**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**

**Техническое задание**

**ЛИСТ УТВЕРЖДЕНИЯ**

**RU.17701729.503100-01 ТЗ 01-1-ЛУ**

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

УТВЕРЖДЕН

RU.17701729. 503100-01 ТЗ 01-1-ЛУ

**Реализация библиотеки гистограммной модели под Apache Spark**

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

**Техническое задание**

**RU.17701729.503100-01 ТЗ 01-1**

**Листов 18**

# 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 «Техническое задание» произведено по требованиям ГОСТ 19.201-78» Техническое задание. Требования к содержанию и оформлению»[7].

**Содержание**

[Аннотация 2](#_Toc498276890)

[1. Введение 5](#_Toc498276891)

[1.1. Наименование разработки 5](#_Toc498276892)

[1.2. Краткая характеристика области применения программного продукта 5](#_Toc498276893)

[2. Основания для разработки 6](#_Toc498276894)

[2.1. Документы, на основании которых ведется разработка 6](#_Toc498276895)

[3. Назначение разработки 7](#_Toc498276896)

[3.1. Функциональное назначение 7](#_Toc498276897)

[3.2. Эксплуатационное назначение 7](#_Toc498276898)

[4. Требования к программе 8](#_Toc498276899)

[4.1. Требования к функциональным характеристикам программы 8](#_Toc498276900)

[4.2. Требование к входным данным 8](#_Toc498276901)

[4.3. Требования к выходным данным 8](#_Toc498276902)

[4.4. Требования к надежности программы 8](#_Toc498276903)

[4.5. Требования квалификация и уровня подготовки пользователя 9](#_Toc498276904)

[4.6. Требования к составу и параметрам технических средств 9](#_Toc498276905)

[4.7. Требования к информационной и программной совместимости 9](#_Toc498276906)

[4.8. Требования к маркировке и упаковке 9](#_Toc498276907)

[4.9. Требования к транспортированию и хранению 9](#_Toc498276908)

[4.9.1 Требования к хранению и транспортировке компакт-дисков (CD) 9](#_Toc498276909)

[4.9.1 Требования к хранению и транспортировке программных документов, предоставляемых в печатном виде 10](#_Toc498276910)

[4.10. Специальные требования 10](#_Toc498276911)

[5. Требования к программной документации 11](#_Toc498276912)

[5.1. Состав программной документации 11](#_Toc498276913)

[6. Технико-экономические показатели 12](#_Toc498276914)

[6.1. Предполагаемая потребность 12](#_Toc498276915)

[6.2. Ориентировочная экономическая эффективность 12](#_Toc498276916)

[6.3. Преимущества разработки по сравнению с отечественными и зарубежными аналогами 12](#_Toc498276917)

[7. Стадии и этапы разработки 13](#_Toc498276918)

[7.1. Этапы разработки 13](#_Toc498276919)

[7.2. Сроки разработки и исполнители 13](#_Toc498276920)

[8. Порядок контроля и приемки 14](#_Toc498276921)

[9. Список использованной литературы 15](#_Toc498276922)

[Приложение 1. Терминология 16](#_Toc498276923)

# 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. Краткая характеристика области применения программного продукта

Разрабатываемую библиотеку можно применить к анализу данных, представленных в виде гистограмм, для определения количественного значения соответствия между данными и запросом пользователя. В качестве запроса могут быть использованы текстовое описание или образец данных. Запрос формируется с использованием функций гистограммной модели и в результате возвращает агрегированные значения, которые пользователь может проанализировать или использовать в последующей обработке.

# Основания для разработки

## 2.1. Документы, на основании которых ведется разработка

Разработка программы ведется на основании приказа Национального исследовательского университета «Высшая Школа Экономики» от 12.12.2018 г. № 2.3-02/1212-01.

# Назначение разработки

## 3.1. Функциональное назначение

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

## 3.2. Эксплуатационное назначение

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. Также пользователь сможет сохранить результаты обработки для последующих манипуляций, например, в распределенной файловой системы HDFS.

