گزارش پروژه نهایی یادگیری ماشین

محمدرضا صیدگر 401422215
مهرانه مقتدایی فر 401422211

پروژه اجرای یک سیستم توصیه‌گر برای فیلم با استفاده از داده‌های موجود در سایت TMDB راه‌اندازی شده است. هدف اصلی این پروژه توصیه‌گری است که بر اساس ویژگی‌های مختلف فیلم‌ها، به کاربران فیلم‌های مناسب را پیشنهاد می‌دهد.
ابتدا، داده‌ها را که از سایت TMDB دریافت می‌کنیم، بررسی می‌کنیم تا بفهمیم کدام ویژگی‌ها برای ما مهم است و می‌خواهیم بر روی آنها مدل ایجاد کنیم. سپس، ویژگی‌های غیرضروری را از داده‌ها حذف می‌کنیم تا فقط ویژگی‌های مورد نیاز برای مدل سیستم توصیه‌گر باقی بمانند.
برای مثال، برخی از ویژگی‌های مهم می‌تواند شامل ژانر فیلم، امتیاز متوسط، مدت زمان فیلم، تعداد رأی‌ها و سال انتشار باشند. این ویژگی‌ها می‌توانند به عنوان ورودی‌های مدل سیستم توصیه‌گر استفاده شوند.

حال در اولین بخش از پیش پردازش دیتا اقدامات زیر انجام شده است:
1. تعیین ستون‌های مورد نیاز: ستون‌هایی که برای تحلیل مورد نیاز هستند، مشخص شده است. این ستون‌ها عبارتند از: 'adult', 'budget', 'genres', 'popularity', 'vote_average','original_language', 'release_date', 'revenue', 'runtime', 'vote_count', 'title' .
2. استخراج ستون‌های مورد نیاز
بنابراین، در این بخش فقط ستون‌های مورد نیاز استخراج و نگهداری شده.

در بخش بعدی استخراج ژانر اصلی فیلم از رشته ژانرها انجام می شود.
به عنوان مثال، در صورتی که رشته ژانر به صورت "[{'id': 12, 'name': 'Adventure'}, {'id': 28, 'name': 'Action'}, {'id': 14, 'name': 'Fantasy'}, {'id': 878, 'name': 'Science Fiction'}]" باشد ، ژانر اصلی "Adventure" را برمی‌گردانیم.
به طور خلاصه، این بخش برای استخراج ژانر اصلی هر فیلم و ایجاد ستون جدید "main_genre" استفاده می‌شود و در نهایت ستون 'genres' هم حذف می شود.

در این مرحله ابتدا تعداد کل داده ها محاسبه شده که برابر با 45466 است. سپس برای هر ردیف و ستون بررسی شده که چه مقدار داده نامعتبر (NaN) دارد. سپس داده های حاوی مقادیر نامعتبر حذف شده است و تعداد داده های باقی‌مانده برابر شد با 44797.

در ادامه ستون‌های 'budget' و 'revenue' که درصد وجود مقادیر صفر بیشتر از 70% دارند، شناسایی شده و به منظور بهبود کیفیت داده‌ها، این دو ستون از دیتاست حذف شده‌اند.

در این بخش داده هایی که در ستون‌های 'runtime' ،'vote_average'، 'popularity' و 'vote_count' مقدار صفر دارند بررسی شده و از دیتاست حذف شده‌اند. در نهایت، تعداد داده های باقی‌مانده پس از حذف با مقدار 41006 برابر است.

در مرحله بعدی یک ستون به نام 'release_date' داریم که در آن تاریخ‌های فیلم‌ها ذخیره شده‌اند.هر تاریخ به سه بخش تقسیم می‌شود: سال، ماه و روز. سپس با استفاده از این تقسیم‌بندی، ستون جدیدی به نام 'release_year' ایجاد می‌شود که فقط شامل سال از تاریخ است. در نهایت، ستون 'release_date' حذف می‌شود تا فقط ستون 'release_year' باقی بماند.
برای مثال، برای تاریخ '30-10-1995'، ستون 'release_year' مقدار '1995' خواهد داشت.

