به نام خداوند جان و خرد

پروژه درس یادگیری ماشین - نیم‌سال دوم ۱۴۰۲-۱۴۰۱

مدرس: سید ابوالفضل مطهری

دانشکده مهندسی کامپیوتر

# پردازش زبان طبیعی - تحلیل احساسات

## مقدمه

داده‌های متنی می‌توانند انتقال‌دهنده احساسات مثبت، منفی یا خنثی باشند. در این پروژه، قصد داریم مدلی را با استفاده از یادگیری ماشین آموزش دهیم که میزان مثبت یا منفی‌بودن یک جمله یا عبارت را تشخیص دهد. شما می‌توانید از هر روشی که در کلاس تدریس شده است، برای طراحی و بهبود عملکرد این مدل استفاده کنید. همچنین مجاز به استفاده از هر نوع کتابخانه‌ای هستید؛ در صورتی که بتوانید نحوه استفاده از آن را به صورت واضح و مشخص در گزارش‌تان بیان کنید.

## دادگان

مجموعه داده موردنیاز پروژه را از لینک زیر دریافت کنید.

https://zaya.io/2zwn4

هر نمونه شامل یک جمله یا عبارت و برچسب متناظر آن است. جمله یا عبارت، نظری درباره یک فیلم است و برچسب می‌تواند یکی از موارد زیر باشد:

- ۰: منفی
- ۱: تا حدی منفی
- ۲: خنثی
- ۳: تا حدی مثبت
- ۴: مثبت

## مراحل انجام پروژه

### ۰. تقسیم‌بندی داده‌ها

در این مرحله، ابتدا داده‌ها را به سه بخش آموزشی[1]، اعتبارسنجی[2] و آزمایشی[3] تقسیم کنید.

### ۱. پیش‌پردازش داده‌ها و استخراج ویژگی

#### رویکرد ۱: استفاده از روش‌های اولیه

در این رویکرد، شما باید ترکیبی از موارد زیر را در نظر بگیرید.

[1] Training
[2] Validation
[3] Testing

- تبدیل حروف بزرگ انگلیسی به حروف کوچک
- حذف برخی علائم مانند علامت سوال، تعجب و ...
- تهیه لیستی از stopword-ها و حذف آن‌ها از جملات و عبارات
- ریشه‌یابی از کلمات
- تهیه لیستی از توکن‌های باقیمانده
- محاسبه بردار متناظر هر جمله با استفاده از روش tf-idf

توجه کنید که ویژگی‌های استخراج‌شده باید بر اساس مجموعه داده آموزشی به دست آیند. به عبارت دیگر، وزن هر توکن بر اساس مجموعه داده آموزشی تعیین می‌شود.

### رویکرد ۲: استفاده از روش‌های نوین

در این رویکرد، شما می‌توانید از یک مدل برای استخراج ویژگی استفاده کنید. یکی از مدل‌های زیر را در نظر بگیرید:

- Word2Vec
- FastText
- GloVe

هر یک از موارد فوق، یکی از مدل‌هایی هستند که در حوزه پردازش زبان طبیعی، برای استخراج ویژگی از کلمات و عبارات متنی به کار می‌روند. شما می‌توانید به دلخواه از هر یک از آن‌ها برای استخراج ویژگی بهره ببرید.

## ۲. آموزش مدل(ها)

از مرحله قبل، شما باید تعدادی بردار و برچسب متناظر آن‌ها را استخراج کرده باشید. بر اساس آنچه در کلاس آموخته‌اید، بهترین مدلی را که می‌توانید، روی داده‌ها آموزش دهید. توجه کنید که اگر مدل شما دارای ابرپارامتر[4] است، باید مقدار ابرپارامترها را بر اساس داده‌های اعتبارسنجی تعیین نمایید.

توجه: پیش از آموزش مدل‌ها، وجود مشکلاتی مانند نامتعادل‌بودن دسته‌ها[5] را بررسی کنید و در صورت لزوم برای رفع آن‌ها راهکاری ارائه نمایید.

[4] Hyperparameter
[5] Class Imbalance

### ۳. ارزیابی مدل‌ها

در این مرحله، بر اساس معیارهای دسته‌بندی از جمله Precision، Recall، F1-Score و Normalized Confusion Matrix، مدل(ها) را ارزیابی نمایید. توجه کنید که ممکن است شما به جای پیش‌بینی برچسب نهایی در یک مرحله، این کار را با استفاده از چند مدل و در چند مرحله انجام دهید؛ به صورتی که برای هر مرحله از یک مدل استفاده کنید. در این صورت برای هر مدل باید ذکر کنید که چه وظیفه‌ای را انجام می‌دهد و معیارهای ارزیابی عملکرد آن را نیز بیان نمایید.

## نحوه نگارش گزارش

پس از انجام پروژه، لازم است یک گزارش نیز تحویل دهید که شامل بخش‌های زیر باشد:

- لیست پکیج‌ها (کتابخانه‌های) استفاده‌شده به همراه علت استفاده از آن‌ها
- نحوه استخراج ویژگی از داده‌ها در رویکرد ۲
- لیستی از مدل‌های به کار گرفته‌شده، هدف از به‌کارگیری آن‌ها و تعداد پارامترهای هر مدل
- نتایج ارزیابی مدل‌ها
- تحلیل نقاط قوت و ضعف مدل
- مقایسه رویکردهای اولیه و نوین، استدلال برای بیان علت کارکرد بهتر یکی از رویکردها
- تلاش‌های شکست‌خورده[۶]، به همراه تحلیل علت شکست و عدم کارایی

نکته ۱: هدف از ارائه گزارش، درک بهتر نحوه عملکرد شماست. لذا **به گزارش‌های ناخوانا، نامرتب، مبهم و یا دارای معایب دیگر که کمکی به درک نحوه انجام پروژه نکنند، نمره‌ای تعلق نمی‌گیرد**. لذا در نوشتن گزارش دقت کنید. همچنین لازم است فونتی مناسب (مثلا یکی از فونت‌های سری B) برای نگارش گزارش انتخاب شود. گزارش شما حداکثر می‌تواند ۱۰ صفحه باشد. لطفا از درج تصاویر به روش‌های غیر آکادمیک (مانند اسکرین‌شات و ...) شدیدا پرهیز کنید.

## نحوه ارزیابی

پروژه در مجموع ۱۰۰ امتیاز دارد که معادل ۱ نمره از ۲۰ نمره پایان‌ترم خواهد بود. البته همانطور که در ابتدای ترم ذکر شد، بنا به شرایطی، این نمره می‌تواند ۲ برابر شود. توجه کنید که کدنویسی تمیز[۷] هم از معیارهای کسب نمره است.

| تقسیم‌بندی داده‌ها | پیش‌پردازش و استخراج ویژگی | آموزش مدل‌ها | ارزیابی مدل‌ها (پیاده‌سازی) | گزارش | Kaggle Competition |
|---|---|---|---|---|---|
| ۱۰ | ۲۰ | ۲۰ | ۱۵ | ۲۰ | ۱۵ |

[۶] برای مثال مدل‌هایی که به کار گرفته شده ولی نتیجه مطلوبی نداشته‌اند

[7] Clean Code