

## Úloha 8 – Analýza nákupního koše

### Zadání – analýza nákupního koše:

Cílem této úlohy je zjistit, které produkty zákazníci kupují společně a z kterých produktů lze vytvořit asociační pravidla. Jedná se o aplikaci úlohy "Vyhledávání častých množin".

### Dataset

- V matlabu načtete soubor `marketBasket.mat` obsahující transakční databázi supermarketu vhodnou k analýze nákupního košíku.

```
load('marketBasket.mat')
```

- V daném souboru se nachází proměnné `"tranDb"` a `"info"`.
- Proměnná `"tranDb"` je transakční databáze v maticové booleanovské formě. Každý řádek je transakce – jeden nákupní košík jednoho zákazníka.
- Každý sloupec je jedna možná položka v košíku (item). Slovní popis těchto položek je v proměnné `"info"`.

### Apriori

- Stáhněte si `aprioriFPM.m`, tato funkce vygeneruje množinu častých položek pro danou transakční databázi algoritmem apriori. Způsob použití funkce `aprioriFPM.m` zjistíte příkazem `"help aprioriFPM"`.
- Najděte množinu častých položek pro vámi zvolenou minimální relativní podporu. Tu odůvodněte.
- Nalezené množiny vypište do souboru pomocí funkce `printFreqSets.m`
- Mezivýsledek: Pro minimální relativní podporu rovnou 0.02 vám má vyjít 1302 častých množin.

### Srovnání položek

- Spusťte `aprioriFPM` pro množinu položek lexikálně srovnanou sestupně podle častosti výskytu a srovnanou vzestupně.
- Porovnejte časy obou běhů.

### Asociační pravidla

- Použijte funkci `associationRules.m` pro vygenerování asociačních pravidel. Způsob použití funkce `aprioriFPM.m` zjistíte příkazem `"help associationRules"`.
- Vygenerujte všechna asociační pravidla pro vámi zvolenou minimální relativní spolehlivost. Tu odůvodněte.
- Mezivýsledek: Pro minimální relativní spolehlivost rovnou 0.7 (a dříve danou minimální relativní podporu 0.02) vám vyjde 480 asociačních pravidel.
- Nalezená pravidla vypište do souboru pomocí funkce `printRules.m`

### Váš protokol by měl obsahovat:

- Časté množiny položek pro vámi vybranou minimální relativní podporu, minimální relativní podporu, kterou jste použili, její zdůvodnění

Metoda k určení minimální relativní podpory:

```
function [ minimum, avg, maximum ] = najdiRelativPodporu( tranDb )
radky = size(tranDb, 1); % zjistim pocet radku
sloupce = size(tranDb, 2); % zjistim pocet sloupce

podpora = zeros(1, sloupce);
for i = 1:sloupce
    podpora(1, i) = mean(tranDb(:, i));
end

minimum = min(podpora);
avg = mean(podpora);
maximum = max(podpora);
```

end

Po spuštění metody jsem získala tyto údaje:

- minimum = 0.0051
- avg = 0.0314
- maximum = 0.1228

Lze předpokládat, že položky mezi minimem a průměrem nebudou velmi časté. Za minimální relativní podporu tak vezmu hodnotu libovolnou nad průměrem – např. 0.033.

- Časy běhu funkce `aprioriFPM` pro obě lexikální řazení, váš komentář k časům

#### ***Funkce pro vzestupné setřídění včetně času:***

```
>>tic; cetnost=aprioriFPM(mujSort(tranDb,0), tranDb,0.033); printFreqSets(cetnost,info,'set1.txt'); toc;
```

Elapsed time is 0.720417 seconds.

Do souboru set1.txt bylo vypsáno celkem 188 položek a s ohledem na jeho rozsáhlost zde uvádím jen prvních 10 položek:

```
(Honey Roasted Peanuts), Confidence = 0.03
(Green Pepper), Confidence = 0.03
(Chardonnay Wine), Confidence = 0.03
(Corn Oil), Confidence = 0.03
(Sliced Chicken), Confidence = 0.03
(Paper Cups), Confidence = 0.03
(Apple Jelly), Confidence = 0.03
(Beef Jerky), Confidence = 0.03
(Liquid Laundry Detergent), Confidence = 0.03
(Hot Chocolate), Confidence = 0.03
```

#### ***Funkce pro sestupné setřídění včetně času:***

```
>>tic; cetnost=aprioriFPM(mujSort(tranDb,1), tranDb,0.033); printFreqSets(cetnost,info,'set2.txt'); toc;
```

Elapsed time is 5.626555 seconds.

