

Хранение и Обработка Больших Объёмов Данных

Антон Горохов

старший разработчик, Яндекс

anton.gorokhov@gmail.com

План лекции

I. Page Rank

- 1) Интуитивная интерпретация
- 2) Матричная запись
- 3) Формула Google
- 4) Вычисление

II. Разное

- 1) Декартово произведение
- 2) Глобальная сортировка
- 3) Частотные ключи

Ранжирование в поиске

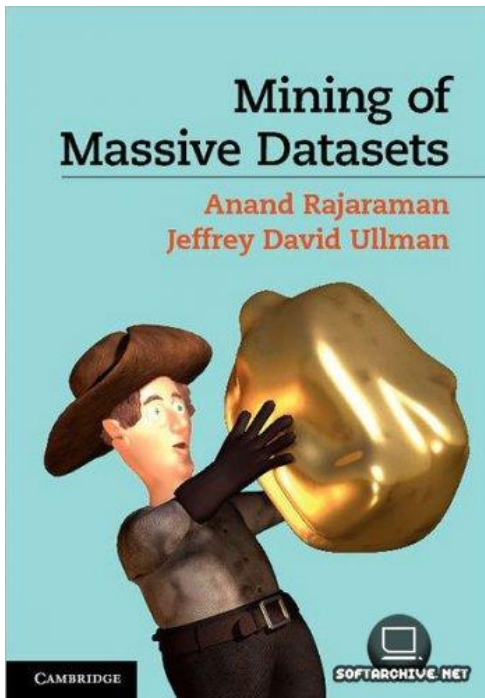
(10 лет назад)

- Документ релевантен запросу, если
 - содержит текст запроса
 - в тегах <title>, <h1>, ..., <meta description="">
 - порядок слов, расстояние между словами
- Ссылочный индекс:
 - ссылка на страницу содержит текст запроса
- Это легко подделать
- PageRank – «авторитетность» страницы

Disclaimer: MMDS

Mining of Massive Dataset

- Книга + слайды: <http://mmds.org/>
- Курс на coursera.org



Далее: MMDS, Chapter 5: Link Analysis

План лекции

I. Page Rank

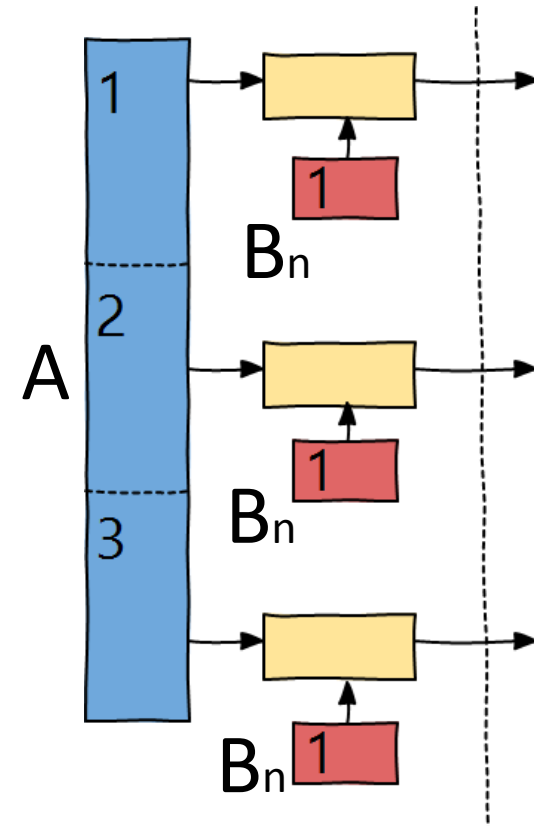
- 1) Интуитивная интерпретация
- 2) Матричная запись
- 3) Формула Google
- 4) Вычисление