در این بخش ابتدا 5 زبان اول با بیشترین تعداد تکرار را پیدا میکنیم که شامل زبان‌های "en" (انگلیسی)، "fr" (فرانسوی)، "ja" (ژاپنی)، "it" (ایتالیایی) و "de" (آلمانی) می‌باشد.
سپس زبان‌های متعلق به لیست بالا را به همان زبان اصلی باقی می‌گذاریم و سایر زبان‌ها را با عنوان 'Other' جایگزین می‌کنیم.
نتیجه آن در ستون جدیدی به نام 'reduced_language' ذخیره می‌شود. در نهایت، ستون 'original_language' حذف می‌شود.

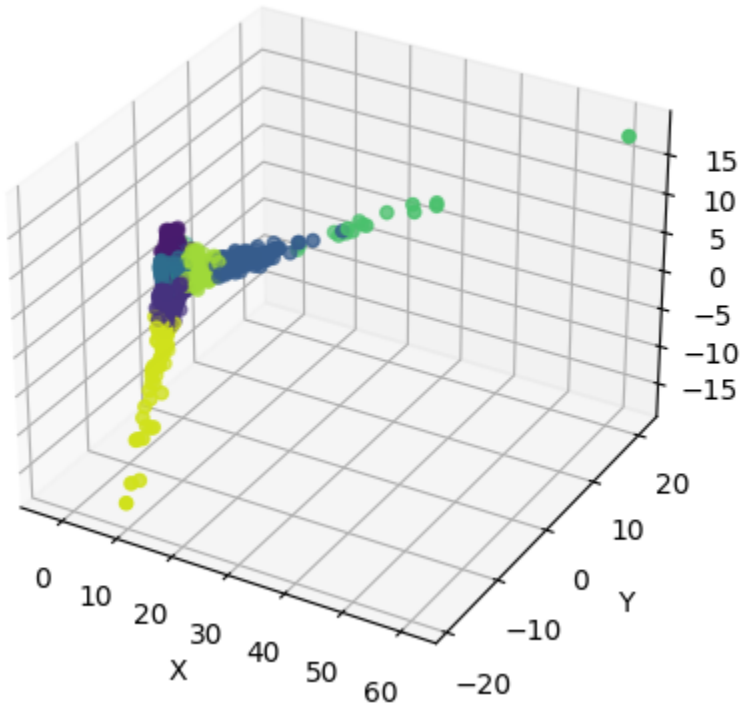
همین اتفاقات برای ستون main_genre هم می افتد. 5 ژانر اول با بیشترین تعداد تکرار شامل ژانرهای "Drama" (درام)، "Comedy" (کمدی)، "Action" (اکشن)، "Documentary" (مستند) و "Horror" (ترسناک) می‌باشد.

در ادامه ستون "adult" حذف شده است. دلیل حذف این ستون این است که بررسی نشان داده که تقریباً تمام مقادیر در این ستون False است و تنها 6 مورد True وجود دارد. این ستون برای تحلیل و پیش‌بینی فیلم‌ها از اهمیت کمی برخوردار است و بنابراین می‌توان آن را حذف کرد تا از حجم داده کاسته شود و پردازش سریع‌تر انجام شود.