Do souboru set2.txt bylo opět vypsáno opět všech 188 položek. Jako příklad uvádím opět jen prvních 10 položek:

```
(Eggs), Confidence = 0.12
(White Bread), Confidence = 0.12
(2pct. Milk), Confidence = 0.11
(Potato Chips), Confidence = 0.10
(98pct. Fat Free Hamburger), Confidence = 0.09
(Hot Dogs), Confidence = 0.09
(Potatoes), Confidence = 0.09
(Sweet Relish), Confidence = 0.09
(Onions), Confidence = 0.08
(Toothpaste), Confidence = 0.08
```

- Asociační pravidla pro vámi vybranou minimální relativní spolehlivost, minimální relativní spolehlivost, kterou jste použili, její zdůvodnění

Pro mé testování jsem si zvolila hodnotu minimální relativní spolehlivosti 0.6. Tu jsem dosadila do následující funkce:

```
printRules(associationRules(cetnost,0.6,tranDb),info,'res.txt');
```

Do souboru res.txt bylo vypsáno následujících 26 položek, tj. asociačních pravidel, která vyhovují zadané podmínce “spolehlivost alespoň 60%”:

```
(Raisins)---> (Eggs), Support = 0.04, Confidence = 0.60
(Sugar Cookies)---> (Eggs), Support = 0.04, Confidence = 0.67
(Salt)---> (Eggs), Support = 0.03, Confidence = 0.64
(Toothpaste)---> (White Bread), Support = 0.05, Confidence = 0.60
(Aspirin)---> (White Bread), Support = 0.04, Confidence = 0.63
(Tomatoes)---> (White Bread), Support = 0.04, Confidence = 0.61
(Sugar Cookies)---> (White Bread), Support = 0.04, Confidence = 0.64
(Canned Tuna)---> (White Bread), Support = 0.03, Confidence = 0.61
(Hamburger Buns)---> (98pct. Fat Free Hamburger),
                        Support = 0.05, Confidence = 0.68
(Hot Dog Buns)---> (Hot Dogs), Support = 0.04, Confidence = 0.71
(Hot Dog Buns)---> (Sweet Relish), Support = 0.04, Confidence = 0.61
(White Bread and 2pct. Milk)---> (Eggs), Support = 0.04, Confidence = 0.70
(Eggs and 2pct. Milk)---> (White Bread), Support = 0.04, Confidence = 0.69
(Eggs and White Bread)---> (2pct. Milk), Support = 0.04, Confidence = 0.65
(White Bread and Potato Chips)---> (Eggs), Support = 0.03, Confidence = 0.66
(Eggs and Potato Chips)---> (White Bread), Support = 0.03, Confidence = 0.70
(Eggs and White Bread)---> (Potato Chips), Support = 0.03, Confidence = 0.61
(2pct. Milk and Potato Chips)---> (Eggs), Support = 0.03, Confidence = 0.75
(Eggs and Potato Chips)---> (2pct. Milk), Support = 0.03, Confidence = 0.70
(Eggs and 2pct. Milk)---> (Potato Chips), Support = 0.03, Confidence = 0.65
(2pct. Milk and Potato Chips)---> (White Bread),
                        Support = 0.03, Confidence = 0.74
(White Bread and Potato Chips)---> (2pct. Milk),
                        Support = 0.03, Confidence = 0.64
(White Bread and 2pct. Milk)---> (Potato Chips),
                        Support = 0.03, Confidence = 0.64
(2pct. Milk and Toothpaste)---> (White Bread),
                        Support = 0.03, Confidence = 0.76
(White Bread and Toothpaste)---> (2pct. Milk),
                        Support = 0.03, Confidence = 0.69
(White Bread and 2pct. Milk)---> (Toothpaste),
                        Support = 0.03, Confidence = 0.64
```

Legenda:

Např. u pravidla č. 1 zjistíme, že množina Raisins + Eggs má spolehlivost 4%, tzn. že přibližně ve 4% transakcí se objeví tyto dva výrobky společně a tak má pravidlo (Raisins)---> (Eggs) má spolehlivost 60%.

Jinými slovy to znamená, že pokud se v nákupním košíku objeví položka Raisins, se spolehlivostí 60% se ve stejném košíku objeví i položka Eggs.