

# Семинарское задание №4

## SimHash

Мосиенко Константин Викторович

2018

Вам дана коллекция симхэшей, записанных в виде текстового файла по одному на строке в десятичной системе исчисления. Необходимо построить график распределения размеров групп полудублей. Кластеризовать на группы можно следующим образом:

- Выбрать первый не обработанный симхэш.
- Найти в коллекции все симхэши, отстоящие от выбранного не более, чем на 3 бита. Это и есть «группа». Обновить статистики.
- Удалить группу из коллекции и перейти к первому пункту, если остались необработанные записи.

Присылайте код и отчёт с графиком и описанием алгоритма, который вы используете для поиска по симхэшам. Я предлагаю реализовать подход с разделением хэшей на четыре части. Не забывайте про логарифмические оси (обе).