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

ОТЧЕТ

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

Студент гр. 8303 _____ Гришин К. И.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2021

Цель работы

Ознакомиться с методами ассоциативного анализа из библиотеки *MLxtend*.

Ход выполнения работы

Загрузка данных

1. Загрузить датасет: https://www.kaggle.com/irfanasrullah/groceries. Данные представлены в виде csv таблицы.
2. Создать Python скрипт. Загрузить данные в датафрейм.
3. Переформировать данные. Установить соответствие – транзакция: список товаров (частичный вывод в табл. 1).

['citrus fruit', 'semi-finished bread', 'margarine', 'ready soups']
['tropical fruit', 'yogurt', 'coffee']
['whole milk']
['pip fruit', 'yogurt', 'cream cheese', 'meat spreads']
…
['ice cream']
['pork', 'beef', 'ice cream', 'rolls/buns', 'newspapers']
['pork', 'tropical fruit', 'other vegetables', 'yogurt', 'semi-finished bread', 'flour', 'margarine', 'artif. sweetener', 'organic products', 'chocolate marshmallow']

Таблица 1. Исходные данные после преобразования.

4. Получить список уникальных товаров
5. Вывод списка уникальных товаров, а также их количество

   В датасете представлено 169 различных товаров.

['vinegar' 'soda' 'baby cosmetics' 'newspapers' 'rice' 'female sanitary products' 'salad dressing' 'preservation products' 'male cosmetics' 'organic sausage' 'coffee' 'herbs' 'frozen vegetables' 'finished products' 'soups' 'roll products' 'nut snack' 'butter' 'red/blush wine' 'pip fruit' 'specialty chocolate' 'frozen potato products' 'white bread' 'root vegetables' 'brandy' 'nuts/prunes' 'bags' 'berries' 'white wine' 'specialty vegetables' 'cocoa drinks' 'cleaner' 'candles' 'organic products' 'liver loaf' 'toilet cleaner' 'snack products' 'frozen fish' 'specialty fat' 'Instant food products' 'soap' 'bottled beer' 'soft cheese' 'popcorn' 'pet care' 'pork' 'semi-finished bread' 'rum' 'pastry' 'sausage' 'seasonal products' 'hygiene articles' 'canned fish' 'liqueur' 'pickled vegetables' 'dish cleaner' 'cream' 'tea' 'chicken' 'chocolate marshmallow' 'citrus fruit' 'meat' 'frozen dessert' 'ready soups' 'fish' 'cream cheese' 'artif. sweetener' 'onions' 'ice cream' 'dog food' 'spices' 'curd cheese' 'skin care' 'pudding powder' 'sliced cheese' 'liquor (appetizer)' 'beef' 'honey' 'make up remover' 'sugar' 'sweet spreads' 'canned vegetables' 'spread cheese' 'packaged fruit/vegetables' 'flower (seeds)' 'waffles' 'whipped/sour cream' 'photo/film' 'rubbing alcohol' 'dishes' 'misc. beverages' 'chewing gum' 'rolls/buns' 'curd' 'sauces' 'sound storage medium' 'instant coffee' 'mustard' 'fruit/vegetable juice' 'frozen chicken' 'flower soil/fertilizer' 'tidbits' 'ham' 'bathroom cleaner' 'other vegetables' 'abrasive cleaner' 'turkey' 'chocolate' 'mayonnaise' 'dental care' 'condensed milk' 'kitchen utensil' 'prosecco' 'bottled water' 'tropical fruit' 'hard cheese' 'hamburger meat' 'kitchen towels' 'grapes' 'beverages' 'hair spray' 'domestic eggs' 'decalcifier' 'frankfurter' 'pasta' 'meat spreads' 'processed cheese' 'softener' 'frozen fruits' 'margarine' 'baby food' 'zwieback' 'sparkling wine' 'salt' 'yogurt' 'whole milk' 'shopping bags' 'UHT-milk' 'light bulbs' 'jam' 'cake bar' 'specialty bar' 'ketchup' 'specialty cheese' 'liquor' 'long life bakery product' 'cling film/bags' 'flour' 'oil' 'candy' 'cat food' 'syrup' 'cooking chocolate' 'salty snack' 'dessert' 'cereals' 'canned beer' 'canned fruit' 'baking powder' 'napkins' 'brown bread' 'butter milk' 'potted plants' 'cookware' 'potato products' 'detergent' 'whisky' 'house keeping products' 'frozen meals']

FPGrowth FPMax

1. Данные преобразованы с помощью TransactionEncoder.
2. Проведен ассоциативный анализ с использованием алгоритма FPGrowth (табл. 2).

	support	itemsets
5	0.255516	(whole milk)
8	0.193493	(other vegetables)
11	0.183935	(rolls/buns)
19	0.174377	(soda)
2	0.139502	(yogurt)
...	...	...
43	0.031012	(onions)
61	0.030605	(rolls/buns, sausage)
44	0.030503	(citrus fruit, whole milk)
42	0.030402	(specialty chocolate)
50	0.030097	(whole milk, pip fruit)
63 rows		

Таблица 2. Результат FPGrowth.

3. Анализ результатов, поиск минимальной и максимальной поддержки для наборов каждой длины

   Алгоритм находит все возможные наборы, которые удовлетворяют минимальной поддержке. Минимальная и максимальная поддержка для разных наборов представлена в таблице 3.

	min	max
1	0.030402	0.255516
2	0.030097	0.074835

Таблица 3. Минимальная и максимальная поддержка для наборов длины 1 и 2 FPGrowth.

4. Проведен аналогичный анализ FPMax.

   Результат работы алгоритма представлен в таблице 4. Минимальная и максимальная поддержка для разных наборов представлена в таблице 5.

   FPMax работает иначе, наборы перебираются таким образом, что один набор не может быть частью другого.

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

Таблица 4. Результат работы FPMax.

	min	max
1	0.030402	0.098526
2	0.030097	0.074835

Таблица 5. Минимальная и максимальная поддержка для наборов длины 1 и 2 FPMax.

5. Сравним результаты работы алгоритмов.

   FPMax отбирает наборы максимальной длины, соответствующие указанному уровню поддержки, поднаборы найденных наборов не указываются. *FPGrowth* в свою очередь, перебирает все возможные наборы, которые удовлетворяют уровню указанной минимальной поддержки.

6. Построим гистограммы для каждого из алгоритмов

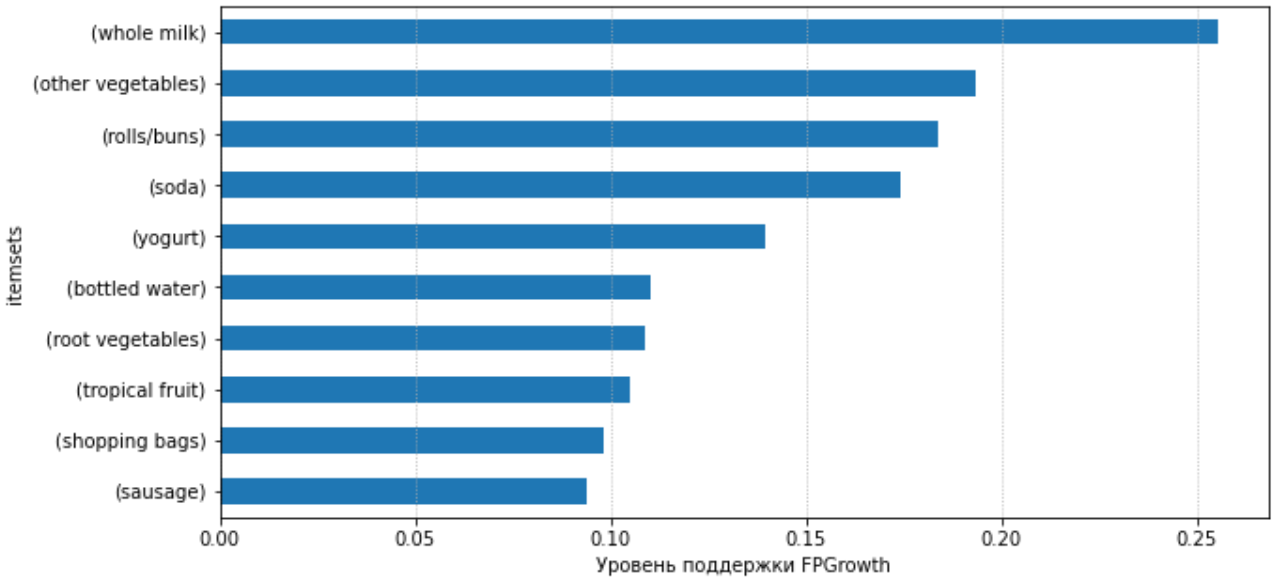

Рисунок 1. Гистограмма первых десяти наборов FPGrowth с лучшей поддержкой.

