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

Студентка гр. 8303 _____ Самойлова А.С.

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, удалив все NaN значения.
4. Получить список уникальных товаров
5. Вывод списка уникальных товаров, а также их количество

   Количество различных товаров — 169:

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

FPGrowth FPMax

1. Данные преобразованы с помощью TransactionEncoder.
2. Проведен ассоциативный анализ с использованием алгоритма FPGrowth:

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

63 rows × 2 columns

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:

	min	max
1	0.030402	0.255516
2	0.030097	0.074835

4. Проведен аналогичный анализ FPMax:

   Результат работы алгоритма:

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

Минимальная и максимальная поддержка для разных наборов:

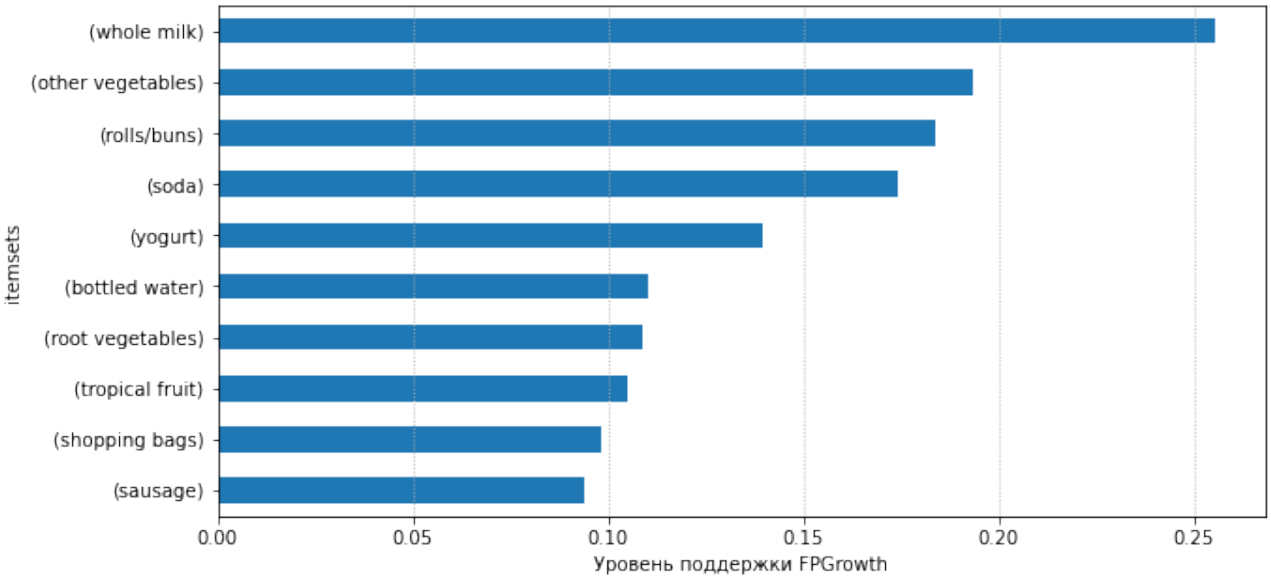
	min	max
1	0.030402	0.098526
2	0.030097	0.074835

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

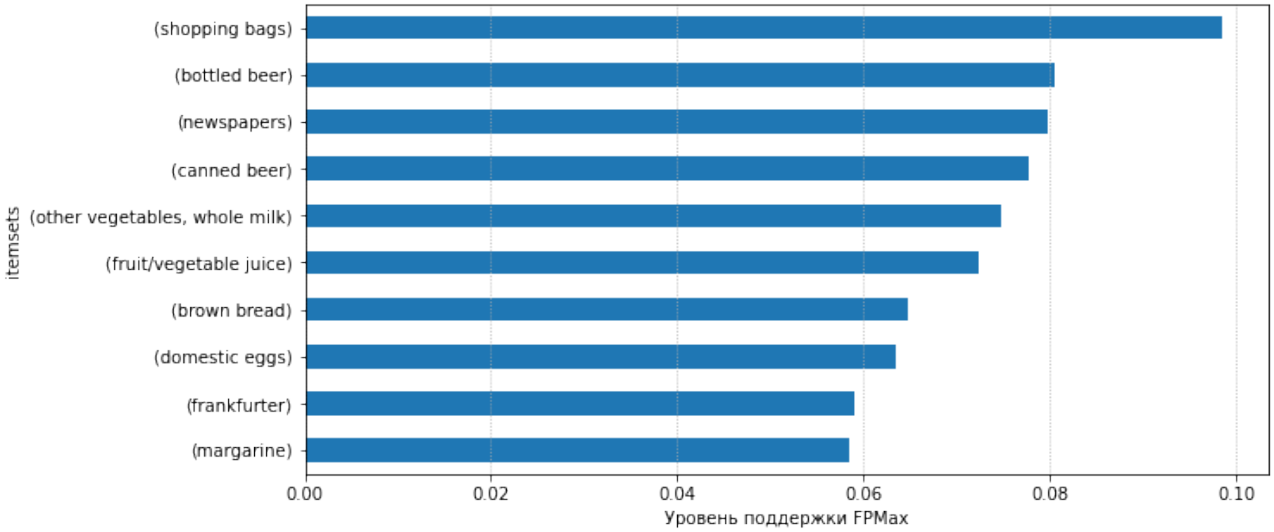
5. Сравним результаты работы алгоритмов.

   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



Гистограмма первых десяти наборов FPGrowth с лучшей поддержкой.



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:

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

Результат FPMax:

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

Минимальная и максимальная поддержка для различных наборов .

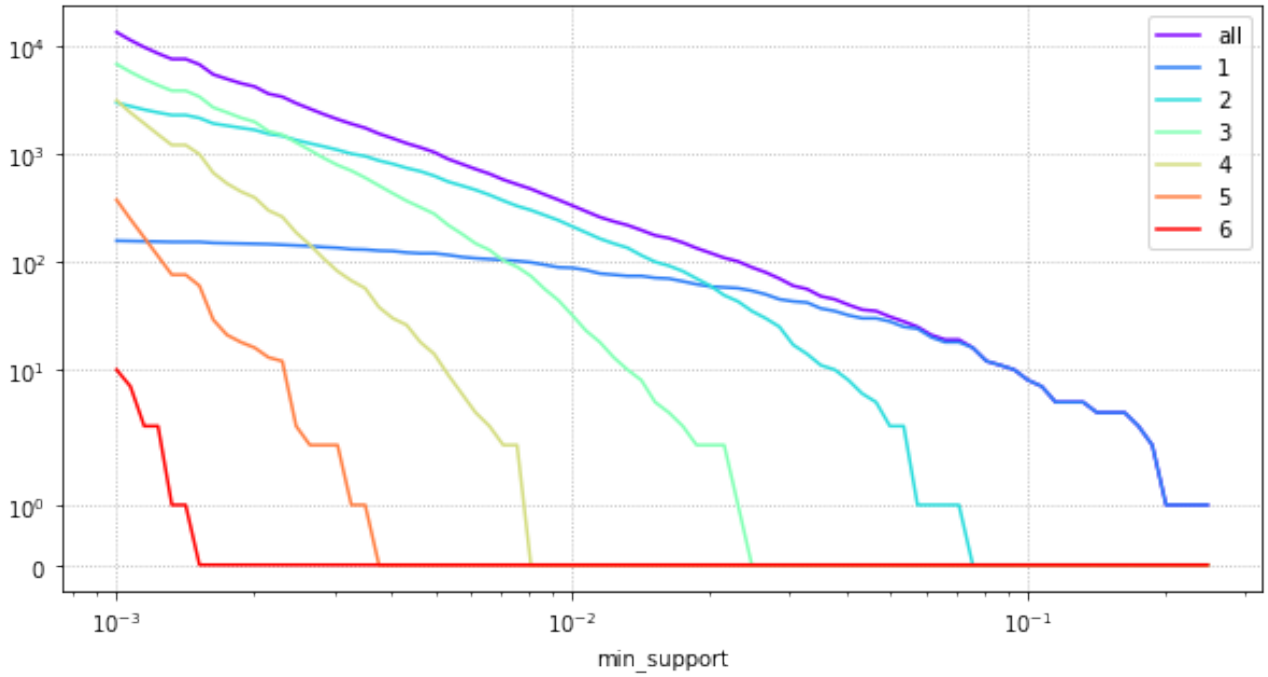
FPGrowth:

