**پیشنهاد فیلم با استفاده از تحلیل وابستگی**

**فاطمه ثقفی نژاد**

**پروژه درس داده کاوی**

**دکتر فهیم**

**فهرست**

چکیده............................................................................................................................................................... 3

معرفی.................................................................................................................................................................3

الگوریتم های تحلیل وابستگی............................................................................................................................. 5

متدولوژی کلی .................................................................................................................................................. 5

پیشنهاد فیلم........................................................................................................................................................ 6

مجموعه داده.......................................................................................................................................................6

پیاده سازی..........................................................................................................................................................6

کارکرد الگوریتم اپریوری................................................................................................................................... 9

استخراج قواعد وابستگی................................................................................................................................... 10

ارزیابی قواعد وابستگی..................................................................................................................................... 14

بحث و بررسی...................................................................................................................................................16

تحلیل نتایج....................................................................................................................................................... 17

نتیجه گیری....................................................................................................................................................... 18

**چکیده:**

در این تحقیق، یک سیستم پیشنهاد دهنده فیلم با استفاده از تحلیل وابستگی ارائه می‌شود. تحلیل وابستگی[[1]](#footnote-1) که هم چنین تحلیل سبد خرید[[2]](#footnote-2) نامیده می‌شود، تکنیکی داده‌کاوی برای برای تعیین اشیائی است که مکررا با هم رخ می‌دهند. مثلا در یک خرید، تعیین اقلامی که مکررا با هم خریداری می‌شوند. تمرکز این تحقیق بر مجموعه داده فیلم‌های مورد علاقه کاربران است. هدف کشف وابستگی فیلم‌ها با بررسی نقدهای کاربران و سپس ایجاد قواعدی برای پیشنهاد فیلم به کاربران با علایق مشابه است. پیاده سازی این سیستم توسط زبان پایتون در محیط جوپیتر نوت بوک و با استفاده از مجموعه داده MovieLens 100K که حاوی 100 هزار رتبه توسط 1000 کاربر روی 1700 فیلم است، و با استفاده از الگوریتم اپریوری[[3]](#footnote-3) انجام شده است.

**معرفی:**

تحلیل وابستگی نوعی داده‌کاوی است که مشابهت بین نمونه‌ها (یا موجودیت‌ها) را بدست می آورد. این مشابهت می تواند مشابهت بین موارد زیر باشد:

* کاربران یک وب سایت، برای فراهم کردن خدمات متنوع یا تبلیغات هدف دار
* اقلام برای فروش به آن کاربران، برای فراهم کردن پیشنهاد فیلم یا محصول

این روش می تواند روابط توامان بین فعالیت‌های صورت گرفته توسط کاربران یا گروه‌های مشخص را آشکار سازد. داده‌های مورد نیاز تحلیل وابستگی غالبا به شکل تراکنش توصیف می‌شوند. این ایده از تراکنش‌های یک فروشگاه نشات می‌گیرد، که در آن اقلامی که با هم خریداری می‌شوند به عنوان شیوه ای برای پیشنهاد محصولات به کاربرانی که ممکن است آن‌ها را بخرند، به کار گرفته می‌شود. در تحلیل وابستگی غالبا نتایج را رتبه بندی می‌کنیم، و n پیشنهاد برتر را انتخاب می‌کنیم.

تحلیل وابستگی در بسیاری از فرآیندهای دیگر نیز کاربرد دارد، از جمله:

* تشخیص تقلب
* بخش بندی مشتریان
* بهینه سازی نرم افزار
* پیشنهاد محصول

دو مفهوم مهم در تحلیل وابستگی عبارتند از: پشتیبان[[4]](#footnote-4) و اطمینان[[5]](#footnote-5). پشتیبان شاخصی است از این که یک مجموعه اقلام[[6]](#footnote-6) چند بار در یک مجموعه داده ظاهر می‌شود. پشتیبان X با توجه به T ، به صورت کسر تراکنش‌های t در مجموعه داده که شامل مجموعه اقلام X است، تعریف می‌شود.

(1)

اطمینان شاخصی است از این که یک قاعده چند بار درست (True) بوده است. مقدار اطمینان یک قاعده () با توجه به مجموعه تراکنش T عبارت است از کسری از تراکنش‌های شامل X که شامل Y نیز هستند.

(2)

**الگوریتم‌های تحلیل وابستگی**

الگوریتم کلاسیک تحلیل وابستگی، الگوریتم اپریوری است. این الگوریتم برای رفع مشکل نمایی ایجاد مجموعه اقلام پرتکرار[[7]](#footnote-7) درون یک مجموعه داده ارائه شده است. وقتی این مجموعه اقلام پرتکرار کشف شدند، ایجاد قواعد وابستگی سر راست است.

در الگوریتم اپریوری، ابتدا مطمئن می‌شویم که یک قاعده پشتیبان کافی درون مجموعه داده دارد. تعریف یک سطح کمینه پشتیبان، پاراکتر کلیدی اپریوری است. برای ساخت یک مجموعه اقلام پرتکرار، اقلام پرتکرار کوچکتر را با هم ترکیب می‌کنیم. برای مجموعه اقلام (A, B) با داشتن یک پشتیبان حداقل 30، هر دوی A و B باید ست کم 30 بار در مجموعه داده تکرار شوند. این خاصیت برای مجموعه‌های بزرگتر بسط داده می‌شود. برای پرتکرار بودن مجموعه اقلام (A, B, C, D) ، مجموعه اقلام (A, B, C) و D نیز باید پرتکرار باشند. این مجموعه اقلام‌های پرتکرار می توانند ساخته شوند و مجموعه اقلامی‌که پرتکرار نیستند، هرگز بررسی نخواهند شد. این کار بشدت پیچیدگی زمانی الگوریتم را کاهش می‌دهد.

دو الگوریتم دیگر برای تحلیل وابستگی، الگوریتم‌های Eclat و FP-growth هستند که موضوع تحقیق ما نیستند.

**متدولوژی کلی**

برای انجام کاوش قاعده وابستگی برای تحلیل وابستگی، ابتدا از الگوریتم اپریوری برای تولید مجموعه اقلام پرتکرار استفاده می‌شود. سپس، قواعد وابستگی (مثلا، اگر یک شخص فیلم الف را پیشنهاد می‌کند، فیلم ب را نیز پیشنهاد خواهد کرد) با آزمودن ترکیب‌های شروط و نتایج درون مجموعه اقلام پرتکرار ایجاد می‌شوند.

