# محبوب‌ترین فیلم ها و سریال های وبسایت ای ام دی بی

سینا طیبی

دانشجوی علوم کامپیوتر دانشگاه شهید بهشتی

sinaapolo969@gmail.com

آریانا آقامحمدی ‌قاضیانی

دانشجوی علوم کامپیوتر دانشگاه شهید بهشتی

cogokeye@gmail.com

دیبا الهی

دانشجوی علوم کامپیوتر دانشگاه شهید بهشتی

[elahidiba82@gmail.com](mailto:elahidiba82@gmail.com)

علی بخشش

دانشجوی علوم کامپیوتر دانشگاه شهید بهشتی

bkhshshly18@gmail.com

چکیده

در این مقاله به تحلیل داده های وب‌سایت IMDb در مورد فیلم ها و سریال ها می‌پردازیم و تاثیرات هرکدام از متغیر های مستقل را بر روی متغییر های اصلی و وابسته تحلیل و بررسی می‌کنیم. متغییر هایی که در جدول داده ها در وب‌سایت Kaggle موجود است، شامل دو بخش متغیر های مستقل و وابسته می‌شوند. متغیر های وابسته و اصلی ما شامل برآورد (Rate) و تعداد رأی Votes)) می‌باشد. به طبع هر چه تعداد رأی دهندگان بیشتر باشد، برآورد از این داده ها از صحت بیشتری برخوردار خواهد بود و معتبرتر است. متغیر هایی مانند میزان خشم مورد نمایش در فیلم ((Violence، ژانر فیلم (Genre)، مدت زمان آن (Duration)، ترسناک بودن آن (Frightening)، میزان برهنگی (Nudity)، استفاده از الفاظ رکیک (Profanity)، نشان دادن الکل و مصرف آن در فیلم (Alcohol)، نوع گواهی آن فیلم (Certificate) و علاوه بر موارد ذکر شده، برای سریال ها تعداد قسمت های آن (Episode) متغیر های مستقل هستند و می‌توانند تأثیرگذار واقع شوند. در ادامه، نوع (Type) به معنی سریال یا فیلم سینمایی بودن آن نیز می‌تواند تاثیرگذار باشد.

کلمات کلیدی : IMDb،Kaggle، برآورد، رأی، خشونت، ژانر، مدت‌ زمان، وحشت، برهنگی، مستقل، وابسته.

۱.مقدمه

برای تحلیل فیلم های سینمایی و سریال های ثبت شده در وب‌سایت IMDb بر اساس نظر بینندگانی که رأی خود را برای هر فیلم ثبت نموده‌اند، در ابتدا داده‌های مورد نیاز را از وب‌سایت Kaggle دریافت می‌کنیم و در اولین مرحله برای قابل استفاده و تحلیل شدن داده ها، آنها را تمیز می‌کنیم و تغییرات مورد نیاز در داده ها و متغیرها را ایجاد می‌کنیم.

در ادامه برای تحلیل داده ها نمودار هایی برای متغیر اصلی و وابسته خود، که برآورد (Rate) است، رسم می‌کنیم و تغییر آن نسبت به تمام متغیرهای مستقل موجود در جدول داده‌ها را رسم و بررسی می‌کنیم. موضوع دیگر قابل بررسی صحت و اعتبار برآورد های موجود است. یکی از راهکارهای مورد استفاده می‌تواند سنجش تعداد رأی‌ها باشد. زیرا به طبع هرچه تعداد رأی‌دهندگان بالاتر باشد، جامعیت برآورد نیز بیشتر است.

در انتها با استفاده از دسته‌بندی داده ها و نتایج قابل برداشت از نمودار ها می‌توانیم نتیجه‌گیری نهایی خود را ثبت کنیم.

بیان پرسش ها

۱. یکی از اصلی‌ترین پرسش‌هایی که در طی این گزارش سعی به پاسخ دادن به آن داریم تأثیر هر کدام از متغیر های ‌مستقل موجود در جدول داده ها برای هر فیلم، بر روی متغیر وابسته و اصلی ما برآورد (Rate) است.

۲. پرسش دیگری که مارا در تحلیل بهتر یاری می‌دهد و نیاز به پاسخ به آن داریم، اعتبار‌سنجی و صحت برآورد موجود در جدول داده‌ها است.

۱-۲.تمیز کردن داده ها

در ابتدا برای شروع تمیزکاری داده ها و آماده‌سازی آنها برای تحلیل، نیاز داریم داده های از دست رفته (Missing value) را با عدد صفر جایگزین کنیم. مقدار داده های از دست رفته در جدول داده ها NaN قرار گرفته است.

در ادامه نیاز داریم مقدار های یکتا را در هر ستون بیابیم و برای هرکدام از آنها مقدرا عددی‌ای در نظر بگیریم که در هنگام تحلیل ساده‌تر باشد.

به عنوان مثال برای ستون ترسناک بودن (Frightening) ستونی برای یکی از متغیر های مستقل ماست، تمام مقادیر موجود در جدول را می‌یابیم. که با استفاده از کتابخانه numpy مقادیر یکتا (Unique values) به شرح زیر خواهد بود :

array([‘Moderate’, ‘None’, ‘Mild’, ‘No Rate’, ‘Severe’], dtype=object)

در این قسمت نوع هر داده (Datatype) از نوع Object است که باید به int64 یا float تبدیل شود.

می‌توان به هر کدام از مقادیر یکتای موجود در ستون مورد نظر یک عدد را اختصاص داد. همچنین برا جداسازی فیلم های سینمایی از سریال هابرای کدام یک عدد صفر و یک انتخاب می‌کنیم.

قسمت بعدی حذف کردن داده هایی است که دوبار تکرار شده‌اند.

در روند تمیزکاری داده ها، برای هر ژانر موجود در جدول داده عددی مانند صفر یا یک برای یک فیلم مشخص اختصاص می‌دهیم. عدد صفر به معنی نبود آن ژانر مشخص در فیلم و عدد یک به معنی وجودیت آن است. پس از این مرحله، ستون پیشینی که به عنوان ژانر در جدول موجود بوده است را حذف می‌کنیم و ستون‌هایی شامل ژانرهارا اضافه و مقادیر صفر و یک را متناسب با فلیم یا سریال مورد بحث در آن جایگذاری می‌کنیم.

در قسمت بعد، به مرتب‌سازی فیلم ها در ردیف‌های جدول داده‌ها از قدیمی‌ترین به جدید‌ترین می پردازیم.

می‌توانیم برا مرتب‌سازی داده ها جدولی بر اساس سال تولید فیلم و سریال ایجاد کنیم، که در آن طول ردیف جدول تعداد نمونه های مورد بررسی برای تاریخ مشخص شده است.

همچنین برای دسته‌بندی ساده تر و آماده‌سازی داده ها برای تحلیل می‌توانیم بر حسب ژانرهای مختلف جدولی شامل برآورد آنها به همرا مدت ‌زمان ایجاد کنیم.

در انتها برای بررسی دقیق‌تر می‌توان فیلتر برآورد را بر روی داده‌های سال‌های مختلف قرار داد و داده‌هایی که برآورد آنها از مقدار مشخصی بالاتر است را در یک جدول جمع‌آوری کرد. در اخر جدولی به صورت محوری از فیلم و سریال های تاریخی مشخص با فیلتر برآورد خواهیم داشت.

پس از مراحل بالا که با استفاده از کتابخانه‌های numpy و pandas صورت گرفته‌است، داده‌های ما آماده تحلیل هستند.