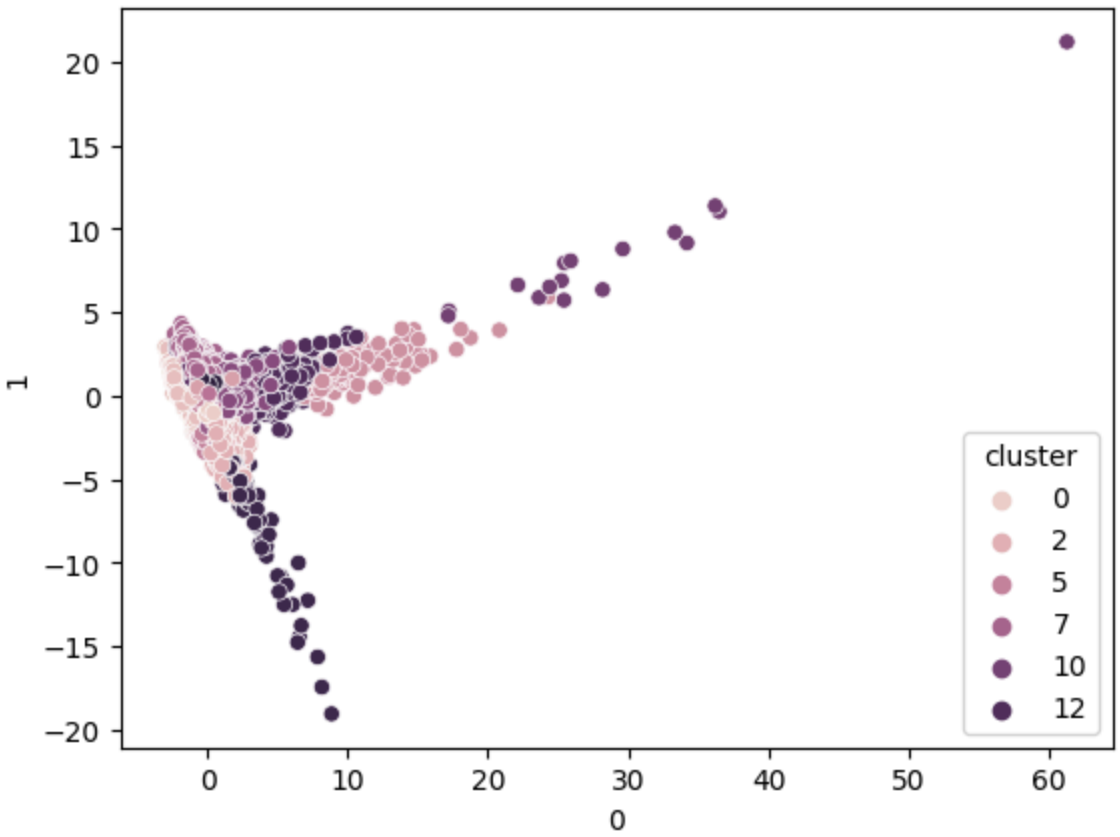
در ادامه تصمیم گرفتیم که بخشی از کد که داده‌های نویزی را حذف می‌کند را کامنت کنیم و از حذف داده‌های پرت خودداری کنیم. این تصمیم اتخاذ شده است زیرا متوجه شدیم که برخی از فیلم‌ها با امتیاز بالا بسیار مهم هستند و حذف آنها ممکن است تاثیر منفی بر روی فرایند پیشنهاد دادن فیلم‌ها داشته باشد. در نتیجه، تصمیم گرفتیم داده‌های نویزی حذف نشوند و در خروجی اطلاعات مهمی مرتبط با فیلم‌های با امتیاز بالا وجود داشته باشد.

در آخرین بخش از پیش پردازش دیتا از `StandardScaler` استفاده می‌کنیم تا این ستون‌ها را نرمال‌سازی کنیم و بازه‌ی مقادیر آنها را بهبود بخشیم.
سپس به دنباله ستون‌های دسته‌بندی می‌رویم. از `OneHotEncoder` استفاده می‌کنیم تا این ستون‌ها را به بردارهای دودویی تبدیل کنیم.
در نهایت، ستون‌های دسته‌بندی اصلی را از داده‌ها حذف می‌کنیم تا داده‌های نرمال‌سازی شده و داده‌های جدید دسته‌بندی شده در یک دیتاست یکپارچه قرار گیرند.