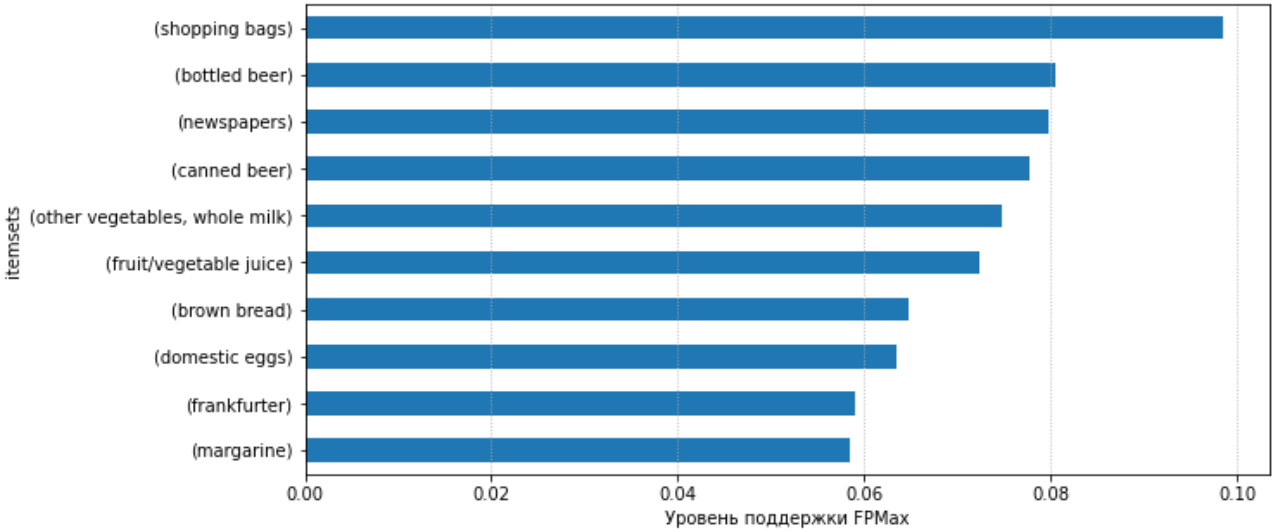

Рисунок 2. Гистограмма первых десяти наборов FPMax с лучшей поддержкой.

7. Преобразуем данные, чтоб они содержали ограниченный набор товаров.
8. Проведен анализ FPGrowth и FPMax для нового набора данных. Результат работы FPGrowth представлен в таблице 6, результат FPMax – в таблице 7.

	support	itemsets		support	itemsets
1	0.182613	(whole milk)	6	0.055923	(tropical fruit)
2	0.149771	(rolls/buns)	14	0.047077	(yogurt, whole milk)
4	0.146721	(other vegetables)	10	0.046162	(canned beer)
7	0.144484	(soda)	16	0.045755	(rolls/buns, whole milk)
0	0.116624	(yogurt)	22	0.038129	(whole milk, root vegetables)
11	0.093645	(shopping bags)	21	0.037011	(other vegetables, root vegetables)
5	0.093238	(bottled water)	13	0.036706	(citrus fruit)
9	0.079715	(root vegetables)	15	0.036706	(other vegetables, yogurt)
8	0.075547	(pastry)	17	0.035282	(other vegetables, rolls/buns)
12	0.063854	(whipped/sour cream)	19	0.032537	(rolls/buns, soda)
3	0.062430	(bottled beer)	20	0.030707	(whole milk, soda)
18	0.060702	(other vegetables, whole milk)			

Таблица 6. Результат работы FPGrowth для ограниченного набора данных.

	support	itemsets		support	itemsets
9	0.093645	(shopping bags)	16	0.045755	(rolls/buns, whole milk)
8	0.093238	(bottled water)	7	0.038129	(whole milk, root vegetables)
5	0.075547	(pastry)	6	0.037011	(other vegetables, root vegetables)
4	0.063854	(whipped/sour cream)	10	0.036706	(other vegetables, yogurt)
3	0.062430	(bottled beer)	0	0.036706	(citrus fruit)
15	0.060702	(other vegetables, whole milk)	14	0.035282	(other vegetables, rolls/buns)
2	0.055923	(tropical fruit)	13	0.032537	(rolls/buns, soda)
11	0.047077	(yogurt, whole milk)	12	0.030707	(whole milk, soda)
1	0.046162	(canned beer)			

Таблица 7. Результат работы FPMax для ограниченного набора данных.

