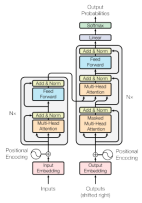

تمرین صفرم
پیش‌پردازش زبان طبیعی

تاریخ بارگزاری تمرین: 13/12/1401
تاریخ تحویل تمرین: 22/12/1401

۱ پرسش‌ها

فرآیند ریشه‌یابی[۱] با لم‌سازی[۲] چه تفاوت‌هایی دارد؟ مثال بزنید.

۲ برنامه‌نویسی

دسته‌بندی[۳] یکی از عملیات‌های مرسوم یادگیری ماشین[۴] به حساب می‌آید. یکی از جنبه‌های این عملیات در زمینه‌ی پردازش زبان طبیعی[۵]، دسته‌بندی متون است. در فایل‌های مربوط به تمرین، کد[۶] و مجموعه‌دادگان[۷] یک برنامه‌ی دسته‌بندی ساده آمده است که براساس الگوریتم دسته‌بندی kنزدیک‌ترین همسایگی[۸] عمل می‌کند. دادگان موجود مجموعه‌ای از نظرات[۹] سایت یوتیوب[۱۰] را تشکیل می‌دهند. هدف این دسته‌بند، تشخیص اسپم[۱۱] بودن یا نبودن این نظرات است. بدون اعمال پیش‌پردازش روی دادگان، این دسته‌بند به صحت[۱۲] 78درصد می‌رسد. مجموعه‌ای از عملیات‌های پیش‌پردازش را به گونه‌ای انتخاب کنید که با اعمال آن‌ها روی دادگان، صحت مدل یاد شده افزایش یابد. در این راستا موارد زیر را در نظر بگیرید:

۱. گام نهایی در عملیات پیش‌پردازش می‌بایست یکی از الگوریتم‌های ریشه‌یابی، لم‌سازی و تصحیح غلط[۱۳] باشد.

۲. عملیات دسته‌بندی بایستی به‌ازای هر کدام از الگوریتم‌های یاد شده، یک‌بار و با مجموعه اَعمال پیش‌پردازش یکسان اجرا گردد.

۳. عملیات‌های پیش‌پردازش بایستی باتوجه به هدف مسئله‌ی دسته‌بندی و کثیفی‌های مجموعه‌دادگان طراحی گردند.

۴. حداقل تعداد پیش‌پردازش ممکن بایستی استفاده گردد.

۵. حداقل بهبود نتیجه صحت 8درصد خواهد بود.

[1] Stemming
[2] Lemmatization
[3] Classification
[4] Machine learning
[5] Natural language processing
[6] Code
[7] Dataset
[8] K nearest neighbours
[9] Comments
[10] Youtube
[11] Spam
[12] Accuracy
[13] Spell correction

پس از انجام اجراهای خواسته شده مشخص کنید کدام یک از الگوریتم‌های ریشه‌یابی، لم‌سازی و تصحیح غلط بهتر عمل می‌کنند. آیا با شخصی‌سازی روند پیش‌پردازش برای هر یک از این الگوریتم‌ها می‌توان به نتیجه‌ی بهتری رسید؟ توضیح دهید. آیا می‌توان با جایگزین کردن الگوریتم دسته‌بندی به نتایج بهتری دست پیدا کرد، الگوریتم پیشنهادی خود را مشخص نمایید و نتایج جدید را گزارش کنید.

۳ خوانش مقاله

مقاله‌ی «On the effectiveness of preprocessing methods when dealing with different levels of class imbalance» [۱] را مورد مطالعه قرار دهید و به پرسش‌های زیر پاسخ دهید:

۱. هدف مقاله را در یک یا دو جمله توضیح دهید.

۲. تفاوت این مقاله با مقالات پیش از خود در چیست؟

۳. روش‌های بیش‌نمونه‌برداری[۱۴] و کم‌نمونه‌برداری[۱۵] که در مقاله آمده است را توضیح دهید. از هر کدام یک مثال بیاورید و به صورت خلاصه بگویید چگونه عمل می‌کنند.

۴. چرا معیار[۱۶] صحت در مورد مجموعه‌دادگان با عدم‌توازن[۱۷]، معیار خوبی نیست؟

۵. چه معیارهایی برای بررسی نتایج الگوریتم‌ها روی مجموعه‌دادگان با عدم‌توازن معرفی شده است، آن‌ها را به طور خلاصه توضیح دهید؟ چرا بهتر هستند؟

۶. آیا مجموعه‌داده فراهم شده در بخش ۲ دارای عدم‌توازن است؟

۷. یکی از معیارهای معرفی شده را به برنامه بخش ۲ اضافه کنید.

۸. نتیجه‌ی مقاله چیست؟

۴ نکات تحویل

۱. پاسخ خود را تحت پوشه‌ای به اسم NLP_NAME_ID و در قالب zip بارگذاری نمایید.

۲. این پوشه می‌بایست حاوی موارد زیر باشد:

- پوشه‌ای با نام code باشد که شامل برنامه‌ی نوشته/تغییر داده شده است.
- پوشه‌ای با نام doc که حاوی داکیومنت‌ها و فایل توضیحات می‌باشد.

۳. لازم به ذکر است که رعایت قوانین نگارشی حائز اهمیت خواهد بود.

مراجع

[1] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol.25, no.1, pp.13–21, 2012.

[14] Over-sampling
[15] Under-sampling
[16] Metric
[17] Imbalance