Семинарское задание №6
SimHash

Мосиенко Константин Викторович

2017

Вам дана коллекция симхэшей, записанных в виде текстового файла по одному на строке в десятеричной системе исчисления. Необходимо построить график распределения размеров групп полудублей. Кластеризовать на группы необходимо 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.

Мне присылайте код и отчёт с графиком и описанием алгоритма, который вы используете для поиска по симхэшам. Не забывайте про логарифмические оси.

Коллекцию можно взять на странице курса: simhash_sorted.txt.bz2.