Рассмотрим минимальную и максимальную поддержку для различных наборов (табл. 8).

	FPGrowth	FPMax
1	[0.036706, 0.182613]	[0.036706, 0.093645]
2	[0.030707, 0.060702]	[0.030707, 0.060702]

Таблица 8. Минимальная и максимальная поддеркжа при ограниченном наборе данных.

9. Построим график зависимости количества наборов от минимальной поддержки.

   Построено четыре графика: FPGrowth со всеми данными на рисунке 3, FPGrowth с выборочными данными – 4, FPMax со всеми данными – 5, FPMax с выборочными данными – 6.

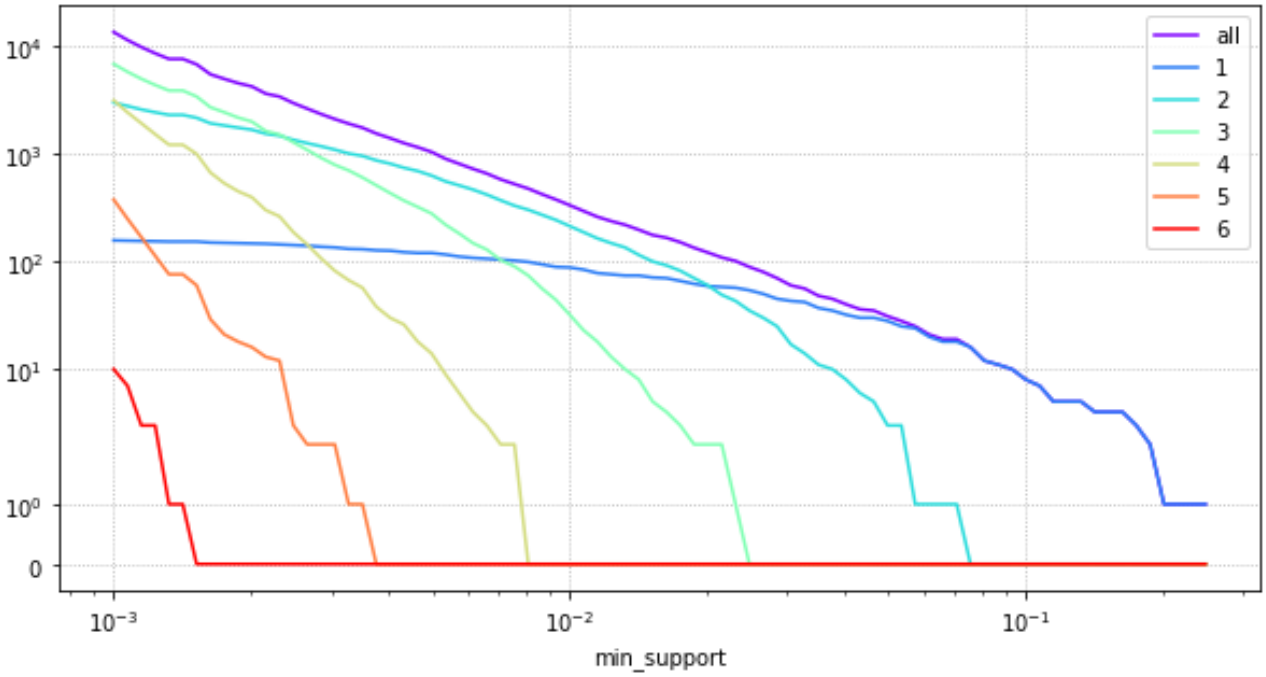

Рисунок 3. Зависимость количества наборов, от минимальной поддержки FPGrowth. Полные данные.

