МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И.УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЁТ

по лабораторной работе №4
по дисциплине «Машинное обучение»
Тема: Ассоциативный анализ

Студент гр. 6304 _____ Корытов П.В.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

1. Цель работы

Ознакомится с методами ассоциативного анализа из библиотеки `MLxtend`.

2. Выполнение

2.1. Загрузка данных

1. Произведена загрузка данных. В наборе данных 9835 транзакций и 169 товаров.
2. Часть списка товаров с их количества приведена на листинге 1.

Листинг 1. Список товаров и количество

```
                       item  number
0               whole milk    2513
1         other vegetables    1903
2               rolls/buns    1809
3                     soda    1715
4                   yogurt    1372
..                     ...     ...
164                   bags       4
165        kitchen utensil       4
166  preservation products       2
167              baby food       1
168   sound storage medium       1

[169 rows x 2 columns]
```

2.2. FPGrowth и FPMax

1. Произведено one-hot кодирование набора данных. Результат на листинге 2.

Листинг 2. Закодированный набор данных

```
      Instant food products  UHT-milk  ...  yogurt  zwieback
0                     False     False  ...   False     False
1                     False     False  ...    True     False
2                     False     False  ...   False     False
3                     False     False  ...    True     False
4                     False     False  ...   False     False
...                     ...       ...  ...     ...       ...
9830                  False     False  ...   False     False
9831                  False     False  ...   False     False
9832                  False     False  ...    True     False
9833                  False     False  ...   False     False
9834                  False     False  ...   False     False

[9835 rows x 169 columns]
```

2. Произведен ассоциативный анализ с использованием алгоритмов FPGrowth и FPMax при уровне поддержки 0.3. Часть результатов на листингах 3 и 4.

Листинг 3. Результаты FPGrowth

```
     support                    itemsets
0   0.255516                (whole milk)
1   0.193493          (other vegetables)
2   0.183935                (rolls/buns)
3   0.174377                      (soda)
4   0.139502                    (yogurt)
..       ...                         ...
58  0.031012                    (onions)
59  0.030605       (rolls/buns, sausage)
60  0.030503  (whole milk, citrus fruit)
61  0.030402       (specialty chocolate)
62  0.030097     (whole milk, pip fruit)

[63 rows x 2 columns]
```

Листинг 4. Результаты FPMax

```
     support                        itemsets
0   0.098526                 (shopping bags)
1   0.080529                  (bottled beer)
2   0.079817                    (newspapers)
3   0.077682                   (canned beer)
4   0.074835  (whole milk, other vegetables)
..       ...                             ...
45  0.031012                        (onions)
46  0.030605           (rolls/buns, sausage)
47  0.030503      (whole milk, citrus fruit)
48  0.030402           (specialty chocolate)
49  0.030097         (whole milk, pip fruit)

[50 rows x 2 columns]
```

3. Определено минимальные и максимальные значения для уровня поддержки для наборов и 1 и 2 объектов для обоих алгоритмов. Результаты на таблицах 1 и 2.

Таблица 1. Значения уровня поддержки для FPGrowth

	1	2
min	0.0304016	0.0300966
max	0.255516	0.0748348

Таблица 2. Значения уровня поддержки для FPMax

	1	2
min	0.0304016	0.0300966
max	0.0985257	0.0748348

Как можно заметить, две таблицы отличаются в ячейке (max, 1) — максимальный уровень поддержки для набора длиной 1. Это связано с тем, что в FPMax набор не может быть частью другого набора большей длины. В данном случае наиболее часто встречающиеся элементы наборов длины 1 вошли в наборы длины 2.

4. Построена гистограмма для топ-10 наиболее часто встречаемых товаров. Результат на рис. 1.

