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

ОТЧЕТ

по лабораторной работе №3
по дисциплине «Машинное обучение»
Тема: Частотный анализ

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

Санкт-Петербург
2021

Цель работы

Ознакомиться с методами частотного анализа из библиотеки *MLxtend*.

Ход выполнения работы

Загрузка данных

1. Загрузить датасет по ссылке: https://www.kaggle.com/acostasg/random-shopping-cart. Данные представлены в виде csv таблицы.
2. Создать Python скрипт. Загрузить данные в датафрейм (табл. 1).

	date	transaction_id	product
0	2000-01-01	1	yogurt
1	2000-01-01	1	pork
2	2000-01-01	1	sandwich bags
3	2000-01-01	1	lunch meat
4	2000-01-01	1	all- purpose
...	...	...	...
22338	2002-02-26	1139	soda
22339	2002-02-26	1139	laundry detergent
22340	2002-02-26	1139	vegetables
22341	2002-02-26	1139	shampoo
22342	2002-02-26	1139	vegetables

Таблица 1. Загруженные данные.

3. Получить список всех ID транзакций. А также посчитать их количество.

   Представлено 1139 транзакций.

4. Получить список всех товаров. А также посчитать их количество.

   Представлено 38 различных товаров.

5. Сформировать датасет для частотного анализа.

   Каждой транзакции сопоставлен список товаров

Подготовка данных

6. Представить данные в виде матрицы с помощью *mlxtend.preporcessing.TransactionEncoder*.
7. Вывод полученного датасета (табл. 2).

	0	1	2	3	4	5	6	7	...	29	30	31	32	33	34	35	36	37
0	+	+	-	+	+	-	-	-	...	+	+	-	-	-	-	+	-	+
1	-	+	-	-	-	+	+	-	...	-	-	-	-	+	+	+	+	+
2	-	-	+	-	-	+	+	-	...	+	+	+	-	+	-	+	-	-
3	+	-	-	-	-	+	-	-	...	-	+	-	-	+	-	-	-	-
4	+	-	-	-	-	-	-	-	...	-	+	+	-	+	+	+	+	+
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1134	+	-	-	+	-	+	+	+	...	+	-	-	+	-	-	-	-	-
1135	-	-	-	-	-	+	+	+	...	+	-	+	-	-	-	+	-	-
1136	-	-	+	+	-	-	-	-	...	+	-	-	+	-	+	+	-	+
1137	+	-	-	+	-	-	+	-	...	+	+	+	+	+	-	+	+	+
1138	-	-	-	-	-	-	-	-	...	-	+	-	-	-	-	+	-	-

Таблица 2. Данные обработанные *mlxtend.preporcessing.TransactionEncoder*. Строки - транзакции, столбцы - товары отсортированные в лексикографическом порядке.

Ассоциативный анализ с использованием алгоритма Apriori

1. Применение алгоритма apriori с минимальной поддержкой 0.3

	support	itemsets		support	itemsets
0	0.374890	(all- purpose)	26	0.367867	(sandwich bags)
1	0.384548	(aluminum foil)	27	0.349429	(sandwich loaves)
2	0.385426	(bagels)	28	0.368745	(shampoo)
3	0.374890	(beef)	29	0.379280	(soap)
4	0.367867	(butter)	30	0.390694	(soda)
5	0.395961	(cereals)	31	0.373134	(spaghetti sauce)
6	0.390694	(cheeses)	32	0.360843	(sugar)
7	0.379280	(coffee/tea)	33	0.378402	(toilet paper)
8	0.388938	(dinner rolls)	34	0.369622	(tortillas)
9	0.388060	(dishwashing liquid/detergent)	35	0.739245	(vegetables)
10	0.389816	(eggs)	36	0.394205	(waffles)
11	0.352941	(flour)	37	0.384548	(yogurt)
12	0.370500	(fruits)	38	0.310799	(vegetables, aluminum foil)
13	0.345917	(hand soap)	39	0.300263	(bagels, vegetables)
14	0.398595	(ice cream)	40	0.310799	(vegetables, cereals)
15	0.375768	(individual meals)	41	0.309043	(vegetables, cheeses)
16	0.376646	(juice)	42	0.308165	(vegetables, dinner rolls)
17	0.371378	(ketchup)	43	0.306409	(dishwashing liquid/detergent, vegetables)
18	0.378402	(laundry detergent)	44	0.326602	(vegetables, eggs)
19	0.395083	(lunch meat)	45	0.302897	(vegetables, ice cream)
20	0.380158	(milk)	46	0.309043	(laundry detergent, vegetables)
21	0.375768	(mixes)	47	0.311677	(vegetables, lunch meat)
22	0.362599	(paper towels)	48	0.331870	(vegetables, poultry)
23	0.371378	(pasta)	49	0.305531	(soda, vegetables)
24	0.355575	(pork)	50	0.315189	(vegetables, waffles)
25	0.421422	(poultry)	51	0.319579	(vegetables, yogurt)

