"نام"

خلاصه

در این پروژه سعی شده است با توجه به دیتا ست داده شده قوانین تجمعی ای (Association Rules) را استخراج کرد که بتواند در یافتن ارتباطات بین محصولات (ProductName) عمل کند.

برای پاکسازی داده ها از روش های مختلفی استفاده شده که در زیر خواهیم دید

نمای داده ها

در این پروژه بنا بر اینکه تمرکز روی استخراج قوانین تجمعی بوده است تنها از ویژگی ProductNames استفاده شده است تا بتوان ارتباط بین محصولات را دریافت (میتوان ویژگی هایی مانند قیمت را نیز گسسته سازی کرد (Binning) و در کنار محصولات استفاده کرد اما با توجه به کابرد رایج ProductNames انجام اینکار کمی بی معنی خواهد بود

StoreName	City	Area	ProductName	QuantitySold	QuantityAvailable	Cost	RetailPrice
National Stores	Ouro Branco	9 Springview Point	Chocolate Bar - Smarties	1	11	4873.07	1500.39
Family Dollar	Gerakaroú	5434 Daystar Circle	Pepper - Red Bell	6	1	3089.77	2095.61
BJ's Wholesale Club	Radoboj	3 Darwin Drive	Chickensplit Half	10	11	4591.63	364.02
Ocean State Job Lot	Al Madīd	684 Bunting Lane	Zucchini - Green	3	1	33.13	2111.21
Ollie's Bargain Outlet	Bo Phloi	50162 John Wall Drive	Cod - Salted, Boneless	9	3	3750.18	934.82
...	...	...	...	...	...	...	...
Walmart	Anserma	7 Roxbury Place	Okra	3	3	2395.99	596.47
T.J. Maxx	Xinglongjie	9476 Morrow Trail	Pork - Smoked Back Bacon	12	12	4646.21	3477.17
Burlington Coat Factory	Isangel	3029 Hollow Ridge Place	Flounder - Fresh	9	6	4739.64	2702.02
Tuesday Morning	Gaya	496 Dapin Hill	Bread - Wheat Baguette	6	6	2311.42	487.96
Big Lots	Sexmoan	99 Granby Place	Soup - Knorr, French Onion	2	2	4712.33	1769.61

تمیز سازی داده ها

از آنجایی که نام محصولات تا حد تثیر گذاری با نام برند آن ها مشخص شده بررسی سطر به سطر سبد ها محصول و جایگذاری نام های برند با خود محصول ممکن نیست اما سعی شده است از کلیاتی تا جای ممکن برای تمیز سازی استفاده شد.

برای تمیز سازی داده ها از روش های زیر به ترتیب استفاده شده است:

- محصولاتی که کمتر تر از 3 حرف (2 یا کمتر) داشته باشند را حذف میکنیم با اینکار تعدادی از سطر ها نیز حذف خواهد شد. (مثال زیر سایز یک محصول را نشان میدهد)

```
X12    Xl
```

- اعداد را در محصولات حذف میکنیم:

```
1%  1.36l  1.5lit  10%  10/20  112con  12  125g  12x16
```

- محصولاتی نیز وجود دارند که هم بشکل مفرد و هم بشکل غیر مفرد نامبرده شده اند (مانند Pork و Porks) این محصلات باید هر دو با نام مفرد خود ظاهر شوند.

- محصولاتی وجود دارند که با پسوند هایی مانند سایز آن ها یا رنگ یک محصول ذکر شده اند.

```
['Large', 'Medium', 'Small', 'Double', 'Yellow', 'Red', 'Blue', 'Green', 'White']
```

- پسوند هایی وجود دارند که با هم مترادف بوده و با یکی کردن نام آن ها به پیچیدگی کمتر مدل منجر میشود. (مانند Organic و Untrimmed).

روش های دیگری را نیز میتوان در پاکسازی داده ها بکار برد (برای مثال استفاده از Regex میتواند بسیار مفید باشد) اما بنا بر اینکه در داده های ما سبد های کالا حداکثر 3 محصول دارند (داده ها میتواند مصنوعی یا Synthetic باشد چرا که عموما در سبد های کالا تعداد محصولات بیشتر خواهد بود. که خود به نفع ما است چرا که در ابتد هدف از این نوع آنالیز داده پیدا کردن محصولاتی است که معمولا با هم خریداری میشوند)

با توجه به توضیح بالا انتظار خواهیم داشت که Support آیتم ست ها کم و آیتم ست های بزرگی نداشته باشیم. در ادامه خواهیم دید که این انتظار به وقوع میپیوندد.

استخراج آیتم ست های معمول

در عکس زیر تعدادی از آیتم ست های رایج همراه ساپورت آن ها آمده است

	support	itemsets
0	0.007	(Appetizer)
1	0.004	(Apple)
2	0.005	(Baby)
3	0.004	(Bacon)
4	0.006	(Bag)
...	...	...
125	0.004	(Wine, Merlot)
126	0.004	(Mix, Sauce)
127	0.004	(Oil, Olive)
128	0.004	(Riesling, Wine)
129	0.004	(Sweet, Shell)

همانطور که میبینیم ایتم ســـت های بزرگتری از 2 آیتم دارای ساپورت نسبتا زیادی نیستند.

برای اینکه قوانین اسـتخراجی دارای معنا باشـند مجبور هسـتیم مینیمم ســـاپورت را عددی پایین مانند (0.003 یعنی حداقل 3 رخداد از 1000 سـطر داده ها) قرار دهیم، البته سـودی در این ساپورت پایین نهفته است که در زیر خواهیم دید.

