Лабораторная работа №6

ИСТОЧНИКИ ДАННЫХ НА PYSPARK

Целью этой лабораторной является получение навыков выгрузки исходных данных и отправки результатов модели в базу данных.

В качестве источника данных по варианту необходимо было реализовать cassandra. Для реализации в будущем следующих лабораторных работ был использован образ spark приложения который позволяет запускать узлы в docker. Ссылка на репозиторий: <https://github.com/big-data-europe/docker-spark>

Был разработан скрипт который отделяет от основного файла данные размером около 300мб. Также разработан отдельный контейнер который служит для записи данных в базу данных Cassandra которая развернута также в докер контейнере. Репозиторий с кодом доступен ниже.

Ссылка на репозиторий git: https://github.com/Saivaks/big\_data\_lab\_6

В качестве модели был выбран алгоритм k средних. Были выбраны некоторые данные из файла а именно значения 'creator', 'pnns\_groups\_1', 'pnns\_groups\_2' для проведения кластеризации. Результат кластеризации был также зафиксирован в базе cassandra.

Также в отчете есть приложения в виде скринов скриптов записи данных и таблиц в базе.

Write\_data.py:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

Таблицы с данными:

Изображение выглядит как текст, снимок экрана, программное обеспечение, меню

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание