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

ОТЧЕТ

по лабораторной работе №3

по дисциплине «Частотный анализ»

Тема: Машинное обучение

Студент гр. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Загрузка данных.

Загрузим данные из csv таблицы, и посмотрим сколько в наборе уникальных покупателей и видов товаров. Результат представлен на рис.1.

```
Кол-во id 1139
Кол-во товаров 38
```

Рисунок 1 – Уникальные покупатели и товары

Подготовка данных.

Так как полученные датасет не пригоден для анализа напрямую, так как каждый список пользователя может содержать разное количество товаров. Поэтому данные надо закодировать так, чтобы их можно было представить в виде матрицы. Для кодированния данных используем TransactionEncoder. После преобразования мы получаем таблицу в которой построчно отмечены списки транзакций каждого id в виде булевых значений, если значение True – транзакция была совершена, если False то не была. Результат представлен на рис. 2.

	all- purpose	aluminum foil	bagels	...	vegetables	waffles	yogurt
0	True	True	False	...	True	False	True
1	False	True	False	...	True	True	True
2	False	False	True	...	True	False	False
3	True	False	False	...	False	False	False
4	True	False	False	...	True	True	True
...	...	...	...	...	...	...	...
1134	True	False	False	...	False	False	False
1135	False	False	False	...	True	False	False
1136	False	False	True	...	True	False	True
1137	True	False	False	...	True	True	True
1138	False	False	False	...	True	False	False

Рисунок 2 – Таблица транзакций

Ассоциативный анализ.

С помощью алгоритма Apriori с уровнем поддержки транзакции 0.3 выведем все транзакции чье появление происходит с частотой 0.3 или более, а значит примерно в каждой третьей операции или чаще. Результат представлен на рис.3

```
34  0.369622                                (tortillas)       1
35  0.739245                               (vegetables)       1
36  0.394205                                  (waffles)       1
37  0.384548                                   (yogurt)       1
38  0.310799                 (vegetables, aluminum foil)       2
39  0.300263                        (bagels, vegetables)       2
40  0.310799                      (vegetables, cereals)       2
41  0.309043                      (vegetables, cheeses)       2
42  0.308165                  (vegetables, dinner rolls)       2
43  0.306409  (vegetables, dishwashing liquid/detergent)       2
44  0.326602                        (vegetables, eggs)       2
45  0.302897                   (vegetables, ice cream)       2
46  0.309043           (vegetables, laundry detergent)       2
47  0.311677                  (vegetables, lunch meat)       2
48  0.331870                     (vegetables, poultry)       2
49  0.305531                        (vegetables, soda)       2
50  0.315189                     (vegetables, waffles)       2
51  0.319579                      (vegetables, yogurt)       2

Process finished with exit code 0
```

Рисунок 3 – Часть таблицы транзакций после применения алгоритма apriori с уровнем поддержки 0.3

Применим тот же алгоритм но ограничим длину набора единицей. Результат на рис. 4.

```
24  0.355575                      (pork)       1
25  0.421422                   (poultry)       1
26  0.367867                (sandwich bags)       1
27  0.349429             (sandwich loaves)       1
28  0.368745                   (shampoo)       1
29  0.379280                      (soap)       1
30  0.390694                      (soda)       1
31  0.373134             (spaghetti sauce)       1
32  0.360843                     (sugar)       1
33  0.378402                (toilet paper)       1
34  0.369622                   (tortillas)       1
35  0.739245                  (vegetables)       1
36  0.394205                   (waffles)       1
37  0.384548                    (yogurt)       1

Process finished with exit code 0
```

Рисунок 4 – Ограничение длины набора единицей

Применим тот же алгоритм но выведем только наборы длины 2. Также выведем количество таких наборов. Результат на рис. 5.

```
     support                                  itemsets  length
38  0.310799              (vegetables, aluminum foil)       2
39  0.300263                    (vegetables, bagels)       2
40  0.310799                   (cereals, vegetables)       2
41  0.309043                   (cheeses, vegetables)       2
42  0.308165              (vegetables, dinner rolls)       2
43  0.306409  (dishwashing liquid/detergent, vegetables)       2
44  0.326602                     (eggs, vegetables)       2
45  0.302897                (vegetables, ice cream)       2
46  0.309043          (laundry detergent, vegetables)       2
47  0.311677                (lunch meat, vegetables)       2
48  0.331870                   (poultry, vegetables)       2
49  0.305531                      (soda, vegetables)       2
50  0.315189                   (vegetables, waffles)       2
51  0.319579                    (yogurt, vegetables)       2

Count of result itemstes =  14
```

Рисунок 5 – Вывод наборов длины 2

Посчитаем количество наборов при различных уровнях поддержки. Начальное значение поддержки 0.05, шаг 0.01. Построем график зависимости

количества наборов от уровня поддержки. Также определим значение уровня поддержки при котором происходит переход к наборам меньшей размерности и отметим их на графике вертикальными линиями. Результат на рис. 6.