II. Разное

- 1) Декартово произведение
- 2) Глобальная сортировка
- 3) Частотные ключи

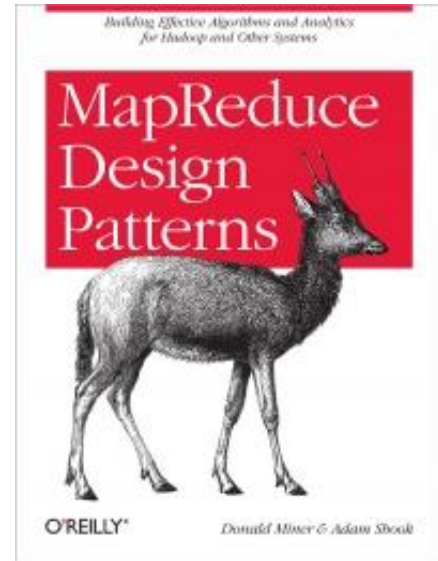
Декартово произведение

- Map-side
 - один из датасетов — на вход маппера
 - N-й блок второго датасета — в память маппера, и так N раз

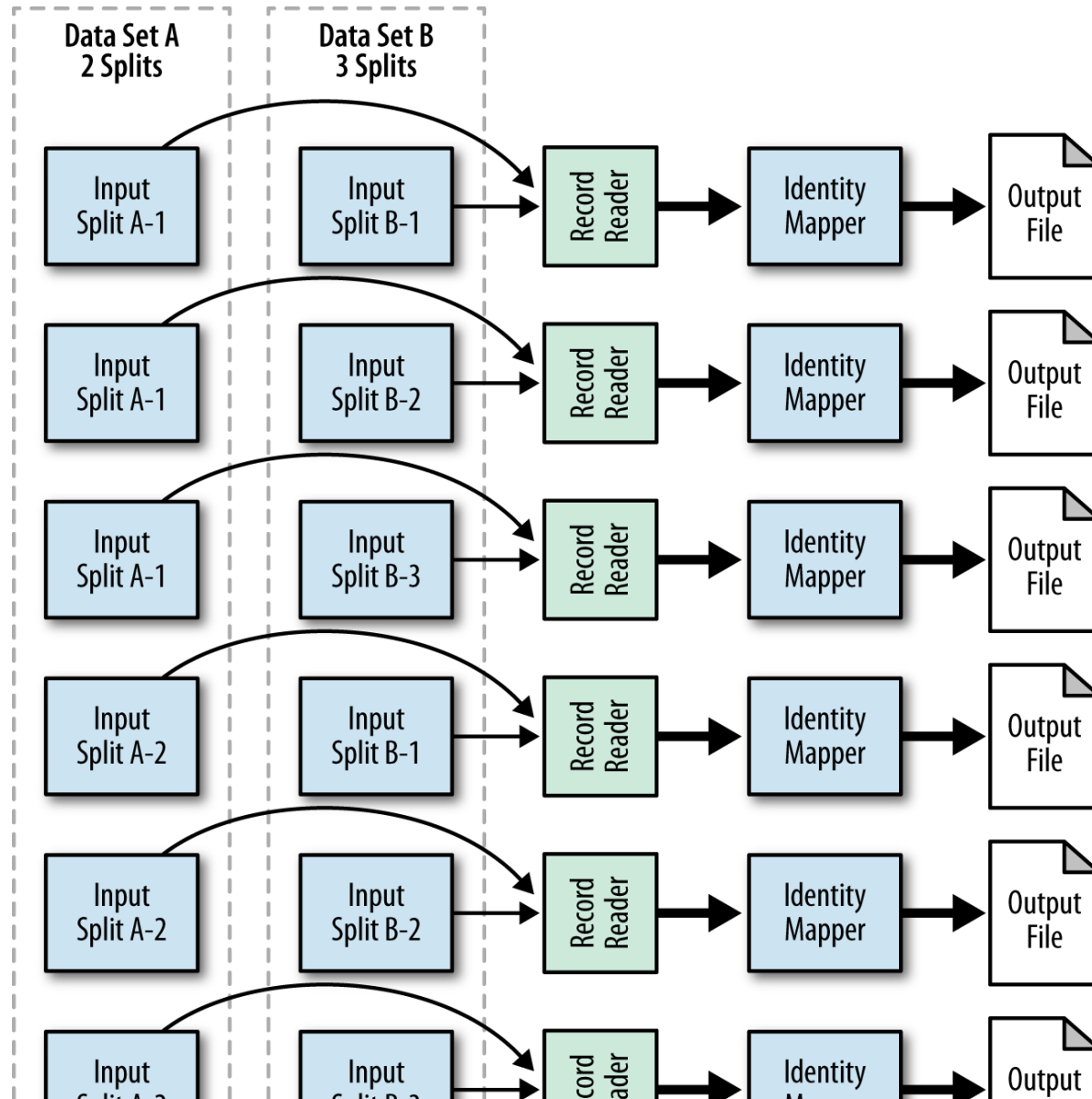


Декартово произведение

- MapReduce Design Patterns, p.128
 - мапперы читают одновременно сплиты из A и B
 - переопределены InputFormat и RecordReader
 - **CartesianInputFormat**
 - `getSplits()` – выдает пары сплитов, по одному из датасета
 - `getRecordReader()` – возвращает `CartesianRecordReader`
 - **CartesianRecordReader** – объединяет сплиты:
key – строка из A, value – строка из B

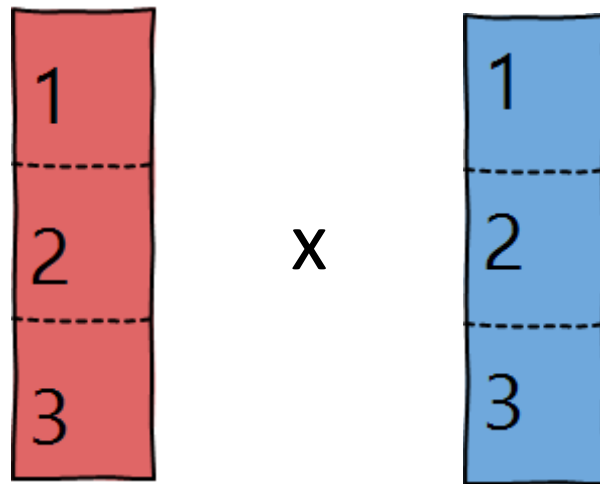


Декартово произведение



Декартово произведение

- Reduce-side – как?



- У нас нет ключа, по которому группируем записи
- Будем ориентироваться на конкретную задачу