1. برای گام اول، الگوریتم اپریوری نیاز دارد به کمترین مقدار پشتیبان که بر اساس آن مجموعه اقلام پرتکرار در نظر گرفته شود. محموعه اقلام کمتر از مقدار حداقل پشتیبان لحاظ نمی‌شوند.
2. در گام دوم، پس از کشف اقلام پرتکرار، قواعد وابستگی بر اساس اطمینان خود آزموده می‌شوند. می توانیم یک سطح اطمینان حداقلی، و تعداد قواعد بازگشتی انتخاب کنیم یا بسادگی همه آن‌ها را برگردانیم و تصمیم بگیریم با آن‌ها چه کنیم.

**پیشنهاد فیلم**

پیشنهاد محصول یک کسب و کار عظیم است. فروشگاه‌های آنلاین با استفاده از این ایده به مشتریان محصولاتی را پیشنهاد می‌دهند که ممکن است بخرند. ارائه پیشنهادات بهتر منجر به فروش بهتر می‌شود. پیشنهاد محصول، از جمله فیلم یا کتاب، برای سال‌ها مورد تحقیق بوده است.

**مجموعه داده**

چندسال پیش، گروهی از پژوهشگران در دانشگاه مینه سوتا، چندین مجموعه داده برای آزمودن الگوریتم‌ها در این حوزه منتشر کرده اند. آن‌ها چندین نسخه از مجموعه داده رتبه دهی فیلم با اندازه‌های مختلف منتشر کرده اند. نسخه مورد استفاده در این پروژه، مجموعه ای از 100 هزار نقد است. این مجموعه با نام MovieLens 100K dataset در وبگاه زیر در دسترس است.

<http://files.grouplens.org/datasets/movielens/ml-100k/>

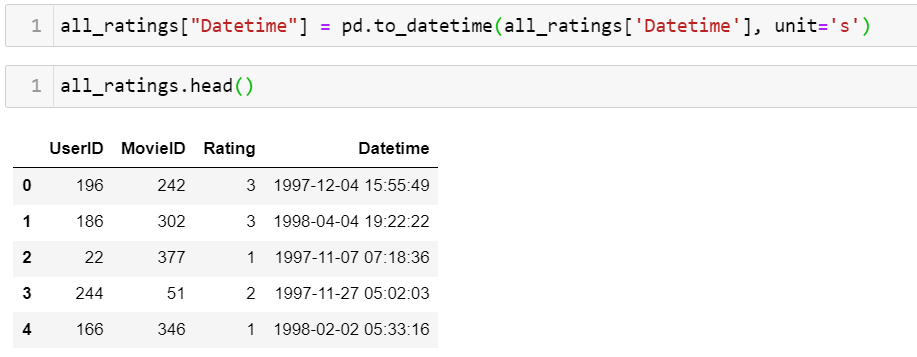
**پیاده سازی**

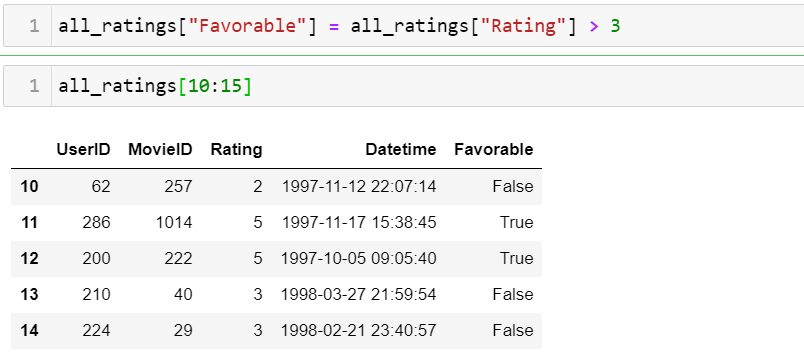
پیاده سازی پیشنهاد فیلم با الگوریتم اپریوری در محیط جوپیتر نوت بوک، به زبان پایتون انجام شده است.

ابتدا مجموعه داده بارگذاری می‌شود. و فایل رتبه‌ها (u.data) خوانده می‌شود.سپس، پنج رکورد اول این فایل با استفاده از کتابخانه pandas نمایش داده می‌شود.



این فایل حاوی چهار ستون (ویژگی) است: شناسه کاربر (UserID) ، شناسه فیلم (MovieID) ، رتبه (rating) ، تاریخ و زمان (Datetime) است. از آن جایی که تاریخ نقدها برای پیشنهاد فیلم در آینده مهم است، تاریخ را به فرمت خواناتر تبدیل می‌کنیم.

هدف ما تولید قواعدی از این دست است: «اگر شخصی این مجموعه از فیلم‌ها را پیشنهاد کند، این فیلم را نیز پیشنهاد خواهد کرد». برای این کار، ابتدا باید تعیین کنیم آیا فرد فیلمی را پیشنهاد خواهد داد. این کار با ایجاد ویژگی جدی بنام مطلوب (Favorable) امکان پذیر است، که در صورتی که فرد نقد مثبتی در مورد یک فیلم داشته باشد، درست (True) است. در اینجا نقد مثبت را به رتبه بالاتر از 3 تعیین می‌کنیم.

در مرحله بعد، مجموعه داده خود را نمونه گیری کرده و یک مجموعه داده آموزش شکل می‌دهیم. با این کار، حجم مجموعه داده مورد بررسی نیز کاهش پیدا می‌کند و سرعت الگوریتم اپریوری افزایش می‌یابد. برای این کار نقدهای 200 کاربر اول را بدست می آوریم. سپس، مجموعه داده ای از تنها نقدهای مثبت درون نمونه ایجاد می‌کنیم. ما در جستجوی نقدهای مثبت کاربر برای مجموعه اقلام خود هستیم. بنابراین، فیلم‌هایی که از هر کاربر رتبه مطلوب دریافت کرده اند را با گروه بندی مجموعه داده بر اساس UserID و تکرار فیلم‌ها در هر گروه محاسبه می‌کنیم. سرانجام، یک DataFrame ایجاد می‌کنیم که فراوانی نقدهای مثبت هر فیلم را نشان می‌دهد.

