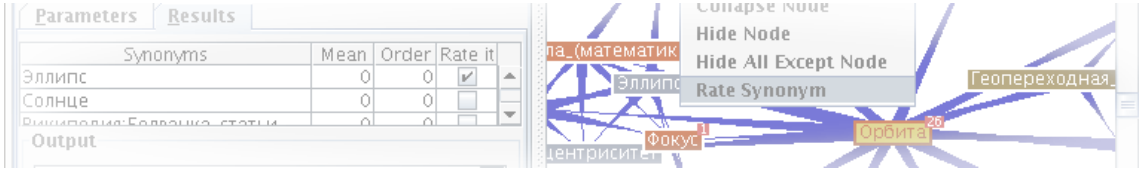


Информационный поиск

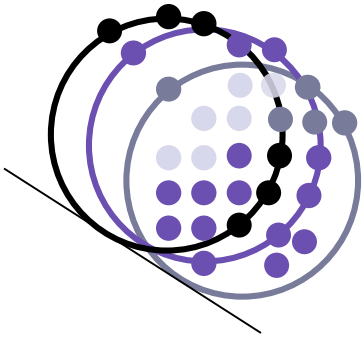
Обработка булевых запросов, пересечение списков словопозиций

Петрозаводский государственный университет

Крижановский Андрей Анатольевич

andrew.krizhanovsky . gmail.com

Информационный поиск

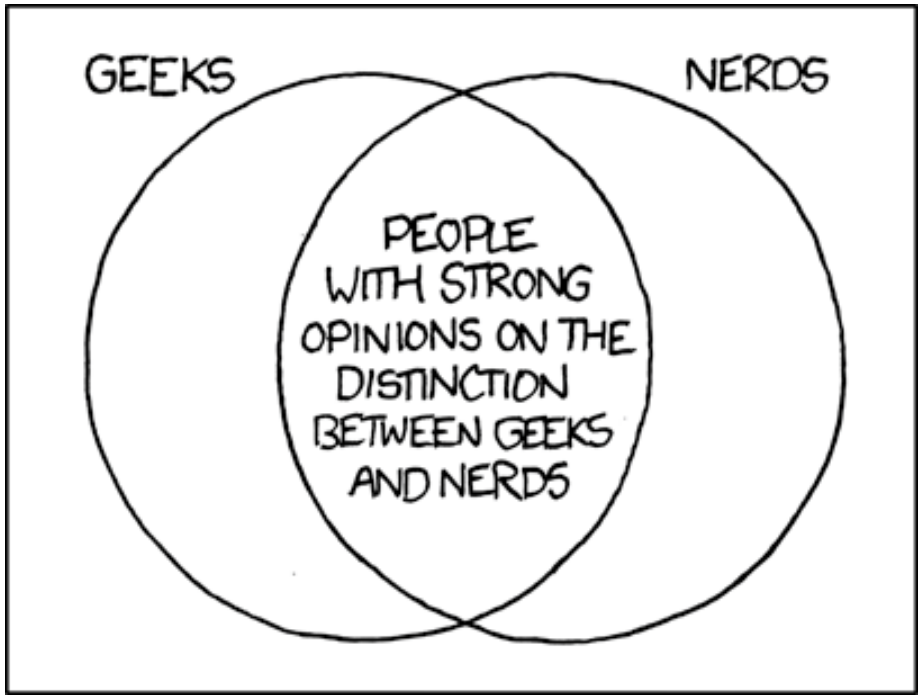


(Information retrieval) — это процесс поиска в большой **коллекции** (хранящейся, как правило, в памяти компьютеров) некоего **неструктурированного** материала ("обычно — документа), удовлетворяющего **информационные потребности**

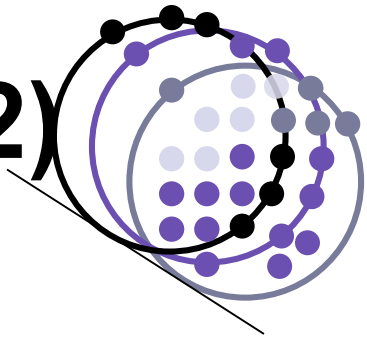
Модель булева поиска (1)

- это модель информационного поиска, в ходе которого можно обрабатывать любой запрос, имеющий вид булева выражения, т.е. выражения, в котором термины используются в сочетании с операциями AND, OR и NOT.

- Тип операции?



Модель булева поиска (2)



- term-document incidence matrix
  - термин (IR) <> слово
  - вектор термина (по строкам)
  - вектор документа (столбец)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Brutus AND Caesar AND NOT Calpurnia

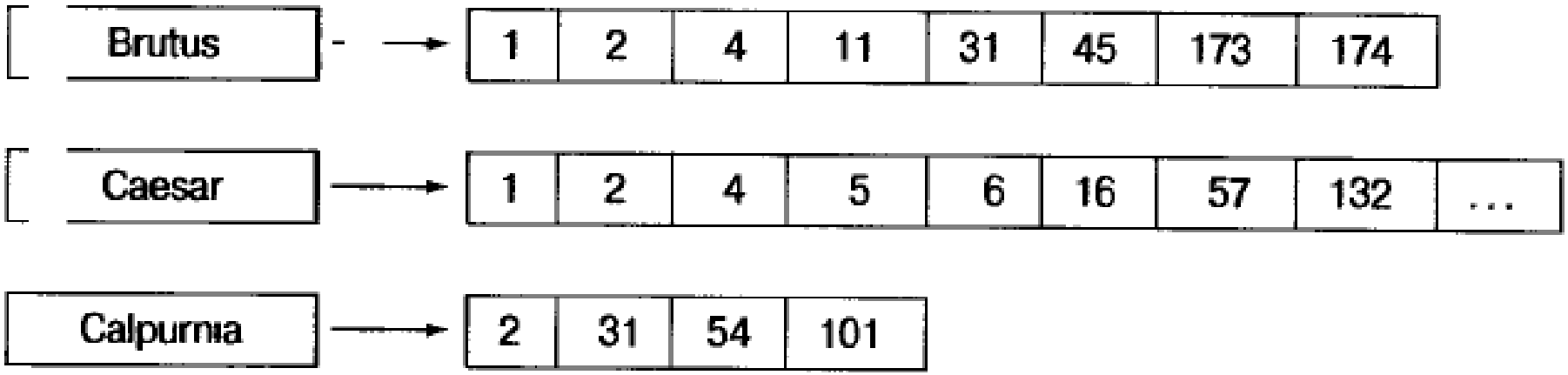
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

110100 AND
110111 AND
NOT 010000

⇒

110100 AND
110111 AND
101111=
100100

(не)координатный инвертированный индекс



∃ термин в документе

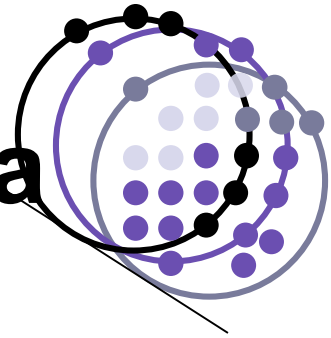
координата термина в этом документе (словопозиция, posting)

Словарь

Словопозиции

Рис. 1.3. Две части инвертированного индекса. Словарь обычно находится в памяти вместе с указателями на каждый список словопозиций, которые хранятся на диске

Этапы построения инвертированного индекса



1. Собираем документы

Цыганы шумною толпой	Для вас, души моей царицы,	В ауле, на своих порогах,

2. Размечаем текст, превращая ∀ документ в список лексем (tokens)

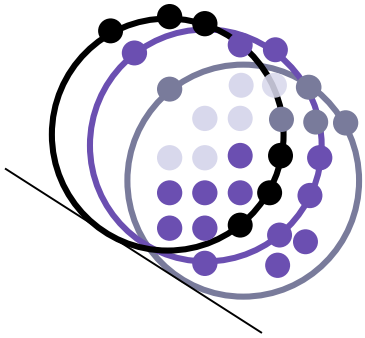
Цыганы	шумною	толпой	…	Для	вас

3. Предварительная лингвистическая обработка

цыган	шумный	толпа	…	для	вас

4. Индексируем документы (рис.)

