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

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Машинное обучение»

Тема: Ассоциативный анализ

Студент гр. 1310 _____ Комаров Д. Е.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2025

Постановка задачи

Цель работы: ознакомление с методами ассоциативного анализа из библиотеки MLxtend.

Выполнение лабораторной работы

Загрузка данных

Для выполнения лабораторной работы используем набор данных «Groceries Market Basket Dataset». Данный набор данных содержит данные о купленных вместе товарах. Переформатируем данные в список списков вместе купленных товаров, удалив все значения NaN. Построим список уникальных товаров. Полученный список представлен на рисунке 1 и содержит 169 элементов.

Рисунок 1 – Уникальные товары

FPGrowth и FPMax

Преобразуем данные в формат, пригодный для дальнейшего анализа. Данный формат представляет собой матрицу, строками которой являются покупателями, а столбцы – товарами. Если покупатель купил товар, то на пересечении строки покупателя и столбца товара будет 1, иначе 0. Фрагмент данных после преобразования представлен на рисунке 2.

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

Рисунок 2 – Данные для дальнейшего анализа

Проведем ассоциативный анализ, используя алгоритм FPGrowth при уровне поддержки 0.03. Результат представлен на рисунке 3.

	support	itemsets	length
0	0.082766	(citrus fruit)	1
1	0.058566	(margarine)	1
2	0.139502	(yogurt)	1
3	0.104931	(tropical fruit)	1
4	0.058058	(coffee)	1
..	...	...	...
58	0.033249	(pastry, whole milk)	2
59	0.047382	(other vegetables, root vegetables)	2
60	0.048907	(whole milk, root vegetables)	2
61	0.030605	(sausage, rolls/buns)	2
62	0.032232	(whipped/sour cream, whole milk)	2

Рисунок 3 – Результат ассоциативного анализа алгоритмом FPGrowth при уровне поддержки 0.03

В результате было получено 63 набора. Минимальный уровень поддержки для наборов длинны 1 был равен 0.03, максимальный 0.26. Для наборов длинны 2 минимальный уровень поддержки был равен 0.03, максимальный 0.07.

Проведем аналогичный анализ алгоритмом FPMax. Данный алгоритм отличается от FPGrowth тем, что находит только максимальные частые наборы элементов (т.е. наборы, которые не имеют надмножества с поддержкой, большей чем минимальный уровень поддержки). Результат представлен на рисунке 4.

	support	itemsets	length
0	0.030402	(specialty chocolate)	1
1	0.031012	(onions)	1
2	0.032944	(hygiene articles)	1
3	0.033249	(berries)	1
4	0.033249	(hamburger meat)	1
5	0.033452	(UHT-milk)	1
6	0.033859	(sugar)	1
7	0.037112	(dessert)	1
8	0.037417	(long life bakery product)	1
9	0.037824	(salty snack)	1
10	0.038434	(waffles)	1
11	0.039654	(cream cheese)	1
12	0.042095	(white bread)	1
13	0.042908	(chicken)	1
14	0.048094	(frozen vegetables)	1
15	0.049619	(chocolate)	1
16	0.052364	(napkins)	1
17	0.052466	(beef)	1
18	0.053279	(curd)	1
19	0.055414	(butter)	1
20	0.057651	(pork)	1
21	0.058058	(coffee)	1
22	0.058566	(margarine)	1
23	0.058973	(frankfurter)	1
24	0.063447	(domestic eggs)	1
25	0.064870	(brown bread)	1
26	0.032232	(whipped/sour cream, whole milk)	2
27	0.072293	(fruit/vegetable juice)	1
28	0.030097	(whole milk, pip fruit)	2
29	0.077682	(canned beer)	1
30	0.079817	(newspapers)	1
31	0.080529	(bottled beer)	1
32	0.030503	(whole milk, citrus fruit)	2
33	0.033249	(whole milk, pastry)	2
34	0.030605	(sausage, rolls/buns)	2
35	0.098526	(shopping bags)	1
36	0.035892	(other vegetables, tropical fruit)	2
37	0.042298	(whole milk, tropical fruit)	2
38	0.047382	(other vegetables, root vegetables)	2
39	0.048907	(whole milk, root vegetables)	2
40	0.034367	(whole milk, bottled water)	2
41	0.034367	(rolls/buns, yogurt)	2
42	0.043416	(other vegetables, yogurt)	2
43	0.056024	(whole milk, yogurt)	2
44	0.032740	(soda, other vegetables)	2
45	0.038332	(soda, rolls/buns)	2
46	0.040061	(soda, whole milk)	2
47	0.042603	(rolls/buns, other vegetables)	2
48	0.056634	(rolls/buns, whole milk)	2
49	0.074835	(whole milk, other vegetables)	2

Рисунок 4 – Результат ассоциативного анализа алгоритмом FPMax при уровне поддержки 0.03

В результате было получено 49 наборов. Минимальный уровень поддержки для наборов длинны 1 был равен 0.03, максимальный 0.09. Для наборов длинны 2 минимальный уровень поддержки был равен 0.03, максимальный 0.07.

Построим гистограмму для 10 наиболее часто встречающихся товаров. Полученная гистограмма представлена на рисунке 5. Каждый столбец является товаром, высота столбца – частота товара.