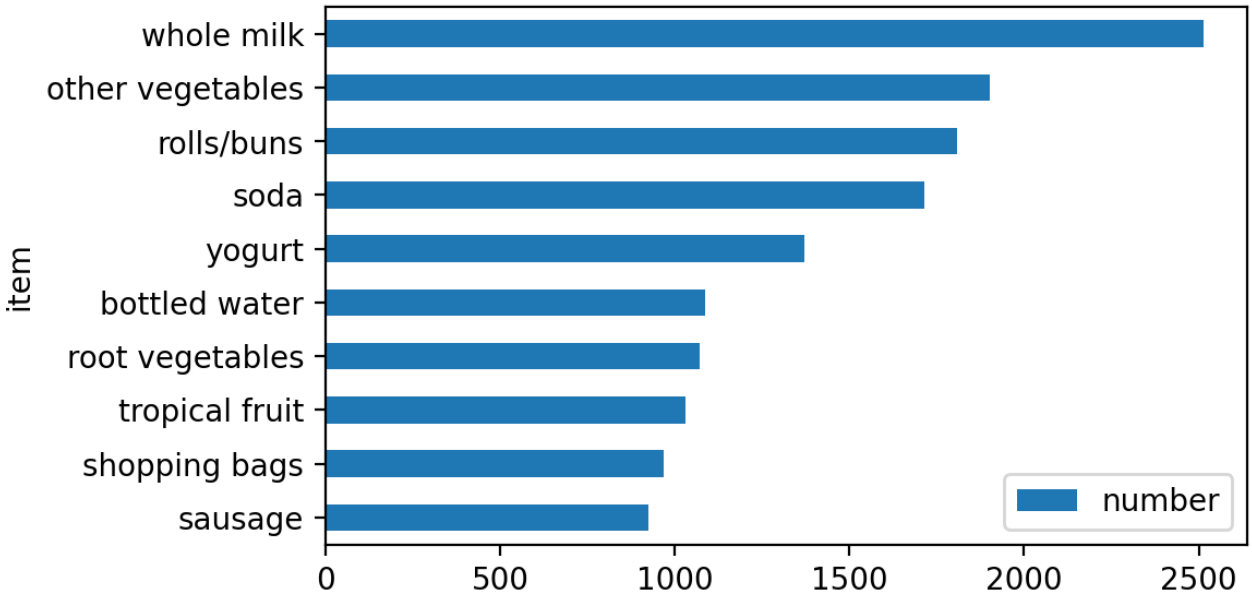

Рисунок 1 – Гистограмма для наиболее часто встречаемых товаров

Товары на гистограмме отвечают наборам длины 1; аналогичный порядок наблюдается на листинге 3 с результатам FPGrowth.

5. Произведено ограничение количества товаров до 16. Часть измененного набора данных представлена на листинге 5.

Листинг 5. Ограниченный набор данных

```
      bottled beer  bottled water  ...  whole milk  yogurt
0            False          False  ...       False   False
1            False          False  ...       False    True
2            False          False  ...        True   False
3            False          False  ...       False    True
4            False          False  ...        True   False
...            ...            ...  ...         ...     ...
9830         False          False  ...        True   False
9831         False          False  ...       False   False
9832         False          False  ...       False    True
9833          True           True  ...       False   False
9834         False          False  ...       False   False

[9835 rows x 16 columns]
```

6. Произведен анализ FPGrowth и FPMax для нового набора данных. Часть результатов представлена на листингах 6 и 7.

Листинг 6. FPGrowth для ограниченного набора данных

```
     support                         itemsets
0   0.255516                     (whole milk)
1   0.193493                 (other vegetables)
2   0.183935                      (rolls/buns)
3   0.174377                            (soda)
4   0.139502                          (yogurt)
..       ...                              ...
29  0.033249               (whole milk, pastry)
30  0.032740            (other vegetables, soda)
31  0.032232  (whipped/sour cream, whole milk)
32  0.030605               (rolls/buns, sausage)
33  0.030503         (whole milk, citrus fruit)

[34 rows x 2 columns]
```

Листинг 7. FPMax для ограниченного набора данных

```
     support                         itemsets
0   0.098526                   (shopping bags)
1   0.080529                    (bottled beer)
2   0.077682                     (canned beer)
3   0.074835    (whole milk, other vegetables)
4   0.057651                           (pork)
..       ...                              ...
17  0.033249               (whole milk, pastry)
18  0.032740            (other vegetables, soda)
19  0.032232  (whipped/sour cream, whole milk)
20  0.030605               (rolls/buns, sausage)
21  0.030503         (whole milk, citrus fruit)

[22 rows x 2 columns]
```

В сравнении с результатами на листингах 3 и 4, количество найденных наборов уменьшилось, но для найденных наборов уровень поддержки не изменился.

