**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** |  |

**БИБЛИОТЕКА HistogramModel**

**Руководство программиста**

**ЛИСТ УТВЕРЖДЕНИЯ**

**RU.17701729.503100-01 33 01-1-ЛУ**

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

УТВЕРЖДЕН

RU.17701729. 503100-01 33 01-1-ЛУ

**БИБЛИОТЕКА HistogramModel**

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

**Руководство программиста**

**RU.17701729.503100-01 33 01-1**

**Листов 19**

**Содержание**

[1. НАЗНАЧЕНИЕ И УСЛОВИЯ ПРИМЕНЕНИЯ БИБЛИОТЕКИ 4](#_Toc511144022)

[1.1. Назначение библиотеки 4](#_Toc511144023)

[1.2. Функции, выполняемые библиотекой 4](#_Toc511144024)

[1.3. Условия, необходимые для выполнения библиотеки 4](#_Toc511144025)

[1.3.1. Требования к составу и параметрам технических средств 4](#_Toc511144026)

[1.3.2. Требования к программным средствам 4](#_Toc511144027)

[2. ХАРАКТЕРИСТИКИ ПРОГРАММЫ 5](#_Toc511144028)

[2.1 Временные характеристики 5](#_Toc511144029)

[2.2 Средства контроля правильности выполнения 5](#_Toc511144030)

[3. ОБРАЩЕНИЕ К БИБЛИОТЕКЕ 6](#_Toc511144031)

[3.1 Создание объекта гистограммной модели 6](#_Toc511144032)

[3.2 Преобразование картинок и текстов в гистограммный вид 6](#_Toc511144033)

[3.3 Формирование элементного высказывания из элементного запроса пользователя 6](#_Toc511144034)

[3.4 Формирование гистограммного высказывания из элементного с учетом гистограммы данных 6](#_Toc511144035)

[3.5 Вычисление значения гистограммного высказывания с применением операций: объединения, пересечения, вычитания, И, ИЛИ, исключающего ИЛИ, КРОМЕ, исключающего КРОМЕ, НЕ 7](#_Toc511144036)

[3.6 Вычисление показателя присутствия элементов из полученного значения гистограммного высказывания 7](#_Toc511144037)

[3.7 Вычисление показателя схожести между данными по их гистограммам 7](#_Toc511144038)

[3.8 Сохранение гистограмм данных после преобразования 7](#_Toc511144039)

[3.9 Сохранение результата вычисления показателя присутствия и схожести 7](#_Toc511144040)

[4. ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ 8](#_Toc511144041)

[5. СООБЩЕНИЯ 9](#_Toc511144042)

[Список использованной литературы 10](#_Toc511144043)

[ПРИЛОЖЕНИЕ 1 11](#_Toc511144044)

# НАЗНАЧЕНИЕ И УСЛОВИЯ ПРИМЕНЕНИЯ БИБЛИОТЕКИ

## Назначение библиотеки

Библиотека предназначена для предоставления реализации гистограммной модели для анализа данных.

## Функции, выполняемые библиотекой

Состав функций, выполняемых библиотекой HistogramModel:

* Создание объекта гистограммной модели
* Преобразование картинок и текстов в гистограммный вид
* Формирование элементного высказывания из элементного запроса пользователя
* Формирование гистограммного высказывания из элементного с учетом гистограммы данных
* Вычисление значения гистограммного высказывания с применением операций[14]: объединения, пересечения, вычитания, И, ИЛИ, исключающего ИЛИ, КРОМЕ, исключающего КРОМЕ, НЕ
* Вычисление показателя присутствия элементов из полученного значения гистограммного высказывания
* Вычисление показателя схожести между данными по их гистограммам
* Сохранение гистограмм данных после преобразования
* Сохранение результата вычисления показателя присутствия и схожести

## Условия, необходимые для выполнения библиотеки

### Требования к составу и параметрам технических средств

Для надёжной и бесперебойной работы библиотеки требуется следующий состав технических средств [10]:

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

### Требования к программным средствам

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

# ХАРАКТЕРИСТИКИ ПРОГРАММЫ

## Временные характеристики

Временные характеристики и алгоритмы работы функций описаны в документе «Пояснительная записка».

