به نام خدا

پروژه اول درس شناسایی الگو – دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت ایران

استاد درس: دکتر مرتضی آنالویی

آموزشیاران: سید محمد پورباقری – پدرام دادخواه

Mhmd.pourbagheri@gmail.com Pedram.Dadkhah1374@gmail.com

موضوع:

تشخیص نظر[1] نویسنده متن

مقدمه

هدف از این پروژه آشنایی عملی با رده‌بندی (classification) در قالب یک مساله پردازش زبان طبیعی است. در این پروژه مجموعه دادگانی (data set) که هر داده عبارت است از متنی شامل چند جمله در نظر گرفته میشود. هر داده نظر کاربری را درباره‌ی فیلم‌ی که دیده است را منعکس میکند. جملاتی که کاربر نوشته نظر مثبت و یا منفی کاربر در مورد فیلم را عرضه می کند. هر داده توسط انسان قرائت شده و مثبت و یا منفی بودن نظر بصورت برچسبی به داده اضافه شده است. مجموعه دادگان به سه قسمت تقسیم میشود. دادگان آموزش، دادگان اعتبارسنجی و دادگان آزمون. در این پروژه شما با استفاده از روشهای مرسوم ابتدا ویژگیهای هر داده را در قالب یک بردار به دست میاورید. سپس با استفاده از دادگان آموزش و اعتبارسنجی اقدام به آموزش یک رده بند با پارامترهای معتبر میکنید. آنگاه رده بند حاصل را برای شناسایی برچسب دادگان آزمون بکار برده و عملکرد آنرا گزارش می نمایید.

مجموعه دادگان

مجموعه دادگان این پروژه IMDB dataset (Sentiment analysis) in CSV format است. نسخه‌ای از این دادگان بر روی گروه تلگرامی درس قرار دارد. همچنین میتوانید این دادگان را از آدرس زیر دریافت نمایید.

https://www.kaggle.com/columbine/imdb-dataset-sentiment-analysis-in-csv-format

(نیاز به VPN دارد)

[1] Sentiment Recognition

مراحل انجام پروژه

پیش پردازش

برای پیش پردازش حتما کلمات توقف[1] و تگ های HTML که ممکن است در متن وجود داشته باشند را حذف کنید. پیش پردازشِ بهتر باعث عملکرد بهتر میشود.

استخراج بردار ویژگی ها

پس از پیش پردازش، استخراج ویژگی صورت میگیرد. روشهای زیادی برای اینکار وجود دارد. در این پروژه هر یک از ۴ روش زیر را بکار میبریم و نتایج نهایی را مقایسه میکنیم. (بکارگیری دو روش انتهایی را به اختیار شما میگذاریم)

- Bag of Words (BOW)
- BERT Embedding
- اختیاری: وزن دهیِ TF-IDF
- اختیاری: Word2Vec

آموزش و تعمیم رده بند

در این مرحله با استفاده از بردار ویژگی های به دست آمده برای دادگان آموزش (Training Set) و دادگان اعتبارسنجی (Validation Set)، رده بند های زیر آموزش میدهید.

- رده بند بیز ساده[2]
- رده بند ماشین بردار پشتیبان[3] (SVM)

رده بند SVM را یکبار بدون کرنل و یکبار با RBF kernel[4] آموزش دهید. پارامتر های C و gamma با استفاده از دادگان اعتبارسنجی به روش جستجوی شبکه ای[5] تنظیم میشوند. برای اطلاعات بیشتر درباره kernel ها و پارامتر های مختلف رده بند SVM لینک زیر را ببینید.

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[1] Stop Words

[2] Naïve Bayes

[3] Support Vector Machine

[4] Radial Basis Function

[5] Grid Search

آزمون رده بند

پس از آموزش هر یک از رده بند ها، عملکرد آن روی دادگان آزمون (Test Set)، آزمایش شود. عملکرد رده بند ها در قالب ماتریس درهم ریختگی[1] ، نمودار ROC[2] و مقدار AUC[3]، دقت[4]، فراخوانی[5] و معیار f1 گزارش و با هم مقایسه و تحلیل میشوند.

شیوه پیاده سازی و موارد تحویلی

برای پیاده سازی این پروژه میتوانید از زبان برنامه نویسی دلخواه خود استفاده کنید (زبان برنامه نویسی پیشنهادی پایتون است). همچنین برای استخراج ویژگی ها و رده بندی میتوانید از کتابخانه های پیش ساخته آن زبان بهره ببرید. برای استخراج ویژگی ها با استفاده از هر یک از روشهای ذکرشده در بالا، میتوانید از کتابخانه های معتبر استفاده نمایید. جزییات انجام پروژه و نتایج حاصله را طی گزارشی ارایه میشود.گزارش نهایی شامل نتایج به تفکیک روش های مختلف استخراج ویژگی و رده بندهای مختلف، به همراه مقایسه و تحلیل آنها است. همچنین کد نرم افزاری کامل نیز به همراه این گزارش در یک فایل فشرده به فرمت

IUSTPR991-StudentFullName-StudentNumber

قرار گرفته و به ایمیل Mhmd.pourbagheri@gmail.com حداکثر تا پایان مهلت پروژه اول (۲۰ آذر) ارسال شود. در جلسه ای که زمان آن برای هر دانشجو اعلام خواهد شد شما گزارش و کدهای خود را شرح داده و به سوالات ما پاسخ میدهید.