7. Построен график изменения количества получаемых правил от уровня поддержки для FPGrowth и FPMax. Результаты на рис. 2 и 3.

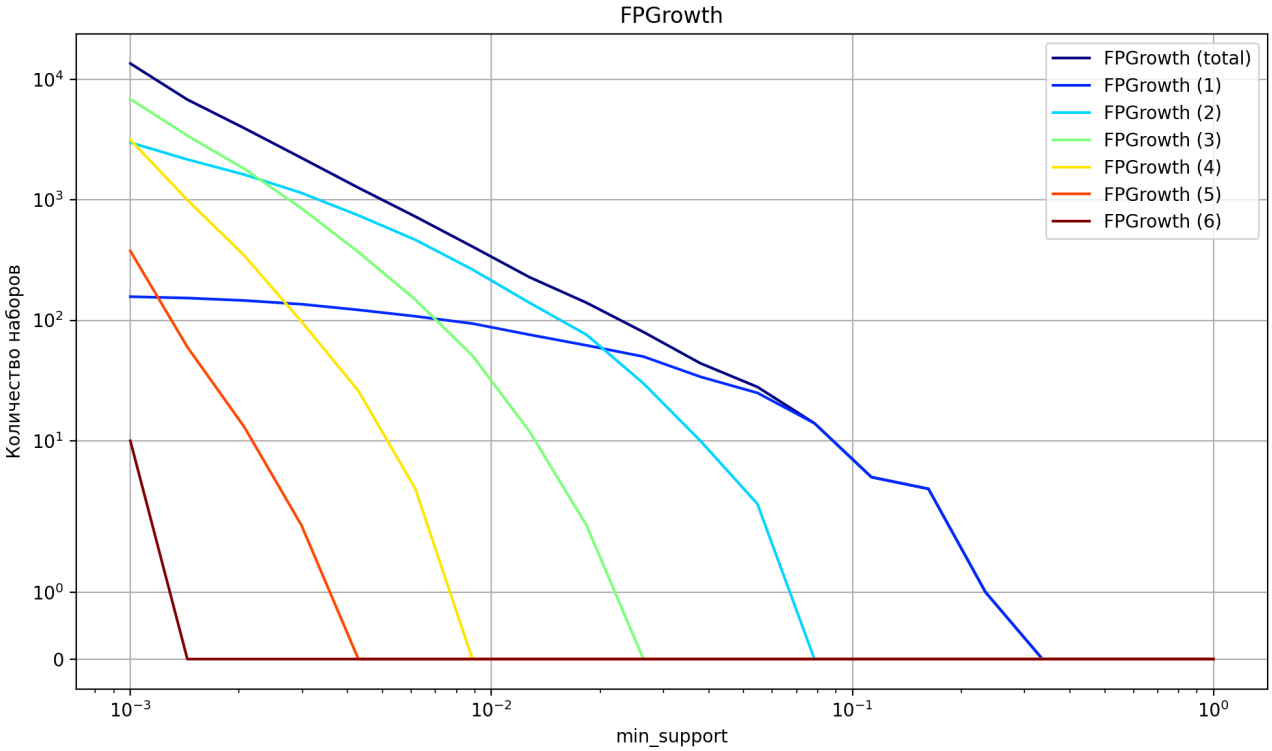

Рисунок 2 – Изменение количества получаемых правил от уровня поддержки для FPGrowth