حال وارد بخش کلاسترینگ میشویم ؛ در این بخش ابتدا الگوریتم K-Means را با تعداد کلاسترهای مورد نظر برابر با 15 و تعداد تکرارهای مجدد (n_init) برابر با 10 می‌سازیم و روی داده‌ها آموزش می‌دهیم. سپس برچسب کلاسترها را برای داده‌ها تخمین می‌زنیم.
تعداد 15 کلاستر انتخاب شده است برای اینکه در هر کلاستر حداقل 6 داده وجود داشته باشد، ابتدا تعداد داده‌های موجود در هر کلاستر را محاسبه می‌کنیم و سپس بررسی می‌کنیم که آیا حداقل یکی از کلاسترها کمتر از 6 داده دارد یا خیر که این در بیشترین $k = 15$ این شرط صادق است.
پس برای اینکه در خروجی 5 فیلم پیشنهاد دهیم، تعداد کلاسترها را 15 انتخاب کرده‌ایم.



این نمایش داده ها و کلاستر بندی آنها در 3 بعد است.



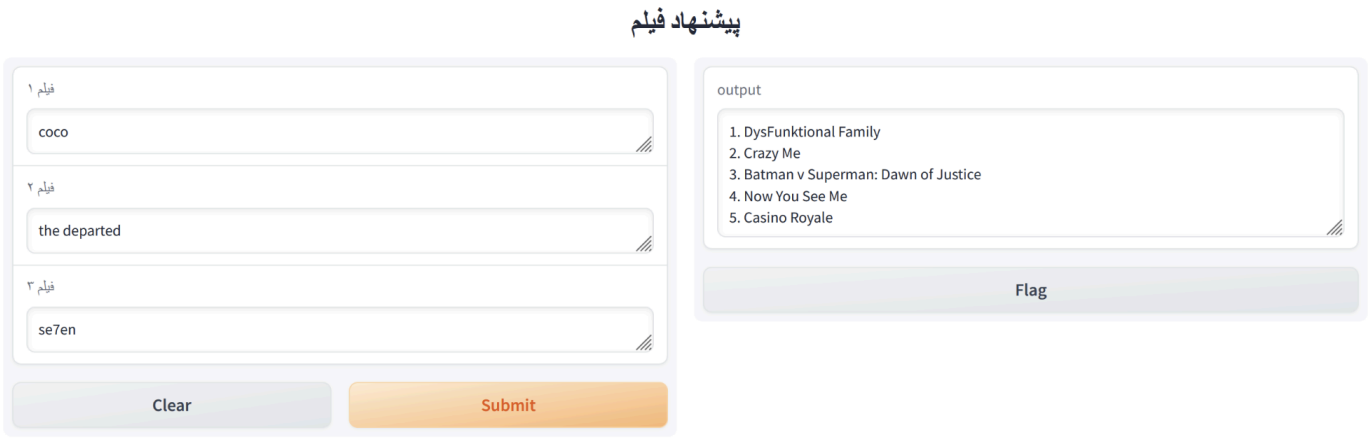
و این هم نمایش داده ها در 2 بعد است.

در آخر یک دیتاست جدید ایجاد می‌شود که دارای ستون‌های 'titles' (عنوان فیلم)، 'clusters' (کلاستر متناظر هر فیلم) است. این دیتاست سپس به فایل 'output.csv' ذخیره می‌شود.
ذخیره کردن این دیتاست به دلیل استفاده بعدی در کد recommender system است. با ذخیره کردن دیتاست به فایل، می‌توانیم در مراحل بعدی از این فایل استفاده کنیم و اطلاعات مربوط به عناوین فیلم‌ها و کلاسترها را در دسترس داشته باشیم.