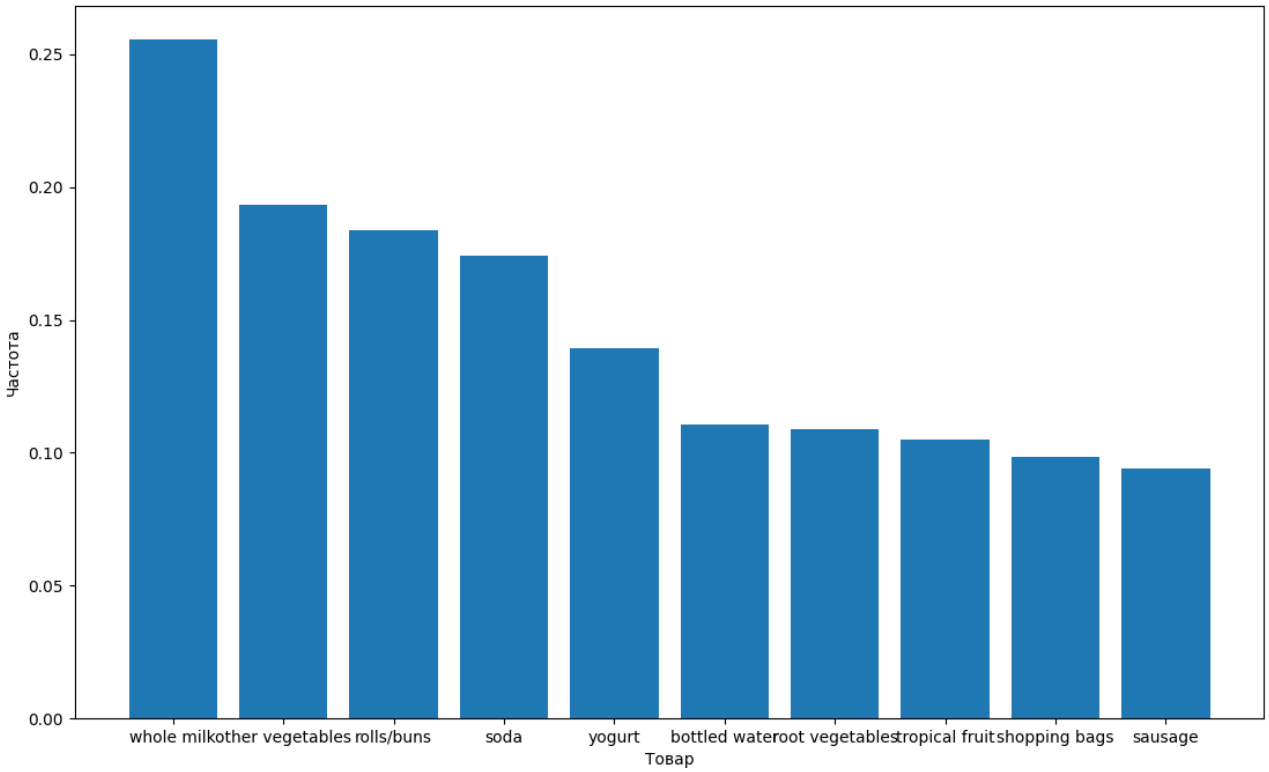

Рисунок 5 – Гистограмма 10 наиболее часто встречающихся товаров

Как можно заметить из диаграммы, изображенной на рисунке 5, наиболее часто встречающиеся товары генерируют наборы с наибольшим уровнем поддержки.

Преобразуем исходный набор данных так, чтобы он содержал ограниченный набор товаров, приведем его к виду, пригодному для дальнейшего анализа и аналогично проанализируем алгоритмами FPGrowth и FPMax.

Проведем ассоциативный анализ, используя алгоритм FPGrowth при уровне поддержки 0.03. Результат представлен на рисунке 6.

	support	itemsets	length
0	0.082766	(citrus fruit)	1
1	0.139502	(yogurt)	1
2	0.104931	(tropical fruit)	1
3	0.255516	(whole milk)	1
4	0.193493	(other vegetables)	1
5	0.183935	(rolls/buns)	1
6	0.080529	(bottled beer)	1
7	0.110524	(bottled water)	1
8	0.174377	(soda)	1
9	0.088968	(pastry)	1
10	0.077682	(canned beer)	1
11	0.093950	(sausage)	1
12	0.098526	(shopping bags)	1
13	0.071683	(whipped/sour cream)	1
14	0.057651	(pork)	1
15	0.030503	(whole milk, citrus fruit)	2
16	0.056024	(yogurt, whole milk)	2
17	0.034367	(yogurt, rolls/buns)	2
18	0.043416	(other vegetables, yogurt)	2
19	0.035892	(other vegetables, tropical fruit)	2
20	0.042298	(tropical fruit, whole milk)	2
21	0.074835	(other vegetables, whole milk)	2
22	0.042603	(other vegetables, rolls/buns)	2
23	0.056634	(whole milk, rolls/buns)	2
24	0.034367	(bottled water, whole milk)	2
25	0.038332	(soda, rolls/buns)	2
26	0.040061	(whole milk, soda)	2
27	0.032740	(other vegetables, soda)	2
28	0.033249	(pastry, whole milk)	2
29	0.030605	(sausage, rolls/buns)	2
30	0.032232	(whipped/sour cream, whole milk)	2

Рисунок 6 – Результат ассоциативного анализа алгоритмом FPGrowth для ограниченного набора товаров при уровне поддержки 0.03

В результате было получено 30 наборов. Минимальный уровень поддержки для наборов длинны 1 был равен 0.05, максимальный 0.26. Для наборов длинны 2 минимальный уровень поддержки был равен 0.03, максимальный 0.07.

