



گزارش عملکرد

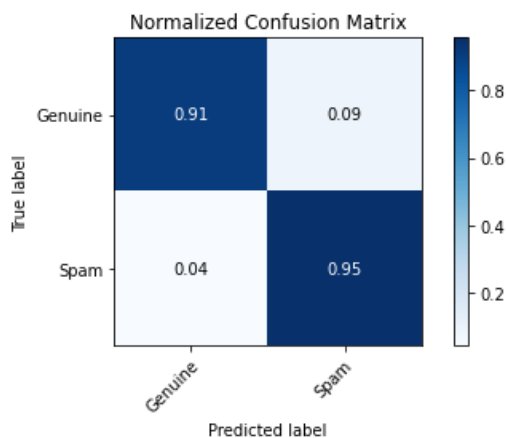
۹۶۱۳۰۲۷

علی علی محمدی

- در فایل فشرده‌ی zip، فایل SpamDetection.ipynb که یک Jupyter Notebook شامل اسکریپت‌هایی به زبان Python وجود دارد.
- این کُد بر روی بستر  توسعه داده شده است و با آن سازگار است.
- آدرس پوشه‌ای که پیکره در آن قرار دارد، در متغیر file_path ذکر شده است. دقت کنید که شیوه‌ی آدرس‌دهی، به نحوی است که با شیوه‌ی آدرس‌دهی در بستر  سازگار است.
- از کتابخانه‌های sklearn، pandas و numpy استفاده شده است.
- برای پیش‌پردازش متن فارسی از کتابخانه‌ی Hazm استفاده شده است.
- نام سلول‌های کُد به همراه کامنت‌های برنامه به وضوح روند کار را تشریح می‌کنند. همچنین نام توابع به صراحت بیان‌گر کاربرد و منظور آن تابع است.
- تمامی مراحل پردازش و تبدیل متن به بُردار، آموزش مدل و محاسبه‌ی معیار فاصله^۱، بدون استفاده از کتابخانه‌های موجود، پیاده‌سازی شده‌اند.
- ✓ گزارش عملکرد دسته‌بند^۲ با استفاده از فاصله‌ی کُسنوسی به شرح زیر است که در آن، برچسب True یعنی آن ایمیل به دسته‌ی اِسپم‌ها تعلق دارد و False به معنای موثق بودن ایمیل است:

	precision	recall	f1-score	support
False	0.95	0.91	0.93	200
True	0.91	0.95	0.93	200
accuracy			0.93	400
macro avg	0.93	0.93	0.93	400
weighted avg	0.93	0.93	0.93	400

- ✓ ماتریس سردرگمی نرمال‌شده‌ی آن نیز به شرح زیر است:



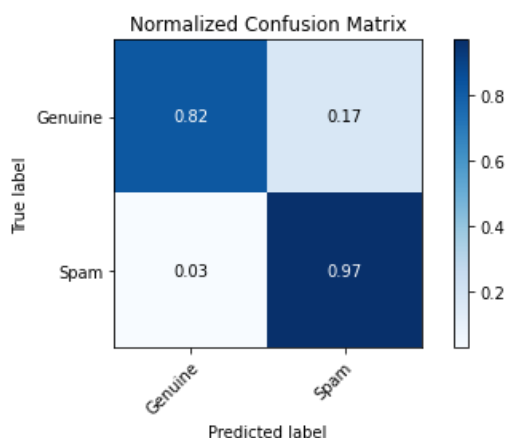
¹ Distance Metric

² Classifier

✓ گزارش عملکرد دسته‌بند با استفاده از فاصله‌ی TF/IDF به شرح زیر است که در آن، برچسب True یعنی آن ایمیل به دسته‌ی اسپم‌ها تعلق دارد و False به معنای موثق بودن ایمیل است:

	precision	recall	f1-score	support
False	0.96	0.82	0.89	200
True	0.85	0.97	0.90	200
accuracy			0.90	400
macro avg	0.91	0.90	0.90	400
weighted avg	0.91	0.90	0.90	400

✓ ماتریس سردرگمی نرمال‌شده‌ی آن نیز به شرح زیر است:



✚ با مقایسه‌ی این دو به این نتیجه می‌رسیم که فاصله‌ی کُسینوسی در تشخیص درست موثق بودن یک ایمیل بهتر عمل می‌کند (مقایسه‌ی سطر اول) در حالی که فاصله‌ی TF/IDF در تشخیص درست اسپم بودن بهتر عمل می‌کند. (مقایسه‌ی سطر دوم)

✚ نکته‌ی جالب توجه در آن است که با افزایش K ، دقت (accuracy) پیشبینی با فاصله‌ی کُسینوسی افزایش می‌یابد در حالی که دقت (accuracy) پیشبینی با فاصله‌ی TF/IDF کاهش می‌یابد.

✚ با توجه به پیچیدگی استفاده از فاصله‌ی TF/IDF و همچنین زمان‌بر بودن آن، استفاده از فاصله‌ی کُسینوسی منطقی‌تر به نظر می‌رسد؛ زیرا هم ساده‌تر پیاده‌سازی می‌شود و هم حجم محاسبات بسیار کمتری دارد و زمان اجرای کم است.

✚ در صورتی که مقدار K کوچک باشد، استفاده از TF/IDF دقت بالاتری دارد و خیلی هم زمان‌بر نیست اما با افزایش K ، دقت فاصله‌ی کُسینوسی بیشتر می‌شود و سرعت رشد دقت آن بیشتر است در حالی که فاصله‌ی TF/IDF لزوماً دقت بالاتری را تضمین نمی‌کند. همچنین به دلیل پیچیدگی بیشتر TF/IDF، استفاده از آن برای K بزرگ، زمان اجرا را بسیار بالا می‌برد اما زمان اجرای برنامه با فاصله‌ی کُسینوسی به مراتب کمتر و همچنان معقول خواهد بود.

در صورت وجود هرگونه سؤال یا ابهام، با ایمیل alialimohammadi@ce.aut.ac.ir در تماس باشید.