# Требования к программе

## 4.1. Требования к функциональным характеристикам программы

Создаваемый программный модуль должен предоставлять программный интерфейс к следующим функциям:

1. Создание объекта 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. Сохранение гистограмм данных после преобразования
9. Сохранение результата вычисления показателя присутствия и схожести

## 4.2. Требование к входным данным

К входным данным программного модуля предъявляются следующие требования:

1. Запрос должен представлять собой корректный запрос с точки зрения гистограммной модели
2. Гистограммное представление должно быть корректным экземпляром класса представления, который может быть получен только в результате работы разрабатываемого модуля, либо иметь возможность приведения к нему (реализованную пользователем)

## 4.3. Требования к выходным данным

Разрабатываемый программный продукт должен возвращать результаты операций в виде коллекции объектов внутренних представлений гистограмм, либо агрегированных значений результатов в случае вычисления схожести. Выходные данные должны иметь возможность сериализации и десериализации с целью последующего сохранения на внешнем носителе или в распределенной файловой системе.

## 4.4. Требования к надежности программы

К программе предъявляются следующие требования надежности:

Программа должна осуществлять проверку корректности введенных пользователем данных и выводить сообщение об ошибке при ее нахождении.

Программа должна корректно завершаться при возникновении ошибок.

## 4.5. Требования квалификация и уровня подготовки пользователя

Пользователь должен иметь базовые навыки программирования с использованием Apache Spark.

Обязательно ознакомление пользователя с руководством оператора.

## 4.6. Требования к составу и параметрам технических 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

## 4.7. Требования к информационной и программной совместимости

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

## 4.8. Требования к маркировке и упаковке

Программный модуль поставляется в составе проекта «Реализация библиотеки гистограммной модели под Apache Spark» на внешнем носителе информации – компакт диске (CD), на котором должны содержаться программная документация, приложение (исполняемые файлы и необходимые для работы программы файлы). Программное изделие должно иметь маркировку с обозначением наименования изделия, темы разработки, фамилии, имени и отчества исполнителей, учебной группы и года выпуска изделия.

## 4.9. Требования к транспортированию и хранению

### 4.9.1 Требования к хранению и транспортировке компакт-дисков (CD)

Программа поставляется заказчику на внешнем носителе информации – компакт-диске (CD). Документация к программе передается как на компакт-диске вместе с программой, так и в печатном виде.

Требования к транспортировке и хранению компакт-дисков с программным обеспечением являются стандартными и должны соответствовать требованиям ГОСТ Р 7.02-2006 [10]:

1. В помещении для хранения компакт-дисков допустимы температура воздуха от 10°С до 20°С и относительная влажность воздуха от 20% до 65%.
2. Максимальные суточные колебания температуры 2°С и относительной влажности воздуха - 5%.
3. Компакт-диски хранят и используют на расстоянии не менее 0,5 м от источников тепла и влаги.
4. Компакт-диски хранят в темноте или при освещении рассеянным светом, не содержащим ультрафиолетовое излучение.
5. Транспортировка производится вертикально в специальных маркированных контейнерах из безопасных материалов с применением мер по предотвращению ударов контейнеров, перемещению и вибрации компакт-дисков внутри контейнеров, а также проникновения влаги, вредных газов, пыли, солнечных лучей и образованию конденсата внутри контейнеров.
6. Очистка компакт-диска производится путем протирания чистым мягким хлопчатобумажным тампоном без длинного ворса, пропитанным водой, этиловым (ГОСТ 18300-87 [11]) или изопропиловым (ГОСТ 9805-84 [12]) спиртом.

### 4.9.1 Требования к хранению и транспортировке программных документов, предоставляемых в печатном виде

Требования к транспортировке и хранению программных документов являются стандартными и должны соответствовать общим требованиям хранения и транспортировки печатной продукции:

1. В помещении для хранения печатной продукции допустимы температура воздуха от10°С до 30°С и относительная влажность воздуха от 30% до 60%.
2. Документацию хранят и используют на расстоянии не менее 0.5 от источников тепла и влаги. Не допускается хранение печатной продукции в помещениях, где находятся агрессивные агенты – растворители, спирт, бензин.
3. Не допускается попадание на документацию агрессивных агентов.
4. Транспортировка производится в специальных контейнерах с применением мер по предотвращению деформации документов внутри контейнеров, а также проникновения влаги, вредных газов, пыли, солнечных лучей и образованию конденсата внутри контейнеров.
5. Программные документы, предоставляемые в печатном виде должны соответствовать общим правилам учета и хранения программных документов, предусмотренных стандартами Единой системы программной документации и соответствовать требованиям ГОСТ 19.602-78 [13].

## 4.10. Специальные требования

Специальные требования не предъявляются.

