

Яндекс

Поиск фраз в документах РСЯ

Дмитрий Агафонов

Я.Субботник, Москва, 18 апреля 2009 года

Яндекс – много разных поисков.

Для пользователей.

А ещё есть внутренние поиски.

Их тоже много.

Сегодня – про один из них.

Подробно!

Наш план

- О чём речь
- Кому это интересно
- Много букв (картинки будут)
- Ваши вопросы
- Мои ответы

О ЧЁМ

Контекстная реклама: клиенты и площадки.

Клиенты дают рекламу.

Площадки дают место.

От клиентов – рекламные материалы. Много.

От площадок – документы в Интернете. Очень много.

Надо связать их эффективно. И быстро.

О ЧЁМ

Процесс привязки большой и сложный.

Одна из задач – найти фразу в документе и определить её релевантность.

Несколько миллионов фраз. Надо найти все!

Время на размышление для одного документа: 20ms.

Как это сделать?

О чём

У нас это решение нашей конкретной задачи.

Общие применения алгоритма:

- Поиск коллекции наборов объектов во множестве объектов
- Определение релевантности данных заданным признакам
- Классификация
- Обратный поиск

Кому интересно

- Разработчикам
- Менеджерам
- Алгоритмистам
- И мне тоже

Тем, кто решал подобную задачу или будет решать.

Поехали!

Что в пути

- Задача
- Решение
 - *Собрать индекс из фраз*
 - *Обработать документы*

Задача

Вроде всё просто

- Есть много фраз
- Есть один документ

Найти фразы в документе.

Посчитать их релевантность.

На входе ⇐

Коллекция фраз. Фраза это:

- *Множество слов*
- *Дополнительные условия*

Документ это:

- *Большое множество слов*
 - *Позиция слова*
 - *Вес слова*
- *Дополнительные условия*

На выходе ⇒

Найденные фразы + релевантность каждой фразы.

Фраза найдена если:

- Документ содержит все слова фразы*
- Слова фразы находятся недалеко друг от друга*
- Дополнительные условия фраз и документа соответствуют*

Порядок слов не важен.

Пример на буквах

Это две фразы:

- a c h
- d f p u

Это документ: a b c d e f g h i j k l m n o p q r s t u v w x y z

Нашли.

Но фраза d f p u слишком размазана по документу.

Решение

Две части решения

- Собрать фразы в индекс
- Обработать документ

А сначала пара особенностей.

«Людям» = «человеку»

Посимвольно «людям» ≠ «человеку».

Есть леммер!

« У слова несколько форм

Основная – Лемма

Иногда не одна

неизвестный японский автор

«людям» (люди, человек)

«человеку» (человек)

Набор лемм слова – каноническая форма слова.

Или далее просто «слово».

Два приёма

- Сделать всё числами
 - *Слова*
 - *Леммы*
 - *Фразы*
 - *Всё остальное*
- Все числа упорядочить

Сборка индекса

На входе ⇐

Коллекция фраз. Фраза это:

- *Множество слов*

Документ это:

- *Большое множество слов*
 - *Позиция слова*
 - *Вес слова*

Этапы сборки индекса

- Подготовить фразы
- Построить индекс

Всего 11 простых действий :-)

Фразы

- Берём слова фраз

«с а b»

«b а»

«a с»

«b e d»

Фразы

- Берём слова фраз
- Считаем частоты слов

«с а b»

«b а»

«а с»

«b e d»

a – 3

b – 3

c – 2

d – 1

e – 1

Фразы

- Берём слова фраз
- Считаем частоты слов
- Переставляем слова во фразах по убыванию частоты

«с а b»

«b а»

«а с»

«b e d»

a – 3

b – 3

c – 2

d – 1

e – 1

«a b c»

«a b»

«а с»

«b d e»

Индекс

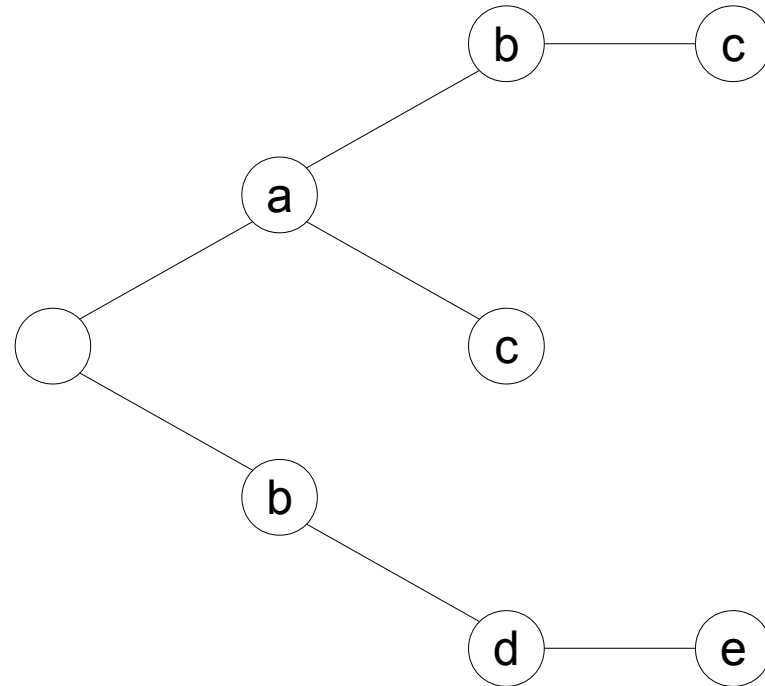
Строим дерево фраз

«a b c»

«a b»

«a c»

«b d e»



Индекс

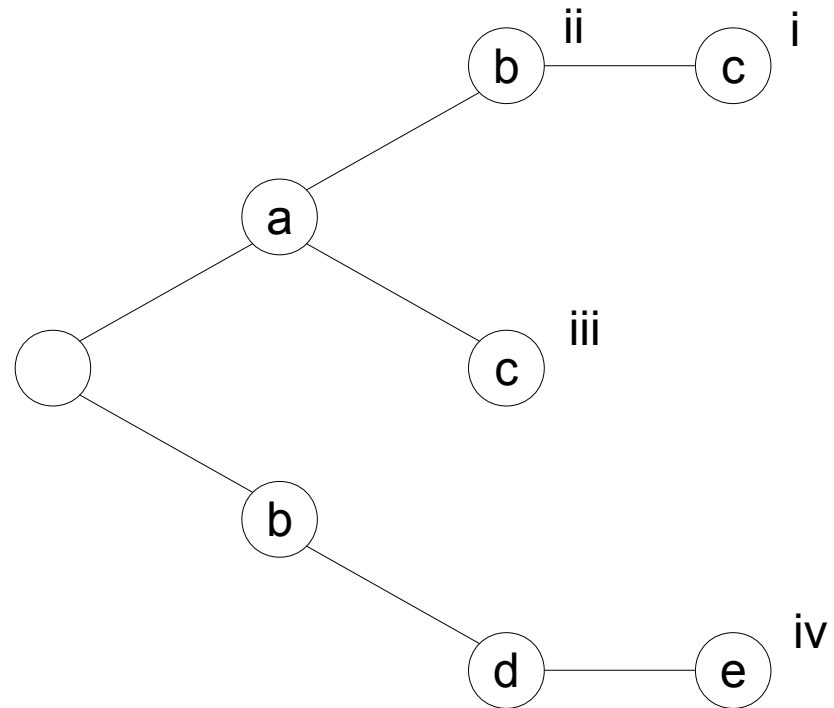
Находим терминальные узлы

i. «a b c»

ii. «a b»

iii. «a c»

