پروژه تحلیل احساسات نقدهای **IMDb**

مقدمه

در این پروژه، شما باید از مجموعه داده‌های نقد فیلم IMDb برای پیش‌بینی احساسات (مثبت یا منفی) نقدها استفاده کنید. هدف این پروژه، آشنایی با روش‌های مهندسی ویژگی، استفاده از مدل‌های یادگیری ماشین و تحلیل نتایج است.

شرح

مجموعه داده مجموعه داده IMDb شامل نقدهای متنی فیلم‌ها به همراه برچسب‌های مثبت و منفی است. هر نقد یک متن دارد که حاوی احساسات مثبت یا منفی درباره فیلم مربوطه است. برچسب‌ها به دو دسته «مثبت» (label=1) و «منفی» (label=0) تقسیم‌بندی شده‌اند.

لینک دانلود مجموعه داده:IMDb Movie Reviews Dataset on Kaggle

مراحل انجام پروژه

1. پیش‌پردازش داده‌ها

- داده‌ها را بارگیری کنید و بررسی کنید که دارای چه ویژگی‌هایی هستند.
- هرگونه نویز را حذف کنید و متن‌ها را به شکل استاندارد تبدیل کنید (مثلاً تبدیل به حروف کوچک، حذف نشانه‌گذاری‌ها، حذف stop words). در صورت تمایل شما می‌توانید از روش‌های پر استفاده در پردازش متن استفاده کنید تا دقت خود را بالا ببرید.

2. تقسیم‌بندی داده‌ها

- مجموعه داده 50,000 نمونه دارد. داده‌ها را به سه بخش تقسیم کنید:
  - مجموعه آموزشی (**Train):** 35,000 نمونه
  - مجموعه اعتبارسنجی (**Validation):** 10,000 نمونه
  - مجموعه آزمون (**Test):** 5,000 نمونه
- فقط از مجموعه‌های آموزشی و اعتبارسنجی برای آموزش مدل و تنظیم هایپرپارامترها استفاده کنید.
- مجموعه آزمون را فقط یک بار برای ارزیابی نهایی مدل به کار ببرید.

3. مهندسی ویژگی‌ها

- روش اول: برای هر متن بررسی کنید که آیا هر کلمه خاص در متن وجود دارد یا خیر (ویژگی‌های دودویی: 0 یا 1).
- روش دوم: تعداد تکرار هر کلمه در متن را شمارش کنید و به عنوان ویژگی ذخیره کنید.
- روش سوم (اختیاری): از دوکلمه‌ای‌ها (bigram) استفاده کنید و همان فرآیند بالا را روی آن‌ها نیز اعمال کنید.
- برای تمامی روش‌ها، تعداد واژگان را محدود کنید (به عنوان مثال 5000 یا 10000 کلمه پرکاربرد را نگه دارید) و تأثیر تعداد واژگان نگه داشته شده را بررسی کنید.

4. مدل‌سازی

- از مدل‌های مختلفی برای پیش‌بینی احساسات استفاده کنید:
  - **K-Nearest Neighbors**
  - **Logistic Regression**
  - **SVM**
- سایر مدل‌هایی که تمایل دارید را نیز آزمایش کنید.

5. ارزیابی و تحلیل

- دقت مدل‌ها را مقایسه کنید و بهترین مدل را انتخاب کنید.
- اثر تعداد واژگان بر عملکرد مدل‌ها را بررسی کنید.
- روش‌های مختلف مهندسی ویژگی را با هم مقایسه کنید.
- تحلیل کنید که چرا یک روش عملکرد بهتری دارد.

6. گزارش و تحویل پروژه

- گزارشی شامل موارد زیر ارائه دهید:
  - توضیح روش‌های پیاده‌سازی‌شده
  - نمودارها و مقایسه عملکرد روش‌های مختلف
  - تحلیل نتایج و ارائه پیشنهادات
- یک Jupyter Notebook یا فایل‌های پایتون شامل کدهای پروژه ارسال کنید.

زمان‌بندی و تحویل

- شما دو هفته فرصت دارید تا این پروژه را انجام دهید. **(3 اسفند)**
- فایل‌های خود را در قالب یک گزارش و یک کد اجرایی ارسال کنید.

موفق باشید!