Проведем аналогичный анализ алгоритмом FPMax. Результат представлен на рисунке 7.

	support	itemsets	length
0	0.057651	(pork)	1
1	0.032232	(whipped/sour cream, whole milk)	2
2	0.077682	(canned beer)	1
3	0.080529	(bottled beer)	1
4	0.030503	(whole milk, citrus fruit)	2
5	0.033249	(pastry, whole milk)	2
6	0.030605	(sausage, rolls/buns)	2
7	0.098526	(shopping bags)	1
8	0.035892	(other vegetables, tropical fruit)	2
9	0.042298	(tropical fruit, whole milk)	2
10	0.034367	(bottled water, whole milk)	2
10	0.034367	(bottled water, whole milk)	2
11	0.034367	(yogurt, rolls/buns)	2
12	0.043416	(other vegetables, yogurt)	2
13	0.056024	(yogurt, whole milk)	2
14	0.032740	(other vegetables, soda)	2
15	0.038332	(soda, rolls/buns)	2
16	0.040061	(whole milk, soda)	2
17	0.042603	(other vegetables, rolls/buns)	2
18	0.056634	(whole milk, rolls/buns)	2
19	0.074835	(other vegetables, whole milk)	2

Рисунок 7 – Результат ассоциативного анализа алгоритмом FPMax для ограниченного набора товаров при уровне поддержки 0.03

В результате было получено 19 наборов. Минимальный уровень поддержки для наборов длинны 1 был равен 0.05, максимальный 0.09. Для наборов длинны 2 минимальный уровень поддержки был равен 0.03, максимальный 0.07.

Как можно заметить после того, как набор товаров был ограничен, количество наборов, полученных в результате работы обоих алгоритмов, также сократилось, однако не сильно, поскольку убранные товары имели малый уровень поддержки, оставшиеся же товары имели достаточно большую поддержку.

Вернемся к исходному набору товаров. Построим график зависимости количества получаемых наборов от минимального уровня поддержки для алгоритма FPGrowth. Для наглядности график был разделен на 2 части. Кривые соответствуют количеству наборов различной длинны. На рисунке 8 представлен данный график на промежутке [0.003;0.05]

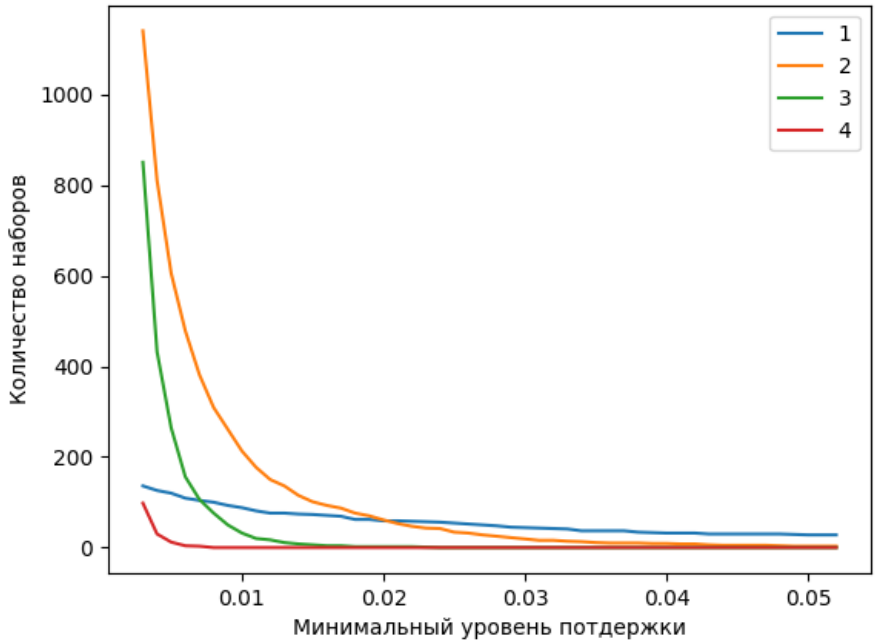

Рисунок 8 – График зависимости количества получаемых наборов от минимального уровня поддержки для алгоритма FPGrowth на промежутке [0.003;0.05]

На рисунке 9 представлен График зависимости график зависимости количества получаемых наборов от минимального уровня поддержки для алгоритма FPGrowth на промежутке [0.05;0.26].

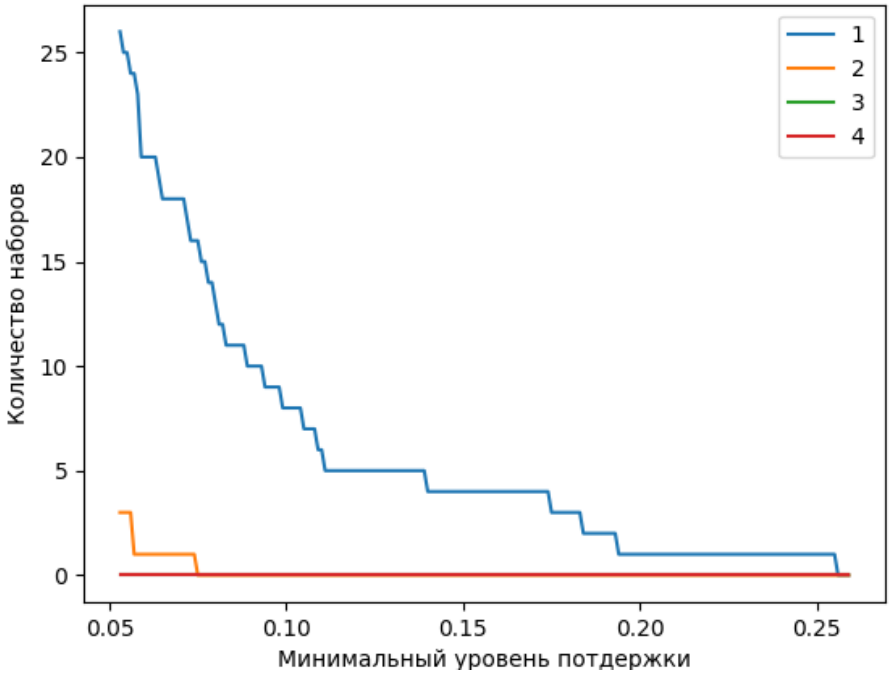

Рисунок 9 – График зависимости количества получаемых наборов от минимального уровня поддержки для алгоритма FPGrowth на промежутке [0.05;0.26]

Из графиков на рисунках 8 и 9 можно сделать вывод, что наборы длинной 2 перестают появляться при минимальном уровне поддержки 0.07, наборы длинны 3 – при минимальном уровне поддержки 0.01, наборы длинны 4 – при минимальном уровне поддержки 0.005. Наборы большей длинны получены не были.

