۲] است (برای هر داده‌ی آموزشی). طبیعی است که بیش‌تر ورودی‌ها برابر با صفر شوند. به عنوان مثال، فرض کنید $V = 1000$ ، و شاخص کلمه $cheap$، $i = 52$ باشد. اگر ایمیل m هفت‌بار کلمه‌ی $cheap$ در آن ظاهر شود، پس اندیس 52 در بردار نـمایش آن، برابر 7 است. برای طبقه‌بندی، شما باید برای هر ایمیل در مجموعه آزمون، شباهتش هر یک از ایمیل‌های مجموعه آموزش را با استفاده از معیار تشابه کسینوسی محاسبه کنید. سپس برای تعیین کلاس ایمیل ، k بالاترین ایمیل‌های مشابه را در نظر بگیرید. معیار تشابه کسینوسی به شرح زیر محاسبه می شود:

$$Sim(A,B) = \frac{AB}{|A||B|} = \frac{x_{1A}x_{1B} + x_{2A}x_{2B} + ...}{\sqrt{x_{1A}^2 + x_{2A}^2 + ...} + \sqrt{x_{1B}^2 + x_{2B}^2 + ...}} \quad (۱)$$

که در آن x_{iA} تعداد کلمه‌ی با اندیس i در ایمیل A است.

۱.۱ محاسبه فاصله با tf-idf

در این بخش، شما باید فاصله کسینوسی را با فاصله tf-idf جایگزین کنید. فرض کنید n ایمیل در مجموعه آموزشی موجود باشد و آن‌ها را به شکل $e_1, e_2, \ldots, e_n$ نمایش دهیم. برای هر ایمیل جدید مانند $\hat{e}$ شامل دنباله کلمات مجزای $w_1, w_2, \ldots, w_T$ میزان شباهت $\hat{e}$ و e_j به شکل زیر تعریف می‌گردد:

$$\text{Score}_{\text{tf-idf}}(\hat{e}, e_j) = \sum_{i=1}^{T} \text{tf}(w_i, e_j)\text{idf}(w_i) \quad (۲)$$

[۱] Spam
[۲] Vocabulary

در عبارت ۲ مقادیر $\text{tf}(w_i, e_j)$ و $\text{idf}(w_i)$ به شکل زیر محاسبه می‌گردند:

$$\text{tf}(w_i, e_j) = \log(\text{count}(w_i, e_j)) + 1 \quad (۳)$$

$$\text{idf}(w_i) = \log(\frac{n}{\text{df}(w_i)}) \quad (۴)$$

در عبارت ۳ منظور از $\text{count}(w_i, e_j)$ تعداد رخداد کلمه w_i در ایمیل e_j است. همچنین، در عبارت ۴ منظور از $\text{df}(w_i)$ تعداد ایمیل‌هایی است که حاوی کلمه w_i هستند.

۲.۱ پیکره

پیکره هرزنامه شریف شامل ۱۰۰۰ ایمیل است که ۵۰۰ مورد آن هرزنامه و ۵۰۰ مورد دیگر ایمیل عادی می‌باشد. هر ایمیل در قالب یک فایل txt ارائه شده است. در این فایل ممکن است علاوه بر متن ایمیل، اطلاعاتی جزئی در مورد فرستنده، گیرنده و تاریخ ارسال ایمیل نیز موجود باشد. کل پیکره در قالب ۴ پوشه زیر قابل دسترس است:

- hamtraining مجموعه ایمیل‌های عادی آموزشی (۳۰۰ نمونه)
- hamtesting مجموعه ایمیل‌های عادی آزمایشی (۲۰۰ نمونه)
- spamtraining مجموعه هرزنامه‌های آموزشی (۳۰۰ نمونه)
- spamtesting مجموعه هرزنامه‌های آزمایشی (۲۰۰ نمونه)

برای دریافت این پیکره به لینک زیر مراجعه کنید:

https://github.com/omidrohanian/Spam-Filtering-For-Persian/tree/master/emails

۳.۱ پیش‌پردازش

قبل از شروع فرآیند یادگیری، باید ابتدا مراحل زیر را روی هر فایل txt اعمال کنید:

۱. حذف کاراکترهای غیر فارسی شامل حروف انگلیسی، اعداد و علائم خاص مانند نقطه، علامت سوال و علامت تعجب

۲. جداسازی کلمات (هر ایمیل به لیستی از کلمات تبدیل شود)

۳. حذف ایست‌واژه‌ها [۳]

۴. ریشه‌یابی کلمات [۴] (هر کلمه با ریشه خود جایگزین گردد)

نکته: برای این مراحل می‌توانید از ماژول‌هایی مانند Hazm یا Parsivar نیز کمک بگیرید. توجه داشته باشید همین مراحل باید روی داده‌های آزمایشی نیز قبل از ورود به مدل اعمال شوند.