**کارکرد الگوریتم اپریوری**

الگوریتم اپریوری بخشی از متدولوژی تحلیل وابستگی ماست و به طور خاص با یافتن اقلام پر تکرار درون مجموعه داده سر و کار دارد. رویه پایه اپریوری، ایجاد مجموعه اقلام کاندید جدید از مجموعه اقلام پر تکرارکشف شده قبلی است. این کاندیداها از نظر پر تکرار بودن بررسی می‌شوند، و سپس الگوریتم به صورت زیر تکرار می‌شود:

گام اول:

ایجاد مجموعه اقلام پر تکرار اولیه به طول 1

گام دوم:

ایجاد مجموعه اقلام کاندید با استفاده از ابر محموعه (k+1) از مجموعه اقلام پر تکرار فعلی

گام سوم:

آزمودن مجموعه اقلام و نگهداشتن تنها مجموعه اقلام پر تکرار

بله

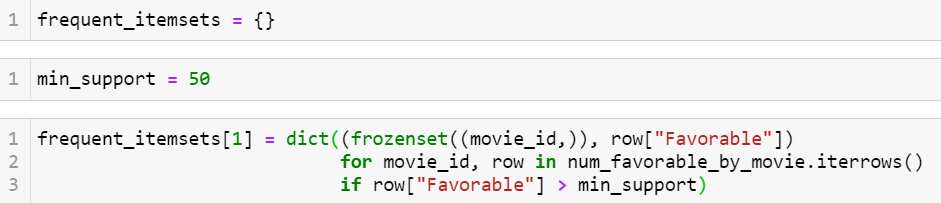
گام چهارم:

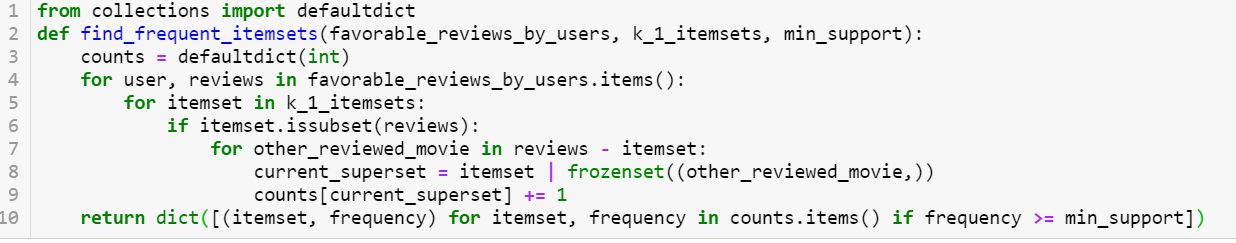
آیا مجموعه اقلام پر تکرار در مرحله سوم کشف شده اند؟

خیر

گام پنجم:

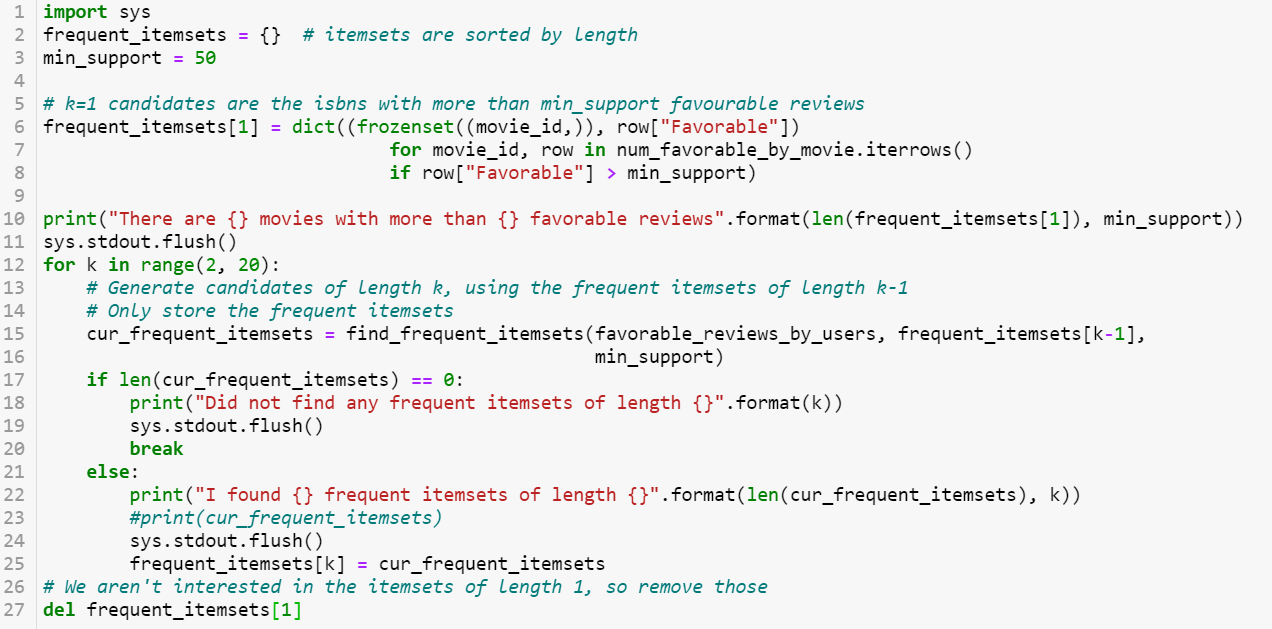
همه مجموعه اقلام پر تکرار را برگردان

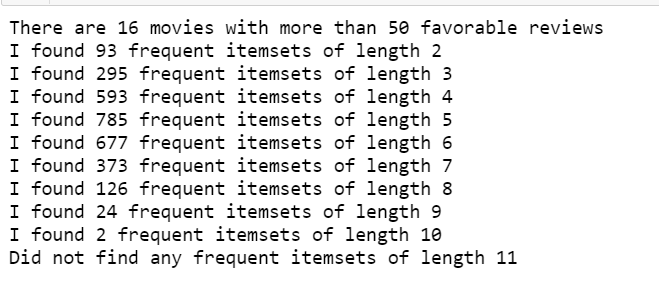
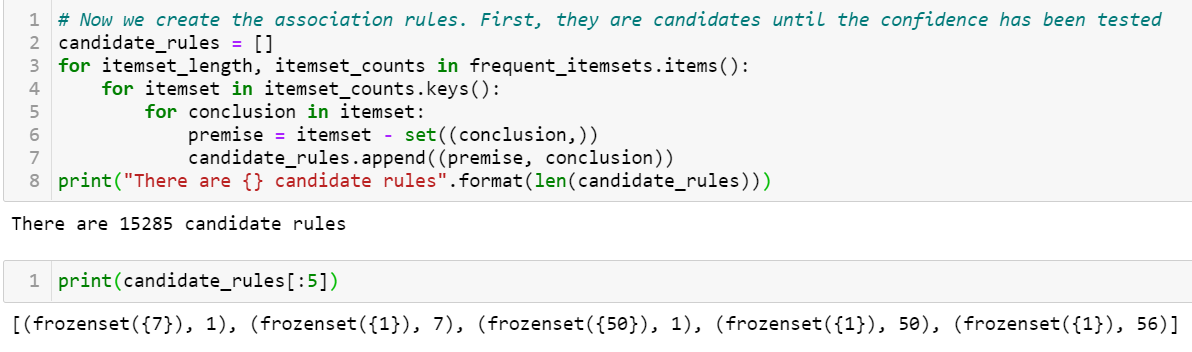
مجموعه اقلام پر تکرار را در دیکشنری با کلید طول مجموعه اقلام ذخیره می‌کنیم. با این کار دسترسی به مجموعه اقلام با طول معین و مجموعه اقلام اخیرا کشف شده، ممکن است. حداقل پشتیبان را نیز برای در نظر گرفتن پر تکرار بودن مجموعه اقلام تعیین می‌کنیم. برای پیاده سازی گام اول، مجموعه اقلامی از تک تک فیلم‌ها ایجاد می‌کنیم و آن را از نظر پر تکرار بودن بررسی می‌کنیم. برای این کار از frozenset برای عملیات سریع مجموعه ای و به عوان کلیدهای دیکشنری شمارش بهره می بریم.