# Требования к программной документации

## 5.1. Состав программной документации

1. «Реализация библиотеки гистограммной модели под Apache Spark». Техническое задание. ГОСТ 19.71-208
2. «Реализация библиотеки гистограммной модели под Apache Spark». Пояснительная записка. ГОСТ 19.404-79
3. «Реализация библиотеки гистограммной модели под Apache Spark». Руководство оператора. ГОСТ 19.505-79
4. «Реализация библиотеки гистограммной модели под Apache Spark». Программа и методика испытаний. ГОСТ 19.301-79
5. «Реализация библиотеки гистограммной модели под Apache Spark». Текст программы ГОСТ 19.401-78

# Технико-экономические показатели

## 6.1. Предполагаемая потребность

Данная программа может быть использована исследователями и аналитиками для анализа данных. Обработка с помощью гистограммной модели поможет сделать пользователю определенные выводы о структуре и содержании исходных данных.

## Ориентировочная экономическая эффективность

В рамках данной работы расчёт экономической эффективности не предусмотрен.

## Преимущества разработки по сравнению с отечественными и зарубежными аналогами

На текущий момент гистограммная модель имеет две реализации С.Ю. Папулина на языках программирования C# и Python. Данный проект позволит расширить охват платформ, на которых модель реализована, и таким образом расширить ее доступность, дать пользователям новый инструмент анализа данных.

# Стадии и этапы разработки

## 7.1. Этапы разработки

1. Техническое задание
2. Обоснование необходимости разработки программы:
3. Постановка задачи
4. Сбор исходных для решения задачи
5. Разработка и утверждение технического задания
6. Определение требований к программе
7. Определение стадий, этапов и сроков разработки программы и документации
8. Согласование и утверждение технического задания
9. Технически проект
10. Разработка технического проекта
11. Разработка алгоритма решения задачи
12. Определение формы представления входных и выходных данных
13. Разработка структуры программы
14. Окончательное определение конфигурации технических средств
15. Утверждение технического проекта
16. Разработка плана мероприятий по разработке и внедрению программы
17. Разработка пояснительной записки
18. Согласование и утверждение технического проекта
19. Рабочий проект
20. Разработка программы
21. Реализация гистограммной модели
22. Отладка программы
23. Разработка программной документации
24. Разработка программной документации в соответствии с требованием ГОСТ 19.101-77
25. Испытания программы
26. Разработка, согласование и утверждение программы и методики испытаний
27. Проведение предварительных приемо-сдаточных испытаний
28. Корректировка программы и программной документации по результатам испытаний
29. Внедрение
30. Подготовка и передача программы
31. Подготовка и передача программы и программной документации для сопровождения

## 7.2. Сроки разработки и исполнители

Разработка программы должна быть закончена к 20 марта 2018 года.

Исполнитель студент группы БПИ152 Данилин Павел Иванович.

# Порядок контроля и приемки

Порядок контроля и приемки программы должны проходит согласно «Программе методики и испытаний» (ГОСТ 19.301-79)

# 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 19.201-78 Техническое задание. 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
14. Папулин С.Ю. Анализ коллекции данных посредством логико-множественного гистограммного представления // Программные продукты и системы. 2014. С. 57-60.
15. Apache Spark [Электронный ресурс] // Apache Spark: [сайт]. [2017]. URL: <https:/​/​spark.apache.org>
16. Cистемные требования для Java [Электронный ресурс] // Java: [сайт]. [2017]. URL: <https:/​/​www.java.com/​ru/​download/​help/​sysreq.xml>
17. Jupyter Notebook [Электронный ресурс] // Jupyter: [сайт]. [2018]. URL: https://jupyter.org/
18. Lightweight Scala kernel for Jupyter / IPython 3 [Электронный ресурс] // Github: [сайт]. [2018]. URL: <https://github.com/jupyter-scala/jupyter-scala>
19. Json4s [Электронный ресурс] // Json4s: [сайт]. [2018]. URL: http://json4s.org/

# Приложение 1. Терминология

**Гистограмма элемента -** модель представления данных, элементы которой упорядочены и соответствуют элементам универсального множества, а их значения указывают на количество этих элементов в данных**.**

**Элементарное высказывание –** высказывание, состоящее из элементов универсального множества и операций между ними.

**Гистограммное высказывание -** высказывание, состоящее из гистограммных элементов и операций между ними.

**Данные** – обрабатываемая гистограммной моделью информация.

**Коллекция –** множество однородных данных.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ | | | | | | | | | |
| Изм. | Номера листов (страниц) | | | | Всего листов (страниц) в докум. | № документа | Входящий № сопроводительного документа и дата | Подпись | Дата |
| измененных | замененных | новых | аннулированных |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |