Министерство науки и высшего образования Российской Федерации

Московский физико-технический институт

Промышленное машинное обучение

Осень

2022

Лабораторная работа №3

ВЗАИМОДЕЙСТВИЕ С ИСТОЧНИКАМИ ДАННЫХ НА SPARK

Цель работы:

Получить навыки выгрузки исходных данных и отправки результатов модели с использованием различных источников данных согласно варианту задания.

Задачи работы:

1. Обеспечить выгрузку данных при каждом запуске модели.
2. Обеспечить загрузку данных сразу по завершении работы модели.
3. Необходимо разработать протокол взаимодействия между моделью и источником данных.
4. Необходимо разработать формат хранения данных исходя из особенностей источника данных.
5. Рекомендуется использование docker контейнеров.

Основное задание:

Разработать на PySpark модель кластеризации на базе алгоритма к-средних. Разрешено использование любых метрик и подходов машинного обучения.

Данные: https://static.openfoodfacts.org/data/openfoodfacts-mongodbdump.tar.gz

https://world.openfoodfacts.org/data

Можно использовать все доступные средства языка Python/Scala.

Схема: модель-источник данных.

Дополнительное задание:

Разработать витрину данных на языке Scala для реализации протокола, единого формата данных. Витрина предназначена для формирования запросов к источнику и отгрузки результатов работы модели. В данном случае модель не взаимодействует с источником данных напрямую, а лишь через витрину данных.

Схема: модель-витрина-источник данных.

Варианты задания

Номер	Источник данных
1	PostgreSQL
2	MySQL
3	Oracle
4	MS SQL Server
5	MongoDB (в случае недоступности использовать Neo4J)
6	Apache HBase
7	Redis
8	Cassandra
9	Greenplum
10	ClickHouse
11	HDFS