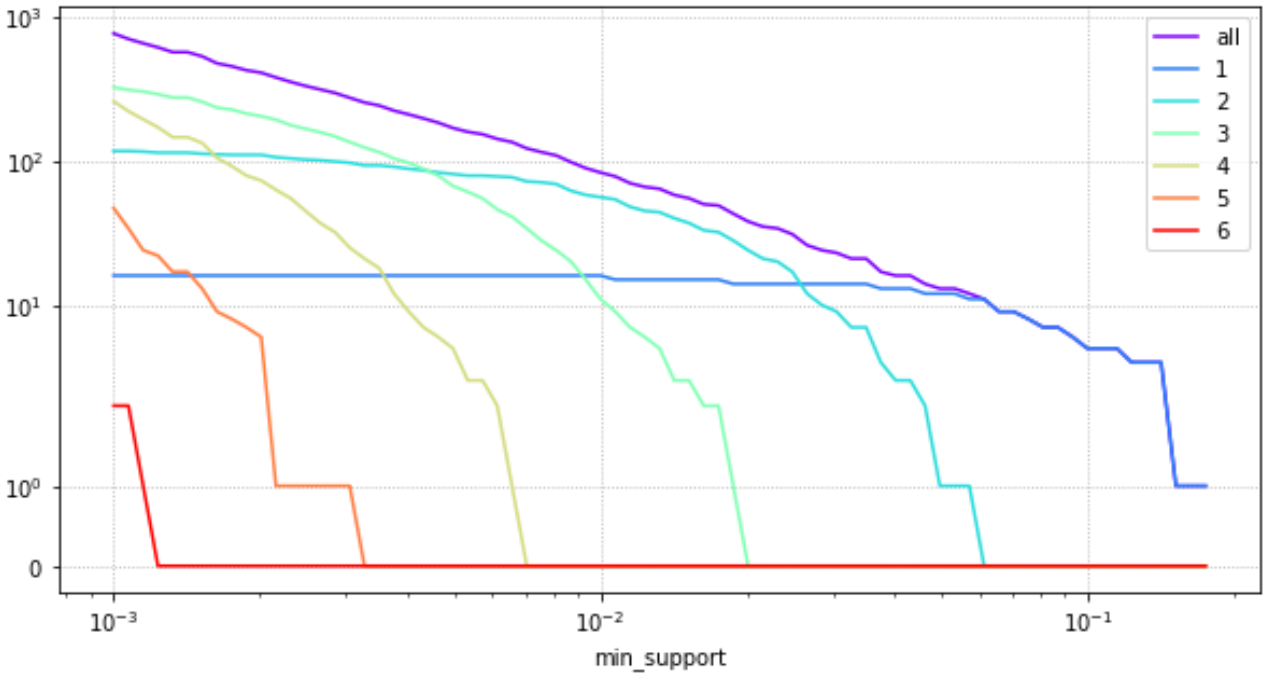

Рисунок 4. Зависимость количества наборов, от минимальной поддержки FPGrowth. Выборочные данные.