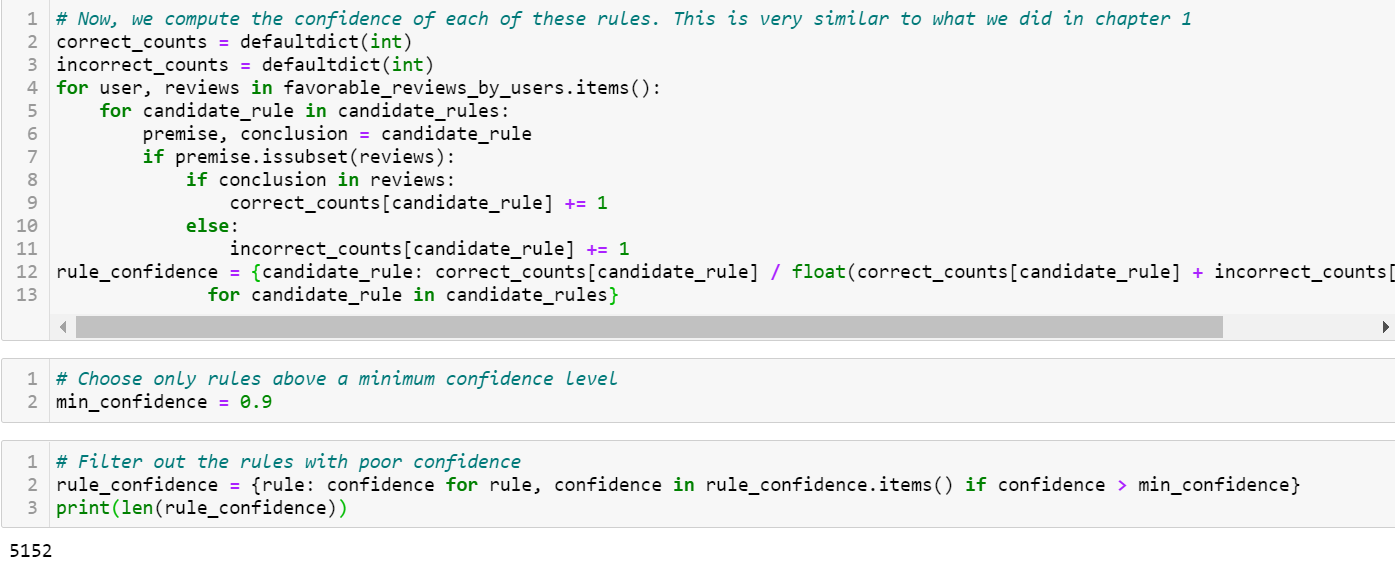
گام دوم و سوم را برای کارایی با ایجاد تابعی که اقلام کشف شده جدید پر تکرار را می‌گیرد، ابر مجموعه‌ها را ایجاد می‌کند، و سپس پر تکرار بودن آن‌ها را بررسی می‌کند، پیاده سازی می‌کنیم.