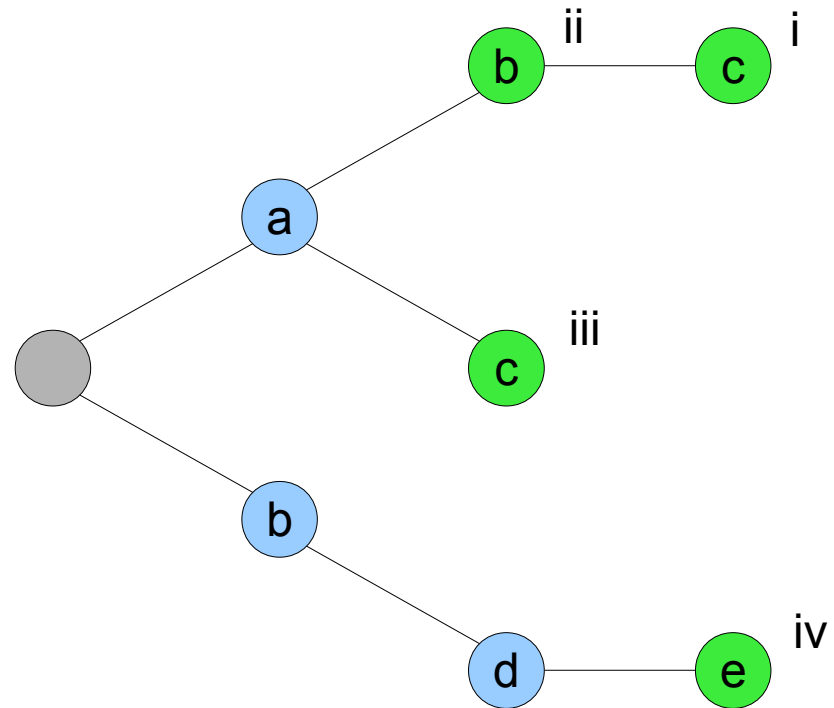
iv. «b d e»



Индекс

Красим узлы по типам

- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»



● корневой

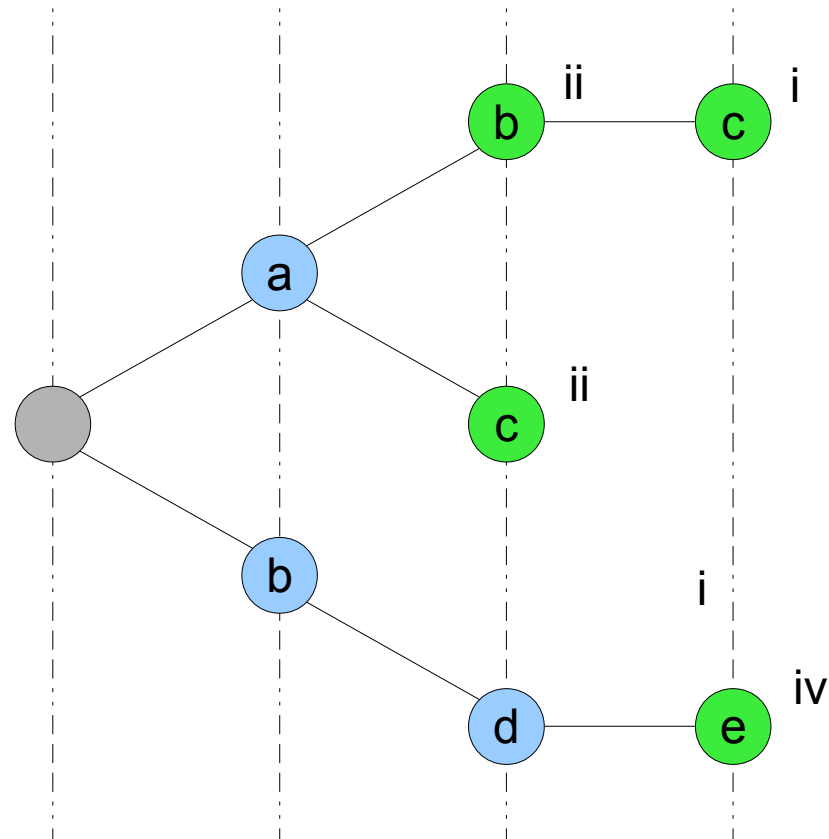
● терминальный

● остальные

Индекс

Проводим вертикальные уровни

- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»



● корневой

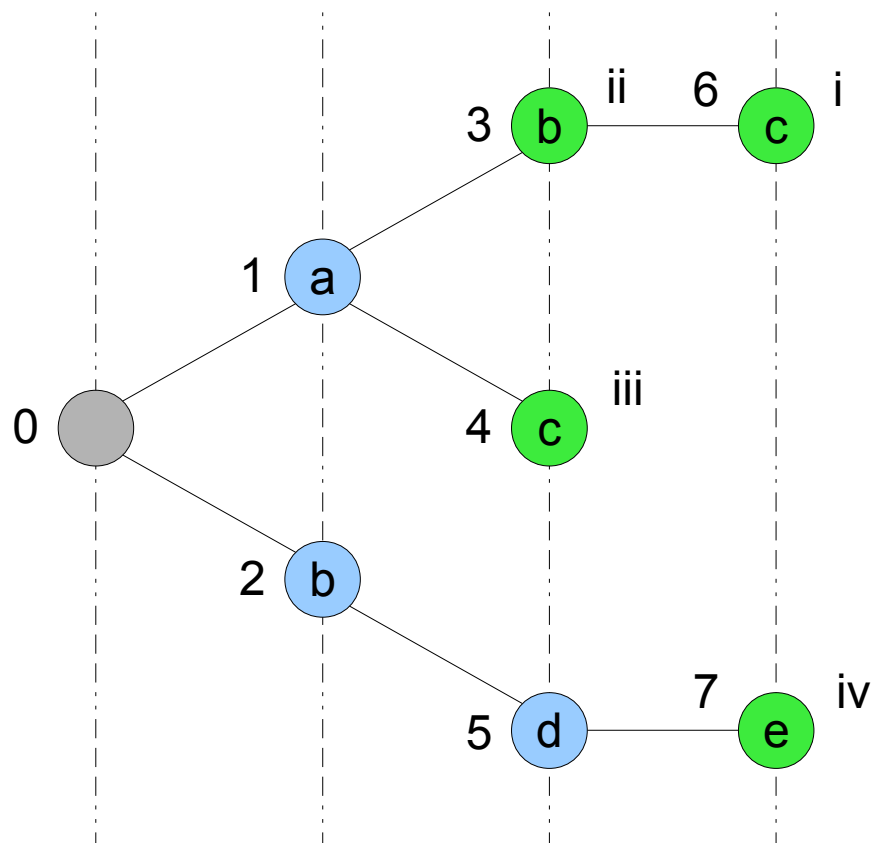
● терминальный

● остальные

Индекс

Нумеруем узлы слева направо и сверху вниз

- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»