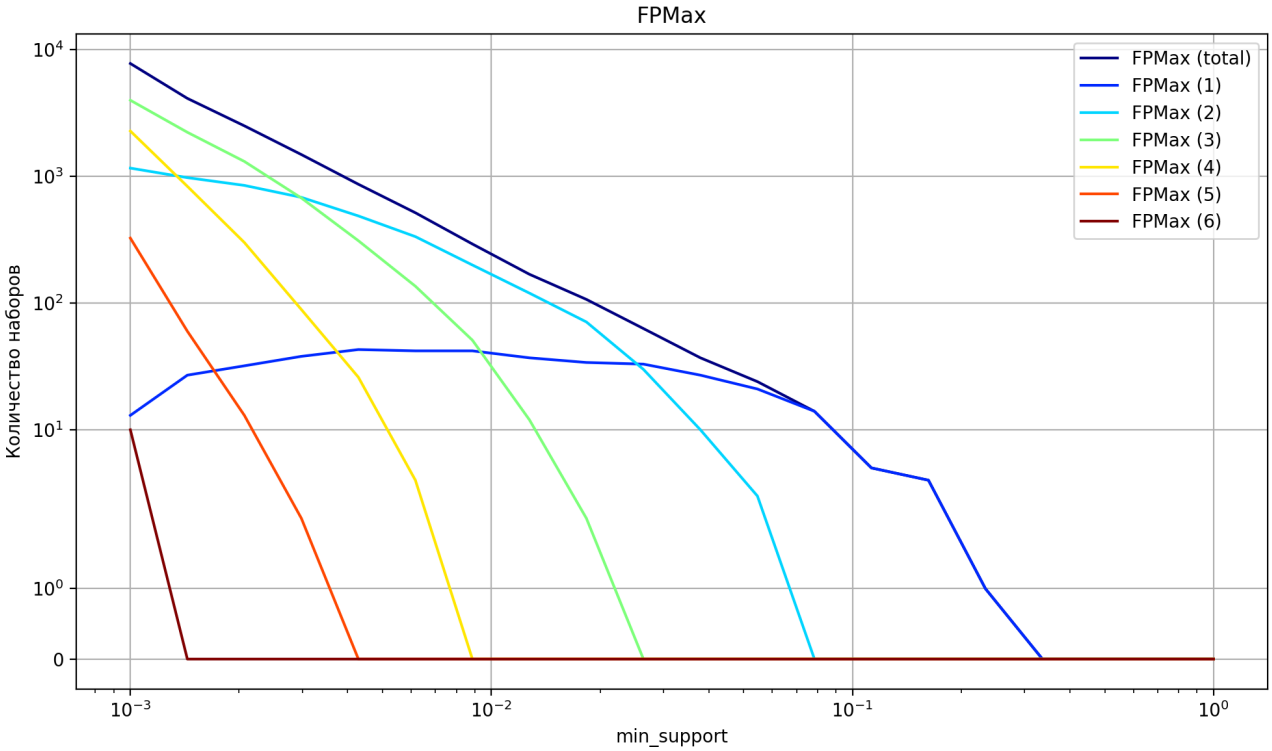

Рисунок 3 – Изменение количества получаемых правил от уровня поддержки для FPMax

Как видно, для алгоритма FPGrowth количество любых наборов уменьшается с уменьшением уровня поддержки. Более длинные наборы исчезают раньше.

Для алгоритма FPMax количество наборов длины 1 с увеличением уровня поддержки увеличивается примерно до уровня 10^{-2}, после чего опять снижается. Это связано с тем, что при небольших уровнях поддержки элементы с большей вероятностью присутствуют в наборах большей длины, что для FPMax исключает их присутствие в наборах меньшей длины.

2.3. Ассоциативные правила

1. Набор данных сформирован так, чтобы размер транзации был $>= 2$.
2. Получены частоты наборов с использованием FPGrowth. Часть результатов на листинге 8.

Листинг 8. Результаты FPGrowth для нового набора данных

```
     support                                itemsets
0   0.241240                                (yogurt)
1   0.185864                          (tropical fruit)
2   0.421869                              (whole milk)
3   0.335079                        (other vegetables)
4   0.296214                              (rolls/buns)
..       ...                                     ...
37  0.059203                    (whole milk, sausage)
38  0.053363              (sausage, other vegetables)
39  0.063834         (whipped/sour cream, whole milk)
40  0.057189   (whipped/sour cream, other vegetables)
41  0.065848                     (whole milk, pastry)

[42 rows x 2 columns]
```

3. Проведен ассоциативный анализ. Результаты в таблице 3.