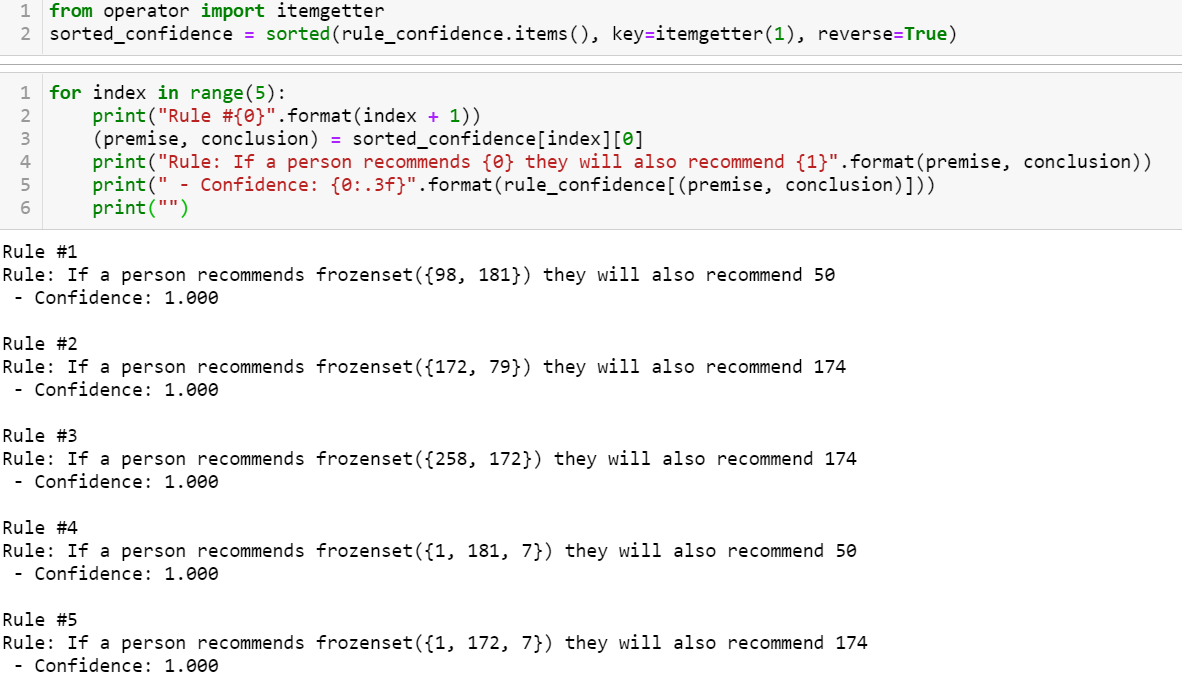
**استخراج قواعد وابستگی**

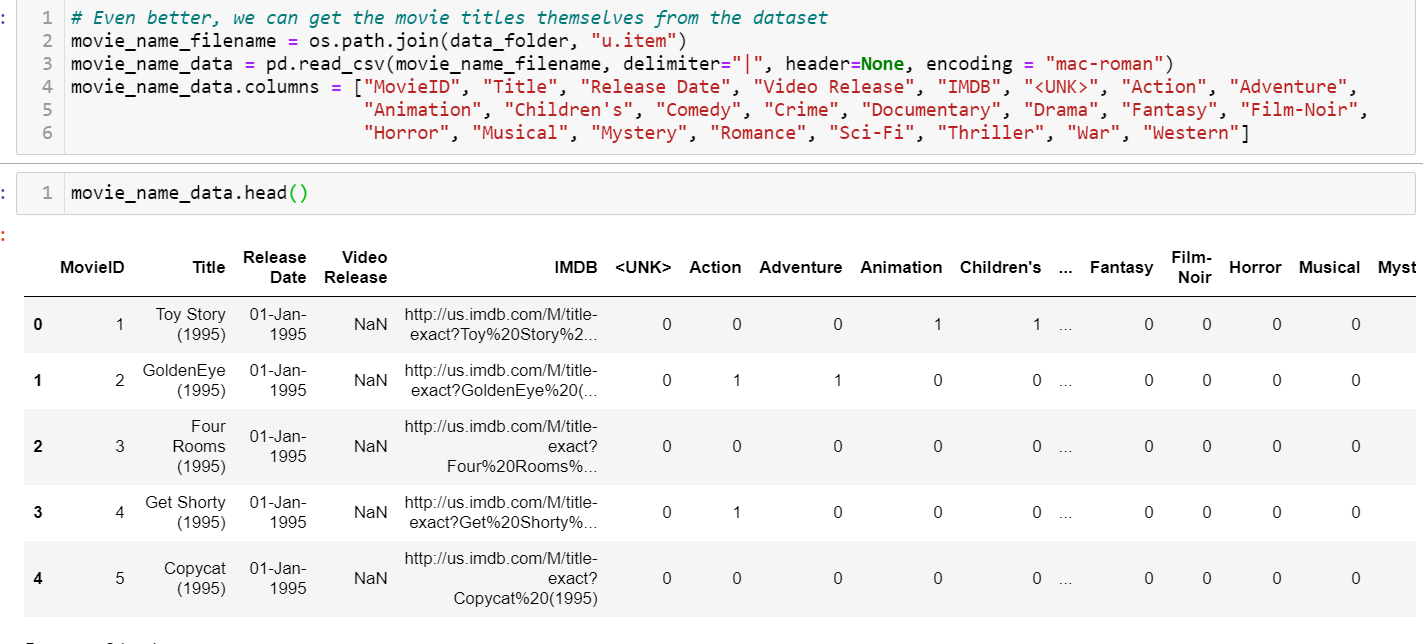
پس از تکمیل الگوریتم اپریوری، لیست مجموعه اقلام پر تکرار بدست می آید. مجموعه اقلام پر تکرار، مجموعه ای از اقلام با حداقل پشتیبان است، در حالی که قاعده وابستگی یک فرضیه و یک نتیجه دارد. داده‌ها برای هر دو یکسان است. برای هر مجموعه اقلام، تعدادی قاعده وابستگی را با قرار دادن هر فیلم به عنوان نتیجه و بقیه فیلم‌ها به عنوان نتیجه، تولید می‌کنیم.

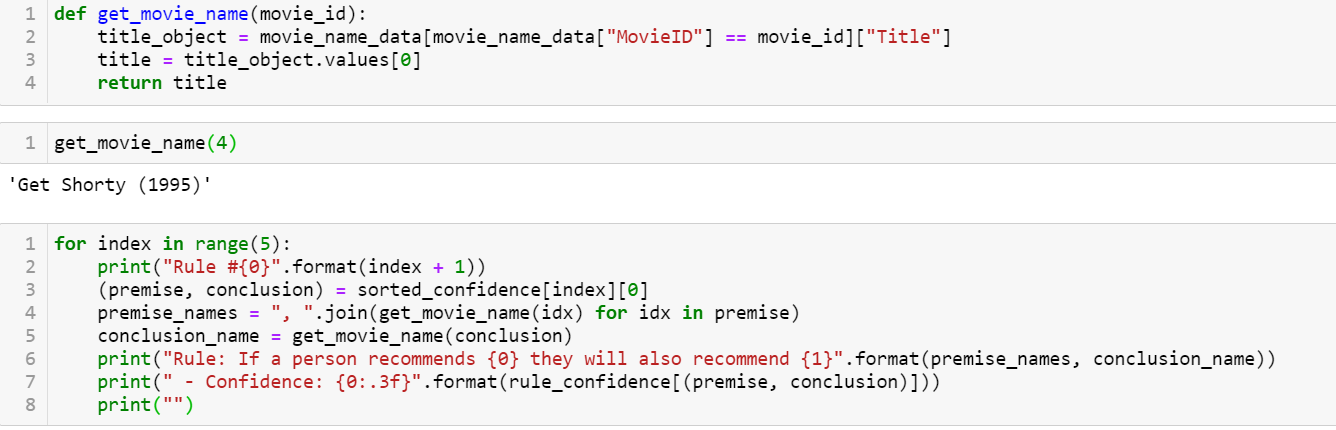
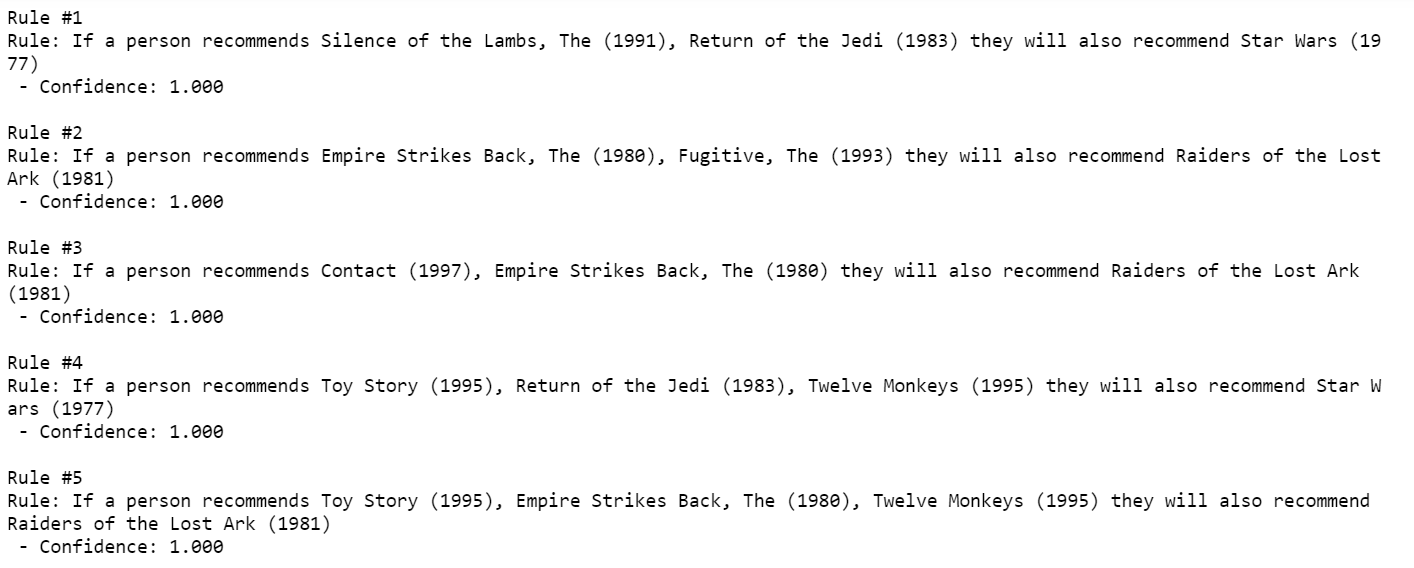
ابتدا لیستی از همه قواعد اقلام پر تکرار با تکرار روی هر مجموعه اقلام پر تکرار کشف شده با هر طول، تولید می‌کنیم. سپس روی هر فیلم در مجموعه اقلام این کار را به عنوان نتیجه تکرار می‌کنیم. با این کار تعداد زیادی قاعده کاندید بر گردانده می‌شود. چند قانون اول را به عنوان نمونه چاپ می‌کنیم.

