КУРСОВая Работа

«Кластеризация точек на scala»

по дисциплине «Языки и средства функционального программирования»

Выполнил

студент гр.3530904/80004 < > Котов И.В

Руководитель <> Лукашин А.А

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2019 г.

Оглавление

[Задание 3](#_Toc27512796)

[Решение 4](#_Toc27512797)

[Демонстрация работы 5](#_Toc27512798)

[Вывод 7](#_Toc27512799)

[Список литературы 8](#_Toc27512800)

# Задание

Кластеризация точек – реализовать считывание точек из файла и отобажение на графике(для построения графика выбрать любую библиотеку). Количество точек от 1 миллиона. Построение в двумерном пространстве.

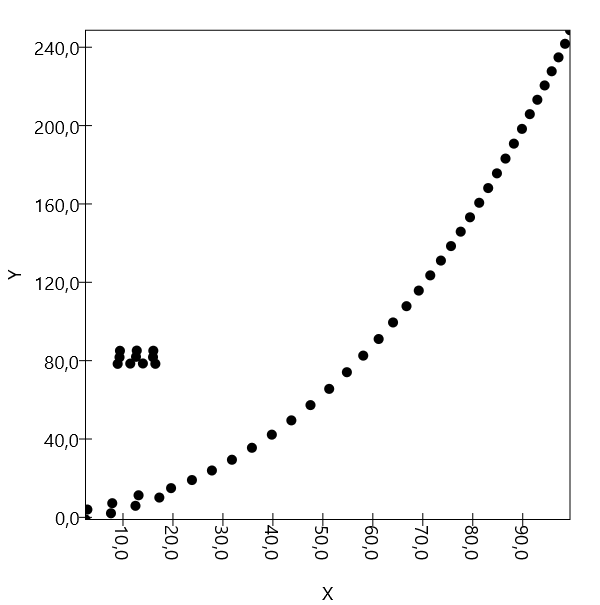
# Решение

При поиске решения данной задачи я нашёл библиотеку mllib, входящую в состав SPARK для scala. Для кластеризации точек я использовал метод K-средних, реализованный в данной библиотеке. Для отображения графика я использовал swiftVis2, который я нашёл в одном из репозитариев на github.

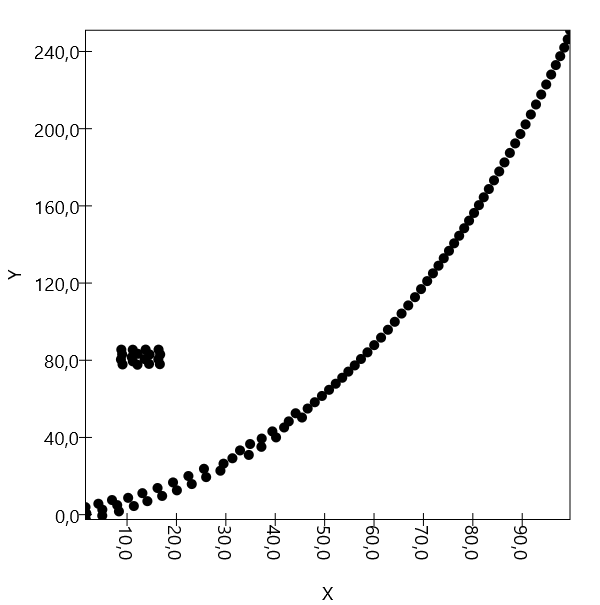
# Демонстрация работы

В качестве рабочего объёма данных сгенерированы точки графика функции 2ого порядка с шумом и некоторое множество точек в окрестности точки (13,83) суммарным количеством 2 миллиона.

Кластеризация на 50 кластеров:



Кластеризация на 100 кластеров:



# Вывод

Я изучил кластеризацию точек методом K-средних, с использованием библиотеки MLLIB, SPARK. Также изучил swiftvis для построения графиков в ходе работы. Для моего специфического объёма данных кластеризация в 50 кластеров показала себя лучше, чем в 100 кластеров с точки зрения сглаживания шумов.

# Список литературы

1. SwiftVis2: Plotting with Spark using Scala https://www.researchgate.net/publication/330845267\_SwiftVis2\_Plotting\_with\_Spark\_using\_Scala
2. Clustering - RDD-based APIhttps://spark.apache.org/docs/latest/mllib-clustering