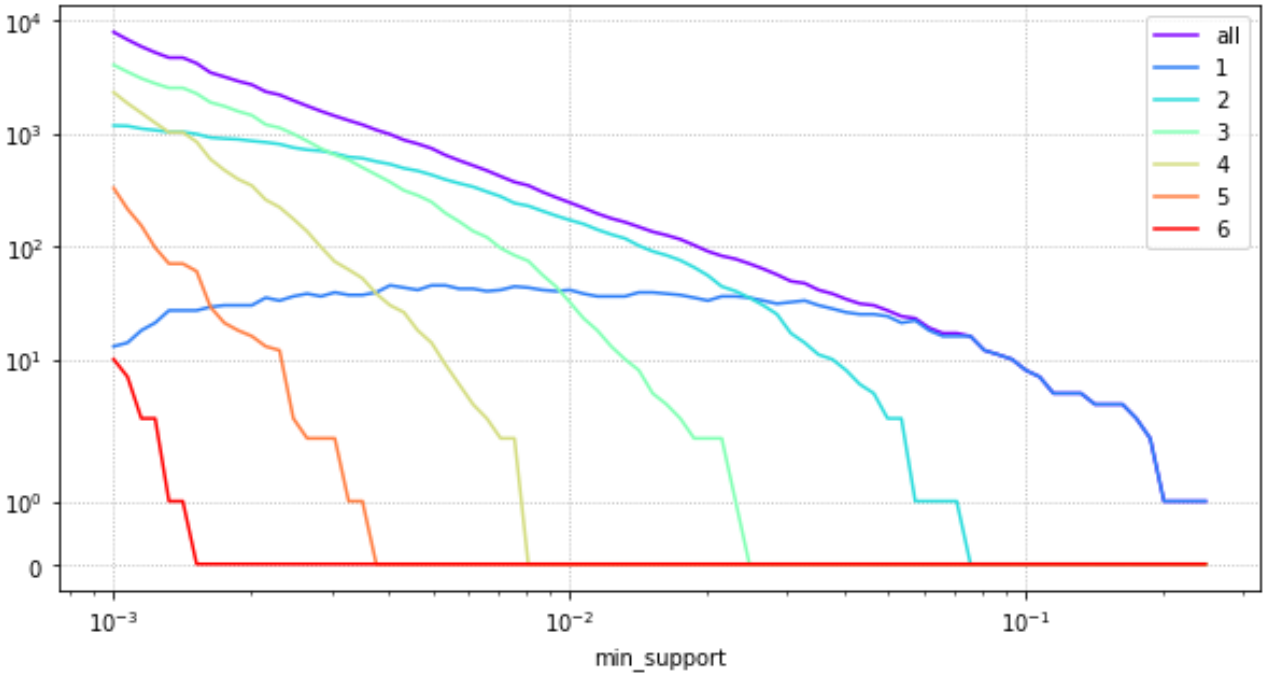

Рисунок 5. Зависимость количества наборов, от минимальной поддержки FPMax. Полные данные.

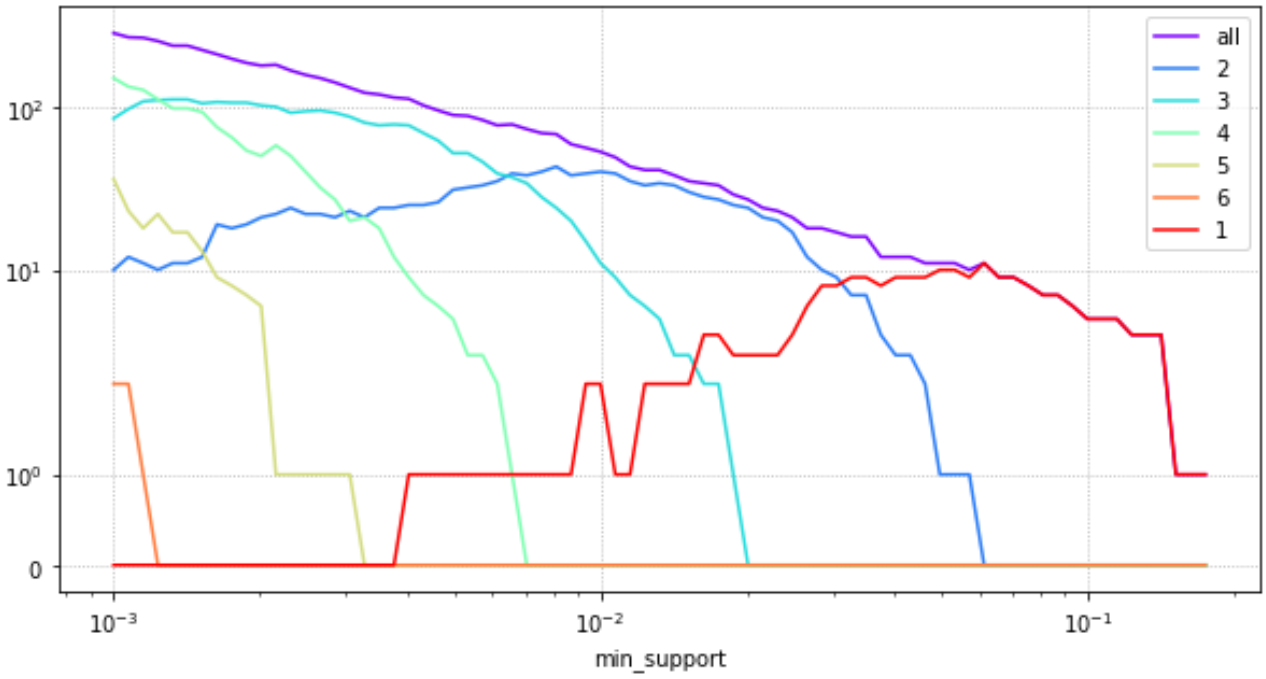

Рисунок 6. Зависимость количества наборов, от минимальной поддержки FPMax. Выборочные данные.

