به نام خدا

پروژه‌ی اول درس شناسایی الگو – دانشکده‌ی مهندسی کامپیوتر – دانشگاه علم و صنعت ایران

استاد درس: دکتر مرتضی آنالویی

تدریس یاران: سید محمد پورباقری – سید حسن طباطبایی

mhmd.pourbagheri@gmail.com Hassan.tbt1989@gmail.com

==========

تحلیل احساسات[1] متون

مقدمه

هدف از این پروژه آشنایی با مفاهیم اولیه‌ی رده‌بندی در قالب یک مسأله‌ی پردازش زبان طبیعی است. در این پروژه یک مجموعه دادگان شامل نظرات کاربران درباره‌ی فیلم‌های موجود، به همراه برچسبی که مشخص‌کننده‌ی مثبت یا منفی بودن آن نظر است، آورده شده‌است. دانشجو باید با استفاده از روش‌های مختلف، بردار ویژگی‌ها را به دست آورده و سپس رده‌بندهای مختلفی را برای پیش‌بینی برچسب نظرات جدید آموزش داده و نتایج را گزارش و با هم مقایسه کند.

مجموعه‌ی دادگان

در این پروژه دانشجویان باید از دیتاست Large Movie Review Dataset که مربوط به دانشگاه استنفورد است، استفاده کنند. نسخه‌ای از این دیتاست بر روی گروه تلگرامی درس قرار خواهد گرفت. همچنین می‌توانید این دیتاست را از آدرس زیر دانلود نمایید:

https://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

شرح پروژه

پیش‌پردازش

در این پروژه تمیزکاری داده‌ها شامل حذف تگ‌های HTML اضافی و تبدیل داده به فرمت مناسب، کاملاً به عهده‌ی دانشجو است.

استخراج بردار ویژگی‌ها

پس از تمیزکاری داده‌ها، استخراج ویژگی‌ها باید به چهار روش مختلف زیر انجام شود:

[1] Sentiment Analysis

- Bag of Words (BOW).
- BERT Embedding. بردار ویژگی‌ها در این روش باید در دو حالت مختلف استفاده شود. حالت اول به دست آوردن بردار ویژگی برای کل جمله و حالت دوم به دست آوردن بردار ویژگی برای هر کلمه از جمله است.
- اختیاری: وزن‌دهی tf-idf به همراه ویژگی n-gram یک بار با حذف کلمات توقف[۲] و یک بار بدون حذف آن‌ها.
- اختیاری: Word2Vec یک بار با حذف کلمات توقف و یک بار بدون حذف آن‌ها.

آموزش رده‌بند

در این مرحله، باید با استفاده از بردار ویژگی‌های به‌دست‌آمده برای داده‌های آموزشی[۳]، رده‌بندهای زیر آموزش داده و نتایج آن‌ها با هم مقایسه شوند.

- رده‌بند بیز ساده[۴].
- رده‌بند ماشین بردار پشتیبان[۵] (SVM).
- رده‌بند درخت تصمیم‌گیری[۶].
- رده‌بند جنگل‌های تصمیم تصادفی[۷].

آزمون رده‌بند

پس از آموزش هر یک از رده‌بندها، باید عملکرد آن رده‌بند روی داده‌ی آزمون[۸]، آزمایش شود. نتایج عملکرد رده‌بندها باید در قالب ماتریس درهم‌ریختگی[۹]، نمودار ROC[۱۰] و مقدار AUC[۱۱]، دقت[۱۲]، فراخوانی[۱۳] و معیار f1 گزارش و با هم مقایسه و تحلیل شوند.

شیوه‌ی پیاده‌سازی و موارد تحویلی

برای پیاده‌سازی این پروژه می‌توانید از زبان برنامه‌نویسی دلخواه خود استفاده کنید (زبان برنامه‌نویسی پیشنهادی پایتون است). همچنین برای استخراج ویژگی‌ها و رده‌بندی می‌توانید از کتابخانه‌های پیش‌ساخته‌ی آن زبان بهره ببرید. برای

---

[2] Stop Words
[3] Training Set
[4] Naïve Bayes
[5] Support Vector Machine
[6] Decision Tree
[7] Random Forest
[8] Test Set
[9] Confusion Matrix
[10] Receiver Operating Characteristic
[11] Area Under Curve
[12] Precision
[13] Recall

استخراج ویژگی‌ها با استفاده از هر یک از روش‌های بالا، می‌توانید از کتابخانه‌های معتبر استفاده نمایید. گزارش نهایی باید شامل نتایج به تفکیک روش‌های مختلف استخراج ویژگی و رده‌بندهای مختلف، به همراه مقایسه و تحلیل آن‌ها باشد. همچنین کد نرم‌افزاری کامل نیز به همراه این گزارش در یک فایل فشرده به فرمت

IUSTPR981-StudentFullName-StudentNumber

قرار گرفته و به ایمیل mhmd.pourbagheri@gmail.com حداکثر تا پایان مهلت پروژه‌ی اول (۱۵ آبان) ارسال شود. همچنین در روز تحویل (که تاریخ آن متعاقباً اعلام می‌شود)، گزارش و کدهای ارسالی، باید توسط دانشجو کاملاً شرح داده و اجرا شوند.

*** دانشجویان عزیز می‌توانند سؤالات خود را از طریق گروه تلگرامی درس به آدرس زیر بپرسند:**

https://t.me/joinchat/BE1gilWcxhl338XwRI7TRg