
به نام خدا

# تمرین اول درس مباحثی در علوم کامپیوتر

# عنوان: تحلیل احساسات جریان داده‌های میکروبلاگ

استاد درس:

دکتر اکبری

موعد تحویل: ۱۳۹۹/۰۸/۱۶

## فهرست مطالب

## هدف

تحلیل احساسات و نیت توییت‌ها، از آن رو که به تشخیص درک کاربران از سازمان‌ها، رویدادها و یا محصولات کمک می‌کند، کار مهمی است. هدف این پروژه، ساخت یک طبقه‌بند[1] برای دسته‌بندی جریان توییت‌های ورودی به کلاس‌های مختلف احساسات است؛ برای مثال، در ساده‌ترین دسته‌بندی، توییت‌ها به دو کلاس مثبت و منفی دسته‌بندی می‌شوند. یک منبع مهم در تجزیه و تحلیل احساسات، واژه‌نامه لغت (دیکشنری) احساسات است که ممکن است بسته به زمان (اصطلاحات جدید ظهور کند) و کلاس‌ها یا جنبه‌ها متفاوت باشد (یک اصطلاح با توجه به کلاس‌ها یا جنبه‌های مختلف، ممکن است دارای احساسات متفاوتی باشد). علاوه بر این‌ها، چالش‌های دیگری نیز وجود دارد که حل آن‌ها می‌تواند پروژه را بهبود ببخشد. از جمله این چالش‌ها می‌توان به تاثیر زیرموضوعات، استفاده از شکلک‌ها، عوامل مختلف اجتماعی و اطلاعات زمانی اشاره کرد. این عوامل باعث بهبود عملکرد سیستم می‌شوند و باید بررسی گردند.

## شرح تمرین

در این پروژه، باید ماژولی را برای تجزیه و تحلیل احساسات جریان‌های ورودی میکروبلاگ‌ها پیاده‌سازی کنید. این ماژول باید شامل عملکردهای زیر باشد:

### طبقه‌بند احساسات اولیه

- باید بتواند بر اساس داده‌های آموزشی[2]، طبقه‌بند(ها) را با استفاده از هر روش مناسب یادگیری ماشین و مبانی ساده فرهنگ لغت پردازش کند. داده‌های توسعه[3] برای تنظیم پارامترها ارائه شده است.
- طبقه‌بند اولیه، فقط با ویژگی‌های متنی، با استفاده از یک مجموعه دیکشنری احساسات ساده و پایه‌ای، آموزش می‌بینند.
- طبقه‌بند اولیه بر روی داده‌های آزمونِ داده‌شده، آزمایش می‌شود. هنگام آزمایش انتظار می‌رود طبقه‌بند توییت‌های ورودی جدید را در سه کلاس مثبت (+۱)، منفی (-۱) و خنثی (صفر) طبقه‌بندی کند.
- برای نشان‌دادن این که طبقه‌بند حاصل‌شده به خوبی کار می‌کند، نمایش نتایج نهایی در یک جدول اهمیت بالایی دارد.

### توابع و عملکردهای بهبوددهنده

- توانایی برداشت اصطلاحات جدید احساسات بر اساس آخرین مجموعه اسناد مرتبط
- توانایی کاوش در اطلاعات اجتماعی و زمانی در تعیین احساسات توییت‌های ورودی
- توانایی تشخیص توییت‌های غیرمرتبط
- بسط بازه‌ی قوت تشخیص احساسات (مثلا بر اساس شدت احساسات، ۱- را به ۵- و ۱+ را به ۵+ بسط دهید)
- استفاده از ویژگی‌های وابسته به حوزه‌ی موضوع، مانند زیرگروه‌ها/جنبه‌ها برای افزایش دقت تجزیه و تحلیل احساسات

[1] Classifier
[2] Training
[3] Development

## داده‌های موجود

مجموعه‌ای از توییت‌های مربوط به تجربه مسافران از پرواز با خطوط هوایی آمریکا در سال ۲۰۱۵ در قالب سه فایل csv به شما داده می‌شود. این فایل‌ها شامل مجموعه‌های آموزشی، توسعه و آزمون هستند. در هر فایل، متن توییت در ستون text و کلاس احساس مربوط به آن در ستون airline_sentiment موجود است. علاوه بر این دو ستون، اطلاعات دیگری نیز در مورد پروازها در فایل موجود است که برخی از آن‌ها در ادامه آورده شده است:

- میزان اطمینان از کلاس توییت در قالب عددی بین ۰ و ۱ در ستون airline_sentiment:confidence
- تاریخ و زمان ایجاد توییت در ستون tweet_created
- منطقه زمانی کاربر در ستون user_timezone
- دلیل نارضایتی از پرواز در ستون negativereason
- نام خط هوایی در ستون airline
- تعداد retweet ها در ستون retweet_count

شما باید از مجموعه‌ی آموزش برای یاد دادن، توسعه برای تنظیم سیستم‌ها/پارامترها و آزمون برای محک مدل و کشیدن جدول نتایج استفاده کنید. توزیع توییت‌ها بر حسب کلاس‌هایشان در ادامه آورده شده است.

| مجموعه | تعداد نمونه‌های مثبت | تعداد نمونه‌های خنثی | تعداد نمونه‌های منفی | مجموع |
|---|---|---|---|---|
| آموزشی | ۱۴۴۷ | ۱۷۹۰ | ۵۵۴۷ | ۸۷۸۴ |
| توسعه | ۴۵۰ | ۶۷۳ | ۱۸۰۵ | ۲۹۲۸ |
| آزمون | ۴۶۶ | ۶۳۶ | ۱۸۲۶ | ۲۹۲۸ |

برای دانلود داده‌ها به لینک زیر مراجعه کنید.

https://drive.google.com/file/d/1yl9-UqLIhy8GPUPmKT6wqqiyMUwPpxXu/view?usp=sharing

## جزئیات پیاده‌سازی

در این تمرین انتظار می‌رود ابتدا داده‌ها را پیش‌پردازش کنید و مراحلی مانند حذف ایست‌واژه‌ها[4]، نرمال‌سازی و حذف کلمات بسیار کوتاه (با طول کمتر از ۳ کاراکتر) را انجام دهید. پس از آماده‌سازی داده‌ها، باید با استفاده از روش $\chi^2$ تاثیرگذارترین کلمات را استخراج کنید.

برای آموزش طبقه‌بند، می‌توانید از SVM، KNN و یا روش Naïve Bayes استفاده کنید. در هر صورت، پس از آموزش طبقه‌بند، باید آن را با معیارهای زیر مورد ارزیابی قرار بدهید:

- Precision
- Average Precision

[4] Stopwords

- Recall
- F1-Score (Micro and Macro)
- Confusion Matrix

سعی کنید با استفاده از Confusion Matrix عملکرد مدل را تحلیل نمایید. یک راه می‌تواند این باشد که درصد اشتباه در هر کلاس را بر اساس این ماتریس محاسبه نمایید.

توجه: توصیه می‌شود برای یادگیری الگوریتم‌های یادگیری، حداقل یکی از آن‌ها (مانند KNN) را خودتان پیاده‌سازی کنید. البته اگر از کتابخانه‌ها هم استفاده نمایید، اشکالی ندارد. در بقیه موارد مانند ارزیابی مدل نیز می‌توانید از کتابخانه‌هایی مانند scikit-learn استفاده کنید.

## معیار ارزیابی پاسخ شما

پاسخ شما به این تمرین بر اساس کیفیت انجام مراحل زیر مورد ارزیابی قرار خواهد گرفت:

- پیش‌پردازش و آماده‌سازی داده‌ها
- انتخاب بهترین ویژگی‌ها با معیار $\chi^2$
- پیاده‌سازی الگوریتم یادگیری (SVM یا KNN یا Naïve Bayes)
- طبقه‌بند اولیه
- توابع و عملکردهای بهبوددهنده
- محاسبه Precision، Recall، F1 و Confusion Matrix
- گزارش

## گزارش کار

موارد زیر را تا قبل از موعد پروژه، باید ارسال کنید:

- گزارشی (حداکثر ۸ صفحه)، شامل ساختار برنامه، جزئیات طبقه‌بند، مراحل آموزش و آزمایش، همراه جدول نتایج آزمایش که نشان‌دهنده تاثیر و عملکرد طبقه‌بند شما است.
- کد مرجع برنامه‌ی شما که به زبان **پایتون ۳** نوشته شده است. برنامه‌ی شما باید به کاربران اجازه دهد توییت‌های جدید را به صورت آنلاین برای آزمایش وارد کنند. لطفا این فایل را در فرمت .py و یا .ipynb ارسال نمایید.

لطفا موارد فوق را از طریق سامانه courses بارگذاری نمایید.

برای مطرح‌کردن سوالات و ابهام‌هایی که دارید می‌توانید از طریق ایمیل‌های زیر با ما در ارتباط باشید.


یاسمن امی
yassi.ommi@gmail.com

آرمان ملک‌زاده
malekzadeh@ieee.org


موفق باشید