FPGrowth менее чувствителен к ограничению данных.

Ассоциативный анализ

1. Сформируем данные так, чтобы размер наборов не был меньше 2.
2. Проведем частотный анализ алгоритмом FPGrowth.
3. Проведем ассоциативный анализ используя association_rules (табл. 9).

	antecedents	consequents	A support	C support	support	conf	lift	leverage	convict
0	(yogurt)	(whole milk)	0.241240	0.421869	0.110954	0.459933	1.090228	0.009183	1.070481
1	(yogurt)	(other vegetables)	0.241240	0.335079	0.085985	0.356427	1.063713	0.005150	1.033172
2	(tropical fruit)	(yogurt)	0.185864	0.241240	0.057994	0.312026	1.293423	0.013156	1.102890
3	(tropical fruit)	(other vegetables)	0.185864	0.335079	0.071083	0.382449	1.141370	0.008804	1.076706
4	(tropical fruit)	(whole milk)	0.185864	0.421869	0.083770	0.450704	1.068352	0.005359	1.052495
5	(other vegetables)	(whole milk)	0.335079	0.421869	0.148208	0.442308	1.048449	0.006849	1.036649
6	(whole milk)	(other vegetables)	0.421869	0.335079	0.148208	0.351313	1.048449	0.006849	1.025026
7	(rolls/buns)	(whole milk)	0.296214	0.421869	0.112163	0.378654	0.897564	-0.012801	0.930450
8	(bottled water)	(whole milk)	0.185461	0.421869	0.068063	0.366992	0.869921	-0.010177	0.913309
9	(bottled water)	(soda)	0.185461	0.267217	0.057390	0.309446	1.158033	0.007832	1.061153
10	(citrus fruit)	(whole milk)	0.146395	0.421869	0.060411	0.412655	0.978159	-0.001349	0.984313
11	(citrus fruit)	(other vegetables)	0.146395	0.335079	0.057189	0.390646	1.165836	0.008135	1.091192
12	(root vegetables)	(other vegetables)	0.196335	0.335079	0.093838	0.477949	1.426378	0.028050	1.273671
13	(root vegetables)	(whole milk)	0.196335	0.421869	0.096859	0.493333	1.169400	0.014031	1.141049
14	(sausage)	(rolls/buns)	0.167539	0.296214	0.060612	0.361779	1.221342	0.010985	1.102730
15	(sausage)	(whole milk)	0.167539	0.421869	0.059203	0.353365	0.837619	-0.011477	0.894062
16	(sausage)	(other vegetables)	0.167539	0.335079	0.053363	0.318510	0.950552	-0.002776	0.975687
17	(whipped/sour cream)	(whole milk)	0.124245	0.421869	0.063834	0.513776	1.217858	0.011419	1.189023
18	(whipped/sour cream)	(other vegetables)	0.124245	0.335079	0.057189	0.460292	1.373683	0.015557	1.232002
19	(pastry)	(whole milk)	0.150624	0.421869	0.065848	0.437166	1.036260	0.002304	1.027179

Таблица 9. Ассоциативные правила исходных данных.

Рассмотрим столбцы полученного датафрейма:

- Antecedent – товар-причина
- Consequent – товар-следствие
- A support (antecedent support) – вероятность появления товара antecedent в транзакции
- C support (consequent support) – вероятность появления товара consequent в транзакции

- Support – шанс появления обоих товаров antecedent и consequent в транзакции
- Conf (confidence) – вероятность появления товара consequent в транзакциях, в которых есть antecedent.
- Lift – показывает отношение совместной вероятности antecedent и consequent к ожидаемой совместной вероятности, если бы они были статистически независимы.
- Leverage – показывает разницу между наблюдаемой вероятностью появления antecedent и consequent и ожидаемой независимой вероятностью.
- Conviction – показывает ожидаемую ошибку. То есть как часто встречается antecedent там, где consequent отсутствует.

4. Определить на основе какой метрики ведется расчет.

   По умолчанию расчет проходит по метре confidence. То есть все наборы подбираются по уровню $min\_treshold > confidence$.