● корневой

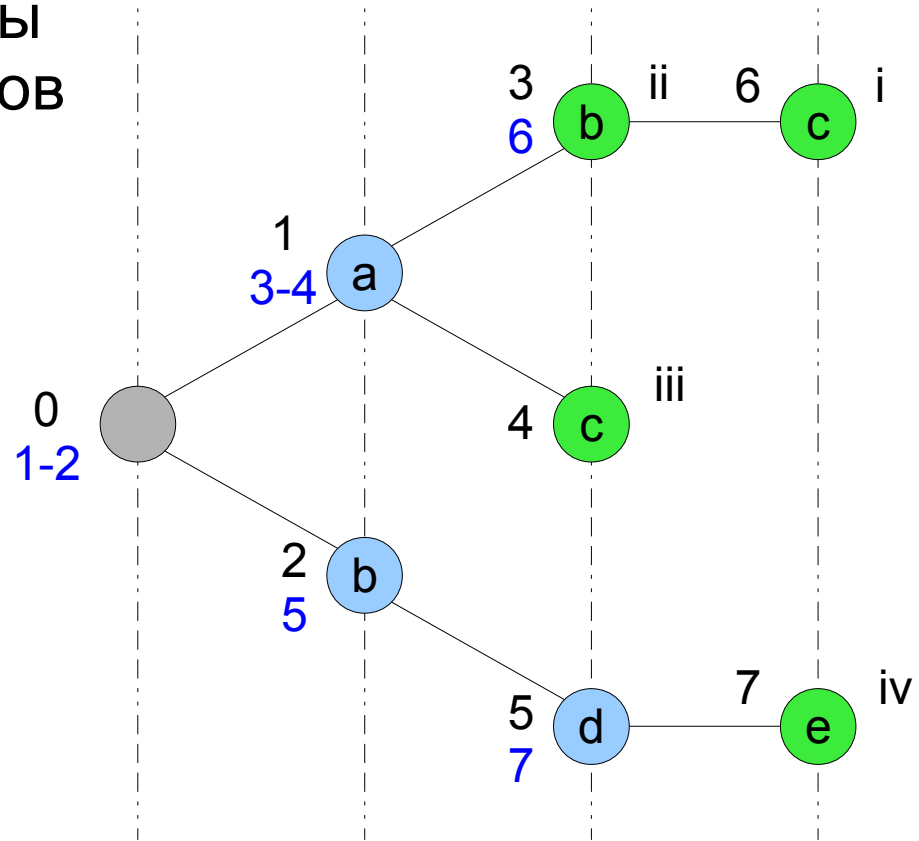
● терминальный

● остальные

Индекс

Запоминаем диапазоны
номеров дочерних узлов

- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»



● корневой

● терминальный

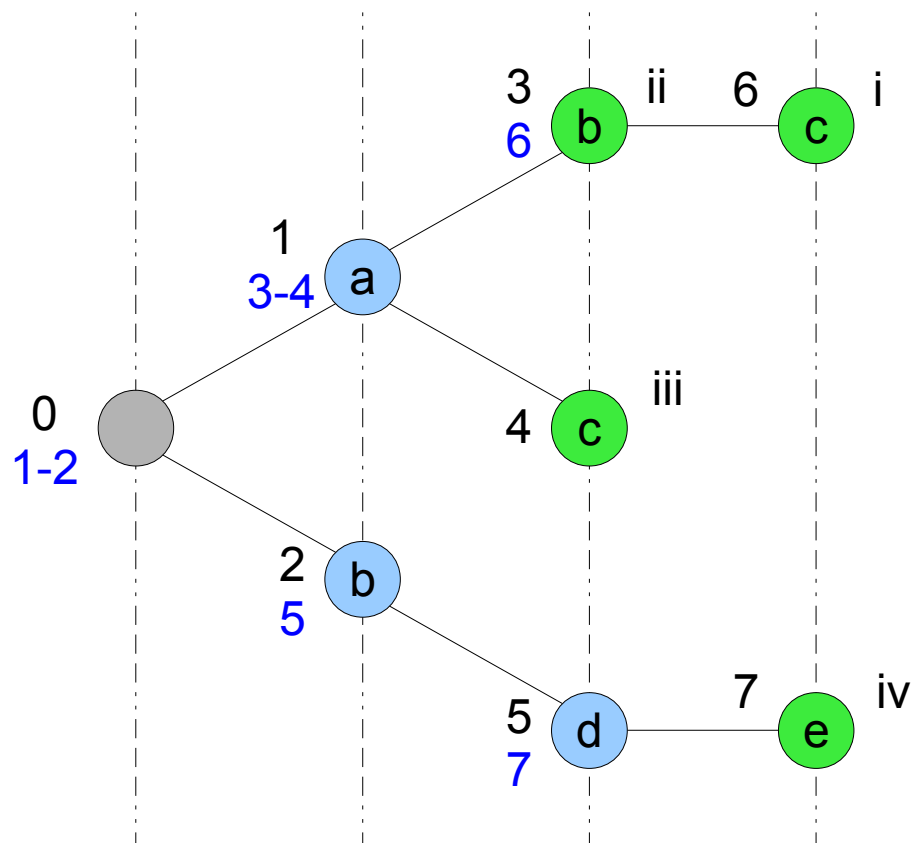
● остальные

Индекс

Делаем таблицу
Слово → Узлы дерева

i.	«a b c»
ii.	«a b»
iii.	«a c»
iv.	«b d e»

