پروژه دوم درس داده کاوی

عنوان :

تحلیل احساسات توئیت کاربران با استفاده از مدل‌های رده‌بندی

استاد:

دکتر حسین رحمانی

پاییز ۱۴۰۲

راهنمای پروژه

- مهلت ارسال پروژه تا **ساعت ۲۳:۵۹ تاریخ ۲۴ / ۰۹ / ۱۴۰۲** است و قابل تمدید نخواهد بود.
- به ازای هر روز تاخیر **۲۵ درصد** از نمره پروژه کسر خواهد شد.
- پاسخ به سوالات این پروژه باید در قالب یک گزارش با **فرمت PDF** یا به همراه **توضیحات فایل نوتبوک (Markdown)** ارائه شود.
- در صورت ارائه گزارش در قالب **توضیحات فایل نوتبوک**، توضیحات باید کامل، جامع و شفاف باشد.
- در صورت ارائه گزارش با **فرمت PDF** ، فایل **کدهای اجراشده** نیز پیوست شود.
- تمامی فایل‌های این پروژه (گزارش و کدها) در قالب یک **فایل فشرده rar یا zip** با نام‌گذاری زیر ارسال شود.

StudentNumber_FirstName_LastName_Prj02.zip

- فایل تمرین را **حتما** در سامانه LMS آپلود نمایید. بدیهی است که تحویل از طریق ایمیل و یا سایر راه‌های ارتباطی قابل پذیرش نخواهد بود.
- **رعایت نکات نگارشی** در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این پروژه **حتما** باید از زبان برنامه‌نویسی **پایتون** استفاده شود.

مباحث تحت پوشش: text mining , classification

هدف پروژه: هدف پروژه ایجاد یک سیستم تحلیل احساسات با استفاده از مجموعه داده‌ای از Twitter است. مجموعهی داده شامل اسناد متنی با برچسب احساس می‌باشد و سیستم قصد دارد هر سند را به عنوان مثبت، منفی، خنثی و نامرتبط رده‌بندی کند. این پروژه شامل پیش پردازش داده‌ها، آموزش مدل، مقایسه و ارزیابی چندین مدل رده‌بندی است.

۱- فایل ورودی:

در زیر یک تفکیک از ستون های مجموعه داده است:

- شناسه توییت: شناسه منحصر به فرد برای هر توییت.
- موجودیت: موجودیت‌هایی در متن که احساساتی درباره‌ی آن‌ها بیان شده است.
- احساس: لحن احساسی بیان شده در توییت، که می تواند مثبت، منفی، خنثی باشد یا اگر به موجودیت مذکور مربوط نباشد نامرتبط است.
- محتوای توییت: متن توییت.

فایل ضمیمه شده دارای سه بخش Twitter-training، Twitter-test و Twitter-validation می باشد. برای آموزش مدل از فایل Twitter-training و برای ارزیابی مدل ها از Twitter-test و برای پیش بینی مدل از Twitter-validation استفاده کنید.

.۲- پیش‌پردازش داده:

مجموعه داده دانلود شده نیاز به تمیز کردن و تبدیل داده‌های متنی برای پیش‌پردازش دارند. وظایفی مانند حذف کاراکترهای خاص، تبدیل متن به حروف کوچک، حذف کلمات توقف و پردازش های Stemming یا Lemmatization (و هرچه که برای پیش پردازش احساس می کنید که نیاز هست، نیز انجام دهید).

۳- انتخاب ویژگی‌ها:

از اسناد متنی پیش‌پردازش شده بردارهای متناظر آن ها را استخراج کنید. روش‌های متداول شامل Bag-of-Words، TF-IDF یا word embedding مانند Word2Vec یا GloVe هستند. داده‌های متنی به نمایش عددی تبدیل شود تا برای مدل هایتان مناسب باشد.

۴- آموزش مدل:

با استفاده از ویژگی‌های استخراج شده و برچسب‌های احساس مربوطه، باید چندین مدل رده‌بندی آموزش داده شود: همانند مدل های Naïve Bayes، SVM و ... . توجه شود که نیاز است حداقل دو مدل استفاده شود.

۵- ارزیابی مدل:

مدل های آموزش دیده را با استفاده از مجموعه داده ی Twitter-test ارزیابی کنید (معیار های ارزیابی هم چون Accuracy, F1, Precision, Confusion Matrix و ....). همچنین با استفاده از Twitter-validation و مدلهای اموزش دیده ی خود برای هر مدل یک فایل اکسل درست کنید و یک ستون اضافه به نام predicted-sentiment در انتهای فایلهای اکسل برچسب پیش‌بینی شده را ذخیره کنید.

۷- نتیجه‌گیری:

الف: بین مدل‌های انتخاب شده بهترین مدل را گزارش کنید و دلیل اینکه چرا این مدل عملکرد بهتری داشته را ذکر کنید.

ب: با توجه به نتایج به دست آمده از Twitter-validation تعداد پیش‌بینی درست احساس را برای هر مدل گزارش کنید. آیا بهترین مدل به درستی تعداد بیشتر را پیش‌بینی کرده است؟ علت را توضیح دهید؟

نکته: هرگونه ابتکار عمل و خلاقیت و تحلیل‌های دیگر مانند استفاده از نمودارها برای تحلیل مجموعه داده و یا مقایسه‌ی مدل‌ها در نحوه‌ی اجرا، نمره ی اضافه خواهد داشت.

موفق باشید.