Таблица 3. Применение apriori с минимальной поддержкой 0.3

2. Применение алгоритма apriori с минимальной поддержкой 0.3 и размером набора равным 1.

	support	itemsets		support	itemsets
0	0.374890	(all- purpose)	19	0.395083	(lunch meat)
1	0.384548	(aluminum foil)	20	0.380158	(milk)
2	0.385426	(bagels)	21	0.375768	(mixes)
3	0.374890	(beef)	22	0.362599	(paper towels)
4	0.367867	(butter)	23	0.371378	(pasta)
5	0.395961	(cereals)	24	0.355575	(pork)
6	0.390694	(cheeses)	25	0.421422	(poultry)
7	0.379280	(coffee/tea)	26	0.367867	(sandwich bags)
8	0.388938	(dinner rolls)	27	0.349429	(sandwich loaves)
9	0.388060	(dishwashing liquid/detergent)	28	0.368745	(shampoo)
10	0.389816	(eggs)	29	0.379280	(soap)
11	0.352941	(flour)	30	0.390694	(soda)
12	0.370500	(fruits)	31	0.373134	(spaghetti sauce)
13	0.345917	(hand soap)	32	0.360843	(sugar)
14	0.398595	(ice cream)	33	0.378402	(toilet paper)
15	0.375768	(individual meals)	34	0.369622	(tortillas)
16	0.376646	(juice)	35	0.739245	(vegetables)
17	0.371378	(ketchup)	36	0.394205	(waffles)
18	0.378402	(laundry detergent)	37	0.384548	(yogurt)

Таблица 4. Применение apriori с минимальной поддержкой 0.3. Выведены наборы размера 1.

3. Применение алгоритма apriori с выводом наборов только размера 2.

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

Таблица 5. Применение apriori с минимальной поддержкой 0.3. Выведены наборы размера 2.

4. Определение графика зависимости количества набора от минимальной поддержки (рис. 1).

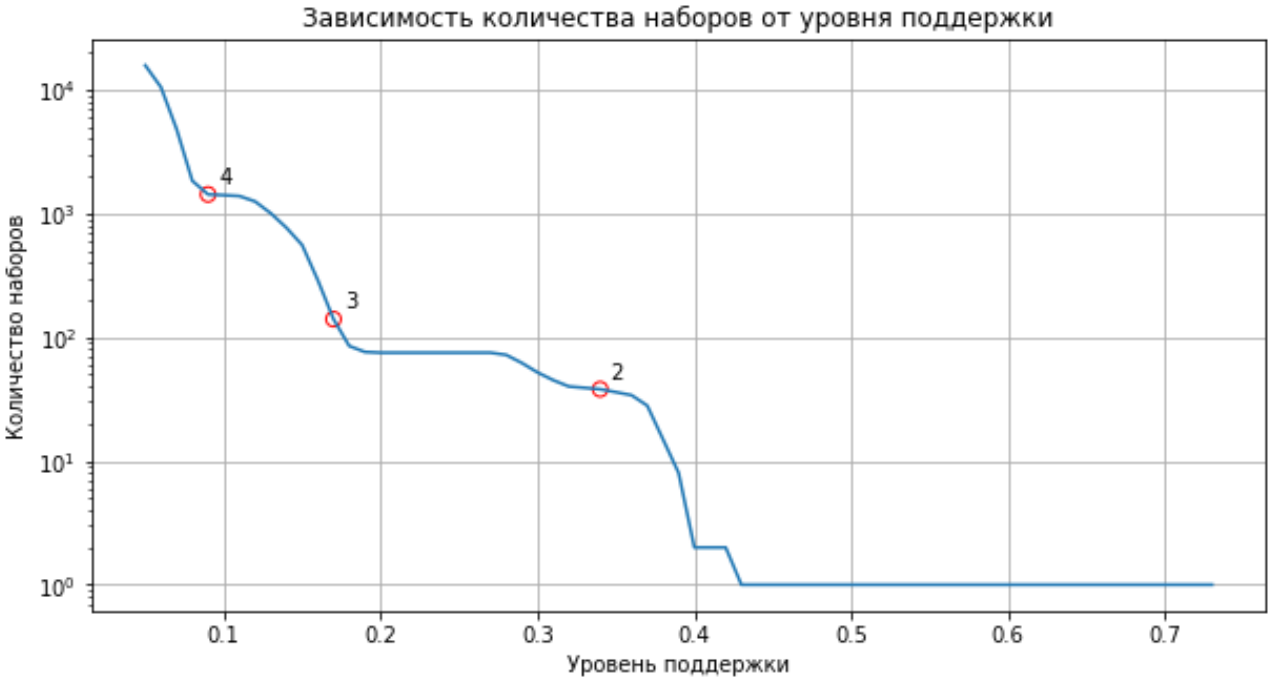

Рисунок 1. Зависимость количества наборов от минимальной поддержки. Отображены точки, на которых перестают генерироваться соответствующие наборы.

5. Определение значение уровня поддержки, при котором уменьшается максимальный размер генерируемых наборов.

   Значения минимальной поддержки, на которой перестают генерироваться наборы соответствующей длины представлены в таблице 6.

Длина набора	Минимальная поддержка
4	0.09
3	0.17
2	0.34

Таблица 6. Минимальная поддержка, при которйо перестают генерироваться наборы указанной длины.

   Данные точки также указаны на графике (рис. 1).

6. Сделана выборка датасета. В каждой транзакции оставлен только тот товар, у которого уровень поддержки выше 0.38

7. Новый датасет представлен в виде матрицы с помощью *mlxtend.preporcessing.TransactionEncoder*.
8. Проведение ассоциативного анализа нового датасета при минимальное поддержке 0.3 (табл. 7).

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

Таблица 7. *Apriori* с поддержкой 0.3 для тех продуктов, у которых поддержка больше 0.38.

9. Проведен ассоциативный анализ для поддержки 0.15. Выведены наборы, которые содержат в себе «*yogurt*» или «*waffles*».

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

Таблица 8. *Apriori* с поддержкой 0.15. Выведены наборы, которые содержат «*yogurt*» или «*waffles*».

10. Сделана выборка датасета. В каждой транзакции оставлен только тот товар, у которого уровень поддержки ниже 0.38. Данные приведены к

матричному виду с помощью *mlxtend.preporcessing.TransactionEncoder*.

11. Проведен анализ apriori с минимальной поддержкой 0.3 для полученного датасета.

	support	itemsets		support	itemsets
0	0.374890	(all- purpose)	12	0.362599	(paper towels)
1	0.374890	(beef)	13	0.371378	(pasta)
2	0.367867	(butter)	14	0.355575	(pork)
3	0.379280	(coffee/tea)	15	0.367867	(sandwich bags)
4	0.352941	(flour)	16	0.349429	(sandwich loaves)
5	0.370500	(fruits)	17	0.368745	(shampoo)
6	0.345917	(hand soap)	18	0.379280	(soap)
7	0.375768	(individual meals)	19	0.373134	(spaghetti sauce)
8	0.376646	(juice)	20	0.360843	(sugar)
9	0.371378	(ketchup)	21	0.378402	(toilet paper)
10	0.378402	(laundry detergent)	22	0.369622	(tortillas)
11	0.375768	(mixes)			

Таблица 9. Apriori с поддержкой 0.3 для тех продуктов, у которых поддержка меньше 0.38.

