Лабораторная работа #5

Модель кластеризации на PySpark

В рамках данной работы было разработано два небольших приложения на spark. Первое считает количество слов в файле. Второе кластеризует данные из датасета (https://world.openfoodfacts.org/data).

Реализация Word Counter (word_count.py)

Реализация KMeans (main.py)

В результате обчуения модели была получена следующая метрика : Silhouette with squared euclidean distance = 0.9999993029872573