## Средства контроля правильности выполнения

Контроль правильности исполнения возможен путем проверки соответствия входных данных библиотеки выходным.

# ОБРАЩЕНИЕ К БИБЛИОТЕКЕ

Для использования библиотеки HistogramModel необходимо подключить ее.

## Создание объекта гистограммной модели

Создание объекта гистограммы осуществляется путем создания объекта класса HistogramImpl вызовом его конструктора или методом объекта HistogramImpl extract:

**val** histogram = *HistogramImpl*(map),

где map (Map[O, Double]) – словарь из множества элементов в количество каждого элемента.

**val** histogram = HistogramImpl.*extract*(source)(atomizer),

где source(S) – объект для разбора,

atomizer (Atomizer[S, O]) – объект, реализующий интерфейс Atomizer для разбора объекта S на элементы O.

## Преобразование картинок и текстов в гистограммный вид

В объекте Implicits находятся соответствующие implicit реализзации интерфейса Atomizer:

import ru.hse.se.ba.danilin.paul.histogram\_model.histogram.Implicits.{imageToPixels, stringToWords}

val histogramText = HistogramImpl.extract(text)

val histogramImage = HistogramImpl.extract(image),

где text, image – соответственно строка и картинка.

## Формирование элементного высказывания из элементного запроса пользователя

Осуществляется путем вызова метода fromString объекта Query:

val query = Query.fromString(stringQuery)(aliases)

stringQuery (String) – строка запроса

aliases (Map[String, Input[E]) – словарь лексем в запросе.

## Формирование гистограммного высказывания из элементного с учетом гистограммы данных

Для получения гистограммного высказывания необходимо вызвать функции, как показано в примере:

val newQuery = Query(q.injectHistogram(histogram))

q (Query) – элементный запрос,

histogram (Histogram[E]) – гистограмма.

## Вычисление значения гистограммного высказывания с применением операций: объединения, пересечения, вычитания, И, ИЛИ, исключающего ИЛИ, КРОМЕ, исключающего КРОМЕ, НЕ

q.execute(histogram)

q (Query) – высказывание

histogram (Histogram[E]) – гистограмма, над которой надо выполнить высказывание.

Также возможно вычисление с использованием интерфейсов Histogram и операций:

val histogram: Histogram[E] = ???

val notHistogram = Not(histogram)

val notHistogram = histogram.not

## Вычисление показателя присутствия элементов из полученного значения гистограммного высказывания

Интерфейс гистограмм позволяет получить присутствие каждого из элементов:

val histogram: Histogram[E] = ???

histogram(element1)

histogram.apply(element1)

где element1 – элемент из множества элементов гистограммы

## Вычисление показателя схожести между данными по их гистограммам

Сравнение схожести двух гистограмм также возможно двумя способами – интерфейсом Histogram и используя объекты операций:

val histogram1: Histogram[E] = ???

val histogram2: Histogram[E] = ???

histogram1.similar(histogram2)

Similar(histogram1, histogram2)

## Сохранение гистограмм данных после преобразования

Сериализация и последующее сохранение осуществляется с помощью json4s[19]:

import io.circe.syntax.\_

import ru.hse.se.ba.danilin.paul.histogram\_model.histogram.Implicits.\_

val json = histogram.asJson

val writer = new PrintWriter(new File(filename))

writer.write(json.spaces2)

writer.close()

## Сохранение результата вычисления показателя присутствия и схожести

Осуществляется средствами языка:

val writer = new PrintWriter(new File(filename))

writer.write(histogram1.similar(histogram2))

writer.close()

# ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ

В качестве входных данных могут быть Map[О, Double], любой объект типа O, для которого будет предоставлена реализация интерфейса Atomizer[O, E] для разбора на элементарные элементы. Выходные данные – результаты операций над гистограммами, сериализованные и сохраненные на диск.

# СООБЩЕНИЯ

Если во время расчетов случается ошибка, библиотека пробрасывает соответствующее исключение.

В случае невозможности разобрать входной запрос, библиотека выбрасывает соответствующее исключение.

# 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]. [2017]. 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 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |