Министерство науки и высшего образования Российской Федерации

Национальный исследовательский университет ИТМО

Инфраструктура больших данных

Весна

2024

Лабораторная работа №5

МОДЕЛЬ КЛАСТЕРИЗАЦИИ НА PYSPARK

Цель работы:

Получить навыки разработки и настройки Spark приложения.

Ход работы:

1. Настроить среду для Spark вычислений: https://sparkbyexamples.com/pyspark/how-to-install-and-run-pyspark-on-windows/
2. Обязательно проверить работоспособность компонентов Spark платформы, запустив примеры (WordCount).
3. Разработать на PySpark модель кластеризации на базе алгоритма k-средних. Разрешено использование любых метрик и подходов машинного обучения.
   Данные: https://static.openfoodfacts.org/data/openfoodfacts-mongodbdump.tar.gz
   https://world.openfoodfacts.org/data
4. Можно использовать все доступные средства языка Python/Scala. Обязательно провести предобработку данных с целью формирования выборки адекватного размера в зависимости от системных ресурсов.

Результаты работы:

1. Отчёт о проделанной работе;
2. Ссылка на репозиторий GitHub;
3. Актуальный дистрибутив модели в zip архиве.

Полезные материалы:

1. https://sparkbyexamples.com/
2. https://spark.apache.org/docs/latest/api/python/
3. https://pythonru.com/biblioteki/pyspark-dlja-nachinajushhih
4. https://habr.com/ru/articles/708468/