در مرحله بعد، اطمینان را برای هر کدام از این قواعد محاسبه می‌کنیم. قواعد برآورده کننده سطح حداقل اطمینان را نگه داشته و قواعد با اطمینان زیر آستانه را نادیده می‌گیریم.

سپس 5 قاعده برتر را با مرتب سازی دیکشنری اطمینان و چاپ نتایج بدست می آوریم. خروجی مورد نظر تنها شناسه فیلم‌ها را چاپ می‌کند که بدون نام فیلم مفید نیست.

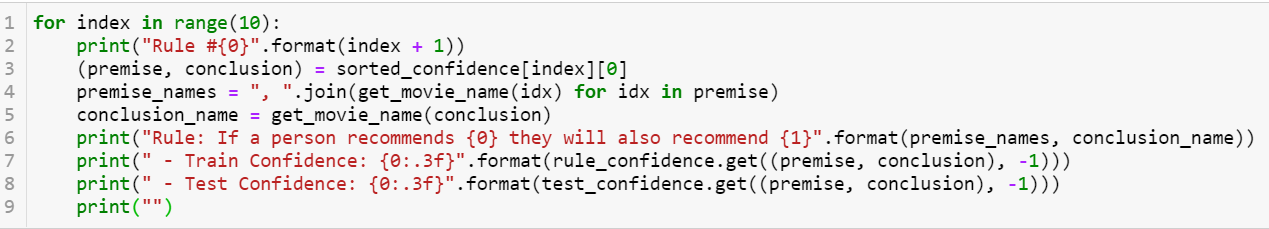
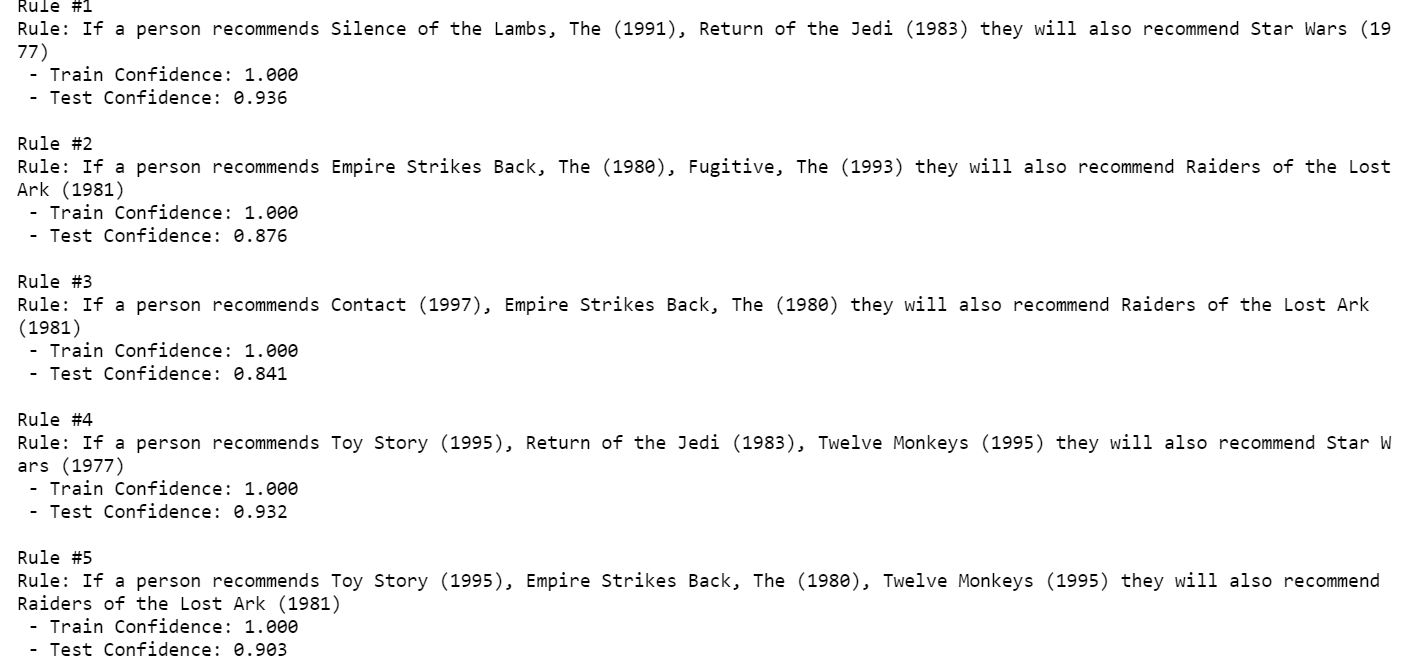
مجموعه داده فایل دیگری بنام u.items دارد که نام فیلم‌ها و شناسه مربوطه و نیز برخی اطلاعات مفید دیگر مثل ژانر را نگهداری می‌کند. این فایل را pandas بارگذاری کرده و عنوان فیلم‌ها را به مجموعه داده اضافه می‌کنیم.

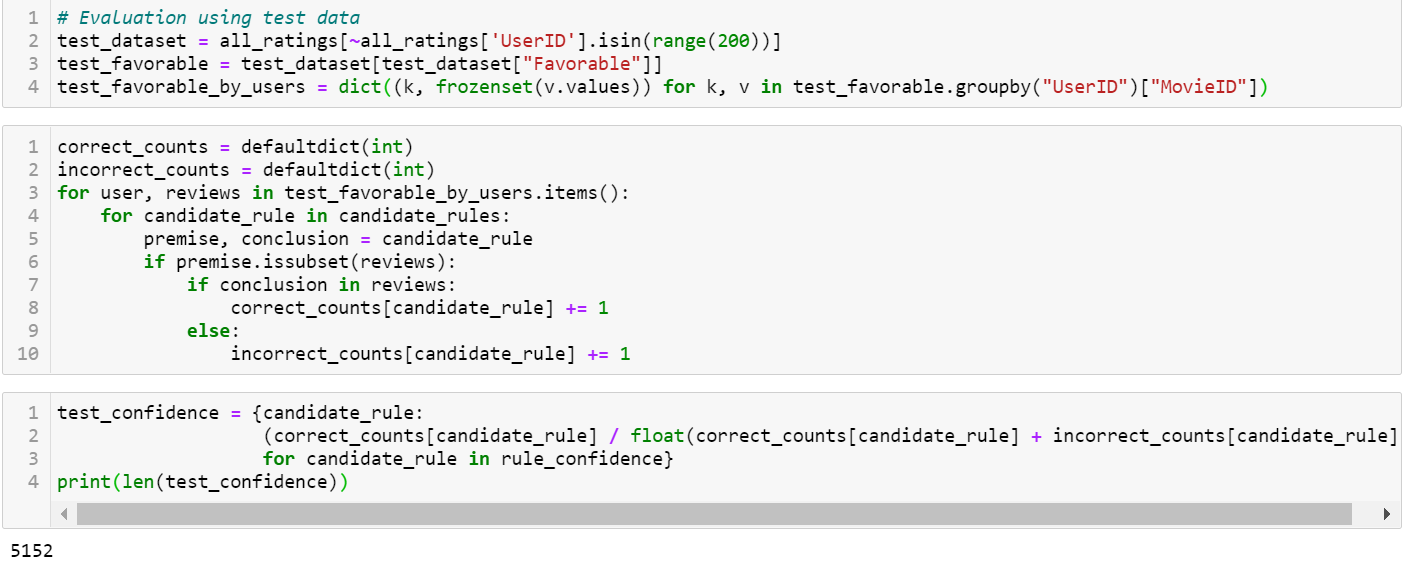
سپس می توان قواعد استخراجی را به فرمت خواناتر درآورد.

**ارزیابی قواعد وابستگی**

می توانیم قواعد وابستگی را با استفاده از مفهوم مشابه برای طبقه بندی ارزیابی کنیم. از یک مجموعه داده آزمون که برای آموزش استفاده نشده است، بهره برده و قواعد کشف شده را بر اساس عملکرد آن‌ها روی مجموعه آزمون ارزیابی می‌کنیم.

برای این کار، اطمینان مجموعه آزمون را محاسبه می‌کنیم، که اطمینان هر قاعده روی مجموعه آزمون است. در این ارزیابی از معیارهای رسمی ارزیابی استفاده نمی‌شود. در این روش، ما به صورت غیر رسمی به قواعد نگاه می‌کنیم تا ببینیم کدام یک قابل اطمینان تر هستند:

1. ابتدا، مجموعه داده آزمون را که شامل همه رکوردهای استفاده نشده در مجموعه آموزش است، استخراج می‌کنیم. 200 کاربر اول را برای مجموعه آموزش استفاده کردیم، و بقیه مجموعه داده را برای آزمون استفاده می‌کنیم. مانند مجموعه آموزش، نقدهای مطلوب را برای هر کاربر در این مجموعه داده نیز بدست می آوریم.
2. سپس نمونه‌های صحیح را که فرض به نتیجه منجر می‌شود، شمارش می‌کنیم.
3. بعد، اطمینان هر قاعده از تعداد صحیح‌ها را محاسبه می‌کنیم و آن‌ها را مرتب می‌کنیم.
4. در نهایت، بهترین قواعد وابستگی را با عناوین فیلم‌ها چاپ می‌کنیم.



می توانیم ببینیم کدام قواعد بر روی داده‌های جدید نادیده عملی تر هستند:

مثلا، قاعده 2 در داده‌های آموزش اطمینان کامل را دارد، ولی در داده‌های آموزش تنها برای 60 درصد از موارد صحیح است. بسیاری از قواعد دیگر در 10 قاعده برتر اطمینان بالا در داده‌های آموزش دارند، و قواعد خوبی برای پیشنهادات هستند.

**بحث و بررسی**

الگوریتم اپریوری الگوریتم ساده ای است مناسب مجموعه داده های کم حجم. این الگوریتم به صورت تکراری، مجموعه اقلام پر تکرار را تولید می کند. اپریوری در گام اول از حداقل مقدار پشتیبان به عنوان سنجه استفاده می کند. و رکوردهای کمتر از آستانه پشتیبان را نادیده می گیرد. در گام بعد، پس از بدست آمدن مجموعه اقلام پر تکرار، یک حداقل اطمینان انتخاب شده و قاعده وابستگی بالای آستانه اطمینان برگردانده می شود.

* **پیش پردازش داده‌ها**

برای قضاوت در مورد این که فردی یک فیلم را به دیگری پیشنهاد خواهد کرد، از مقادیر رتبه هایی که کاربران به فیلم ها داده اند، استفاده می شود. این رتبه ها از 1 تا 5 (عدد یا ستاره) هستند. عدد یا ستاره بالاتر به میزان محبوبیت بیشتر فیلم از نظر کاربر است. در این جا عدد مطوب رتبه بالای 3 (یعنی 4 و 5) در نظر گرفته می شود و کلیه رکوردهای با رتبه 1 ، 2 و 3 نادیده گرفته می شوند.

سپس 200 رکورد برتر از این مجموعه به عنوان مجموعه آموزش، برای آموزش الگوریتم اپریوری استفاده می شوند. و تعداد پیشنهادات برای هر فیلم محاسبه می شوند.

الگوریتم اپریوری به ترتیب:

- ایجاد یک مجموعه اقلام پر تکرار اولیه

- ایجاد مجموعه پسین از مجموعه اقلام پر تکرار جاری، و تعیین این که آیا این مجموعه پسین خود نیز پر تکرار است، و حذف غیر پر تکرارها

- برگرداندن مجموعه اقلام پر تکرار

روی مجموعه داده اجرا می شود.

در نهایت ارزیابی عملکرد الگوریتم روی بقیه مجموعه داده صورت می گیرد و با عملکرد آن روی مجموعه آموزش مقایسه می شود.