   Пусть A — антецент, C — консеквент. Значения столбцов следующие:

   - `antecedent support`, `consequent support` — поддержка антецедента (support(A)) и консеквента (support(C)).
   - `support` — поддержка набора из антецента и консеквента

$$\text{support}(A \to C) = \text{support}(A \cup C) \tag{2.1}$$

Таблица 3. Результаты ассоциативного анализа

	antecedents	consequents	ant. support	cons. support	support	confidence	lift	leverage	conviction
0	[’yogurt’]	[’whole milk’]	0.24	0.42	0.11	0.46	1.09	0.01	1.07
1	[’yogurt’]	[’other vegetables’]	0.24	0.34	0.09	0.36	1.06	0.01	1.03
2	[’tropical fruit’]	[’yogurt’]	0.19	0.24	0.06	0.31	1.29	0.01	1.10
3	[’tropical fruit’]	[’other vegetables’]	0.19	0.34	0.07	0.38	1.14	0.01	1.08
4	[’tropical fruit’]	[’whole milk’]	0.19	0.42	0.08	0.45	1.07	0.01	1.05
5	[’whole milk’]	[’other vegetables’]	0.42	0.34	0.15	0.35	1.05	0.01	1.03
6	[’other vegetables’]	[’whole milk’]	0.34	0.42	0.15	0.44	1.05	0.01	1.04
7	[’rolls/buns’]	[’whole milk’]	0.30	0.42	0.11	0.38	0.90	-0.01	0.93
8	[’bottled water’]	[’whole milk’]	0.19	0.42	0.07	0.37	0.87	-0.01	0.91
9	[’bottled water’]	[’soda’]	0.19	0.27	0.06	0.31	1.16	0.01	1.06
10	[’citrus fruit’]	[’whole milk’]	0.15	0.42	0.06	0.41	0.98	-0.00	0.98
11	[’citrus fruit’]	[’other vegetables’]	0.15	0.34	0.06	0.39	1.17	0.01	1.09
12	[’root vegetables’]	[’other vegetables’]	0.20	0.34	0.09	0.48	1.43	0.03	1.27
13	[’root vegetables’]	[’whole milk’]	0.20	0.42	0.10	0.49	1.17	0.01	1.14
14	[’sausage’]	[’rolls/buns’]	0.17	0.30	0.06	0.36	1.22	0.01	1.10
15	[’sausage’]	[’whole milk’]	0.17	0.42	0.06	0.35	0.84	-0.01	0.89
16	[’sausage’]	[’other vegetables’]	0.17	0.34	0.05	0.32	0.95	-0.00	0.98
17	[’whipped/sour cream’]	[’whole milk’]	0.12	0.42	0.06	0.51	1.22	0.01	1.19
18	[’whipped/sour cream’]	[’other vegetables’]	0.12	0.34	0.06	0.46	1.37	0.02	1.23
19	[’pastry’]	[’whole milk’]	0.15	0.42	0.07	0.44	1.04	0.00	1.03

- `confidence` — вероятность увидеть консеквент в транзакции, содержащей антецедент.

$$\text{confidence}(A \to C) = \frac{\text{support}(A \to C)}{\text{support}(A)} \tag{2.2}$$

- `lift` — насколько чаще $A \to C$ встречаются вместе, чем если бы они были статистически независимы. Для независимых A, C lift будет равен 1.

$$\text{lift}(A \to C) = \frac{\text{support}(A \to C)}{\text{support}(A)\text{support}(C)} \tag{2.3}$$

- `leverage` — разница между частотой появления $A \to C$ вместе и частотой, ожидаемой при статистической независимости A, C. Для независимых A, C leverage будет равен 0.

$$\text{leverage}(A \to C) = \text{support}(A \to C) - \text{support}(A)\text{support}(C) \tag{2.4}$$

- `conviction` — зависимость консеквента от антецедента.

$$\text{conviction}(A \to C) = \frac{1 - \text{support}(A \to C)}{1 - \text{confidence}(A \to C)} \tag{2.5}$$

  В случае абсолютной зависимости $1 - \text{confidence}(A \to C) = 0$ и $\text{conviction}(A \to C) = \infty$. Для независимых A, C conviction равен 1.