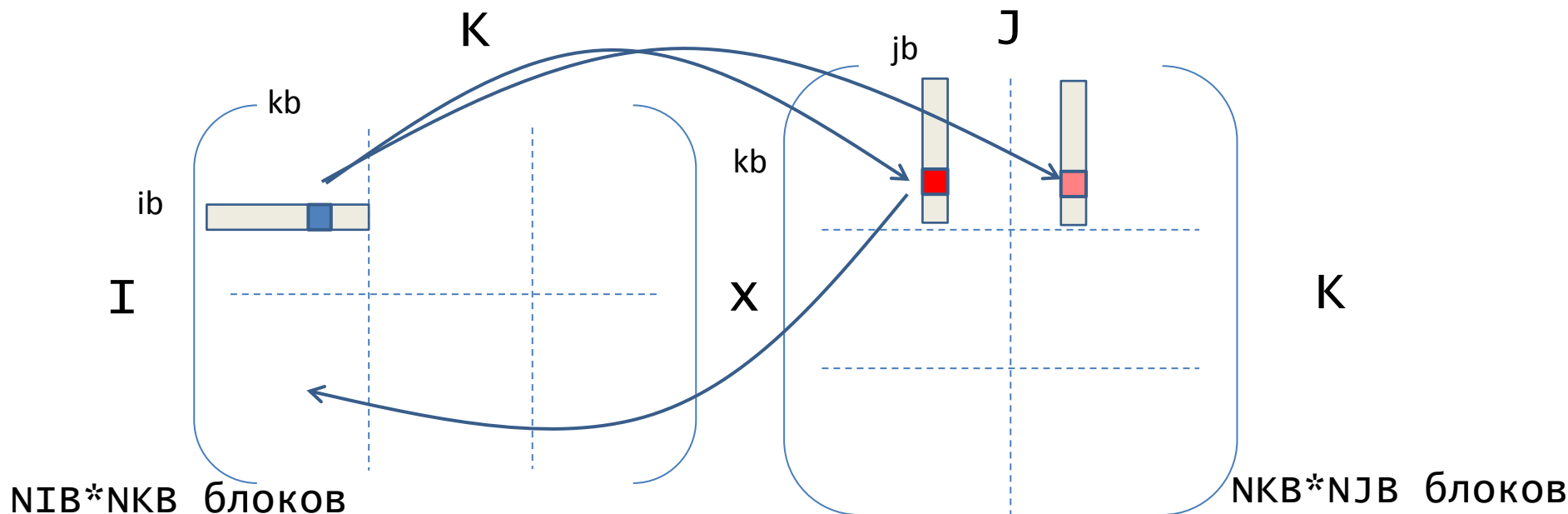
Перемножение матриц

<http://www.norstad.org/matrix-multiply/>

$$\begin{array}{c}
 \text{I} \quad \begin{array}{c} \text{K} \\ \left(\begin{array}{cc|cc} A_{11} & A_{12} & A_{13} & \\ \hline A_{21} & A_{22} & A_{23} & \end{array} \right) \end{array} \times \begin{array}{c} \text{J} \\ \left(\begin{array}{cc|cc} B_{11} & B_{12} & & \\ \hline B_{21} & B_{22} & & \\ \hline B_{31} & B_{32} & & \end{array} \right) \end{array} \quad \text{K} \\
 \\
 = \left(\begin{array}{cc|cc} A_{11}B_{11}+A_{12}B_{21}+A_{13}B_{31} & & & \dots \\ \hline & & & \\ & \dots & & \\ & & & \dots \end{array} \right)
 \end{array}$$

Перемножение матриц

<http://www.norstad.org/matrix-multiply/>



mapper размножает элементы,
по одному на блок в соседней:

$(ib, kb, jb, 0) \rightarrow a[i, k]$

$(ib, kb, jb, 1) \rightarrow b[k, j]$

$\underbrace{\hspace{1.5cm}}$
partitioner

reducer1:

Вход: блок из A, блок из B

Выход: блок $C[ib, kb, jb]$

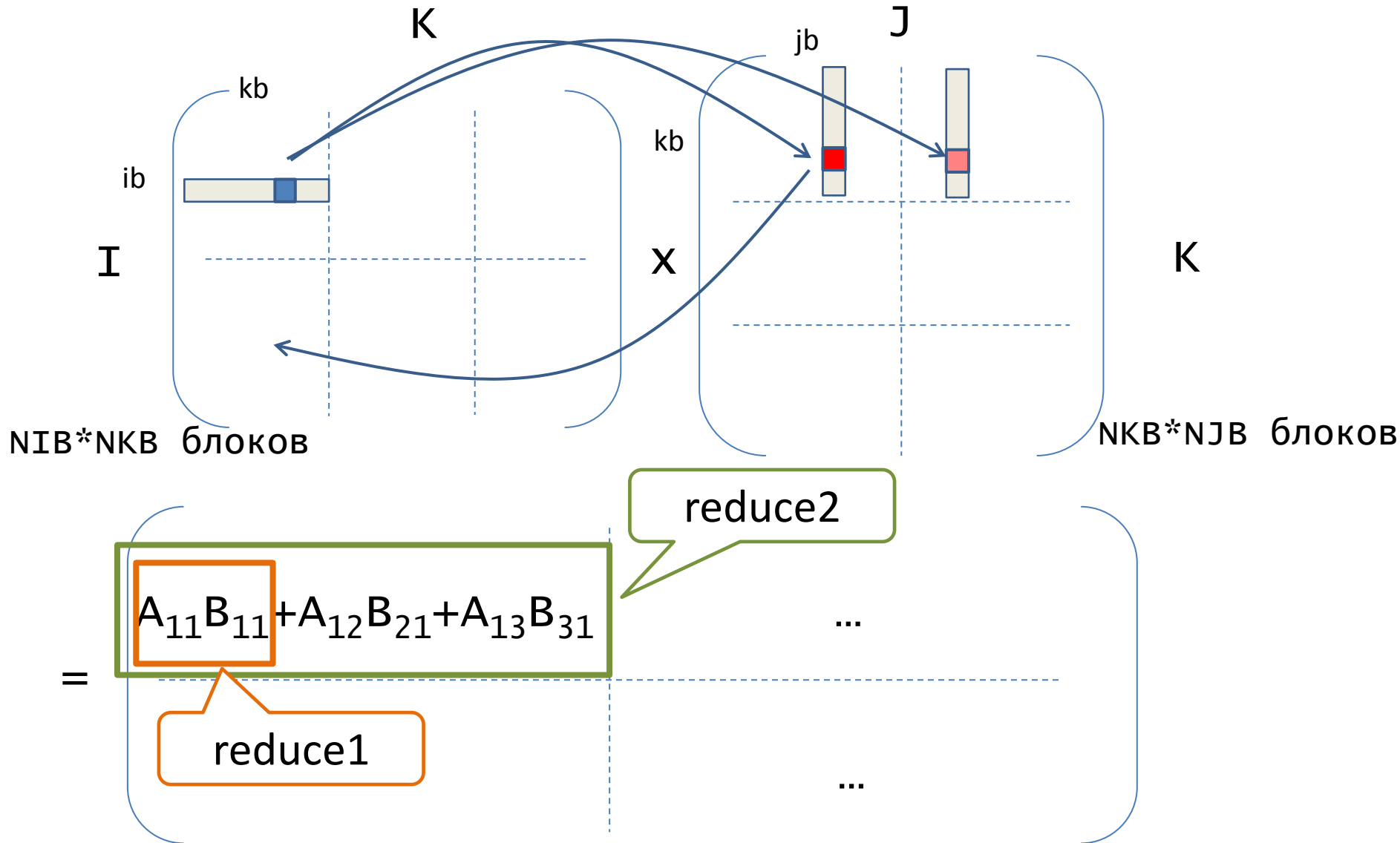
reducer2:

