# Семинарское задание №8
# Вероятностный поиск SimHash

Мосиенко Константин Викторович

2017

Для выполнения задания вам понадобится набор документов из задания про обход википедии. Если его нет - попросите у товарищей.

- Подсчитайте симхэш для каждого документа из коллекции.
- Для каждого документа с помощью точного алгоритма, делящего сигнатуры на 4 части, вычислите размер группы его полудублей на расстоянии $\leq 3$ бита.
- Реализуйте алгоритм вероятностного поиска полудублей, делающий конфигурируемое количество $k$ дополнительных обращений в табличку со всеми симхэшами.
- Постройте график полноты поиска полудублей в зависимости от $k$: для заданного $k$, какой процент от общего количества полудублей находится за $k$ обращений.
- Какое необходимо выбрать $k$, чтобы найти 30%, 50%, 80% полудублей.

Мне присылайте код и отчёт с графиком и описанием вашей реализации алгоритма с указанием выбранных констант.