[۳] Stopwords
[۴] Stemming

۴.۱ محاسبه مهم‌ترین کلمات

پس از پیش‌پردازش داده‌ها، لطفا با استفاده از یکی از معیارهای χ^2 و یا information gain مهم‌ترین کلماتی را که در تشخیص نوع ایمیل نقش دارند مشخص کنید. برای این کار می‌توانید تعداد کلمات را معادل ۲۰۰ یا ۵۰۰ در نظر بگیرید.

۱.۴.۱ تکرار آزمایش‌ها با مهم‌ترین کلمات (امتیازی)

پس از محاسبه مهم‌ترین کلمات، بقیه کلمات را از پیکره حذف کنید و آزمایش‌ها را تکرار نمایید. نتیجه را با قسمت قبل مقایسه کنید.

۵.۱ مقایسه با الگوریتم Naive Bayes (امتیازی)

پس از تبدیل ایمیل‌ها به لیستی از کلمات، فرض کنید یک ایمیل به شکل زیر در آمده است:

$$w_1 w_2 w_3 \dots w_n$$

که در آن w_i ها هر کدام یک کلمه هستند. شما باید احتمال تعلق این ایمیل به مجموعه هرزنامه‌ها و مجموعه ایمیل‌های عادی را محاسبه کنید. به عبارت دیگر به دنبال مقادیر زیر هستیم:

$$P(\text{spam}|w_1 w_2 \dots w_n) \quad (۵)$$

$$P(\text{ham}|w_1 w_2 \dots w_n) \quad (۶)$$

مثلا اگر مقدار عبارت ۶ بزرگتر از عبارت ۵ باشد، ایمیل به مجموعه ایمیل‌های عادی تعلق دارد. توجه کنید که برای دریافت نمره مربوط به این قسمت، حتما باید کل الگوریتم Naive Bayes را خودتان پیاده‌سازی کنید و استفاده از ماژول‌های آماده مجاز نیست. در صورت پیاده‌سازی این قسمت، باید نهایتا کارایی آن را با الگوریتم KNN از طریق معیارهای ارزیابی که در ادامه ذکر می‌گردند مقایسه کنید.

۶.۱ معیارهای ارزیابی

برای این تمرین از شما انتظار می‌رود سیستمی را که پیاده‌سازی کرده‌اید، با استفاده از معیارهای Precision ، Recall و F1-Score ارزیابی نمایید. همچنین ماتریس سردرگمی [۵] را نیز بدست آورده و گزارش نمایید.

۲ معیارهای تصحیح

پاسخ شما به این تمرین بر اساس موارد زیر ارزیابی می‌شود:

- پیاده‌سازی KNN توسط شما
- پیش‌پردازش و آماده‌سازی داده‌ها
- محاسبه فاصله کسینوسی و امتیاز بر حسب tf-idf

[۵] Confusion Matrix

- پیاده‌سازی الگوریتم Naive Bayes و مقایسه خروجی دو مدل (امتیازی)
- محاسبه مهم‌ترین کلمات
- تکرار آزمایش‌ها با استفاده از مهم‌ترین کلمات (امتیازی)
- ارزیابی مدل‌ها
- گزارش شامل تحلیل شما از نحوه عملکرد سیستم

۳ نکاتی از تمرین قبل

در پروژه‌های ارسال تمرین قبل، اشکالات رایج و ساده‌ای وجود داشت که اشاره‌ی کوتاهی به آن‌ها می‌کنیم، تا در پروژه‌های بعدی عملکرد بهتری داشته باشیم. :)

- برای خوانا و تـمیز بودن کد، می‌توانید از کامنت‌های واضح و یا قطعه‌های مارک‌دان [۶] در نوت‌بوک‌هایتان استفاده کنید.
- انـجام این نوع پروژه‌ها بدون تـحلیل خروجی و نتیجه‌های به‌دست‌آمده، لطفی ندارد. لطفا این موارد را به گزارش‌هایتان اضافه کنید.

۴ نحوه ارسال پاسخ

لطفا از این تمرین یک نوت‌بوک اصلی داشته باشید که نشان‌دهنده مسیر اصلی اجرای کد شما باشد. اگر می‌خواهید بخشی از کدهایتان را در فایل دیگری ذخیره کنید، می‌توانید با استفاده از دستور import از آن در این نوت‌بوک بهره ببرید. از شما انتظار می‌رود موارد زیر را در قالب یک فایل zip ارسال نمایید:

- یک نوت‌بوک (فایل با پسوند ipynb)
- یک فایل pdf که شامل گزارش شما است.
- هر گونه فایل دیگری با پسوند py که از آن در نوت‌بوک استفاده کرده‌اید.

مهلت ارسال پاسخ: جمعه ۱۴ آذرماه ۱۳۹۹
مهلت ارسال با تاخیر (۱۰ درصد کسر نمره به ازای هر روز): دوشنبه ۱۷ آذرماه ۱۳۹۹

ارتباط با ما

کانال تلگرام: https://t.me/autsocialmedia

ایمیل:

یاسمن امی
yassi.ommi@gmail.com

آرمان ملک‌زاده
malekzadeh@ieee.org

موفق باشید!

[۶] برای این‌کار می‌توانید نوع هر cell در نوت‌بوک را به markdown تغییر دهید