Министерство науки и высшего образования Российской Федерации

**Федеральное государственное автономное образовательное учреждение высшего образования**

**«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(национальный исследовательский университет)**

Высшая школа электроники и компьютерных наук

Кафедра «Системное программирование»

**ОТЧЁТ**

по лабораторной работе

«Поиск частых наборов»

|  |  |
| --- | --- |
|  | Выполнил  Студент группы КЭ-404  \_\_\_\_\_\_\_\_\_\_\_\_\_ Р.В. Биленко  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2022 г. |
|  | Проверил  \_\_\_\_\_\_\_\_\_\_\_\_\_ Я.А. Краева  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_2022 г. |

Челябинск 2022

ОГЛАВЛЕНИЕ

[1 ФОРМУЛИРОВКА ЗАДАНИЯ 3](#_Toc272770587)

[2 ГИПЕРССЫЛКИ НА КАТАЛОГ РЕПОЗИТОРИЯ С ИСХОДНЫМИ ТЕКСТАМИ, НАБОРАМИ ДАННЫХ И СОПУТСВУЮЩИМИ МАТЕРИАЛАМИ 4](#_Toc783233073)

[3 РЕЗУЛЬТАТЫ ВИЗУАЛИЗАЦИИ 5](#_Toc1593273493)

[4 ПОЯСНЕНИЯ РАСКРЫВАЮЩИЕ СМЫСЛ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ 7](#_Toc1322072667)

# 1 ФОРМУЛИРОВКА ЗАДАНИЯ

Заданием данной лабораторной работы является разработка программы, которая осуществляет поиск частых наборов объектов в заданном наборе данных с помощью алгоритма Apriori.

Список результирующих наборов должен содержать как наборы, так и значение поддержки для каждого набора. Параметрами программы являются набор, порог поддержки и способ упорядочивания результирующего списка наборов (по убыванию значения поддержки или лексикографическое).

# 2 ГИПЕРССЫЛКА НА КАТАЛОГ РЕПОЗИТОРИЯ С ИСХОДНЫМИ ТЕКСТАМИ, НАБОРАМИ ДАННЫХ И СОПУТСВУЮЩИМИ МАТЕРИАЛАМИ

Репозиторий github с кодами и датасетами:

<https://github.com/romanbilenko174/L1-Analysis>

В данной работе были использованы два набора данных.

1. Датасет покупок в супермаркете (Groceries-dataset): <https://www.kaggle.com/heeraldedhia/groceries-dataset>
2. Датасет «Lastfm»: <https://www.biz.uiowa.edu/faculty/jledolter/DataMining/lastfm.csv>

Также была импортирована библиотека apyori (простая реализация алгоритма Apriori для python) – Apyori: <https://github.com/ymoch/apyori>

# 3 РЕЗУЛЬТАТЫ ВИЗУАЛИЗАЦИИ

Для датасета «Groceries-dataset» визуализация зависимости времени работы и количества наборов от значения поддержки продемонстрированы на рисунках 1 и 2 соответственно.

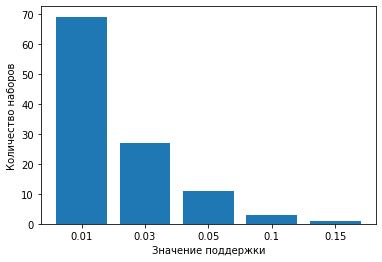


Рисунок 1 – визуализация зависимостей для «Groceries-dataset»

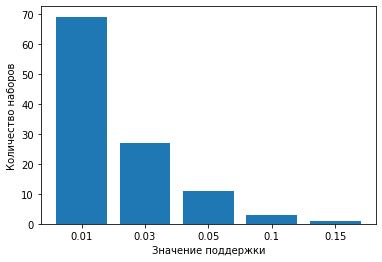


Рисунок 2 – визуализация зависимостей для «Groceries-dataset»

Для датасета «Lastfm» результаты визуализации зависимости времени работы и количества наборов от значения поддержки представлены на рисунках 3 и 4 соответственно.

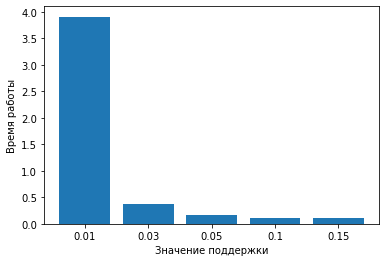


Рисунок 3 – визуализация зависимостей для «Lastfm»

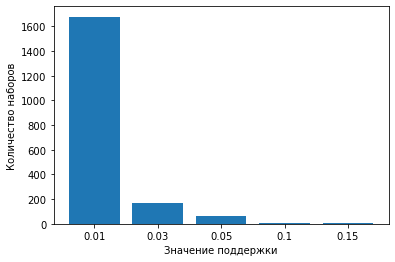


Рисунок 4 – визуализация зависимостей для «Groceries-dataset»

# 4 ПОЯСНЕНИЯ РАСКРЫВАЮЩИЕ СМЫСЛ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

В результате полученных результатов можно утверждать, что чем меньше значение поддержки, тем большее количество множеств мы получим, это связано с тем что такому условию будут удовлетворять большее число наборов, среди всех данных. Из чего можно сделать вывод, что если мы хотим узнать лишь самые частые наборы, то и значение поддержки необходимо устанавливать больше.

Также по представленным графикам видно, что при уменьшении значения задержки, сильно увеличивается количество наборов, что сильно сказывается на времени работы программы. Ведь поскольку при нахождении наборов, удовлетворяющих условиям, выполняется добавление значений в массив, что не может не сказываться на быстродействии. При этом сам алгоритм apyori работает достаточно быстро. Из чего можно сделать вывод для того чтобы программа и алгоритм работали достаточно быстро, необходимо задавать адекватное значение поддержки, которое не было бы слишком малым для конкретного набора данных, потому что это приведет к падению скорости работы. Например, для датасета «Lastfm» правильнее сразу начинать со значения задержки 0.01 и выше.