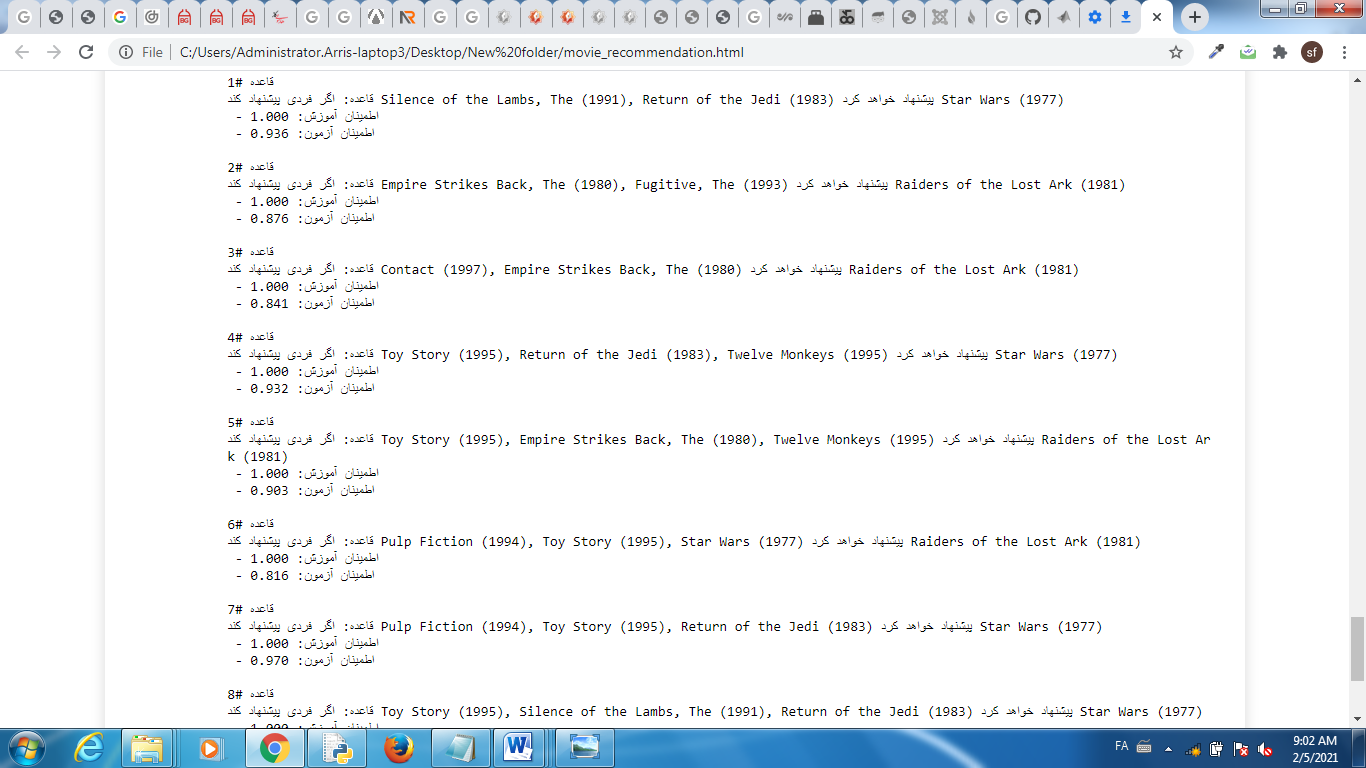
نتایج ارزیابی حاکی از آن است که کارایی الگوریتم متوسط به بالاست و روی مجموعه داده فوق موثر عمل کرده است. گر چه ارزیابی صورت گرفته یک ارزیابی رسمی بر اساس معیارهای دقت و سرعت معمول نیست.

الگوریتم اپریوری گرچه ساده و پیاده سازی آن آسان است. اما نیازمند بررسی قواعد کاندیدای زیادی است، برای پایگاه داده های بزرگ مناسب نیست و تنها یک آستانه حداقل پشتیبان را می پذیرد. مساله بعد، مجموعه داده است. مجموعه داده مووی لنز از پیش تعریف شده و تمیز است ولی همه مجموعه داده ها این گونه نیستند و برخی از آن ها نیاز به تمیزکاری و تبدیلات قبل از اعمال الگوریتم هستند.

برای آموزش و تست الگوریتم نیز از روش ساده تقسیم به داده های آموزش و داده های آزمون استفاده شد که می توان از روش های پیشرفته تر مثل Cross Validation برای آموزش و آزمون الگوریتم سود برد.

**تحلیل نتایج**

قسمتی از قواعد در شکل زیر آورده شده است:



در پیاده سازی، حداقل آستانه پشتیبان 50 انتخاب شده است (min\_support =50)، یعنی اقلامی پرتکرار لحاظ می شوند که در بیش از یک چهارم تراکنش ها رخ داده باشند. بر اساس نتایج اقلام پر تکرار: مجموعه اقلام پرتکرار با طول 2 تا 10 بدست می آید. برای این مجموعه اقلام پر تکرار، 15285 قاعده کاندیدا می شود. برای فیلتر کردن این قواعد، از سطح اطمینان قاعده که در اینجا بالاتر از 0.9 است (سطح اطمینان بین 0 و 1 است)، استفاده می شود. از این تعداد، 5152 قاعده باقی می ماند. برای مثال قاعده یک می گوید:

اگر فردی سکوت بره ها و بازگشت جدی را پیشنهاد کند، با اطمینان 1 (100%) جنگ ستارگان را نیز پیشنهاد می دهد.

همین فرآیند روی مجموعه آزمون انجام می شود و مثلا قاعده یک به صورت زیر بدست می آید:

اگر فردی سکوت بره ها و بازگشت جدی را پیشنهاد کند، با اطمینان 0.936 (93.6%) جنگ ستارگان را نیز پیشنهاد می دهد.

این مورد نشان می دهد که الگوریتم در مرحله آزمون روی داده های نادیده خوب عمل کرده است. اما برای همه قوانین دقت در این حد نیست.

**نتیجه‌گیری**

در این پروژه، تحلیل وابستگی به منظور پیشنهاد فیلم بر اساس مجموعه بزرگی از نقدها صورت گرفت. این کار در دو مرحله صورت گرفت. ابتدا، مجموعه اقلام پر تکرار در داده ها با استفاده از الگوریتم اپریوری پیدا شدند. سپس، قواعد وابستگی برای این مجموعه اقلام ایجاد شد.

با توجه اندازه مجموعه داده، استفاده از الگوریتم اپریوری ضروری بود. ولی برای مجموعه داده‌های بزرگ تر الگوریتم‌های هوشمندتری لازم است. آموزش را روی زیرمجموعه از داده‌ها برای یافتن قواعد وابستگی انجام دادیم، و سپس این قواعد را روی مجموعه آزمون بررسی کردیم. نتایج بدست آمده، نشان از عملکرد خوب الگوریتم داشت.

1. affinity analysis [↑](#footnote-ref-1)
2. Market Basket Analysis [↑](#footnote-ref-2)
3. Apriori algorithm [↑](#footnote-ref-3)
4. support [↑](#footnote-ref-4)
5. confidence [↑](#footnote-ref-5)
6. itemset [↑](#footnote-ref-6)
7. frequent itemsets [↑](#footnote-ref-7)