Ассоциативные правила

Сформируем набор данных из определенных товаров и так, чтобы размер транзакции был 2 и более, приведем их к виду, пригодному для дальнейшего анализа и получим частоты наборов, используя алгоритм FPGrowth при минимальном уровне поддержки 0.05. Результат представлен на рисунке 10.

	support	itemsets
0	0.247758	(yogurt)
1	0.191241	(tropical fruit)
2	0.427737	(whole milk)
3	0.339103	(other vegetables)
4	0.305318	(rolls/buns)
5	0.116580	(bottled beer)
6	0.191032	(bottled water)
7	0.149739	(citrus fruit)
8	0.274870	(soda)
9	0.085089	(canned beer)
10	0.172263	(sausage)
11	0.171220	(shopping bags)
12	0.126590	(whipped/sour cream)
13	0.101773	(pork)
14	0.154745	(pastry)
15	0.114911	(yogurt, whole milk)
16	0.056100	(yogurt, soda)
17	0.070490	(yogurt, rolls/buns)
18	0.089051	(other vegetables, yogurt)
19	0.060063	(tropical fruit, yogurt)
20	0.073618	(tropical fruit, other vegetables)
21	0.086757	(tropical fruit, whole milk)
22	0.050469	(tropical fruit, rolls/buns)
23	0.153493	(other vegetables, whole milk)
24	0.087383	(other vegetables, rolls/buns)
25	0.116163	(whole milk, rolls/buns)
26	0.050886	(bottled water, other vegetables)
27	0.070490	(bottled water, whole milk)
28	0.059437	(bottled water, soda)
29	0.062565	(whole milk, citrus fruit)
30	0.059228	(other vegetables, citrus fruit)
31	0.078624	(soda, rolls/buns)
32	0.082169	(whole milk, soda)
33	0.067153	(other vegetables, soda)
34	0.062774	(sausage, rolls/buns)
35	0.061314	(sausage, whole milk)
36	0.055266	(other vegetables, sausage)
37	0.050469	(shopping bags, soda)
38	0.050261	(whole milk, shopping bags)
39	0.066111	(whipped/sour cream, whole milk)
40	0.059228	(whipped/sour cream, other vegetables)
41	0.068196	(whole milk, pastry)

Рисунок 10 – Наборы для ассоциативного анализа

Проведем ассоциативный анализ полученных наборов. Поскольку результат имеет много столбцов, вывод был поделен ни 2 рисунка. Результат представлен на рисунке 11.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(yogurt)	(whole milk)	0.247758	0.427737	0.114911	0.463805	1.084322
1	(yogurt)	(other vegetables)	0.247758	0.339103	0.089051	0.359428	1.059936
2	(tropical fruit)	(yogurt)	0.191241	0.247758	0.060063	0.314068	1.267638
3	(tropical fruit)	(other vegetables)	0.191241	0.339103	0.073618	0.384951	1.135203
4	(tropical fruit)	(whole milk)	0.191241	0.427737	0.086757	0.453653	1.060589
5	(whole milk)	(other vegetables)	0.427737	0.339103	0.153493	0.358849	1.058230
6	(other vegetables)	(whole milk)	0.339103	0.427737	0.153493	0.452645	1.058230
7	(rolls/buns)	(whole milk)	0.305318	0.427737	0.116163	0.380464	0.889482
8	(bottled water)	(whole milk)	0.191032	0.427737	0.070490	0.368996	0.862669
9	(bottled water)	(soda)	0.191032	0.274870	0.059437	0.311135	1.131938
10	(citrus fruit)	(whole milk)	0.149739	0.427737	0.062565	0.417827	0.976832
11	(citrus fruit)	(other vegetables)	0.149739	0.339103	0.059228	0.395543	1.166439
12	(sausage)	(rolls/buns)	0.172263	0.305318	0.062774	0.364407	1.193532
13	(sausage)	(whole milk)	0.172263	0.427737	0.061314	0.355932	0.832128
14	(sausage)	(other vegetables)	0.172263	0.339103	0.055266	0.320823	0.946093
15	(whipped/sour cream)	(whole milk)	0.126590	0.427737	0.066111	0.522241	1.220938
16	(whipped/sour cream)	(other vegetables)	0.126590	0.339103	0.059228	0.467875	1.379741
17	(pastry)	(whole milk)	0.154745	0.427737	0.068196	0.440701	1.030307

Рисунок 11 – Результат ассоциативного анализа

Продолжение результата ассоциативного анализа представлено на рисунке 12.