Сумма $C[NIB, NKB, NJB]$

по $NKB=0 \dots NK-1$

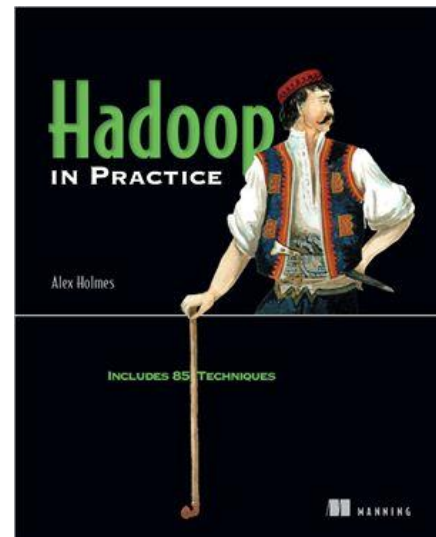
Перемножение матриц

<http://www.norstad.org/matrix-multiply/>



Глобальная сортировка

- `job.setNumReduceTasks(1)`
- несколько reducer'ов, разделение на диапазоны
 - InputSampler
 - Получает семпл ключей
 - Записывает partition file – файл с диапазонами
 - TotalOrderPartitioner – разделяет на основе partition file
 - «Hadoop in Practice», p.222
 - «HDG», p.272



Частотный ключ

- Задача: статистика сайтов

- koshkiclub.ru
- ovasuaritma.com
- intek.by.ru

.....

- yandex.ru

Неравномерное распределение ключей по reducer'ам.

Частотный ключ

- Задача: статистика сайтов

- koshkiclub.ru
- ovasuaritma.com
- intek.by.ru

.....

- yandex.ru-QF6
- yandex.ru-8FD
- yandex.ru-84G
- ...

Добавляем
соль

«Монстр»

yandex.ru

Обработка в 2 этапа: с солью и без.

Как обнаружить
«монстров»?

Передача параметров в streaming

```
$ hadoop jar hadoop-streaming.jar  
    -Dparam=value  
    -input in_dir -output out_dir  
    -mapper mapper.py  
    -file mapper.py
```

Через переменные окружения:

```
mapper.py:  
import os  
value = os.environ["param"]
```

Попробуйте:

```
print "; ".join(os.environ)
```


Вопросы?

I. Page Rank

- 1) Интуитивная интерпретация
- 2) Матричная запись
- 3) Формула Google
- 4) Вычисление

II. Разное

- 1) Декартово произведение
- 2) Глобальная сортировка
- 3) Частотные ключи