12. Написано правило вывода только тех наборов, в которых есть хотя бы два товара, начинающиеся на «*s*»

	support	itemsets		support	itemsets
675	0.137840	(sandwich bags, sandwich loaves)	1351	0.115013	(vegetables, sandwich bags, sandwich loaves)
676	0.146620	(sandwich bags, shampoo)	1352	0.122915	(vegetables, sandwich bags, shampoo)
677	0.158911	(soap, sandwich bags)	1353	0.129939	(soap, vegetables, sandwich bags)
678	0.162423	(soda, sandwich bags)	1354	0.129061	(soda, vegetables, sandwich bags)
679	0.147498	(sandwich bags, spaghetti sauce)	1355	0.123793	(vegetables, sandwich bags, spaghetti sauce)
680	0.131694	(sandwich bags, sugar)	1356	0.113257	(vegetables, sandwich bags, sugar)
686	0.150132	(shampoo, sandwich loaves)	1361	0.129061	(vegetables, shampoo, sandwich loaves)
687	0.158033	(soap, sandwich loaves)	1362	0.132572	(soap, vegetables, sandwich loaves)
688	0.141352	(soda, sandwich loaves)	1363	0.121159	(soda, vegetables, sandwich loaves)
689	0.150132	(sandwich loaves, spaghetti sauce)	1364	0.122915	(vegetables, sandwich loaves, spaghetti sauce)
690	0.136962	(sugar, sandwich loaves)	1365	0.121159	(vegetables, sugar, sandwich loaves)
696	0.151010	(soap, shampoo)	1370	0.124671	(soap, vegetables, shampoo)
697	0.150132	(soda, shampoo)	1371	0.128183	(soda, vegetables, shampoo)
698	0.139596	(shampoo, spaghetti sauce)	1372	0.117647	(vegetables, shampoo, spaghetti sauce)
699	0.147498	(sugar, shampoo)	1373	0.122037	(vegetables, sugar, shampoo)
705	0.174715	(soap, soda)	1378	0.141352	(soap, vegetables, soda)
706	0.160667	(soap, spaghetti sauce)	1379	0.136962	(soap, vegetables, spaghetti sauce)
707	0.154522	(soap, sugar)	1380	0.127305	(soap, vegetables, sugar)
713	0.167691	(soda, spaghetti sauce)	1385	0.138718	(soda, vegetables, spaghetti sauce)
714	0.162423	(soda, sugar)	1386	0.136084	(soda, vegetables, sugar)
720	0.144864	(sugar, spaghetti sauce)	1391	0.124671	(vegetables, sugar, spaghetti sauce)

Таблица 10. Apriori с поддержкой 0.1. Выведены наборы, которые содержат хотя бы два продукта, начинающиеся на «s»

13. Написано правило для вывода наборов с поддержкой от 0.1 до 0.25.

	support	itemsets
38	0.157155	(all- purpose, aluminum foil)
39	0.150132	(bagels, all- purpose)
40	0.144864	(beef, all- purpose)
41	0.147498	(butter, all- purpose)
42	0.151010	(all- purpose, cereals)
...	...	...
1401	0.135206	(vegetables, waffles, toilet paper)
1402	0.130817	(vegetables, yogurt, toilet paper)
1403	0.121159	(tortillas, vegetables, waffles)
1404	0.130817	(tortillas, vegetables, yogurt)
1405	0.146620	(vegetables, waffles, yogurt)

Таблица 11. Частичный вывод наборов, у которых поддержка находится в промежутке [0.1, 0.25].

Вывод

В ходе лабораторной работы были изучен алгоритм частотного анализа *Apriori* из библиотеки *MLxtend*.

Apriori позволяет выделить наиболее частые наборы в выборках данных.

Для работы с алгоритмом *Apriori* было необходимо применить преобразование транзакций с помощью функции *mlxtend.preprocessing.TransactionEncoder*.

Было проведено исследование алгоритма *Apriori* на тестовых данных. Параметр минимальной поддержки позволяет указать условие для отбора данных.