|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ **Информатика, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ и системы  
 управления**

КАФЕДРА **Компьютерные системы и сети (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/07 Интеллектуальные системы анализа,**

**обработки и интерпретации больших данных**

**Отчет**

|  |  |
| --- | --- |
| **по лабораторной работе №** | 10 |

**Название:**

Spark

**Дисциплина:** Языки программирования для работы с большими данными

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент | ИУ6-22М |  |  | А.А. Морозова |
|  | (Группа) |  | (Подпись, дата) | (И.О. Фамилия) |
|  |  |  |  |  |
| Преподаватель |  |  |  | П.В. Степанов |
|  |  |  | (Подпись, дата) | (И.О. Фамилия) |

**Цель работы** — получение навыков работы со Spark с использованием языка Java.

1) Выбрать любой датасет на kaggle.com.

2) Cделать 10 выборок данных по выбранной предметной области.

Был выбран датасет по ссылке - <https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database>.

Для запуска лабораторной работы был использован фреймворк для автоматизации сборки проектов Maven. Конфигурация представлена в листинге 1.

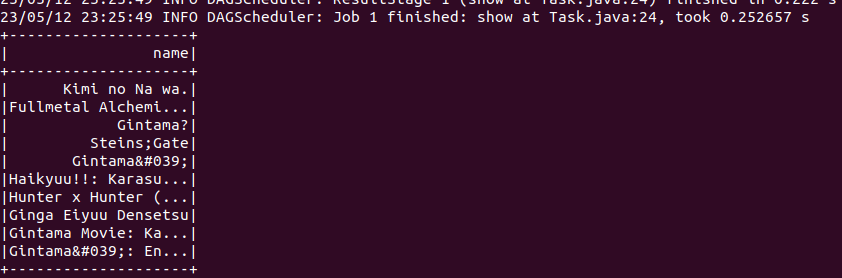
Листинг 1 – pom.xml

|  |
| --- |
| <project>  <groupId>edu.moroz </groupId>  <artifactId>task</artifactId>  <modelVersion>4.0.0</modelVersion>  <name>Task</name>  <packaging>jar</packaging>  <version>1.0</version>  <repositories>  </repositories>  <dependencies>  <dependency>  <groupId>org.apache.spark</groupId>  <artifactId>spark-core\_2.11</artifactId>  <version>2.4.6</version>  </dependency>  <dependency>  <groupId>org.apache.spark</groupId>  <artifactId>spark-sql\_2.11</artifactId> <!-- matching Scala version -->  <version>2.4.6</version> <!-- matching Spark Core version -->  </dependency>  </dependencies>  </project> |

Сборка была произведена и запущена командами mvn package и mvn -e exec:java -Dexec.mainClass="Task". Часть кода в запросами представлена в листинге 2.

Листинг 2 – Запросы к датасету

|  |
| --- |
| spark.sql("select name from anime limit 10").show();  spark.sql("select \* from anime where lower(genre) like '%drama%'").show();  spark.sql("select \* from anime where episodes > 20 limit 10").show();  spark.sql("select \* from anime where members > 100000 and episodes > 24").show();  spark.sql("select \* from anime order by rating desc limit 5").show();  spark.sql("select \* from anime order by rating asc limit 5").show();  spark.sql("select \* from anime where lower(genre) like '%action%' order by rating desc limit 10").show();  spark.sql("select avg(rating) from anime where lower(genre) like '%mecha%'");  spark.sql("select avg(rating) from anime where lower(genre) like '%ecchi%'");  spark.sql("select avg(episodes) from anime where lower(name) like '%jojo%'");List<String> sublist = myList.stream().skip(1).limit(2).collect(Collectors.toList()); |

Часть результатов выполнения представлена на рисунке 1. 



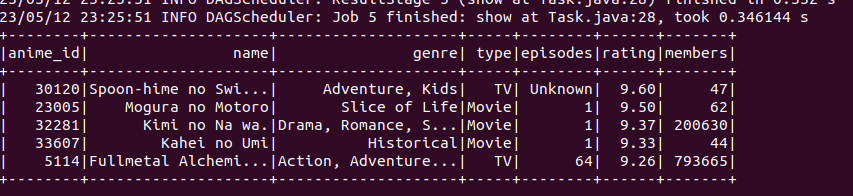


Рисунок 1 – Результаты выполнения SQL запросов

Полный код задания размещен в репозитории по ссылке – https://github.com/moroz-matros/BDL.

**Вывод** — в ходе работы были получены навыки работы со Spark в языке Java.