ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

Департамент программной инженерии

|  |  |
| --- | --- |
| СОГЛАСОВАНО  Научный руководитель  старший преподаватель  департамента больших данных и информационного поиска  факультета компьютерных наук  С.Ю. Папулин  «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 г. | УТВЕРЖДЕНО  Академический руководитель  образовательной программы  «Программная инженерия»  профессор департамента программной инженерии, канд. техн. наук  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.В. Шилов  «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 г. |

|  |  |
| --- | --- |
| Подп. и дата |  |
| Инв. № дубл. |  |
| Взам. инв. № |  |
| Подп. и дата |  |
| Инв. № подл |  |

**Реализация библиотеки гистограммной модели под Apache Spark**

**Пояснительная записка**

**ЛИСТ УТВЕРЖДЕНИЯ**

**RU.17701729.503100-01 81 01-1-ЛУ**

|  |  |
| --- | --- |
|  | Исполнитель  студент группы БПИ152  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ /П.И. Данилин/  «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 г. |

УТВЕРЖДЕН

RU.17701729.503100-01 81 01-1-ЛУ

**Реализация библиотеки гистограммной модели под Apache Spark**

|  |  |
| --- | --- |
| Подп. и дата |  |
| Инв. № дубл. |  |
| Взам. инв. № |  |
| Подп. и дата |  |
| Инв. № подл |  |

**Пояснительная записка**

**RU.17701729.503100-01 81 01-1**

**Листов 23**

# **Аннотация**

В данном программном документе приведено техническое задание для «Реализация библиотеки гистограммной модели под Apache Spark». Данный программный продукт предназначен для организации вычислений, основанных на гистограммной модели. Он даст возможность пользователю провести анализ массива данных с точки зрения упомянутой модели на кластере компьютеров, совместимых с Apache Spark.

Оформление программного документа «Пояснительная записка» произведено по требованиям ГОСТ 19.404-79 «Пояснительная записка. Требования к содержанию и оформлению»[7].

**Содержание**

[**Аннотация** 3](#_Toc510909288)

[**1.** **Введение** 5](#_Toc510909289)

[**1.1. Наименование разработки** 5](#_Toc510909290)

[**1.2. Документы, на основании которых ведется разработка** 5](#_Toc510909291)

[**2.** **Назначение и область применения** 6](#_Toc510909292)

[**2.1. Назначение программы** 6](#_Toc510909293)

[**2.2. Область применения программы** 6](#_Toc510909294)

[**3.** **Технические характеристики** 7](#_Toc510909295)

[**3.1. Постановка задачи для разработки программы** 7](#_Toc510909296)

[**3.2. Описание алгоритмов программы** 7](#_Toc510909297)

[**3.2.1. Описание алгоритма создания гистограммы** 7](#_Toc510909298)

[**3.3. Описание и обоснование выбора метода организации входных и выходных данных** 7](#_Toc510909299)

[**3.3.1. Описание метода организации входных и выходных данных** 7](#_Toc510909300)

[**3.3.2. Обоснования выбора метода организации входных и выходных данных** 7](#_Toc510909301)

[**3.4. Описание и обоснование выбора состава технических и программных средств** 8](#_Toc510909302)

[**3.4.1. Состав технических и программных средств** 8](#_Toc510909303)

[**4.** **Ожидаемые технико-экономические показатели** 9](#_Toc510909304)

[**1.1.** **Предполагаемая потребность** 9](#_Toc510909305)

[**5.** **Список использованной литературы** 10](#_Toc510909306)

[**Приложение 1** 11](#_Toc510909307)

[**Терминология** 11](#_Toc510909308)

[**Приложение 2** 12](#_Toc510909309)

[**Описание и функциональное назначение полей и методов** 12](#_Toc510909310)

# **Введение**

## **1.1. Наименование разработки**

Наименование программного продукта – «Реализация библиотеки гистограммной модели под Apache Spark».

Наименование программного продукта на английском языке – «Spark Library for Logic-Set Histogram».

## **1.2. Документы, на основании которых ведется разработка**

Разработка программы ведется на основании приказа Национального исследовательского университета «Высшая Школа Экономики» от ХХ.ХХ.2017 г. № ХХХ «ХХХХ».

# **Назначение и область применения**

## **2.1. Назначение программы**

Программный продукт предназначен для организации распределенных вычислений на основе гистограммной модели с использованием Apache Spark.

## **2.2. Область применения программы**

Данная разработка позволит анализировать большие массивы данных на основе гистограммной модели на компьютерных кластерах, совместимых с Apache Spark, например, под управлением менеджера кластера Hadoop Yarn. Она даст аналитикам и исследователям еще один инструмент обработки коллекций данных с эксплуатацией мощностей кластера.

# **Технические характеристики**

# **Постановка задачи для разработки программы**

Задача для разработки программы состоит в создании программной реализации гистограммной модели в виде библиотеки для Apache Spark. Библиотека должна предоставлять следующие функции:

1. Инициализацию элементного состава гистограммной модели
2. Преобразование определённых типов данных (например, изображений) в гистограммный вид
3. Формирование элементного высказывания из элементного запроса пользователя
4. Формирование гистограммного высказывания из элементного с учетом гистограммы данных
5. Вычисление значения гистограммного высказывания с применением операций[14]: объединения, пересечения, вычитания, И, ИЛИ, исключающего ИЛИ, КРОМЕ, исключающего КРОМЕ, НЕ
6. Вычисление показателя присутствия элементов из полученного значения гистограммного высказывания
7. Вычисление показателя схожести между данными по их гистограммам
8. Применение вычислений для коллекции данных
9. Сохранение гистограмм данных после преобразования
10. Сохранение результата вычисления показателя присутствия и схожести

## **Описание алгоритмов программы**

## **Описание алгоритма создания гистограммы**

Алгоритм представлен далее (рис. 1):

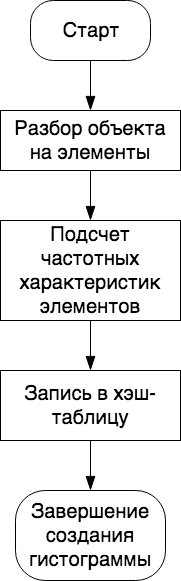


Рисунок 1. Алгоритм создания гистограммы

Сложность: О(количество элементов в объекте)

## **Описание алгоритма вычисления операции И**

Алгоритм представлен далее (рис. 2):

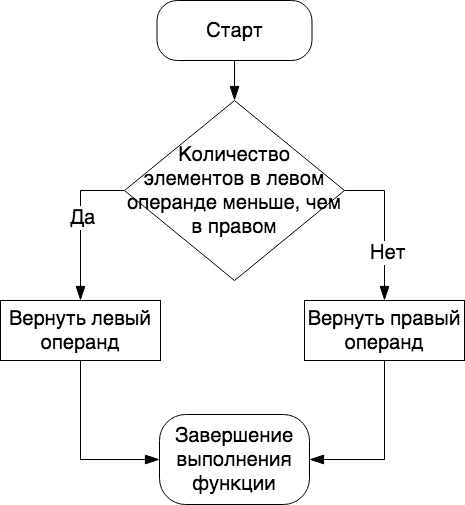


Рисунок 2. Алгоритм вычисления операции И

Сложность: О(Количество различных элементов в гистограммах)

## **Описание алгоритма вычисления операции КРОМЕ**

Алгоритм представлен далее (рис. 3):

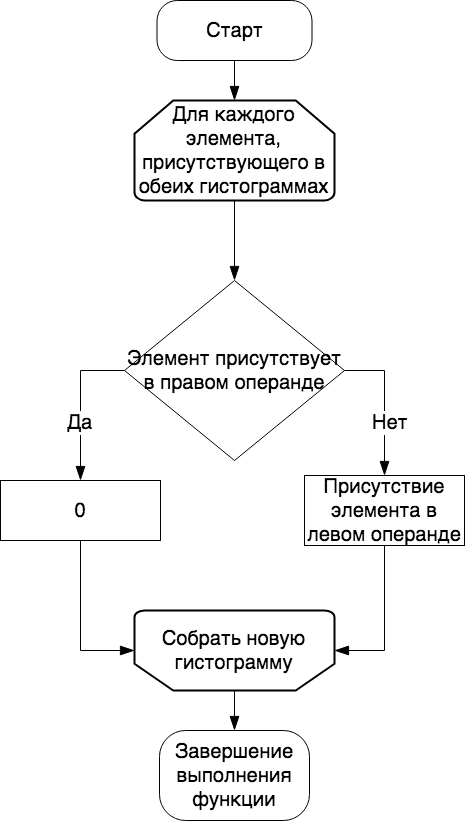


Рисунок 3. Алгоритм вычисления операции КРОМЕ

Сложность: О(Количество различных элементов в гистограмме)

## **Описание алгоритма вычисления операции Вычитание**

Аналогично операции КРОМЕ (см. п. 3.2.3).

## **Описание алгоритма вычисления операции Пересечение**

Алгоритм представлен далее (рис. 4):

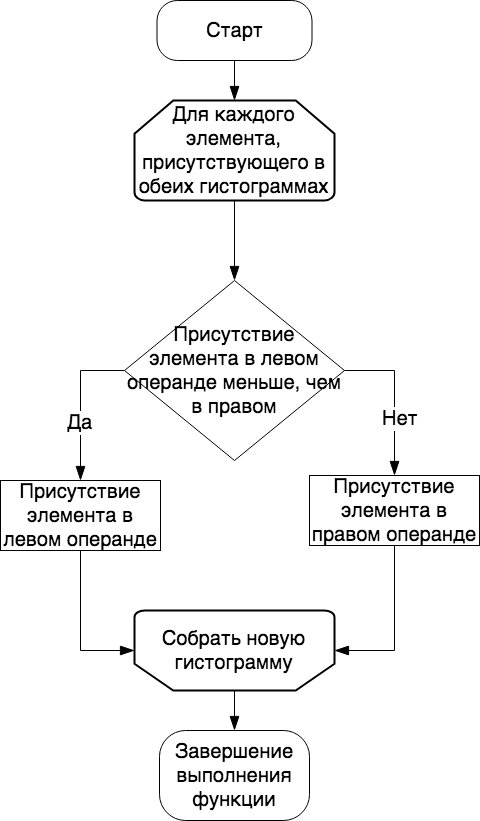


Рисунок 4. Алгоритм вычисления операции Пересечение

Сложность: О(Количество различных элементов в гистограмме)

## **Описание алгоритма вычисления операции НЕ**

Алгоритм представлен далее (рис. 5):



Рисунок 5. Алгоритм вычисления операции НЕ

Сложность: О(операции КРОМЕ)

## **Описание алгоритма вычисления операции Объединение**

Алгоритм представлен далее (рис. 6):

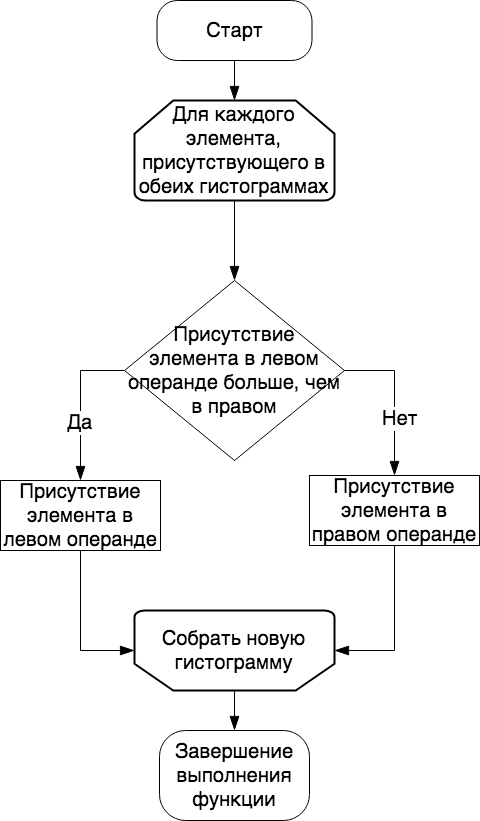


Рисунок 6. Алгоритм вычисления операции Объединение

Сложность: О()

## **Описание алгоритма вычисления операции ИЛИ**

Аналогично операции Объединение (см. п. 3.2.7)

## **Описание алгоритма вычисления операции Схожесть**

Алгоритм представлен далее (рис. 7):

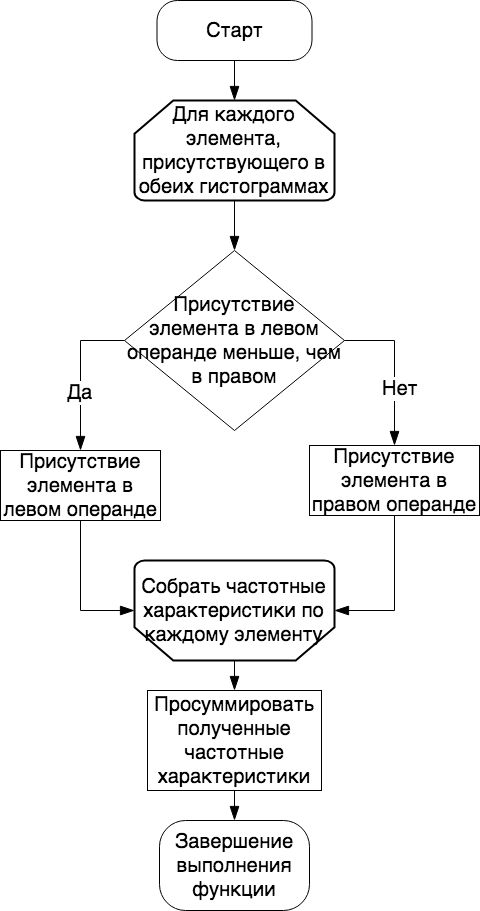


Рисунок 7. Алгоритм вычисления операции Схожесть

Сложность: О()

## **Описание алгоритма вычисления операции ИСКЛЮЧАЮЩЕЕ КРОМЕ**

Алгоритм представлен далее (рис. 8):



Рисунок 8. вычисления операции ИСКЛЮЧАЮЩЕЕ КРОМЕ

Сложность: О()

## **Описание алгоритма вычисления операции ИСКЛЮЧАЮЩЕЕ ИЛИ**

Алгоритм представлен далее (рис. 9):

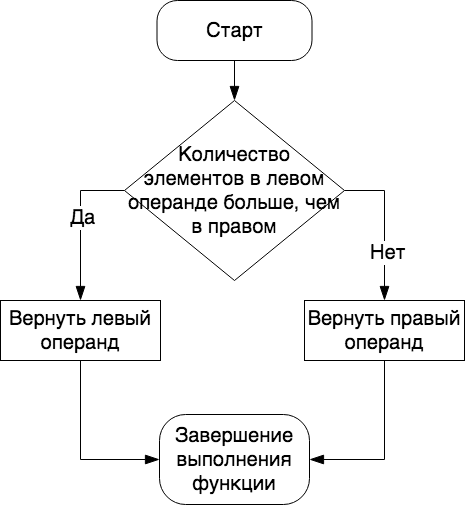


Рисунок 9. Алгоритм вычисления операции ИСКЛЮЧАЮЩЕЕ ИЛИ

Сложность: О()

## **3.3. Описание и обоснование выбора метода организации входных и выходных данных**

### **3.3.1. Описание метода организации входных и выходных данных**

Входные данные должны представлять собойлюбые объекты, для которых был реализован интерфейс Atomizer для разбора объекта на элементы и предоставлен объект, реализующий данный интерфейс.

Выходные данные программы представляют собой объекты, реализующие интерфейс Histogram. Также возможна JSON сериализация полученных гистограмм.

### **3.3.2. Обоснования выбора метода организации входных и выходных данных**

Входные данные представляют собой реализацию минимально необходимого для создания гистограммы объектов.

Выходные данные представляют собой реализацию поставленной задачи в виде требуемых объектов гистограмм в бинарном для JVM формате, либо в JSON сериализованном виде.

## **3.4. Описание и обоснование выбора состава технических и программных средств**

### **3.4.1. Состав технических и программных средств**

Для эффективной работы срограммой предъявляются следующие требования к составу и параметрам технических средств:

1. 8-ядерный процессор
2. Оперативная память не менее 8 Гб
3. 4 жестких диска со свободным объемом не менее 500Гб
4. Стабильное интернет соединение, желательно 10 Гбит/сек
5. Монитор
6. Клавиатура
7. Мышь

Для корректной работы программного продукта требуется установленный программный комплекс Apache Spark версии 2.2.0 со всеми зависимостями.

**3.4.2. Обоснование выбора технических и программных средств**

Выбор программных средств связан с постановленными задачами при разработке.

Технические требования связаны с рекомендованными характеристиками для эффективной работы Apache Spark, необходимого по постановке задачи.

# **Ожидаемые технико-экономические показатели**

## **Предполагаемая потребность**

* 1. **Ориентировочная экономическая эффективность**

# **Список использованной литературы**

1. ГОСТ 19.101-77 Виды программ и программных документов. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
2. ГОСТ 19.102-77 Стадии разработки. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
3. ГОСТ 19.103-77 Обозначения программ и программных документов. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
4. ГОСТ 19.104-78 Основные надписи. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
5. ГОСТ 19.105-78 Общие требования к программным документам. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
6. ГОСТ 19.106-78 Требования к программным документам, выполненным печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
7. ГОСТ 19.404-79 Пояснительная записка. Требования к содержанию и оформлению. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
8. ГОСТ 19.603-78 Общие правила внесения изменений. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
9. ГОСТ 19.604-78 Правила внесения изменений в программные документы, выполненные печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001.
10. ГОСТ Р 7.02-2006 Консервация документов на компакт-дисках. Общие требования. – М.: ИПК Издательство стандартов, 2006
11. ГОСТ 18300-87 Спирт этиловый ректификованный технический. Технические условия. – М.:ИПК Издательство стандартов, 1997
12. ГОСТ 9805-84. Спирт изопропиловый. Технические условия. – М.: ИПК Издательство стандартов, 1984.
13. ГОСТ 19.602-78 Правила дублирования, учета и хранения программных документов, выполненных печатным способом. //Единая система программной документации. – М.: ИПК Издательство стандартов, 2001

# **Приложение 1**

# **Терминология**

# **Приложение 2**

# **Описание и функциональное назначение полей и методов**

Таблица 2.1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ** | | | | | | | | | |
| Изм. | Номера листов (страниц) | | | | Всего  листов  (страниц)  в докум. | № докум. | Входящий № сопроводитель-ного документа и дата | Подпись | Дата |
| изме-ненных | заме-  ненных | новых | анну-  лиро-  ванных |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |