Отчет по первой лабораторной

В рамках первой работы требовалось реализовать два алгоритма поиска ассоциативных правил – Apriori и FpGrowth. Алгоритмы были реализованы на языке Python.

Сравнение работы алгоритмов проводились на датасетах, взятых из источника: https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery . Данные были предоставлены из пекарни, всего было проверено 4 файла, содержащих 1000, 5000, 25000 и 75000 транзакции. Всего в транзакциях участвовало 50 следующих товаров:

1. Шоколадный торт (цена 8.95)
2. Лимонный торт (8.95)
3. Торт «Казино» (15.95)
4. Торт «Опера» (15.95)
5. Клубничный торт (11.95)
6. Трюфельный торт (15.95)
7. Шоколадный эклер (3.25)
8. Кофейный эклер (3.5)
9. Ванильный эклер (3.25)
10. Торт «Наполеон» (13.49)
11. Миндальный тарт (3.75)
12. Яблочный пирог (5.25)
13. Яблочный тарт (3.25)
14. Абрикосовый тарт (3.25)
15. Ягодный тарт (3.25)
16. Ежевичный тарт (3.25)
17. Черничный тарт (3.25)
18. Шоколадный тарт (3.75)
19. Вишневый тарт (3.25)
20. Лимонный тарт (3.25)
21. Кленовый тарт (3.75)
22. Печенье “Ganache” (1.15)
23. Печение “Gongolais” (1.15)
24. Клюквенное печенье (1.09)
25. Лимонное печенье (0.79)
26. Шоколадная меринга (1.25)
27. Ванильная меринга (1.15)
28. Марципановое печенье (1.25)
29. Печенье “Tuile” (1.25)
30. Печенье с грецким орехом (0.79)
31. Миндальный круассан (1.45)
32. Яблочный круассан (1.45)
33. Абрикосовый круассан (1.45)
34. Сырный круассан (1.75)
35. Шоколадный круассан (1.75)
36. Абрикосовая трубочка (1.15)
37. Яблочная трубочка (1.15)
38. Миндальная трубочка (1.15)
39. Миндальный пирог (1.95)
40. Черничная трубочка (1.15)
41. Лимонный лимонад (3.25)
42. Клюквенный лимонад (3.25)

43. Апельсиновый сок (2.75)
44. Зеленый чай (1.85)
45. Вода в бутылке (1.80)
46. Горячий кофе (2.15)
47. Шоколадный кофе (2.45)
48. Ванильный фраппучино (3.85)
49. Вишневая сода (1.29)
50. Эспрессо (1.85)

После запуска алгоритмов на представленных датасетах получились следующие результаты:

1) При запуске не самом маленьком наборе данных (1000 транзакций) получилось, что самыми популярными товарами являются марципановое печенье (№28), печенье “Ganache” (№22) и клубничный торт (№5). Но на остальных датасетах наиболее встречающимися товарами оказались шоколадный эклер (№7), марципановое печенье (№28), и вода в бутылке (№45). Кажется, что это вполне отвечает действительности, т.к. включает в себя несколько не самых дорогих, но при этом почти всеми любимых, позиций. Каждый из этих товаров встречается пример в 1/10 всех покупок.
2) На маленьком датасете – получилось, что чаще всего покупают вместе яблочный пирог (№12), миндальный круассан (№31) и абрикосовую трубочку (№36). Довольно разнообразный набор, думаю, что вполне подходит для людей, который заходят в пекарню по дороге домой или в гости и покупают что-то вкусное. Но на остальных датасетах выявилась другая закономерность – торт «Казино» (№3), шоколадный тарт (№18) и шоколадный круассан (№35). Похоже, что рядом с пекарней живут большие сладкоежки, который покупают кучу шоколада ☺ Но такое обилие его все-таки смущает и кажется, что датасет основан не совсем на реальных данных.

Сравнение времени работы алгоритмов

Параметры minSupport = 0.1, minConfidence = 0.68 (часто встречающихся множеств = ~3)

	Apriori	FpGrowth
1000 транзакций	0.078	0.088
5000 транзакций	0.46	0.388
20000 транзакций	1.83	1.54
75000 транзакций	5.966	6.25

Параметры minSupport = 0.04, minConfidence = 0.68 (часто встречающихся множеств = ~60)

	Apriori	FpGrowth
1000 транзакций	0.56	0.132
5000 транзакций	3.038	0.616
20000 транзакций	15.924	2.548
75000 транзакций	62.878	9.643

Видно, что в тех случаях, когда заданы слишком высокие значения параметров, разницы во времени работы алгоритмов практически нет и можно использовать любой из них. Но в тех случаях, когда большое число множеств разных размеров имеют больше заданного значения поддержки, алгоритм fpgrowth показывает в разы большую скорость работы