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

Студентка гр. 8303 _____ Самойлова А.С.

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:

	0	1	2
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

22343 rows × 3 columns

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

```
unique_id = list(set(all_data[1]))
print(len(unique_id))
```

```
1139
```

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

```
items = list(set(all_data[2]))
print(len(items))
```

```
38
```

5. Сформировать датасет для частотного анализа.

   Каждой транзакции сопоставлен список товаров

	0	1	2	3	4	5	6	7	8	9	...	24	25	26	
0	yogurt	pork	sandwich bags	lunch meat	all-purpose	flour	soda	butter	vegetables	beef	...	None	None	None	
1	toilet paper	shampoo	hand soap	waffles	vegetables	cheeses	mixes	milk	sandwich bags	laundry detergent	...	None	None	None	
2	soda	pork	soap	ice cream	toilet paper	dinner rolls	hand soap	spaghetti sauce	milk	ketchup	...	spaghetti sauce	pork	vegetables	ch
3	cereals	juice	lunch meat	soda	toilet paper	all-purpose	None	None	None	None	...	None	None	None	
4	sandwich loaves	pasta	tortillas	mixes	hand soap	toilet paper	vegetables	vegetables	paper towels	vegetables	...	all-purpose	soda	yogurt	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1134	sugar	beef	sandwich bags	hand soap	paper towels	paper towels	all-purpose	beef	fruits	coffee/tea	...	beef	cereals	juice	p
1135	coffee/tea	dinner rolls	lunch meat	spaghetti sauce	pasta	vegetables	cereals	dinner rolls	soap	milk	...	None	None	None	
1136	beef	lunch meat	eggs	poultry	vegetables	tortillas	beef	beef	individual meals	dishwashing liquid/detergent	...	vegetables	pork	None	
1137	sandwich bags	ketchup	milk	poultry	cheeses	soap	toilet paper	yogurt	beef	waffles	...	None	None	None	
1138	soda	laundry detergent	vegetables	shampoo	vegetables	None	None	None	None	None	...	None	None	None	

1139 rows × 34 columns

Подготовка данных

6. Представить данные в виде матрицы с помощью *mlxtend.preporcessing.TransactionEncoder*.

7. Вывод полученного датасета:

	all-purpose	aluminum foil	bagels	beef	butter	cereals	cheeses	coffee/tea	dinner rolls	dishwashing liquid/detergent	...	shampoo	soap	soda	spaghetti sauce	sugar	toilet paper	tor
0	True	True	False	True	True	False	False	False	True	False	...	True	True	True	False	False	False	
1	False	True	False	False	False	True	True	False	False	True	...	True	False	False	False	False	True	
2	False	False	True	False	False	True	True	False	True	False	...	True	True	True	True	False	True	
3	True	False	False	False	False	True	False	False	False	False	...	False	False	True	False	False	True	
4	True	False	False	False	False	False	False	False	True	False	...	False	False	True	True	False	True	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1134	True	False	False	True	False	True	True	True	True	True	...	True	True	False	False	True	False	
1135	False	False	False	False	False	True	True	True	True	True	...	False	True	False	True	False	False	
1136	False	False	True	True	False	False	False	False	True	True	...	True	True	False	False	True	False	
1137	True	False	False	True	False	False	True	False	False	False	...	False	True	True	True	True	True	
1138	False	False	False	False	False	False	False	False	False	False	...	True	False	True	False	False	False	

1139 rows × 38 columns

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	length
0	0.374890	(all- purpose)	1
1	0.384548	(aluminum foil)	1
2	0.385426	(bagels)	1
3	0.374890	(beef)	1
4	0.367867	(butter)	1
5	0.395961	(cereals)	1
6	0.390694	(cheeses)	1
7	0.379280	(coffee/tea)	1
8	0.388938	(dinner rolls)	1
9	0.388060	(dishwashing liquid/detergent)	1
10	0.389816	(eggs)	1
11	0.352941	(flour)	1
12	0.370500	(fruits)	1
13	0.345917	(hand soap)	1
14	0.398595	(ice cream)	1
15	0.375768	(individual meals)	1
16	0.376646	(juice)	1
17	0.371378	(ketchup)	1
18	0.378402	(laundry detergent)	1
19	0.395083	(lunch meat)	1
20	0.380158	(milk)	1
21	0.375768	(mixes)	1
22	0.362599	(paper towels)	1
23	0.371378	(pasta)	1
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
38	0.310799	(vegetables, aluminum foil)	2
39	0.300263	(bagels, vegetables)	2
40	0.310799	(vegetables, cereals)	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	(vegetables, eggs)	2
45	0.302897	(vegetables, ice cream)	2
46	0.309043	(laundry detergent, vegetables)	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

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

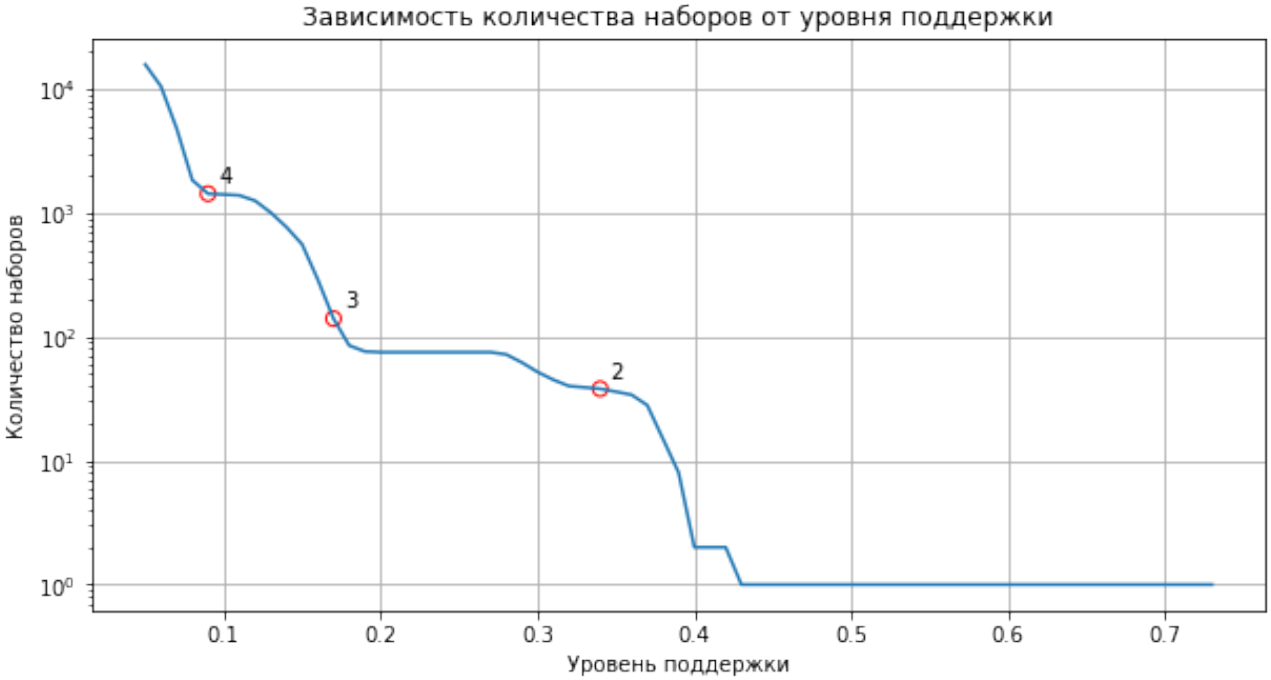
	support	itemsets
0	0.374890	(all- purpose)
1	0.384548	(aluminum foil)
2	0.385426	(bagels)
3	0.374890	(beef)
4	0.367867	(butter)
5	0.395961	(cereals)
6	0.390694	(cheeses)
7	0.379280	(coffee/tea)
8	0.388938	(dinner rolls)
9	0.388060	(dishwashing liquid/detergent)
10	0.389816	(eggs)
11	0.352941	(flour)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

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

	support	itemsets	length
38	0.310799	(vegetables, aluminum foil)	2
39	0.300263	(bagels, vegetables)	2
40	0.310799	(vegetables, cereals)	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	(vegetables, eggs)	2
45	0.302897	(vegetables, ice cream)	2
46	0.309043	(laundry detergent, vegetables)	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

```
Count of result itemstes =  14
```

4. Определение графика зависимости количества набора от минимальной поддержки:



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:

Длина набора	Минимальная поддержка
4	0.09
3	0.17
2	0.34

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*.

	aluminum foil	bagels	cereals	cheeses	dinner rolls	dishwashing liquid/detergent	eggs	ice cream	lunch meat	milk	poultry	soda	vegetables	waffles	yogurt
0	True	False	False	False	True	False	False	True	True	False	False	True	True	False	True
1	True	False	True	True	False	True	False	False	False	True	False	False	True	True	True
2	False	True	True	True	True	False	True	True	True	True	True	True	True	False	False
3	False	False	True	False	False	False	False	False	True	False	False	True	False	False	False
4	False	False	False	False	True	False	True	False	False	True	True	True	True	True	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1134	False	False	True	True	True	True	False	True	False	False	True	False	False	False	False
1135	False	False	True	True	True	True	True	False	True	True	True	False	True	False	False
1136	False	True	False	False	True	True	True	False	True	False	True	False	True	False	True
1137	False	False	False	True	False	False	False	False	False	True	True	True	True	True	True
1138	False	False	False	False	False	False	False	False	False	False	False	True	True	False	False

1139 rows × 15 columns

8. Проведение ассоциативного анализа нового датасета при минимальное поддержке 0.3:

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

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

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

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*.

	all-purpose	beef	butter	coffee/tea	flour	fruits	hand soap	individual meals	juice	ketchup	...	pasta	pork	sandwich bags	sandwich loaves	shampoo	soap	spaghetti sauce	su
0	True	True	True	False	True	False	False	False	False	False	...	False	True	True	False	True	True	False	F
1	False	False	False	False	False	False	True	True	False	False	...	False	False	True	False	True	False	False	F
2	False	False	False	False	False	False	True	False	False	True	...	False	True	False	True	True	True	True	F
3	True	False	False	False	False	False	False	False	True	False	...	False	False	False	False	False	False	False	F
4	True	False	False	False	True	False	True	True	False	False	...	True	True	False	True	False	False	True	F
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1134	True	True	False	True	False	True	True	False	True	False	...	False	True	True	False	True	True	False	
1135	False	False	False	True	False	False	True	True	False	False	...	True	False	False	False	False	True	True	F
1136	False	True	False	False	False	False	True	True	True	False	...	False	True	False	False	True	True	False	
1137	True	True	False	False	False	False	False	False	False	True	...	False	False	True	False	False	True	True	
1138	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	True	False	False	F

1139 rows × 23 columns

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

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

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

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

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

1331 rows × 2 columns

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.