4. Проведено построение ассоциативных правил для каждой из метрик. Среднее значение, медиана и СКО для каждой метрики представлены в таблице 4.

Таблица 4. Статистический анализ метрик

	min_threshold	mean	median	mse
support	0.08	0.102019	0.0953484	0.0210117
confidence	0.35	0.417044	0.412655	0.053484
lift	1.16	1.26685	1.22134	0.0974603
leverage	0.01	0.0155331	0.0135937	0.00606334
conviction	1.05	1.11696	1.09696	0.068747

5. Построен граф для заданного анализа. Результат на рис. 4.

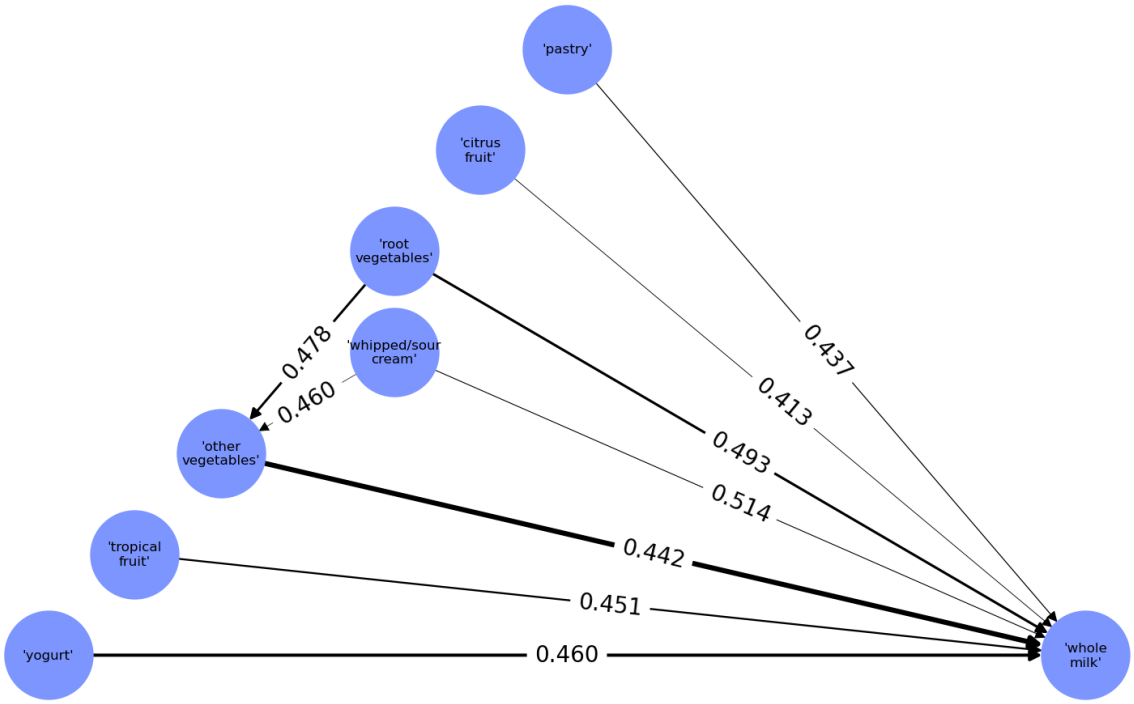

Рисунок 4 – Граф

Из графа можно сделать выводы вроде: если в транзакции есть предметы вроде yougurt, tropical fruits и т.п., то с высокой вероятностью также окажется whole milk.

6. Произведена визуализация тех же результатов с помощью матрицы. Результат на рис. 5