5. Произведен расчет по каждой метрики так, чтоб выводилось не менее 10 правил.
6. Посчитаны математическое ожидание, медиана и СКО для каждой метрики (табл. 10).

	count	mean	std	median
support	52.0	0.074685	0.022549	0.066955
confidence	52.0	0.289579	0.103683	0.264439
lift	52.0	1.042997	0.183264	1.056081
leverage	12.0	0.015533	0.006063	0.013594
conviction	52.0	1.017200	0.083993	1.022851

Таблица 10. Количество, математическое ожидание, СКО и медиана каждой метрики.

7. Построим граф для анализа по метрике confidence с минимальным значением 0.4 (рис. 7).

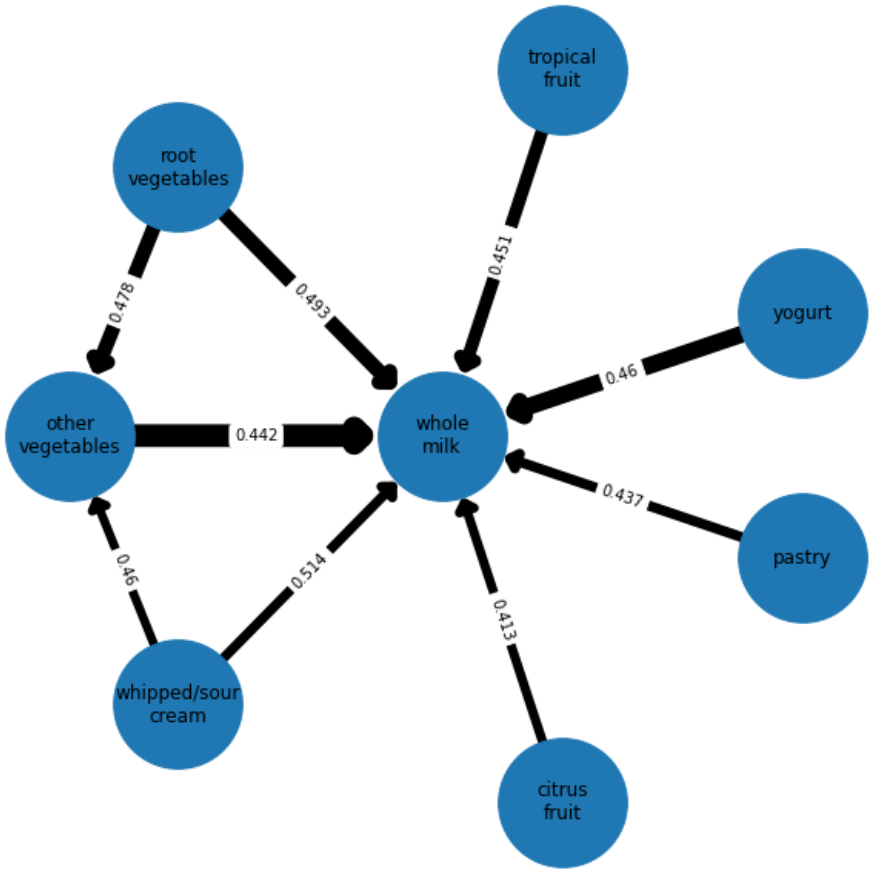

Рисунок 7. Граф, описывающий правила ассоциаций.

Первым выводом из графика можно сделать то, что товар «whole milk» является консеквентом для остальных товаров, чаще всего «whole milk» берется вместе с овощами. Также можно заметить, что «root vegetables» и «whipped/sour cream» берутся вместе с «other vegetables», из чего можно сделать вывод, что отделы с молочной продукцией и овощами стоит разместить рядом друг с другом.

Значение confidence для каждой пары примерно одинаково.

8. Так же подобные данные можно визуализировать с помощью окрашенных цветом матриц смежности или инцидентности.

Вывод

В ходе лабораторной работы были изучены алгоритм частотного анализа *FPGrowth* и *FPMax* из библиотеки *MLxtend*.

FPGrowth также как и *Apriori* позволяет выделить наиболее частые наборы в выборках данных. *FPMax* преследует несколько иную цель, задачей алгоритма является выделение наборов наибольшей длины, исключая при этом возможные подмножества.

Проведено исследование алгоритмов на тестовых данных. Для проведения частотного анализа данные были предварительно обработаны функцией *TransactionEncoer*.

С помощью функции *association_rules* был проведен ассоциативный анализ данных, выделены ассоциативные правила между различными товарами.

Построен граф, визуализирующий найденные ассоциативные правила.