تعدادی از محصـولاتی که در زیر میبینیم نام های غریبی داشـته به همین دلیل توضیح مختصری از این محصولات میبینیم:

- Aborio: نوعی از برنج (شرقی)
- Knor: نام برندی از طعم دهنده سوپ (این آیتم با محصول Soup ارتباط زیادی دارد)
- Oasis: نوعی نوشیدنی الکلی
- Merlot: نوعی نوشیدنی الکلی
- Magnotta: نوعی نوشیدنی الکلی
- Riesling: نوعی نوشیدنی الکلی
- Rose: نوعی نوشیدنی الکلی
- Organic: منظور محصولات سبزیجاتی است (این محصولات با تعدادی از انواع گوشت مانند گوشت پرندگان و گاو زیاد خریداری شده است)
- ...

antecedents	consequents	antecedent support	consequent support	support	confidence
(Rice)	(Aborio)	0.007	0.003	0.003	0.428571
(Aborio)	(Rice)	0.003	0.007	0.003	1.000000
(Apple)	(Juice)	0.004	0.011	0.003	0.750000
(Bacon)	(Pork)	0.004	0.015	0.004	1.000000
(Bagel)	(Presliced)	0.006	0.003	0.003	0.500000
(Presliced)	(Bagel)	0.003	0.006	0.003	1.000000
(Base)	(Chicken)	0.006	0.013	0.003	0.500000
(Tail)	(Beef)	0.006	0.021	0.004	0.666667
(Cane)	(Beets)	0.003	0.005	0.003	1.000000
(Beets)	(Cane)	0.005	0.003	0.003	0.600000
(Beets)	(Organic)	0.005	0.011	0.003	0.600000
(Blk)	(Tray)	0.003	0.004	0.003	1.000000
(Tray)	(Blk)	0.004	0.003	0.003	0.750000
(Blush)	(Sauce)	0.003	0.020	0.003	1.000000
(Roll)	(Bread)	0.007	0.020	0.003	0.428571
(Pancake)	(Cake)	0.004	0.007	0.004	1.000000
(Cake)	(Pancake)	0.007	0.004	0.004	0.571429
(Campbells)	(Soup)	0.006	0.022	0.004	0.666667
(Cane)	(Organic)	0.003	0.011	0.003	1.000000
(Chili)	(Pepper)	0.006	0.019	0.003	0.500000
(Chili)	(Soup)	0.006	0.022	0.003	0.500000
(Coffee)	(Cream)	0.007	0.012	0.003	0.428571
(Pasta)	(Dry)	0.013	0.021	0.007	0.538462
(Sherry)	(Dry)	0.004	0.021	0.003	0.750000
(Individual)	(Muffin)	0.009	0.008	0.004	0.444444
(Muffin)	(Individual)	0.008	0.009	0.004	0.500000
(Knorr)	(Soup)	0.005	0.022	0.005	1.000000
(Strawberry)	(Lemonade)	0.006	0.012	0.003	0.500000
(Salad)	(Longos)	0.003	0.006	0.003	1.000000
(Longos)	(Salad)	0.006	0.003	0.003	0.500000
(Magnotta)	(Wine)	0.005	0.066	0.004	0.800000
(Merlot)	(Wine)	0.004	0.066	0.004	1.000000
(Oasis)	(Mix)	0.006	0.015	0.003	0.500000
(Olive)	(Oil)	0.007	0.014	0.004	0.571429
(Tenderloin)	(Organic)	0.006	0.011	0.003	0.500000
(Paste)	(Primerba)	0.007	0.004	0.003	0.428571
(Primerba)	(Paste)	0.004	0.007	0.003	0.750000
(Squash)	(Pattypan)	0.005	0.003	0.003	0.600000
(Pattypan)	(Squash)	0.003	0.005	0.003	1.000000
(Quail)	(Whole)	0.004	0.016	0.003	0.750000
(Riesling)	(Wine)	0.004	0.066	0.004	1.000000

(Tenderloin)	(Veal)	0.006	0.013	0.003	0.500000
(Cane, Beets)	(Organic)	0.003	0.011	0.003	1.000000
(Organic, Beets)	(Cane)	0.003	0.003	0.003	1.000000
(Cane, Organic)	(Beets)	0.003	0.005	0.003	1.000000
(Beets)	(Cane, Organic)	0.005	0.003	0.003	0.600000
(Cane)	(Organic, Beets)	0.003	0.003	0.003	1.000000
(Veal, Organic)	(Tenderloin)	0.003	0.006	0.003	1.000000
(Veal, Tenderloin)	(Organic)	0.003	0.011	0.003	1.000000
(Tenderloin, Organic)	(Veal)	0.003	0.013	0.003	1.000000
(Tenderloin)	(Veal, Organic)	0.006	0.003	0.003	0.500000

اگر به میزان اعتماد (confidence) توجه کنیم میبینیم که با این حال که Support محصولات به نسبط عرف آنچه که باید در مسائل مشابه باشد کمتر است اما با درجه اطمینان بسیار خوبی میتوان گفت قوانین بسیار قابل اعتماد هستند.

نتیجه گیری

مهمترین نتیجه ای که از این تمرین حاصل میشود این است که برای آنکه قوانین استخراج شده نه تنها قابل اعتماد بلکه قابل استفاده باشند (عموما برای پیشنهادات خرید های اینترنتی از این نوع آنالیز داده ای استفاده میشود) باید در داده های خود سبد های خرید نسبتا بزرگی داشته باشیم در غیر این صورت درجه اعتماد بسیار بالا و Support پایین خواهد بود (البته در مورد این گونه داده ها میتوان از Support نسبی استفاده کرد)

تمیز سازی داده ها بسیار پر اهمیت بوده و احتمال رخداد اشتباه در ثبت داده ها توسط انسان بسیار رایج است.

همچنین گویا نوشیدنی های الکلی بسیار در فروشگاه های عمومی پر طرفدار هستند :D