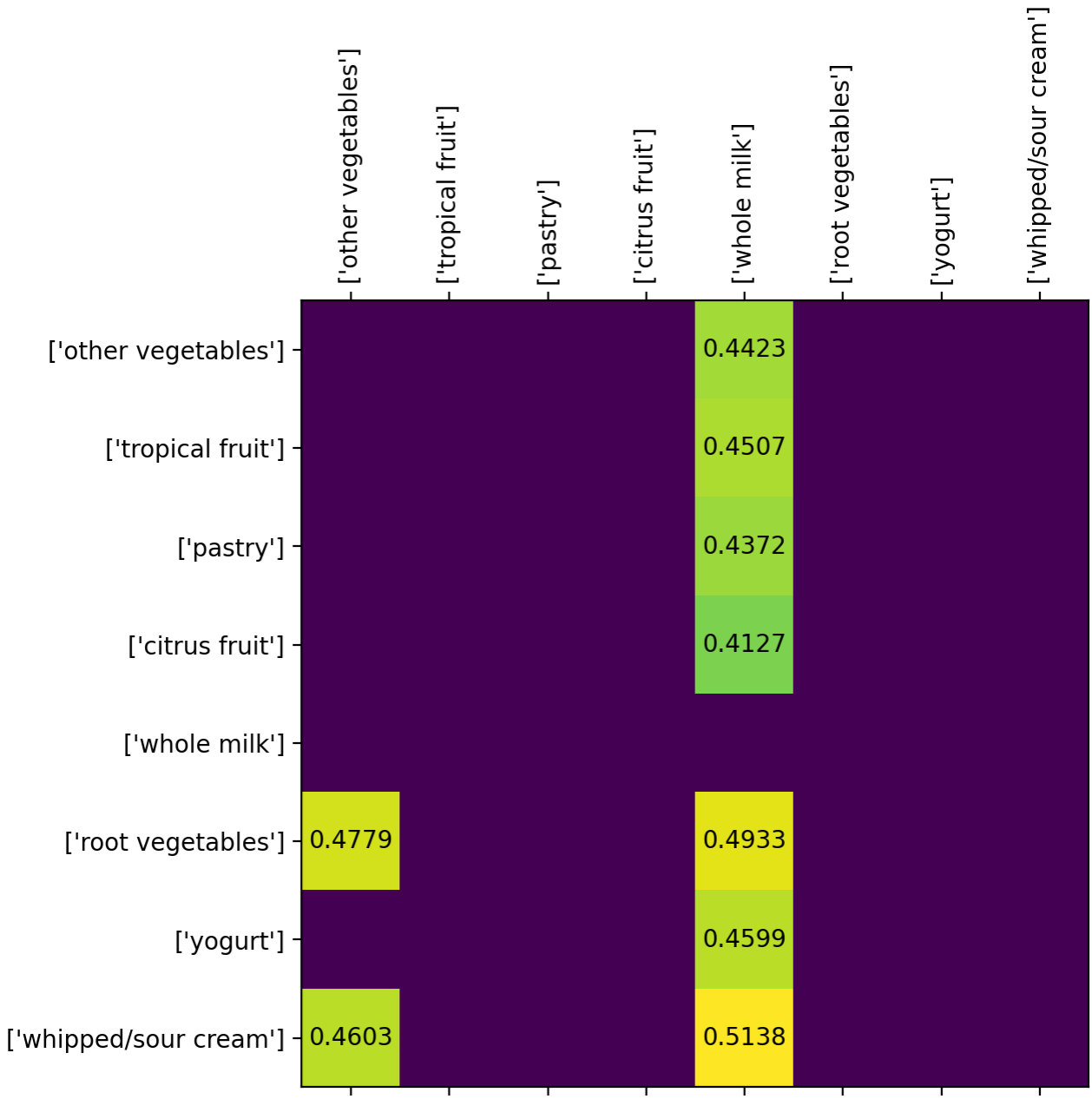

Рисунок 5 – Матрица

3. Выводы

Произведено знакомство с методами ассоциативного анализа из библиотеки `MLxtend`.

Установлено, что применение алгоритма FPGrowth к рассматриваемому набору данных с уровнем поддержки 10^{-4} требует примерно 12 ГБ оперативной памяти.

Отличие результатов между FPGrowth и FPMax объясняется тем, что в FPMax набор не может быть частью другого набора большей длины. Этим же вызван рост числа наборов длины 1 для FPMax при небольших (до 10^{-2}) уровнях поддержки, тогда как в FPGrowth наблюдается только снижение.

Ограничение товаров в наборе данных изменило количество наборов в результатах алгоритмов, но не изменило уровень поддержки для оставшихся наборов.