	min	max
1	0.036706	0.182613
2	0.030707	0.060702

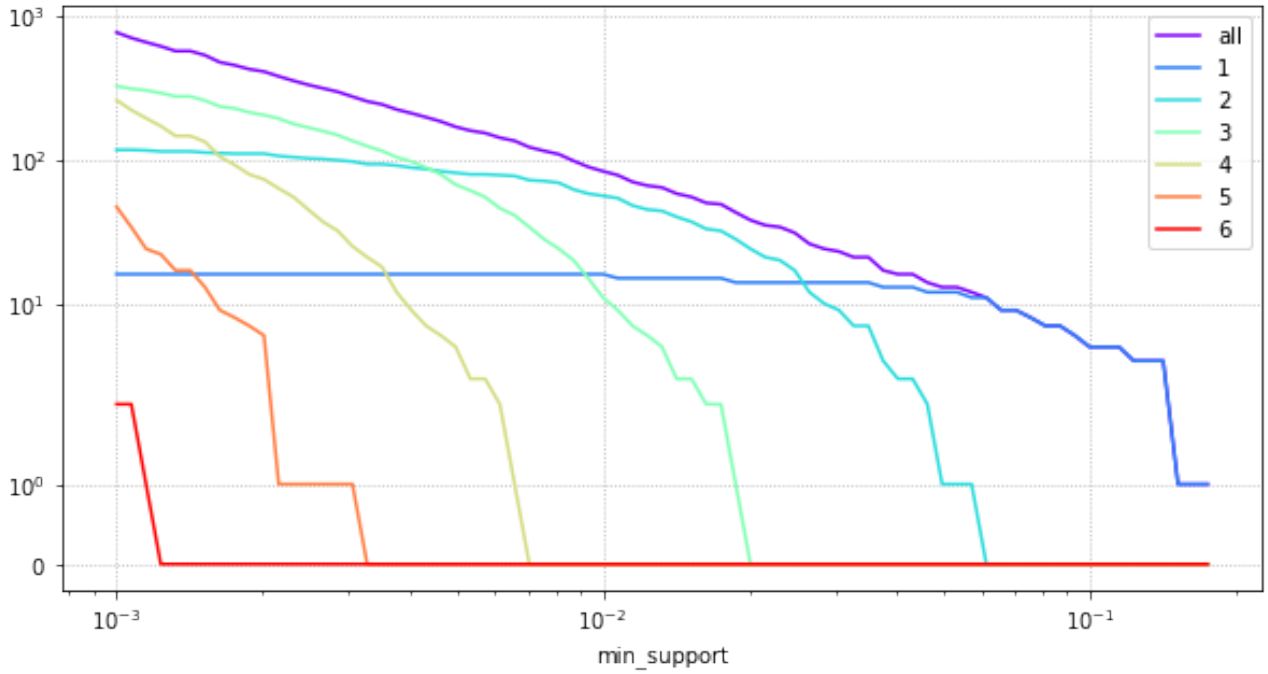
FPMax:

	min	max
1	0.036706	0.093645
2	0.030707	0.060702

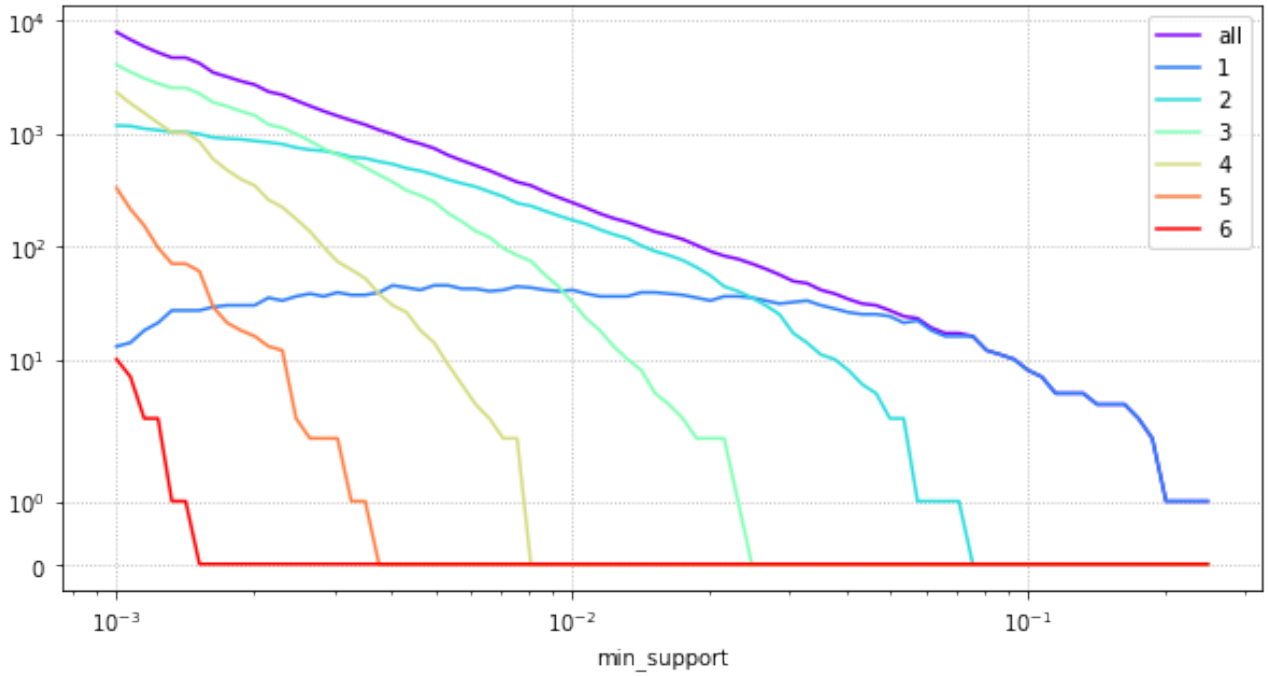
9. Построим график зависимости количества наборов от минимальной поддержки.



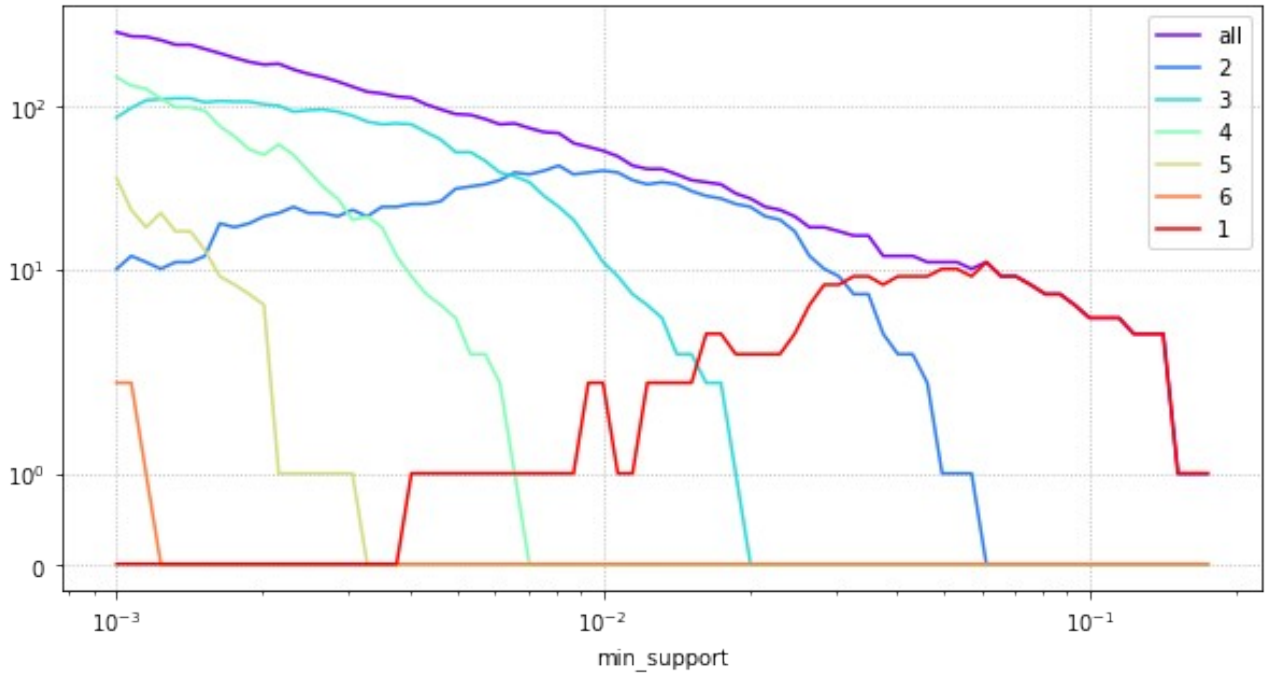
Зависимость количества наборов, от минимальной поддержки FPGrowth. Полные данные.