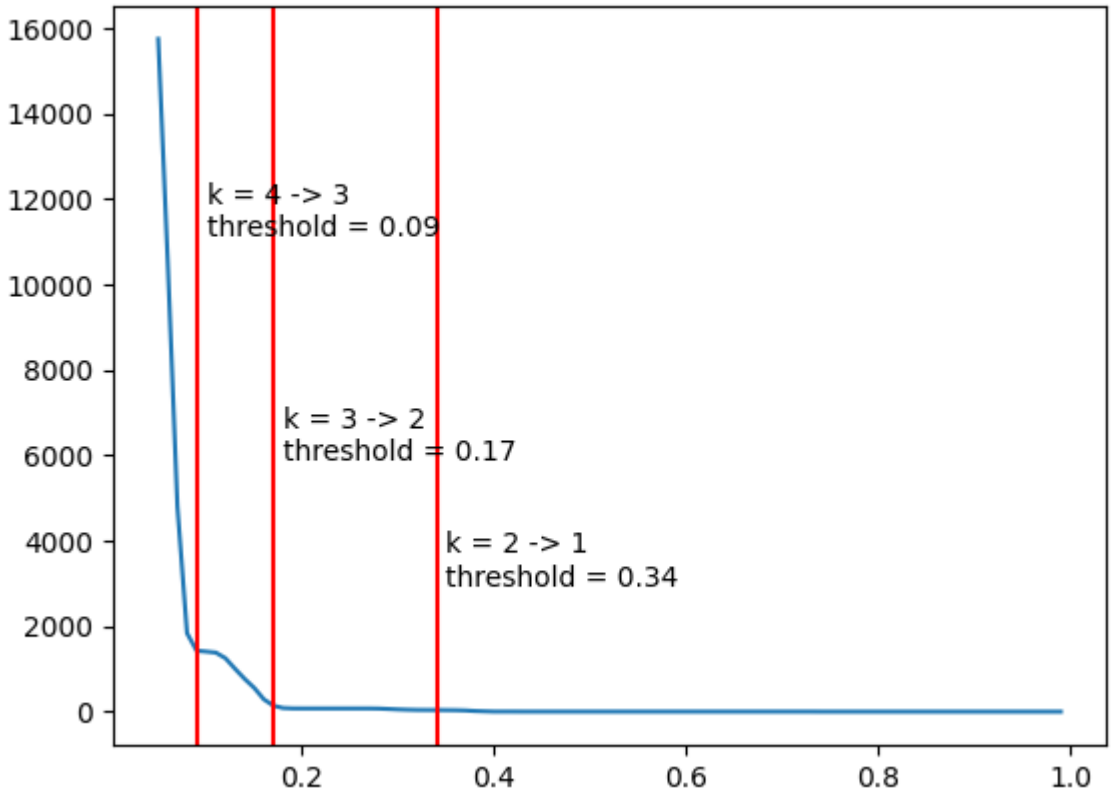

Рисунок 7 – График зависимости количества наборов от значения уровня поддержки

Построим датасет только из тех элементов, которые попадают в наборы размером 1 при уровне поддержки 0.38. Далее приведем полученный датасет к формату, который можно обработать и проведем ассоциативный анализ при уровне поддержки 0.3 для нового датасета. Результаты на рис. 8.

```
     support                                  itemsets  length
0   0.384548                          (aluminum foil)       1
1   0.385426                                 (bagels)       1
2   0.395961                                (cereals)       1
3   0.390694                                (cheeses)       1
4   0.388938                           (dinner rolls)       1
5   0.388060          (dishwashing liquid/detergent)       1
6   0.389816                                  (eggs)       1
7   0.398595                             (ice cream)       1
8   0.395083                           (lunch meat)       1
9   0.380158                                  (milk)       1
10  0.421422                               (poultry)       1
11  0.390694                                  (soda)       1
12  0.739245                             (vegetables)       1
13  0.394205                               (waffles)       1
14  0.384548                                (yogurt)       1
15  0.310799              (vegetables, aluminum foil)       2
16  0.300263                    (vegetables, bagels)       2
17  0.310799                   (cereals, vegetables)       2
18  0.309043                   (vegetables, cheeses)       2
19  0.308165                (vegetables, dinner rolls)       2
20  0.306409  (vegetables, dishwashing liquid/detergent)       2
21  0.326602                      (vegetables, eggs)       2
22  0.302897                 (vegetables, ice cream)       2
23  0.311677                (lunch meat, vegetables)       2
24  0.331870                   (vegetables, poultry)       2
25  0.305531                      (vegetables, soda)       2
26  0.315189                   (vegetables, waffles)       2
27  0.319579                    (vegetables, yogurt)       2

Process finished with exit code 0
```

Рисунок 8 – Анализ нового датасета

Различия с анализом старого датасета состоят в том, что в данном случае анализировались только транзакции, частота которых была равна из превышала значение 0.38 старого датасета, поэтому все наборы состоят только из них. Сходство же в том, что идентичные наборы имеют идентичную частоту в обоих датасетах.