	antecedents	consequents	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
0	(yogurt)	(whole milk)	1.0	0.008936	1.067266	0.103377	0.204985	0.063026	0.366227
1	(yogurt)	(other vegetables)	1.0	0.005036	1.031728	0.075171	0.178886	0.030753	0.311018
2	(tropical fruit)	(yogurt)	1.0	0.012681	1.096671	0.261056	0.158503	0.088149	0.278246
3	(tropical fruit)	(other vegetables)	1.0	0.008768	1.074543	0.147263	0.161187	0.069372	0.301024
4	(tropical fruit)	(whole milk)	1.0	0.004956	1.047435	0.070636	0.163009	0.045287	0.328241
5	(whole milk)	(other vegetables)	1.0	0.008446	1.030798	0.096155	0.250255	0.029878	0.405747
6	(other vegetables)	(whole milk)	1.0	0.008446	1.045505	0.083260	0.250255	0.043524	0.405747
7	(rolls/buns)	(whole milk)	1.0	-0.014433	0.923696	-0.151722	0.188303	-0.082607	0.326020
8	(bottled water)	(whole milk)	1.0	-0.011222	0.906908	-0.164428	0.128566	-0.102648	0.266897
9	(bottled water)	(soda)	1.0	0.006928	1.052646	0.144084	0.146229	0.050013	0.263686
10	(citrus fruit)	(whole milk)	1.0	-0.001484	0.982978	-0.027138	0.121507	-0.017317	0.282049
11	(citrus fruit)	(other vegetables)	1.0	0.008451	1.093373	0.167819	0.137864	0.085399	0.285102
12	(sausage)	(rolls/buns)	1.0	0.010179	1.092966	0.195896	0.151332	0.085059	0.285004
13	(sausage)	(whole milk)	1.0	-0.012369	0.888513	-0.195962	0.113821	-0.125476	0.249638
14	(sausage)	(other vegetables)	1.0	-0.003149	0.973085	-0.064403	0.121171	-0.027659	0.241900
15	(whipped/sour cream)	(whole milk)	1.0	0.011963	1.197805	0.207185	0.135412	0.165140	0.338400
16	(whipped/sour cream)	(other vegetables)	1.0	0.016301	1.241995	0.315117	0.145716	0.194844	0.321268
17	(pastry)	(whole milk)	1.0	0.002006	1.023178	0.034801	0.132603	0.022653	0.300068

Рисунок 12 – Результат ассоциативного анализа (продолжение)

Рассмотрим результат подробнее. В результате ассоциативного анализа было получено 17 правил. Колонка *antecedents* отвечает за предпосылки для ассоциативного правила, колонка *consequents* – за следствия из правила. Колонки *antecedent support* и *consequents support* отвечают за уровень поддержки набора-предпосылки и набора-следствия соответственно.

Остальные колонки называются метриками. Метрика *support* отвечает за уровень поддержки для данного правила и вычисляется по формуле

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

Метрика *confidence* вычисляется по формуле

$$confidence(A \to C) = \frac{\text{support}(A \to C)}{\text{supp}ort\ (A)}.$$

Метрика *lift* вычисляется по формуле

$$lift(A \to C) = \frac{\text{confidence}(A \to C)}{\text{supp}ort\ (C)}.$$

Метрика *leverage* вычисляется по формуле

$$leverage(A \to C) = \text{support}(A \to C) - \text{support}(A) * \text{support}(C).$$

Метрика *convinction* вычисляется по формуле

$$convinction(A \to C) = \frac{1 - support(C)}{1 - confidence(A \to C)}.$$

Метрика *zhangs_metric* вычисляется по формуле

$$zhangs\_metric(A \to C) = \frac{confidence(A \to C) - confidence(A' \to C)}{Max[confidence(A \to C), confidence(A' \to C)]}.$$

Метрика *jacard* вычисляется по формуле

$$jacard(A \to C) = \frac{\text{support}(A \to C)}{\text{supp}ort(A) + \text{supp}ort(C) - \text{support}(A \to C)}.$$

Метрика *certainty* вычисляется по формуле

$$certainty\ (A \to C) =$$

$$= \begin{cases} \frac{confidence(A \to C) - support(C)}{1 - support(C)}, confidence(A \to C) > support(C) \\ \frac{confidence(A \to C) - support(C)}{support(C)}, confidence(A \to C) \geq support(C) \end{cases}.$$

Метрика *kulczynski* вычисляется по формуле

$$kulczynski\ (A \to C) = \frac{1}{2}\left(\frac{\text{support}(A \to C)}{\text{supp}ort(A)} + \frac{\text{support}(A \to C)}{\text{supp}ort(C)}\right).$$

Правила, получаемые в результате ассоциативного анализа, отсеиваются по значению какой-либо из метрик. Минимальное значение метрики, которое должно иметь правило, получаемое в результате ассоциативного анализа, задается параметром *min_threshold*. По умолчанию используется метрика

confidence. Протестируем ассоциативный анализ с использованием других метрик. Значение *min_threshold* будем выбирать так, чтобы выводилось не менее 10 правил. На рисунке 13 представлен результат ассоциативного анализа для метрики *support* и значения *min_threshold*=0.08.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(whole milk)	(yogurt)	0.427737	0.247758	0.114911	0.268649	1.084322
1	(yogurt)	(whole milk)	0.247758	0.427737	0.114911	0.463805	1.084322
2	(yogurt)	(other vegetables)	0.247758	0.339103	0.089051	0.359428	1.059936
3	(other vegetables)	(yogurt)	0.339103	0.247758	0.089051	0.262608	1.059936
4	(whole milk)	(tropical fruit)	0.427737	0.191241	0.086757	0.202828	1.060589
5	(tropical fruit)	(whole milk)	0.191241	0.427737	0.086757	0.453653	1.060589
6	(whole milk)	(other vegetables)	0.427737	0.339103	0.153493	0.358849	1.058230
7	(other vegetables)	(whole milk)	0.339103	0.427737	0.153493	0.452645	1.058230
8	(rolls/buns)	(other vegetables)	0.305318	0.339103	0.087383	0.286202	0.843997
9	(other vegetables)	(rolls/buns)	0.339103	0.305318	0.087383	0.257688	0.843997
10	(whole milk)	(rolls/buns)	0.427737	0.305318	0.116163	0.271575	0.889482
11	(rolls/buns)	(whole milk)	0.305318	0.427737	0.116163	0.380464	0.889482
12	(whole milk)	(soda)	0.427737	0.274870	0.082169	0.192101	0.698882
13	(soda)	(whole milk)	0.274870	0.427737	0.082169	0.298938	0.698882

Рисунок 13 – Результат ассоциативного анализа для метрики *support* и значения *min_threshold*=0.08

На рисунке 14 представлен результат ассоциативного анализа для метрики *lift* и значения *min_threshold*=1.1.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(tropical fruit)	(yogurt)	0.191241	0.247758	0.060063	0.314068	1.267638
1	(yogurt)	(tropical fruit)	0.247758	0.191241	0.060063	0.242424	1.267638
2	(tropical fruit)	(other vegetables)	0.191241	0.339103	0.073618	0.384951	1.135203
3	(other vegetables)	(tropical fruit)	0.339103	0.191241	0.073618	0.217097	1.135203
4	(soda)	(bottled water)	0.274870	0.191032	0.059437	0.216237	1.131938
5	(bottled water)	(soda)	0.191032	0.274870	0.059437	0.311135	1.131938
6	(citrus fruit)	(other vegetables)	0.149739	0.339103	0.059228	0.395543	1.166439
7	(other vegetables)	(citrus fruit)	0.339103	0.149739	0.059228	0.174662	1.166439
8	(sausage)	(rolls/buns)	0.172263	0.305318	0.062774	0.364407	1.193532
9	(rolls/buns)	(sausage)	0.305318	0.172263	0.062774	0.205601	1.193532
10	(whipped/sour cream)	(whole milk)	0.126590	0.427737	0.066111	0.522241	1.220938
11	(whole milk)	(whipped/sour cream)	0.427737	0.126590	0.066111	0.154559	1.220938
12	(whipped/sour cream)	(other vegetables)	0.126590	0.339103	0.059228	0.467875	1.379741
13	(other vegetables)	(whipped/sour cream)	0.339103	0.126590	0.059228	0.174662	1.379741

Рисунок 14 – Результат ассоциативного анализа для метрики *lift* и значения *min_threshold*=1.1

На рисунке 15 представлен результат ассоциативного анализа для метрики *leverage* и значения *min_threshold*=0.008.

	antecedents	consequents	antecedent support	consequent support	support	leverage	conviction
0	(whole milk)	(yogurt)	0.427737	0.247758	0.114911	0.008936	1.028565
1	(yogurt)	(whole milk)	0.247758	0.427737	0.114911	0.008936	1.067266
2	(tropical fruit)	(yogurt)	0.191241	0.247758	0.060063	0.012681	1.096671
3	(yogurt)	(tropical fruit)	0.247758	0.191241	0.060063	0.012681	1.067562
4	(tropical fruit)	(other vegetables)	0.191241	0.339103	0.073618	0.008768	1.074543
5	(other vegetables)	(tropical fruit)	0.339103	0.191241	0.073618	0.008768	1.033026
6	(whole milk)	(other vegetables)	0.427737	0.339103	0.153493	0.008446	1.030798
7	(other vegetables)	(whole milk)	0.339103	0.427737	0.153493	0.008446	1.045505
8	(citrus fruit)	(other vegetables)	0.149739	0.339103	0.059228	0.008451	1.093373
9	(other vegetables)	(citrus fruit)	0.339103	0.149739	0.059228	0.008451	1.030197
10	(sausage)	(rolls/buns)	0.172263	0.305318	0.062774	0.010179	1.092966
11	(rolls/buns)	(sausage)	0.305318	0.172263	0.062774	0.010179	1.041967
12	(whole milk)	(whipped/sour cream)	0.427737	0.126590	0.066111	0.011963	1.033082
13	(whipped/sour cream)	(whole milk)	0.126590	0.427737	0.066111	0.011963	1.197805
14	(whipped/sour cream)	(other vegetables)	0.126590	0.339103	0.059228	0.016301	1.241995
15	(other vegetables)	(whipped/sour cream)	0.339103	0.126590	0.059228	0.016301	1.058245

Рисунок 15 – Результат ассоциативного анализа для метрики *leverage* и значения *min_threshold*=0.008

На рисунке 16 представлен результат ассоциативного анализа для метрики *convinction* и значения *min_threshold*=1.04.

	antecedents	consequents	antecedent support	consequent support	support	leverage	conviction
0	(yogurt)	(whole milk)	0.247758	0.427737	0.114911	0.008936	1.067266
1	(yogurt)	(tropical fruit)	0.247758	0.191241	0.060063	0.012681	1.067562
2	(tropical fruit)	(yogurt)	0.191241	0.247758	0.060063	0.012681	1.096671
3	(tropical fruit)	(other vegetables)	0.191241	0.339103	0.073618	0.008768	1.074543
4	(tropical fruit)	(whole milk)	0.191241	0.427737	0.086757	0.004956	1.047435
5	(other vegetables)	(whole milk)	0.339103	0.427737	0.153493	0.008446	1.045505
6	(bottled water)	(soda)	0.191032	0.274870	0.059437	0.006928	1.052646
7	(citrus fruit)	(other vegetables)	0.149739	0.339103	0.059228	0.008451	1.093373
8	(sausage)	(rolls/buns)	0.172263	0.305318	0.062774	0.010179	1.092966
9	(rolls/buns)	(sausage)	0.305318	0.172263	0.062774	0.010179	1.041967
10	(whipped/sour cream)	(whole milk)	0.126590	0.427737	0.066111	0.011963	1.197805
11	(whipped/sour cream)	(other vegetables)	0.126590	0.339103	0.059228	0.016301	1.241995
12	(other vegetables)	(whipped/sour cream)	0.339103	0.126590	0.059228	0.016301	1.058245

Рисунок 16 – Результат ассоциативного анализа для метрики *convinction* и значения *min_threshold*=1.04

Расчитаем среднее значение, медиану и среднеквадратичное отклонение для каждой из метрик. Результаты представлены в таблице 1.

Таблица 1 – Медиана, среднее значение и СКО метрик

Метрика	Среднее значение	Медиана	СКО
support	0.08	0.07	0.03
confidence	0.40	0.38	0.06
lift	1.08	1.06	0.14
leverage	0.00	0.01	0.01
convinction	1.04	1.05	0.09
zhangs_metric	0.07	0.09	0.15
jacard	0.16	0.15	0.04
certainty	0.03	0.04	0.08
kulczynski	0.31	0.30	0.05

Построим граф для ассоциативного анализа с параметром *min_threshold*=0.4 и метрикой *confidence*. Полученный граф представлен на рисунке 17 Вершины обозначают наборы, а ребра направлены от антецедента к консеквенту. Ширина ребер пропорциональна уровню поддержки правила, а подпись – значению *confidence*.

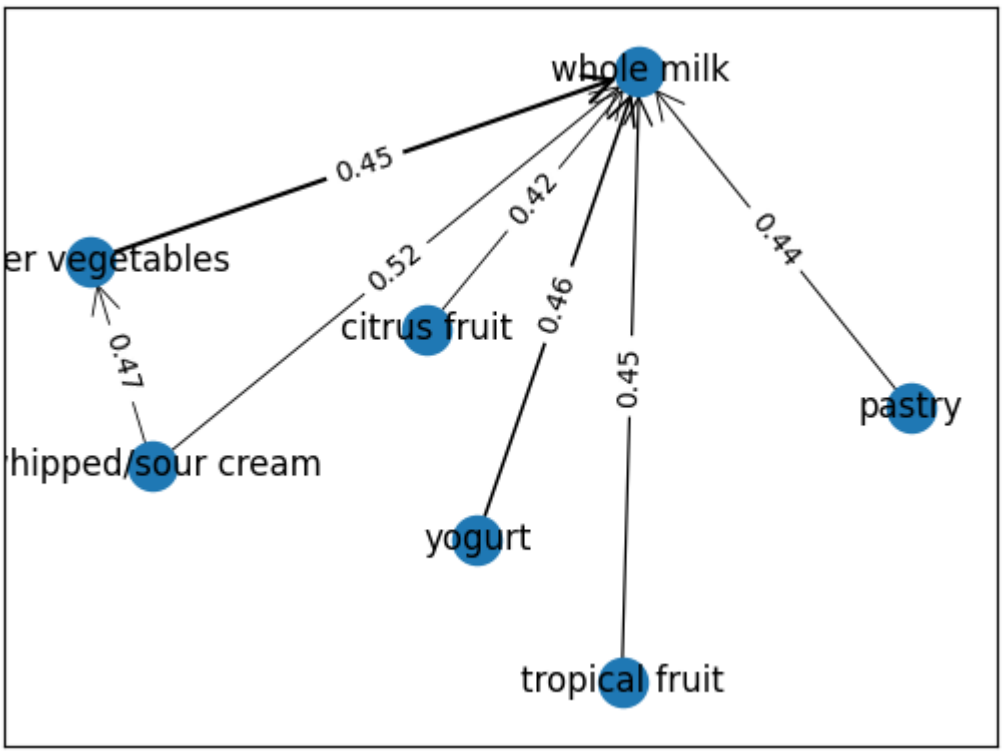

Рисунок 17 – Граф ассоциативных правил

Полученный граф дает наглядное представление ассоциативных правил. Из полученного графа можно сделать вывод, что правило с антецедентом *other vegetables* и консеквентом *whole milk* имеет наибольший уровень поддержки, а правило с антецедентом w*hipped/sour cream* и консеквентом *whole milk* имеет наибольшее значение метрики *confidence*. Также из графа видно, что консеквентом большинства правил является *whole milk*. В качестве альтернативными способами визуализации являются радиальные и линейные дендограммы.

Выводы

В ходе выполнения лабораторной работы было проведено ознакомление с методами ассоциативного анализа из библиотеки MLxtend.

Были изучены алгоритмы FPGrowth и FPMax для поиска часто встречающихся наборов. Отличие данных алгоритмов в том, что FPGrowth ищет все наборы с уровнем поддержки больше минимального, а FPMax только те, которые не имеют надмножеств с поддержкой, большей чем минимальный

уровень. Также для алгоритма FPGrowth был получен график зависимости находимых наборов от минимального уровня поддержки.

Был проведен ассоциативный анализ полученных наборов. Анализ проводился с использованием различных метрик и для различного их минимального значения. Также для каждой метрики были рассчитаны среднее значение, медиана, среднеквадратичное отклонение. Также полученные ассоциативные правила были наглядно отображены на графе.

Код программы, написанной для выполнения лабораторной работы представлен в приложении А.

ПРИЛОЖЕНИЕ А

КОД ПРОГРАММЫ

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from mlxtend.preprocessing import TransactionEncoder
import mlxtend.frequent_patterns as fp
import networkx as nx

all_data = pd.read_csv('groceries - groceries.csv')
print(all_data)
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,str)] for row in
np_data]

unique_items = set()
for row in np_data:
    for elem in row:
        unique_items.add(elem)
print(unique_items)
print(len(unique_items))

te = TransactionEncoder()
te_ary = te.fit(np_data).transform(np_data)
data = pd.DataFrame(te_ary, columns=te.columns_)
print(data)

result = fp.fpgrowth(data, min_support=0.03, use_colnames = True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
print(result)
r1 = result[result['length'] == 1]
print(min(r1['support']))
print(max(r1['support']))
r2 = result[result['length'] == 2]
print(min(r2['support']))
print(max(r2['support']))

result = fp.fpmax(data, min_support=0.03, use_colnames = True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
print(result)
r1 = result[result['length'] == 1]
print(min(r1['support']))
print(max(r1['support']))
r2 = result[result['length'] == 2]
print(min(r2['support']))
print(max(r2['support']))

result=fp.fpgrowth(data, min_support=0.03, use_colnames = True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
result = result[result['length'] == 1]
result=result.sort_values(by='support',ascending=False)[:10]
x=[list(i)[0] for i in result['itemsets']]
y=[s for s in result['support']]
plt.bar(x,y)
plt.xlabel('Товар')
plt.ylabel('Частота')
plt.show()

np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,str) and elem in items]
for row in np_data]
```

```
te = TransactionEncoder()
te_ary = te.fit_transform(np_data)
data2 = pd.DataFrame(te_ary, columns=te.columns_)
print(data2)

result = fp.fpgrowth(data2, min_support=0.03, use_colnames = True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
print(result)
r1 = result[result['length'] == 1]
print(min(r1['support']))
print(max(r1['support']))
r2 = result[result['length'] == 2]
print(min(r2['support']))
print(max(r2['support']))

result = fp.fpmax(data2, min_support=0.03, use_colnames = True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
print(result)
r1 = result[result['length'] == 1]
print(min(r1['support']))
print(max(r1['support']))
r2 = result[result['length'] == 2]
print(min(r2['support']))
print(max(r2['support']))

maxn=4
X=[sup/1000 for sup in range(50,260)]
Y=[]
for i in range(maxn):
    Y.append([])
for x in X:
    result= fp.fpgrowth(data,x, use_colnames = True)
    result['length'] = result['itemsets'].apply(lambda x: len(x))
    for i in range(maxn):
        Y[i].append(len(result[result['length'] == i+1]))
for i in range(maxn):
    plt.plot(X[:50],Y[i][:50])
plt.legend(range(1,maxn+1))
plt.xlabel("Минимальный уровень поддержки")
plt.ylabel("Количество наборов")
plt.show()

for i in range(maxn):
    plt.plot(X[50:],Y[i][50:])
plt.legend(range(1,maxn+1))
plt.xlabel("Минимальный уровень поддержки")
plt.ylabel("Количество наборов")
plt.show()

items = ['whole milk', 'yogurt', 'soda', 'tropical fruit', 'shopping
bags','sausage',
'whipped/sour cream', 'rolls/buns', 'other vegetables', 'root','vegetables',
'pork', 'bottled water', 'pastry', 'citrus fruit', 'canned beer', 'bottled
beer']
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,str) and elem in items]
for row in np_data]
np_data = [row for row in np_data if len(row) > 1]
te = TransactionEncoder()
te_ary = te.fit(np_data).transform(np_data)
data = pd.DataFrame(te_ary, columns=te.columns_)
result = fp.fpgrowth(data, min_support=0.05, use_colnames = True)
```

```
print(result)
rules = fp.association_rules(result, min_threshold = 0.3)
r1=rules.copy()
r2=rules.copy()
r1=r1.drop(['representativity',  'leverage',  'conviction',  'zhangs_metric',
  'jaccard',  'certainty',  'kulczynski'],axis=1)
r2=r2.drop(['antecedent    support',    'consequent    support',    'support',
'confidence', 'lift' ],axis=1)
print(r1)
print(r2)

rules = fp.association_rules(result, min_threshold = 0.08, metric='support')
r1=rules.copy()
r1=r1.drop(['representativity',  'leverage',  'conviction',  'zhangs_metric',
  'jaccard',  'certainty',  'kulczynski'],axis=1)
print(r1)

rules = fp.association_rules(result, min_threshold = 1.1, metric='lift')
r1=rules.copy()
r1=r1.drop(['representativity',  'leverage',  'conviction',  'zhangs_metric',
  'jaccard',  'certainty',  'kulczynski'],axis=1)
print(r1)

rules = fp.association_rules(result, min_threshold = 0.008, metric='leverage')
r1=rules.copy()
r1=r1.drop(['representativity',  'lift',  'confidence',  'zhangs_metric',   '
jaccard',  'certainty',  'kulczynski'],axis=1)
print(r1)

rules = fp.association_rules(result, min_threshold = 1.04, metric='conviction')
r1=rules.copy()
r1=r1.drop(['representativity',  'lift',  'confidence',  'zhangs_metric',   '
jaccard',  'certainty',  'kulczynski'],axis=1)
print(r1)

rules = fp.association_rules(result, min_threshold = 0.3)
for r in rules:
    if(isinstance(rules[r][0], float)):
                        print(r+":    mean:    %.2f    median    %.2f    std
%.2f"%(rules[r].mean(),rules[r].median(),rules[r].std(ddof=1)))

rules = fp.association_rules(result, min_threshold = 0.4, metric='confidence')
print(rules)
G=nx.DiGraph()
for i,r in rules.iterrows():
     G.add_edges_from([(list(r['antecedents'])[0],list(r['consequents'])[0],{'
support': r['support'],'confidence': "%.2f"%r['confidence']})])
pos = nx.spring_layout(G)
nx.draw_networkx_nodes(G, pos)
nx.draw_networkx_labels(G, pos)
sup = [G[u][v].get('support', 1)*10 for u, v in G.edges()]
nx.draw_networkx_edges(
    G, pos,
    edgelist=G.edges(),
    width=sup,
    arrows=True,
    arrowsize=25,
    arrowstyle='->')
nx.draw_networkx_edge_labels(G, pos, nx.get_edge_attributes(G, 'confidence'))
plt.show()
```