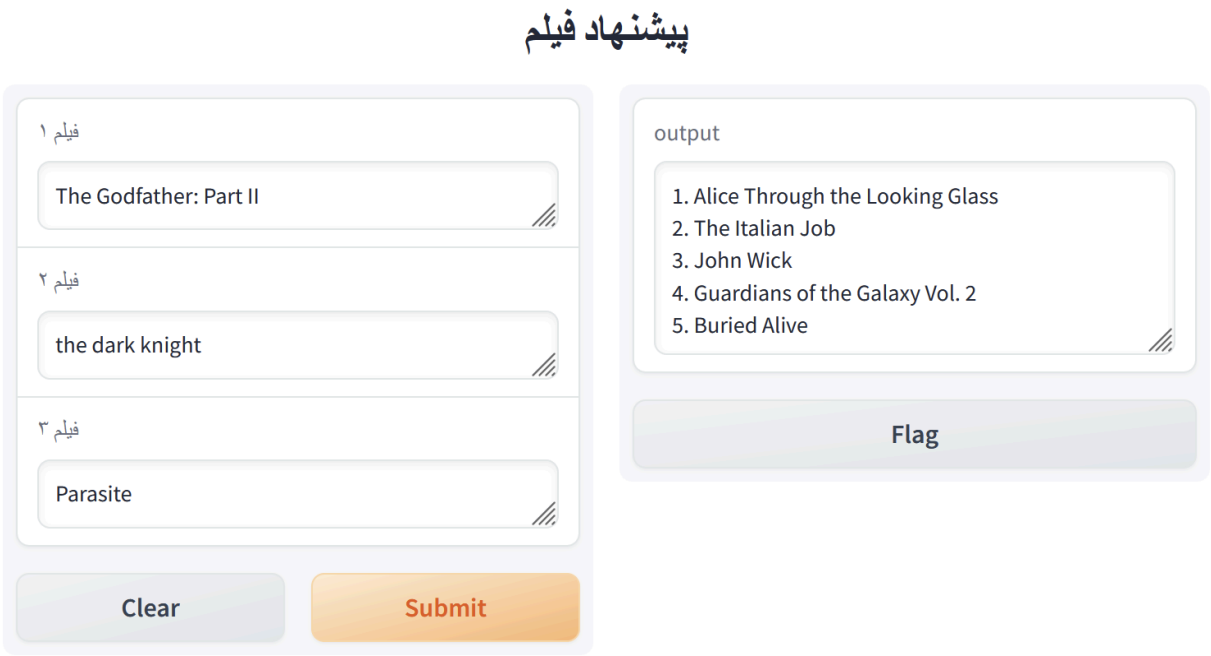
Recommender System ما یک سیستم است که با دریافت سه فیلم از کاربر، فیلم‌های پیشنهادی را بر اساس کلاسترهای مشابه در دیتاست ارائه می‌دهد.
در ابتدا دیتاست از فایل 'output.csv' خوانده می‌شود. سپس عناوین سه فیلم ورودی توسط کاربر به صورت عنوان با اولین حرف هر کلمه بزرگ (title case) تبدیل می‌شوند.
سپس با استفاده از عنوان فیلم‌ها، کلاسترهای مربوطه را پیدا می کنیم.
در ادامه، فیلم‌هایی که در هر کلاستر مشابه وجود دارند را به صورت تصادفی انتخاب می‌کنیم و این عملیات برای سه کلاستر مختلف صورت می‌گیرد.
در نهایت، لیستی از فیلم‌های پیشنهادی را با استفاده از شماره‌بندی و عنوان فیلم‌ها ساخته میشود.

در بخش نمایش گرافیکی (interface)، سه ورودی تکست با برچسب "فیلم ۱"، "فیلم ۲" و "فیلم ۳" ایجاد می‌شود.
با اجرای این کد، یک رابط کاربری تحت وب برای سیستم پیشنهاد فیلم ایجاد می‌شود و کاربر می‌تواند سه فیلم را وارد کرده و فیلم‌های پیشنهادی را دریافت کند.

نمونه اجرا1:



در این مثال، با ورودی فیلم‌های "The Departed"، "Coco" و "Se7en"، سیستم پیشنهاد دهنده ۵ فیلم را از کلاسترهای مختلف انتخاب می‌کند و در خروجی نمایش می‌دهد.
نمونه اجرا2:



نمونه اجرا3:

پیشنهاد فیلم

فیلم ۱

inside out

فیلم ۲

scarface

فیلم ۳

The Inglorious Bastards

Clear

Submit

output

1. Girlfight
2. The Scarlet and the Black
3. Moulin Rouge!
4. Contact
5. Project Almanac

Flag