Проведем ассоциативный анализ при уровне поддержки 0.15 для нового датасета. Выведем все наборы размер которых больше 1 и в котором есть 'yogurt' или 'waffles'. Результат на рис. 9.

```
      support                                itemsets
27   0.169447                (aluminum foil, waffles)
28   0.177349                 (aluminum foil, yogurt)
40   0.159789                       (bagels, waffles)
41   0.162423                        (bagels, yogurt)
52   0.160667                     (cereals, waffles)
53   0.172081                      (cereals, yogurt)
63   0.172959                     (cheeses, waffles)
64   0.172081                      (cheeses, yogurt)
73   0.169447                  (dinner rolls, waffles)
74   0.166813                   (dinner rolls, yogurt)
82   0.175593  (dishwashing liquid/detergent, waffles)
83   0.158033   (yogurt, dishwashing liquid/detergent)
90   0.169447                        (eggs, waffles)
91   0.174715                         (eggs, yogurt)
97   0.172959                   (ice cream, waffles)
98   0.156277                    (ice cream, yogurt)
103  0.184372                  (lunch meat, waffles)
104  0.161545                   (lunch meat, yogurt)
108  0.167691                         (milk, yogurt)
111  0.166813                    (poultry, waffles)
112  0.180860                     (poultry, yogurt)
114  0.177349                       (soda, waffles)
115  0.167691                        (yogurt, soda)
116  0.315189                  (vegetables, waffles)
117  0.319579                   (vegetables, yogurt)
118  0.173837                      (yogurt, waffles)
119  0.152766      (vegetables, aluminum foil, yogurt)
128  0.157155                (eggs, vegetables, yogurt)
130  0.157155        (lunch meat, vegetables, waffles)
131  0.152766             (poultry, vegetables, yogurt)

Process finished with exit code 0
```

Рисунок 9 – Вывод с ограничениями по именам и размеру набора

Построем датасет, из тех элементов, которые не попали в новый датасет и приведем его к удобному для анализа виду. Проведем анализ apriori для полученного датасета. Напишем правило, для вывода всех наборов, в которых хотя бы два элемента начинаются на 's'. Затем напишем правило, для вывода всех наборов, для которых уровень поддержки изменяется от 0.1 до 0.25. Результаты на рис. 10 и 11.

```
      support                                          itemsets
248   0.137840                 (sandwich bags, sandwich loaves)
249   0.146620                         (sandwich bags, shampoo)
250   0.158911                            (sandwich bags, soap)
251   0.147498                 (sandwich bags, spaghetti sauce)
252   0.131694                           (sandwich bags, sugar)
...        ...                                             ...
4574  0.030729      (toilet paper, sandwich loaves, sugar, soap)
4575  0.030729  (sandwich loaves, toilet paper, soap, tortillas)
4576  0.032485      (toilet paper, spaghetti sauce, sugar, soap)
4577  0.032485  (spaghetti sauce, toilet paper, soap, tortillas)
4578  0.031607            (toilet paper, sugar, soap, tortillas)

[998 rows x 2 columns]

Process finished with exit code 0
```

Рисунок 10 – Вывод наборов в котором хотя бы два элемента начинаются с ‘s’

```
      support                          itemsets
23    0.144864             (beef, all- purpose)
24    0.147498           (butter, all- purpose)
25    0.146620       (coffee/tea, all- purpose)
26    0.142230            (flour, all- purpose)
27    0.150132           (fruits, all- purpose)
..         ...                              ...
271   0.151888  (spaghetti sauce, toilet paper)
272   0.148376     (tortillas, spaghetti sauce)
273   0.151888            (sugar, toilet paper)
274   0.147498               (tortillas, sugar)
275   0.156277        (tortillas, toilet paper)

[253 rows x 2 columns]

Process finished with exit code 0
```

Рисунок 11 – Вывод с уровнем поддержки от 0.1 до 0.25

Выводы

В ходе выполнения данной лабораторной работы было произведено знакомство с частотным анализом. Были произведены трансформации транзакций с помощью TransactionEncoder. Было проведено исследование алгоритма apriori библиотеки MLxtend на тестовых данных.