a	1
b	2, 3
c	4, 6
d	5
e	7



● корневой

● терминальный

● остальные

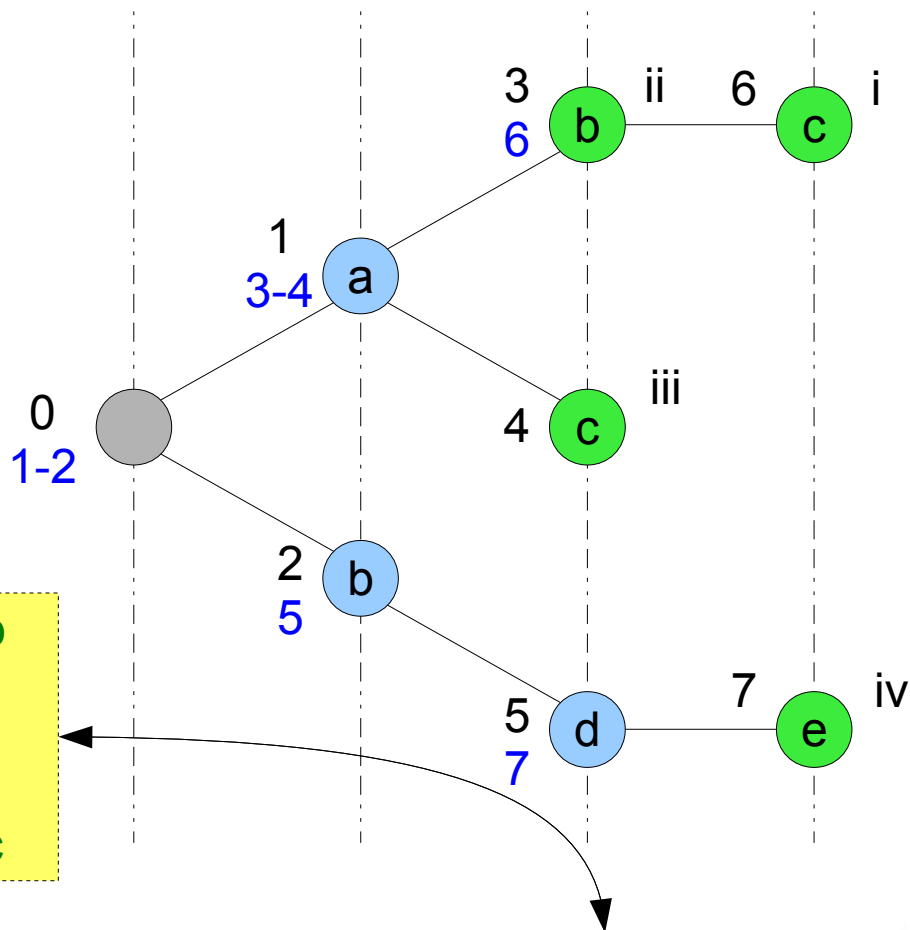
Индекс

Делаем таблицу
Лемма → Слова

- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»

a	1
b	2, 3
c	4, 6
d	5
e	7

α	a, b
β	e
γ	c, d
δ	b
ε	a, c



«людям»	(люди, человек)
«человеку»	(человек)

люди →	(люди, человек)
человек →	(человек), (люди, человек)

Индекс готов!

Теперь документ

Что делать с документом

- Подготовить документ
- Найти терминальные узлы в индексе
- Проверить расстояния между словами фраз
- Посчитать релевантность

ЕСТЬ КТО ЖИВОЙ? :-)

Что делать с документом

- Подготовить документ
- Найти терминальные узлы в индексе
- Проверить расстояния между словами фраз
- Посчитать релевантность

ГОТОВИМ ДОКУМЕНТ

На входе ⇐

Коллекция фраз. Фраза это:

- *Множество слов*

Документ это:

- *Большое множество слов*
 - *Позиция слова*
 - *Вес слова*

ГОТОВИМ ВОООБЩЕ

«людям» (люди, человек)

«человеку» (человек)

Надо чтобы «людям» из документа = «человеку» в индексе. Поэтому не подходит поиск в индексе по словам из документа. Надо искать через леммы.

Преобразуем документ в таблицу:

Лемма → Список словопозиций с весами

Оригинал документа больше не нужен.

ГОТОВИМ ДЛЯ ИНДЕКСА

В узлах индекса слова, а не леммы.

лемма слова

α	a, b
β	e
γ	c, d
δ	b
ϵ	a, c

Преобразуем документ в другую таблицу:

Слово в индексе → Словопозиции и веса

Мы перевели документ в пространство слов индекса.
Теперь всё найдём!

Первая таблица *Лемма → Список словопозиций с весами* МОЖЕТ понадобиться для другого индекса.

ИЩЕМ В ИНДЕКСЕ

(особая поисковая магия)

ПОИСК

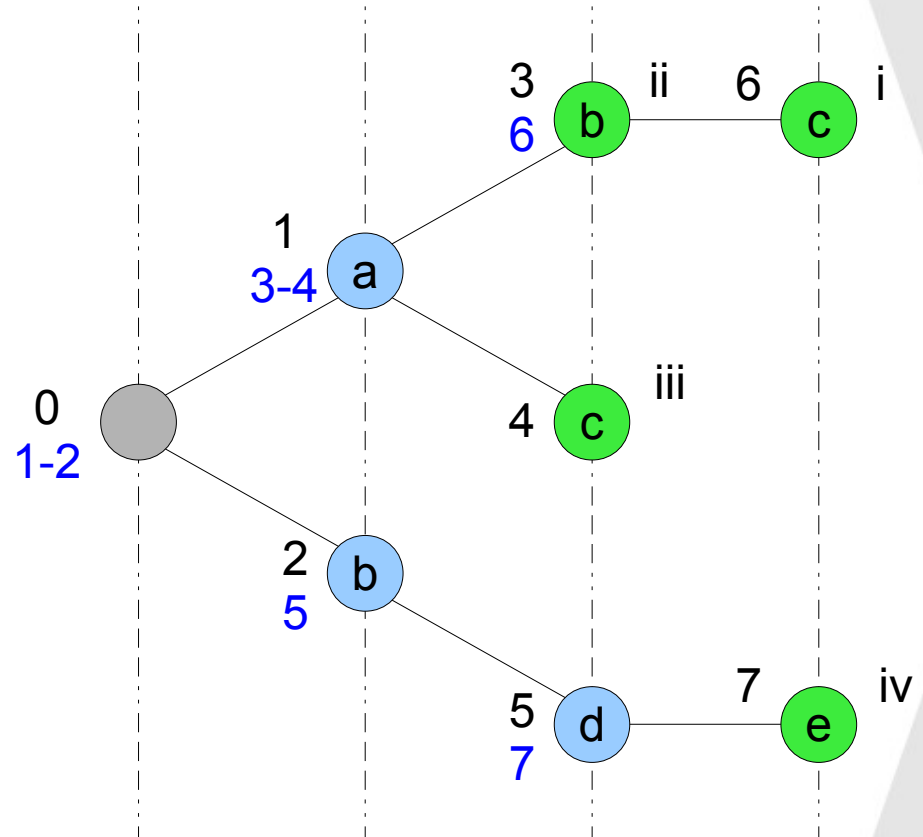
Это – разреженный интервал поиска™



Это документ

a	1:5	2:7	9:4
b	3:9	8:4	
c	5:6		
e	4:3	7:1	

слова, позиции и веса



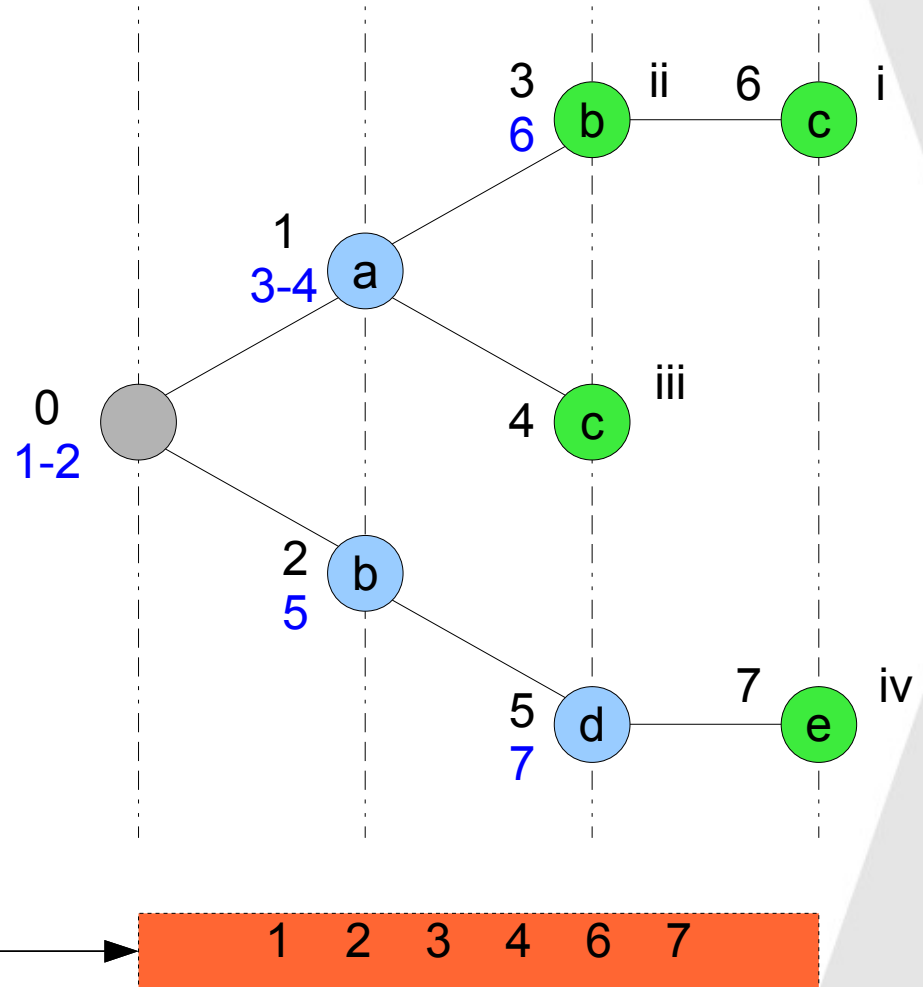
ПОИСК

Берём таблицу слова →
узлы индекса

a	1
b	2, 3
c	4, 6
d	5
e	7

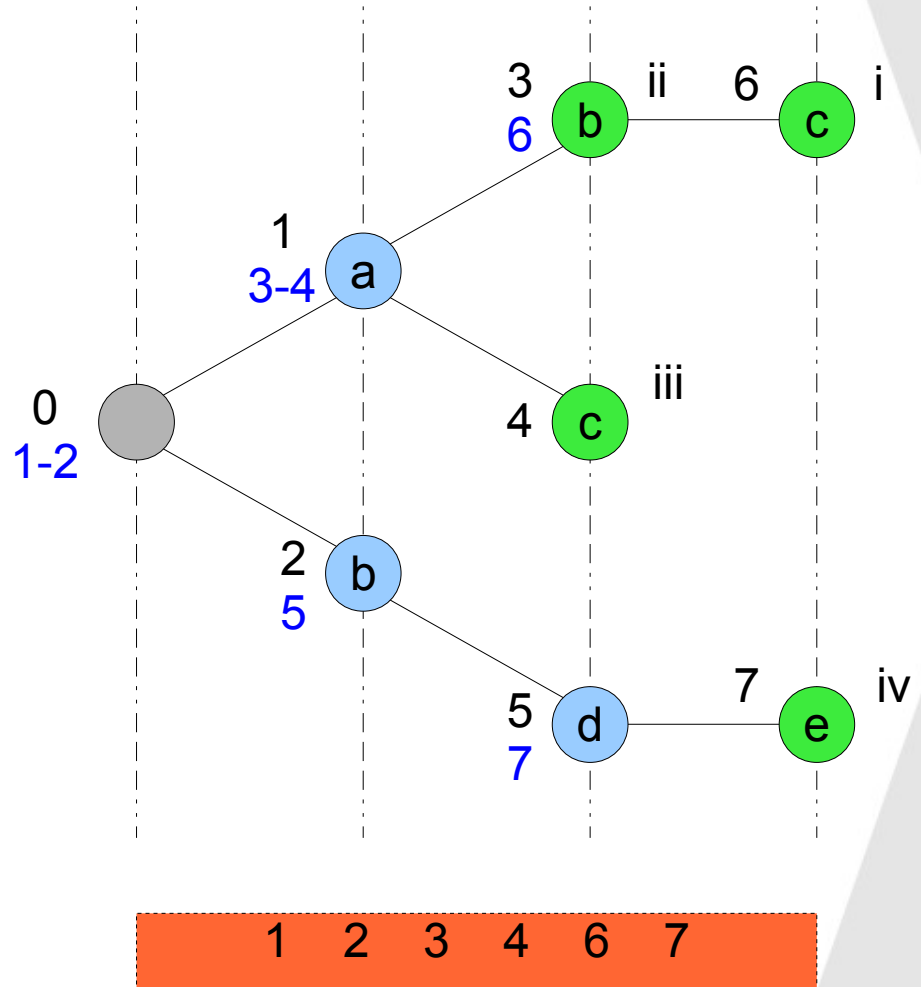
Делаем из документа
список узлов индекса

a	1:5	2:7	9:4
b	3:9	8:4	
c	5:6		
e	4:3	7:1	



ПОИСК

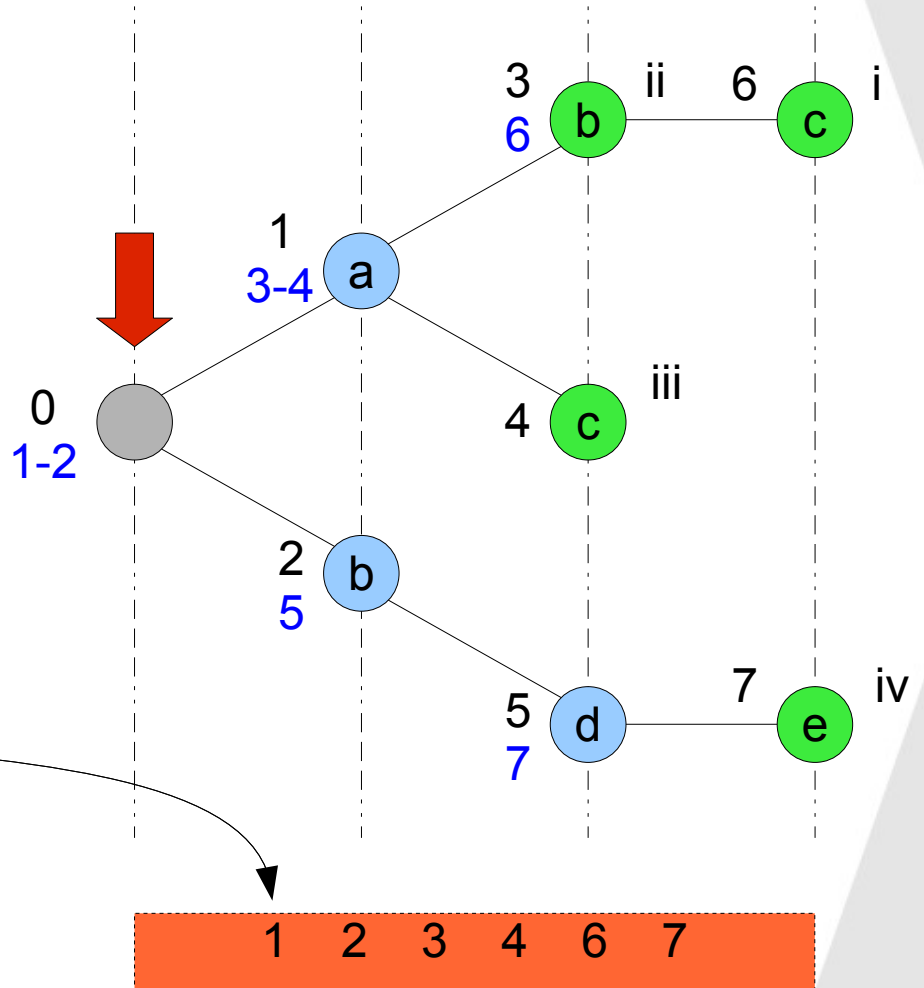
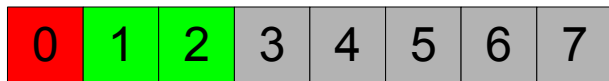
1. Начинаем в корне, с пустым интервалом поиска.
2. В каждом узле:
 - Обрезаем интервал от начала и до номера текущего узла включительно.
 - Если есть диапазон номеров дочерних узлов, добавляем его в интервал.
 - Если узел терминальный, добавляем его фразы в список найденных.
3. Находим нижнюю границу пересечения интервала с документом. Это следующий узел.



Пример

Пришли в узел 0

Добавляем в пустой интервал диапазон 1-2

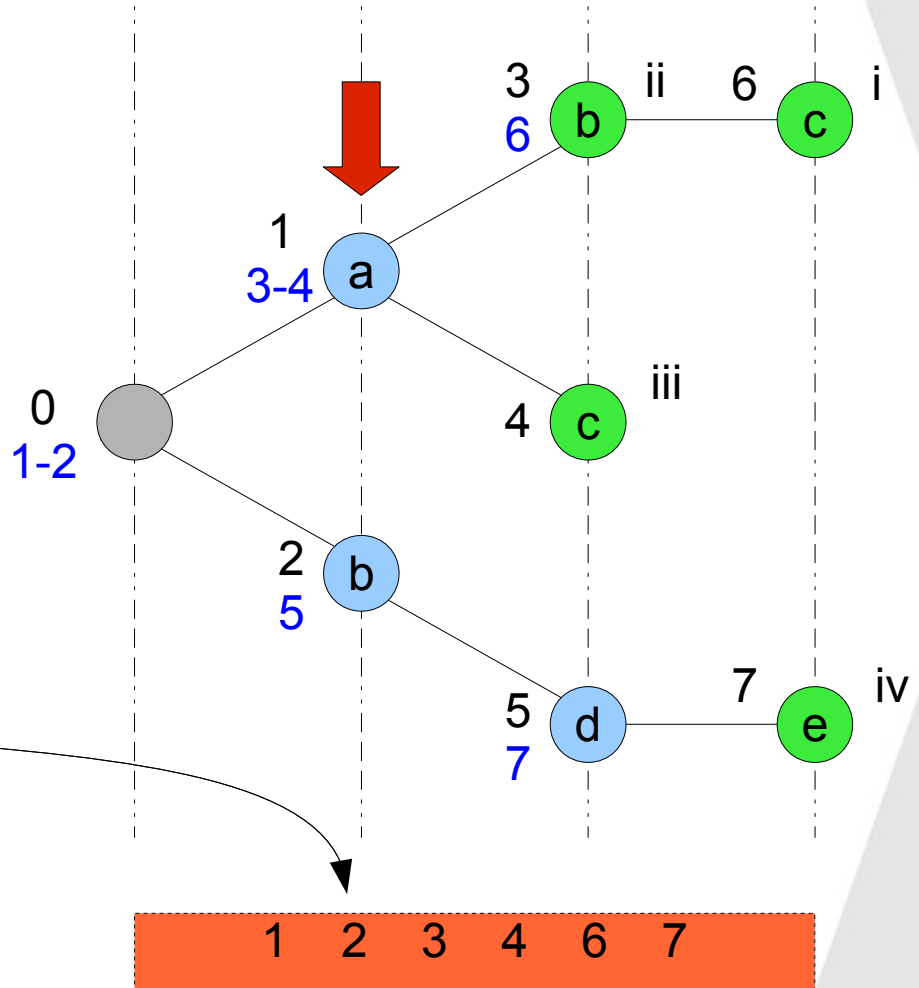


Пример

Пришли в узел 1

Удаляем из интервала 1

Добавляем диапазон 3-4

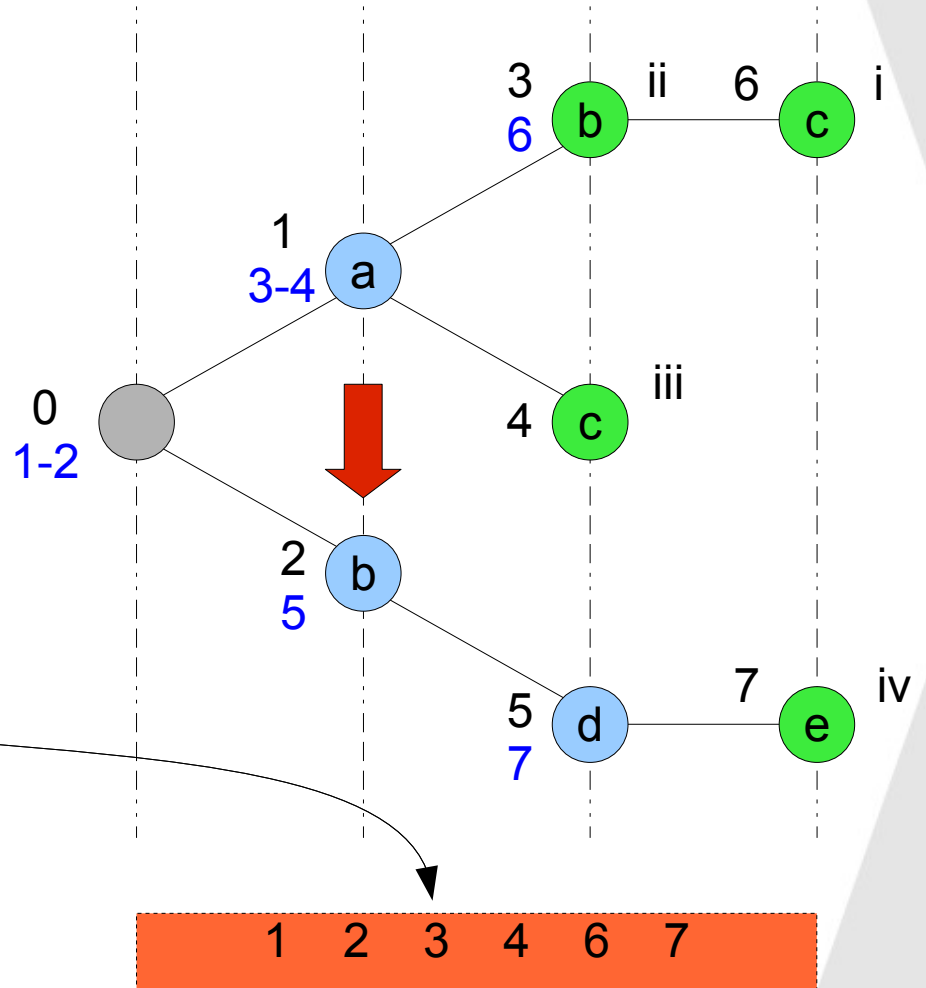


Пример

Пришли в узел 2

Удаляем из интервала 2

Добавляем 5



Пример

Пришли в узел 3

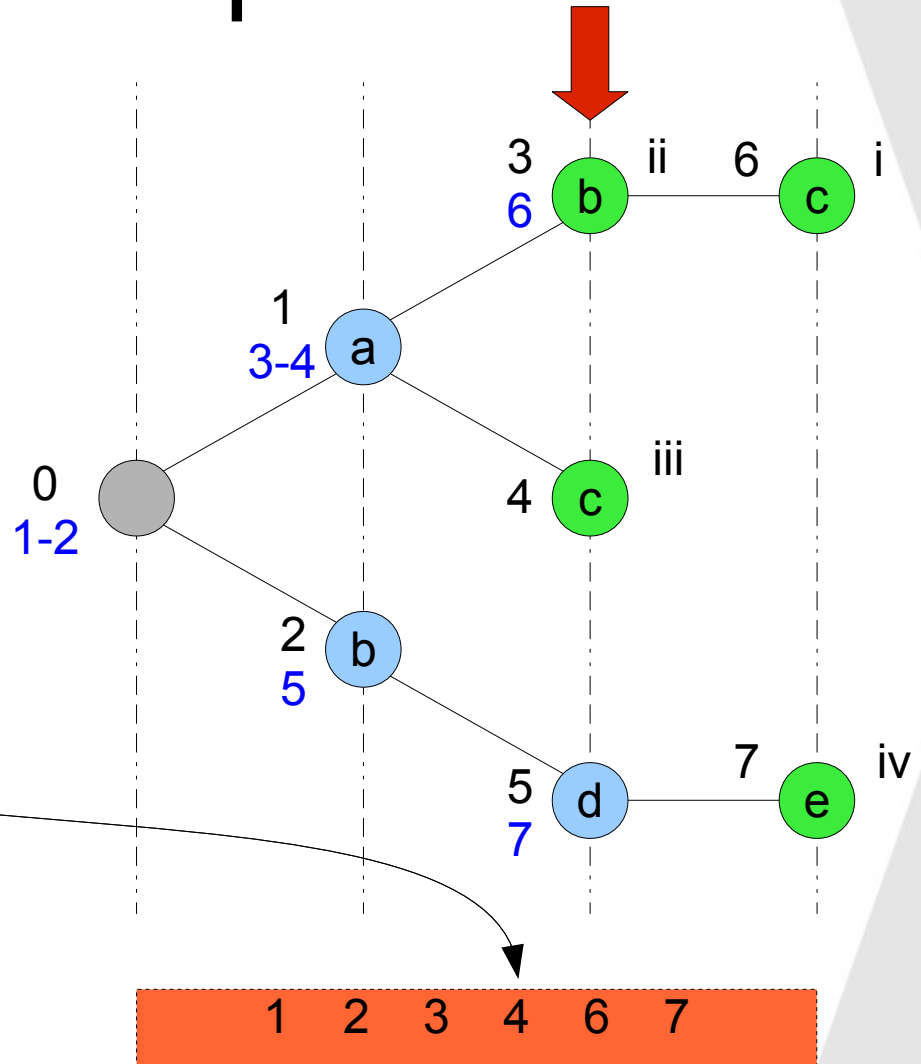
Удаляем из интервала 3

Добавляем 6

Нашли фразу «a b»!



- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»



Пример

Пришли в узел 4

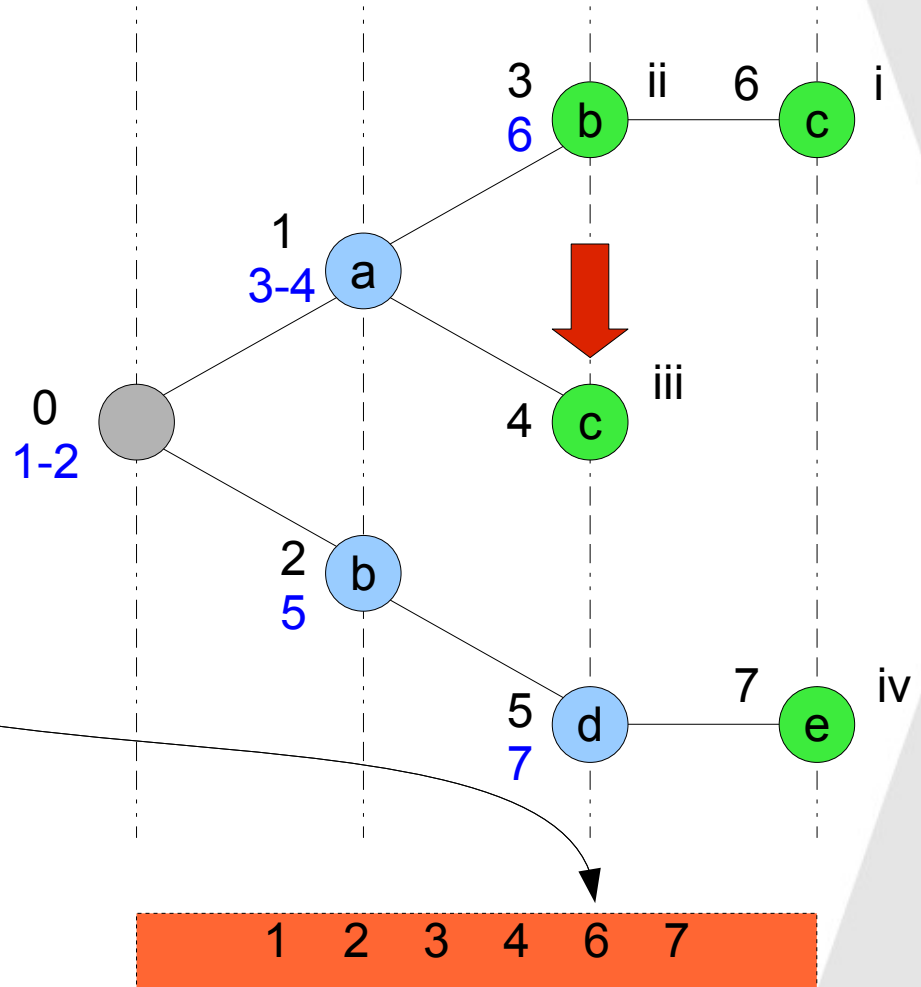
Удаляем из интервала 4

Добавлять нечего

Нашли фразу «a c»!



- i. «a b c»
- ii. «a b»
- iii. «a c»
- iv. «b d e»



Пример

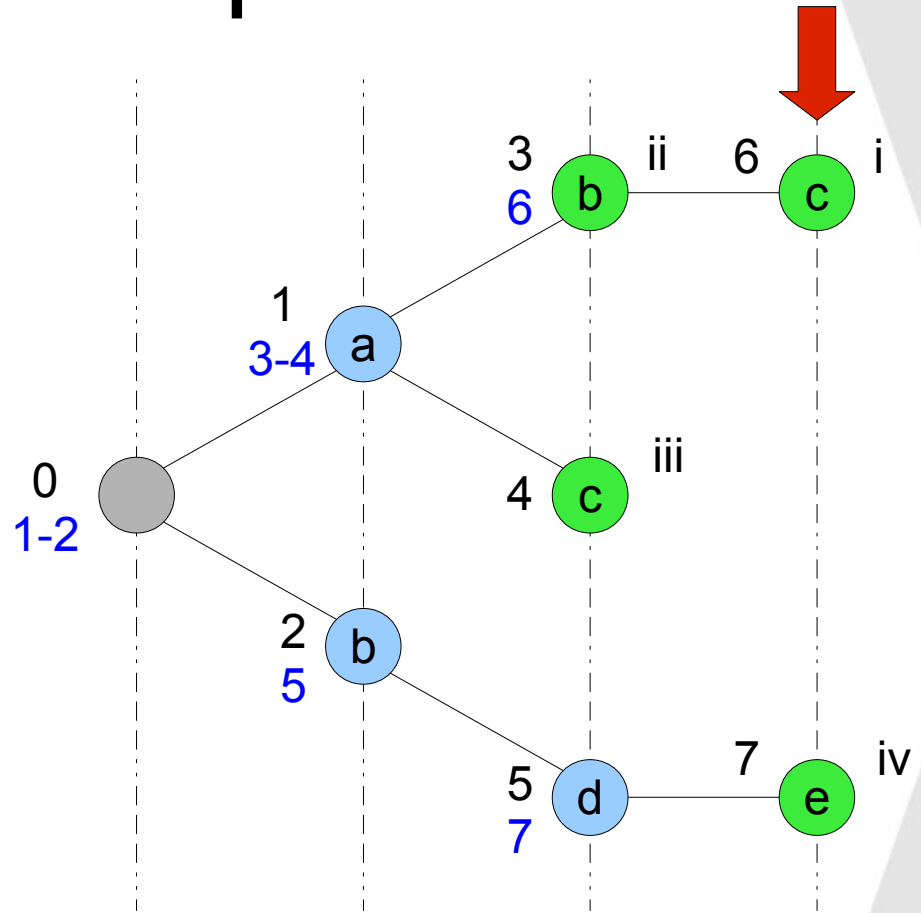
Пришли в узел 6

Удаляем из интервала 5-6

Добавлять снова нечего

Нашли фразу «a b c»!

Всё!



i. «a b c»

ii. «a b»

iii. «a c»

iv. «b d e»



Заметки к поиску

- Двигаемся по двум упорядоченным спискам чисел – высокая скорость поиска.



- В узлах можно проверять соответствие дополнительных условий фраз, «проходящих» через узел, и условий документа – можно не заходить в лишние ветки.
- Для более сложных индексов и документов интервал поиска действительно становится разреженным, т. е. зелёные и серые участки вперемешку.

Наш план

- О чём речь
- Кому это интересно
- Много букв (картинки будут)
- Ваши вопросы
- Мои ответы

Осталось чуть-чуть

- Проверить расстояния между словами найденных фраз
- Посчитать релевантность

Пример на буквах

Это две фразы:

- a c h
- d f p u

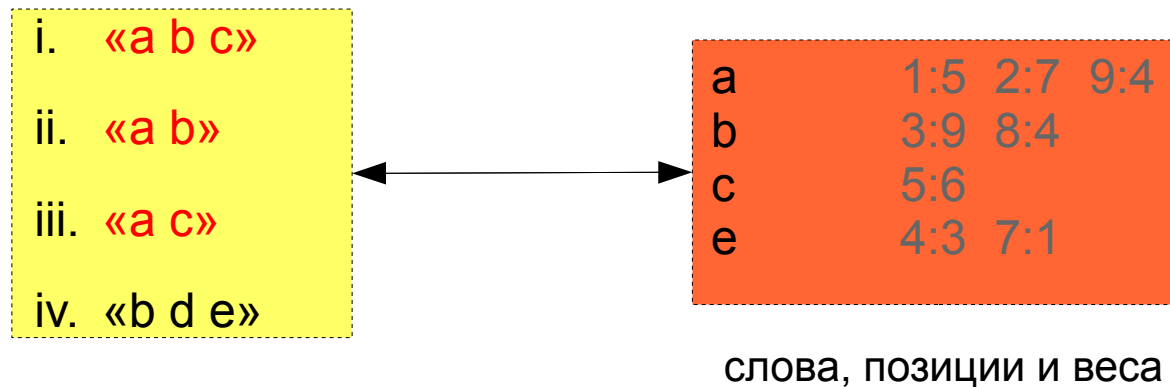
Это документ: a b c d e f g h i j k l m n o p q r s t u v w x y z

Нашли.

Но фраза d f p u слишком размазана по документу.

Проверка расстояний

Мы знаем слова найденных фраз и их позиции в документе.



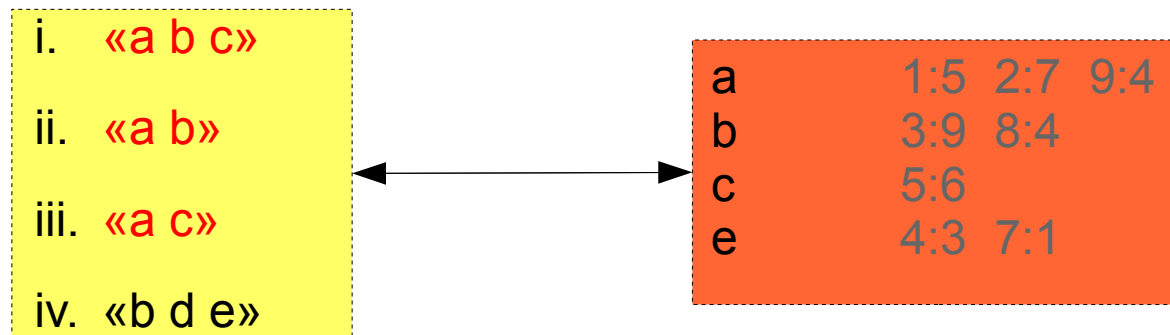
Делаем список словопозиций.

Привязываем позиции к фразам.

Проверяем за один проход.

Релевантность

У слов найденных фраз есть веса в документе.



слова, позиции и веса

Вычислить релевантность фраз можно, например, по TF-IDF. Или другим методом.

Вопросы?



Дмитрий Агафонов

Старший разработчик

111033, Россия, Москва,
ул. Самокатная, д. 1, стр. 21.

+7 (495) 739-00-00

+7 (495) 739-70-70 — факс

daga@yandex-team.ru