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



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



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:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
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

Рассмотрим столбцы полученного датафрейма:

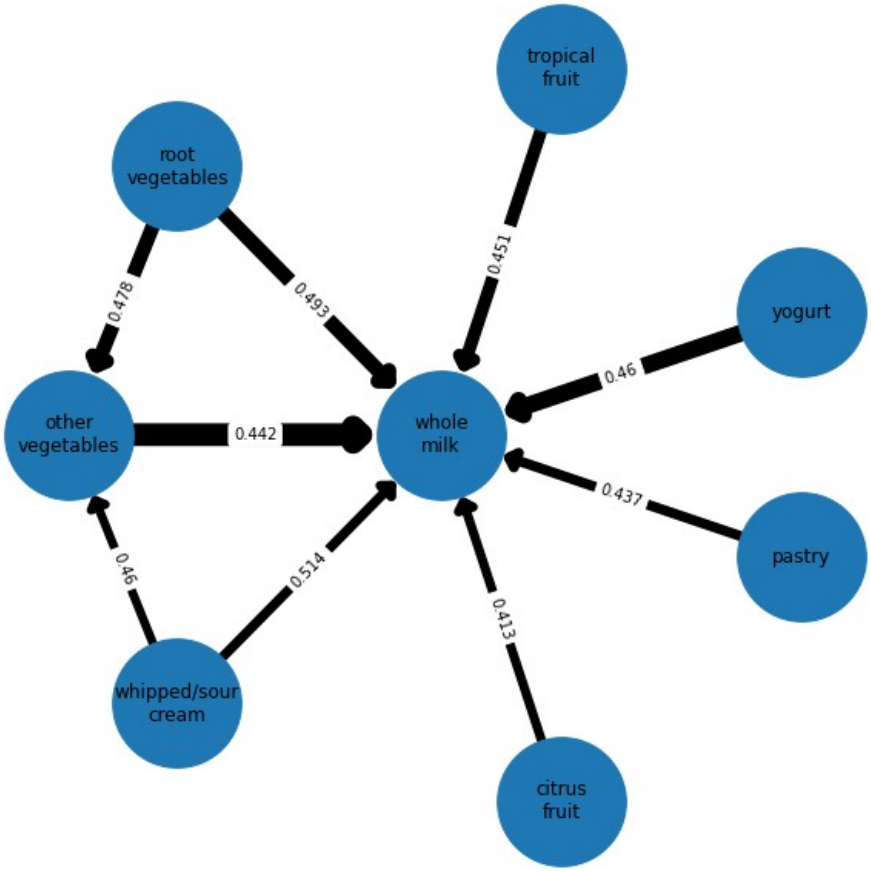
- Antecedent – товар-причина
- Consequent – товар-следствие
- A support (antecedent support) – вероятность появления товара antecedent в транзакции
- C support (consequent support) – вероятность появления товара consequent в транзакции
- Support – шанс появления обоих товаров antecedent и consequent 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
   $$min_{tres}hold > confidence.$$
5. Произведен расчет по каждой метрики так, чтоб выводилось не менее 10 правил.
6. Посчитаны математическое ожидание, медиана и СКО для каждой метрики:

	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

7. Построим граф для анализа по метрике confidence с минимальным значением 0.4:



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 *FP-Growth* 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.