Построение инв. индекса



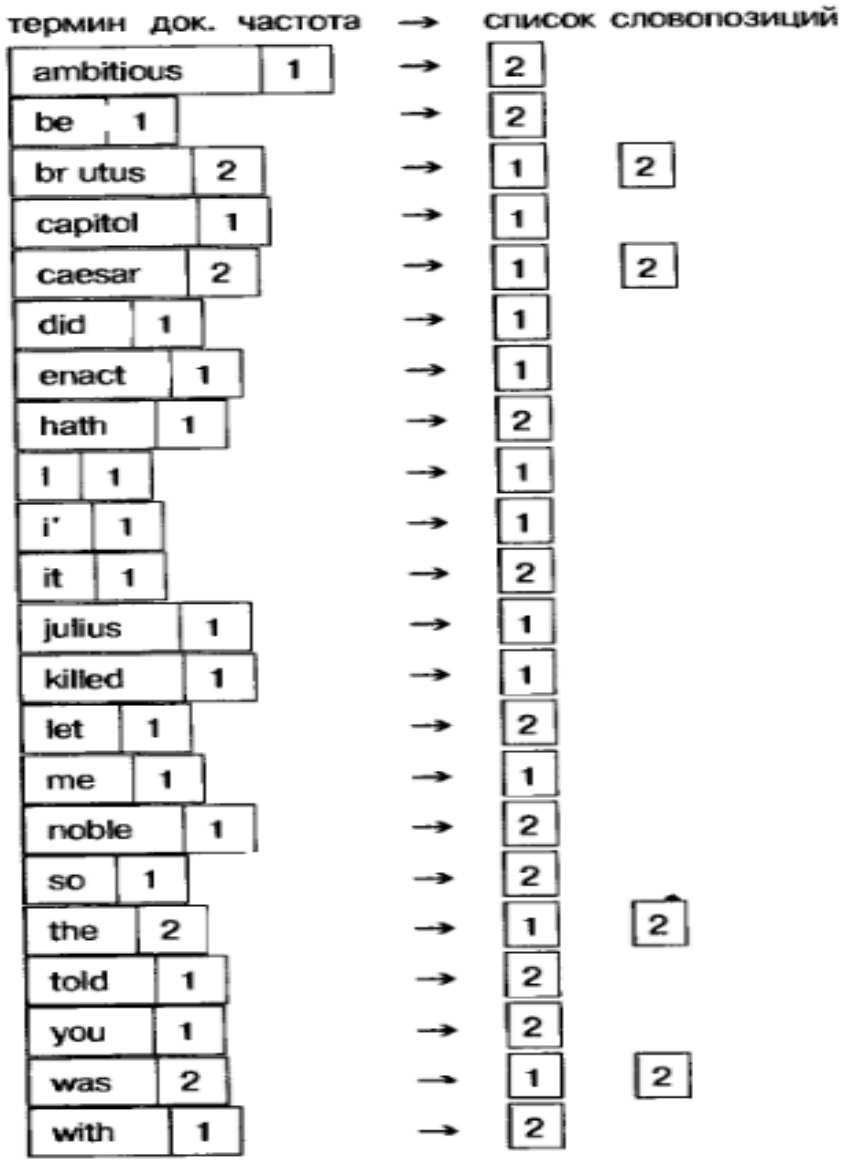
- docID – идентификатор документа
- ∀ документ, на входе:
  - список нормализованных лексем для каждого документа, т.е. список пар «термин-docID»

термин	docID
I	1
did	
enact	
julius	
caesar	
I	
was	1
killed	1
i'	
the	
capitol	
br utus	
killed	
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
br utus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

sort

термин	docID
ambitious	2
be	2
br utus	1
br utus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	
i'	
it	
julius	
killed	
killed	
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

документная частота
(document frequency)

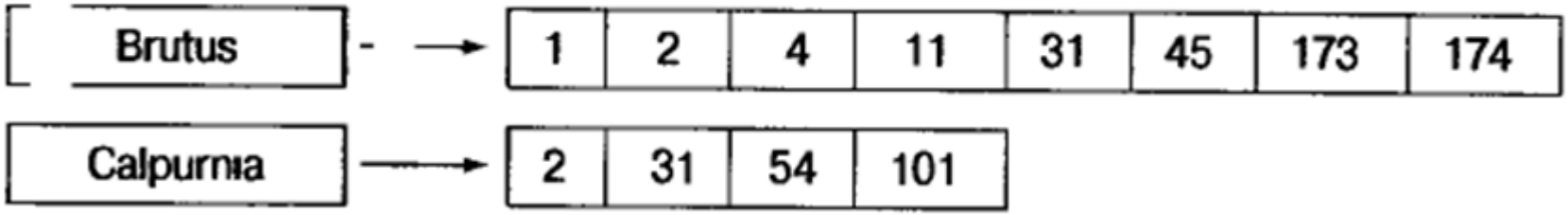


Группировка по слову, по документы

Этапы обработки запроса:

Brutus AND Calpurnia

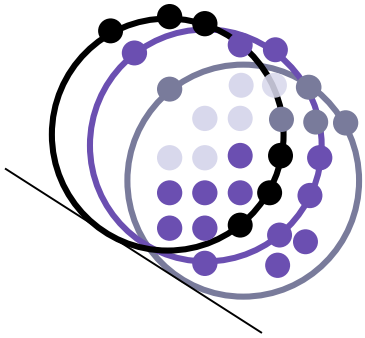
1. Обнаруживаем термин Brutus в словаре
   1. Находим список его словопозиций
2. Обнаруживаем термин Calpurnia в словаре
   1. Находим список его словопозиций
3. Находим пересечение этих двух списков



2, 31

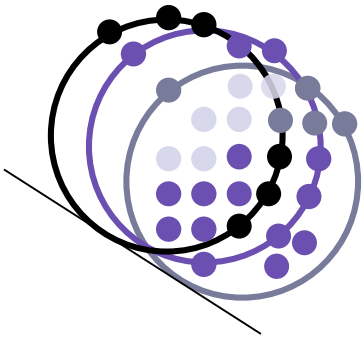
Д.б. эффективное пересечение, v

Алгоритм слияния (merge algorithm)



- Пересечение списков словопозиций
- Требование:
  - **Все** словопозиции д.б. упорядочены по одному критерию, ex.: сортировка числовых идентификаторов документов docID.
- На входе:
  - Указатели на два списка
- Сложность
  - Проход за t=O(N),
    - N – количество словопозиций в списках

Алгоритм пересечения двух списков словопозиций p1 и p2

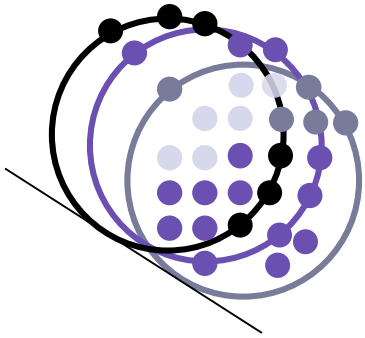


```
INTERSECT(p1, p2)
 1  answer ← ⟨⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then Add(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

zɐˈtɕem

МФА: [zɐˈtɕem]

Оптимизация запроса

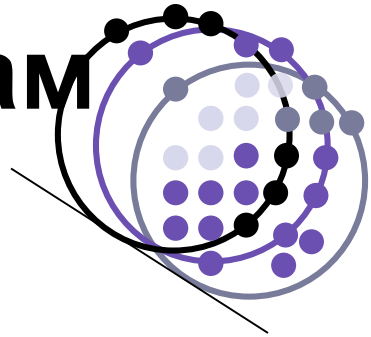


- - выбор такого способа организации обработки запроса, чтобы можно было *минимизировать общий объём работы*, которую должна выполнить система.

  (Brutus OR Caesar) AND NOT Calpurnia

  - Порядок доступа к спискам словопозиций

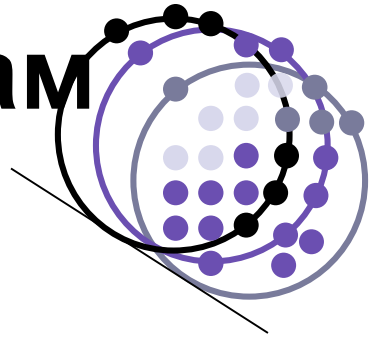
Порядок доступа к спискам словопозиций (1)



- Ex.: **Brutus AND Caesar AND Calpurnia**

1. Найти списки словопозиций для каждого термина
2. Применить к ним операцию AND.

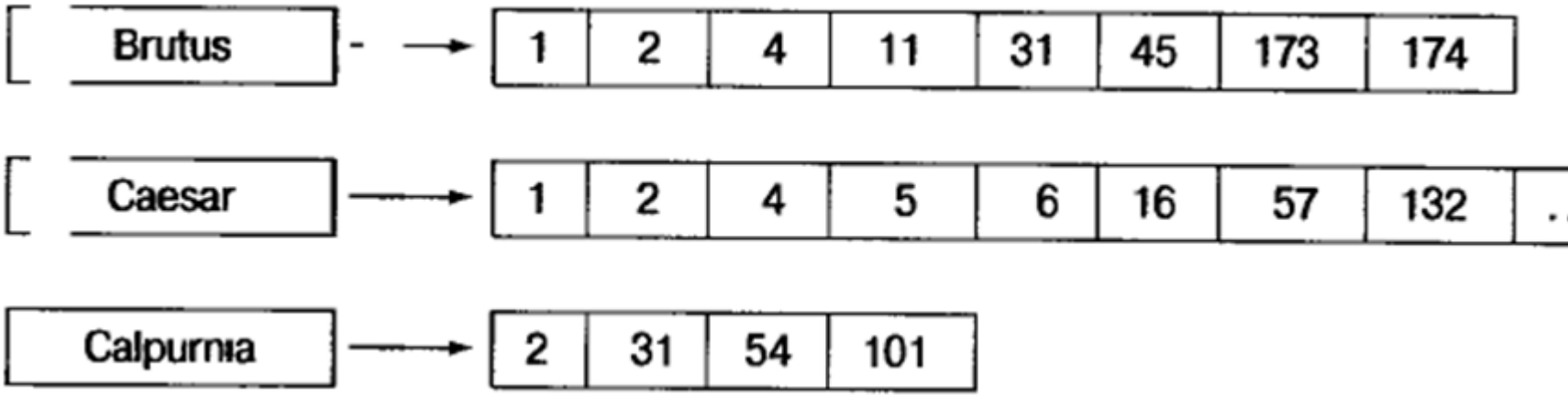
Порядок доступа к спискам словопозиций (2)



- Ex.: **Brutus AND Caesar AND Calpurnia**
- На практике (эвристика):
  - Обработка в порядке возрастания частоты употребления термина;
  - Если начать с пересечения двух наименьших списков словопозиций…

Порядок доступа к спискам словопозиций (3)

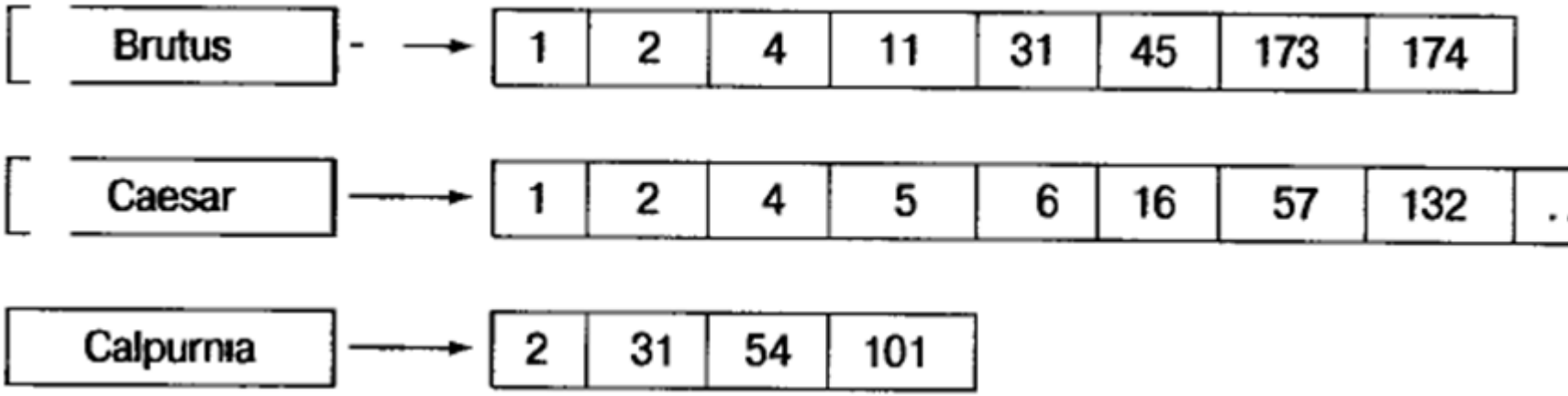
- Ex.: **Brutus AND Caesar AND Calpurnia**



- **На практике…?**

Порядок доступа к спискам словопозиций (3)

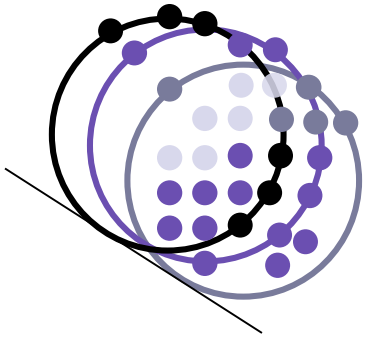
- Ex.: **Brutus AND Caesar AND Calpurnia**



- **На практике…?**

(Calpurnia AND Brutus) AND Caesar

Оптимизация запроса

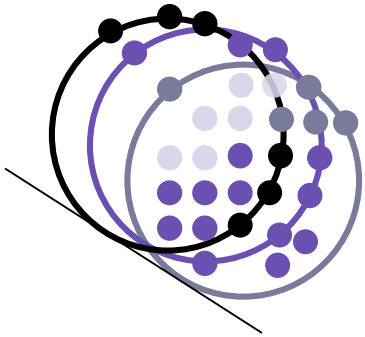


- Подсчитывать частоту терминов в словаре
  - Не нужен доступ к словопозициям
- Общий вид:

(madding AND crowd) AND

(ignoble OR strife) AND (killed OR slain)

Оптимизация запроса:
Общий вид

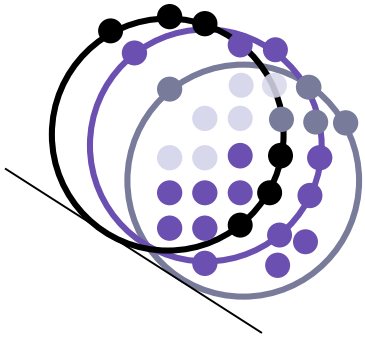


(madding AND crowd) AND

(ignoble OR strife) AND (killed OR slain)

- Определить частоту термина
- Сложить частоты для оператора OR.
- Дальше – снова – в порядке ↗ частоты каждого дизъюнктного термина.

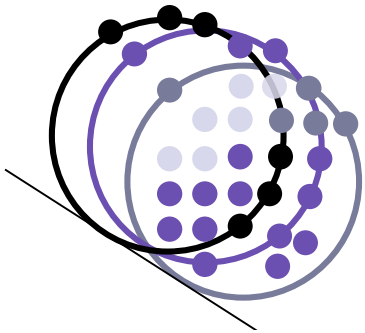
Задача (1)



Упражнение 1.4 [*]. Можно ли для каждого из запросов, указанных ниже, выполнить операции пересечения за время $O(x+y)$, где x и y - длины списков словопозиций для терминов Brutus и Caesar? Если нельзя, то каких результатов следует ожидать?

1. Brutus AND NOT Caesar
2. Brutus OR NOT Caesar

Задача (2)



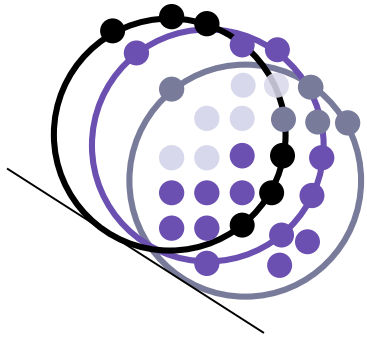
Упражнение 1.7 [*]. Порекомендуйте порядок обработки следующего запроса.

1. (tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)

Будем считать, что список словопозиций имеет следующий вид.

Термин	Размер списка словопозиций
eyes	213312
kaleidoskope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Литература



- **Маннинг** К., **Рагхаван** П., **Шютце** Х. **Введение в информационный поиск**. — Вильямс, 2011. — ISBN 978-5-8459-1623-5. http://rutracker.org/forum/viewtopic.php?t=3887364

Спасибо за внимание!

http://vk.com/imath_petrsu

“Интернет-математика в ПетрГУ”