



ТЕХНОСФЕРА

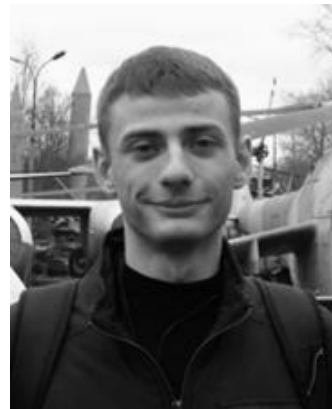
Введение в информационный поиск

Сергукова Юлия,
программист отдела инфраструктуры проекта
Поиск@Mail.Ru

Преподаватели



Юлия
Сергукова



Олег
Сафонов



Андрей
Мурашев



Евгений
Чернов



Дмитрий
Соловьев



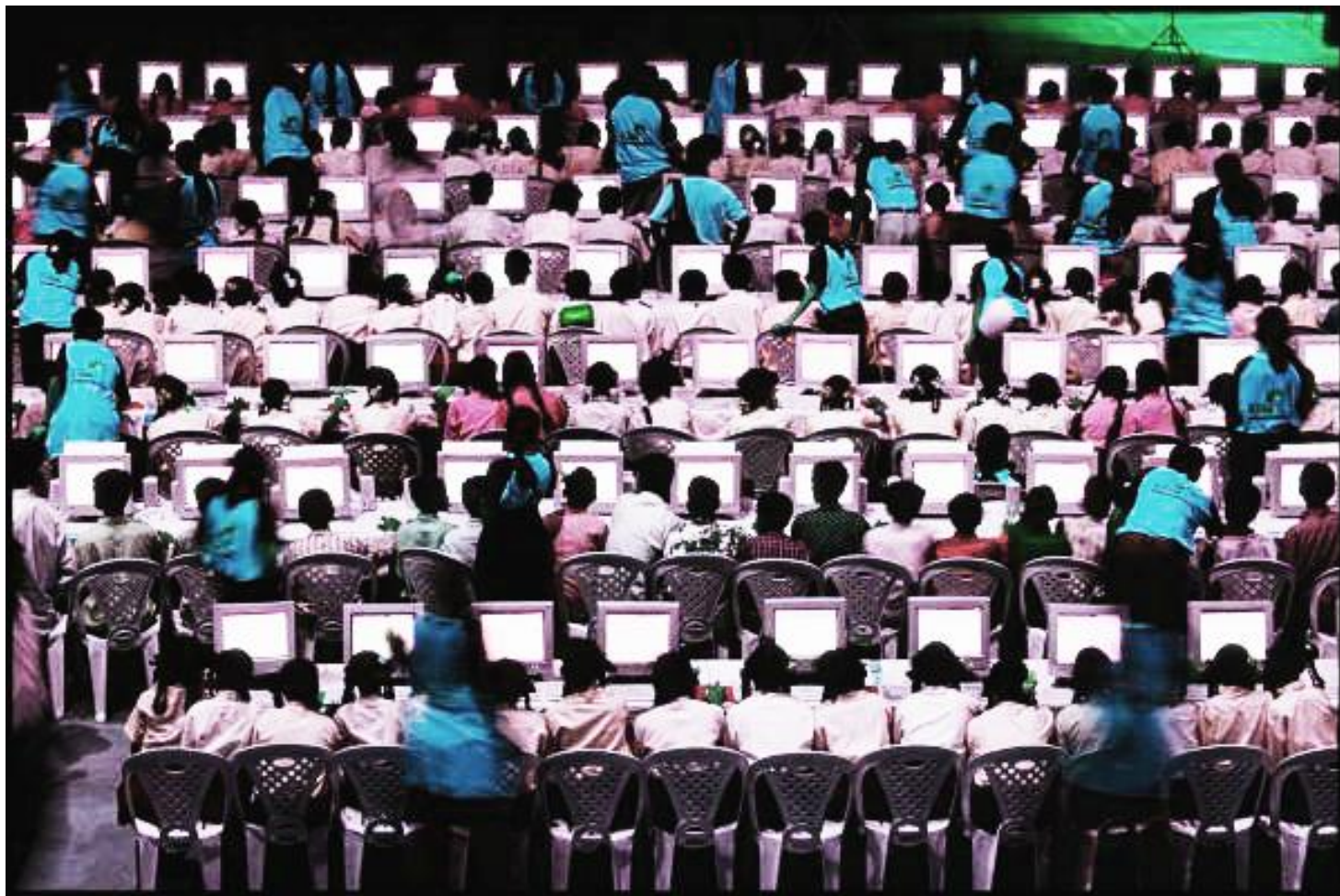
Владимир
Гулин



Михаил
Плеханов

Что такое Поиск?

Что такое Поиск?

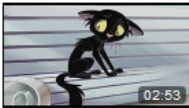



Поиск@Mail.Ru


- Текстовый поиск
- Картинки
- Видео
- Приложения
- Новости
- Музыка
- Ответы
- Обсуждения

поиск@mail.ru черный кот x

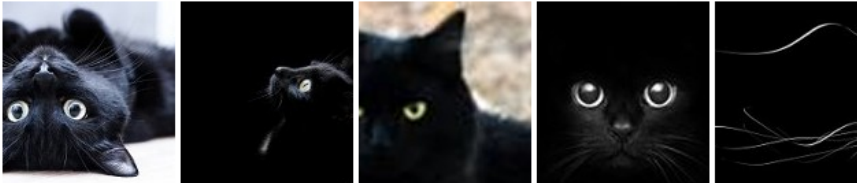
Интернет Картинки Видео Приложения Новости Ответы

 **Черный кот (Тамара Миансарова) — Youtube.com**
youtube.com/watch?v=QCeG1hy19sg
tabelta, 4 625 575 просмотров
Чёрный кот песня Юрия Саульского на стихи Михаила Танича. Написана в конце 1963 года Первое исполнение: Тамара Миансарова Видеоклип, в основе которого фрагме...

 **Сериал Черный кот Black Cat смотреть онлайн бесплатно!**
seasonvar.ru/serial-5006-CHernyj-kot
★★★★☆ 255 голосов
Черный кот – это не представитель семейства кошачьих, который вечно норовит перебежать нам дорогу.

 **Черный кот - смотреть онлайн аниме бесплатно все серии подряд в...**
onlinemultfilmy.ru/chernyj-kot
Тайная организация под названием «Хронос» из аниме «Черный кот» держит под своим контролем немалую часть экономики планеты, она стремится поддерживать...

Картинки



Технические детали

3500

серверов

15Pb

места в
Hadoop'е

8kM

документов в
индексе

50kM

документов в
Hbase'е

100

разработчиков

200

ассессоров

9

отделов:

ранжирование, рекомендации,
инфраструктура, анализ запросов,
лингвистика, вертикали, фронтенд,
качество, десктоп и мобильные системы

О курсе

Основан на **Introduction to information retrieval (Stanford)**

<https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

Читаем более 10 лет

Основной упор – на практическое применение

О курсе

Основан на **Introduction to information retrieval (Stanford)**

<https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>

Читаем более 10 лет

Основной упор – на практическое применение

Делайте домашки!

Структура курса

$\frac{1}{2}$: 7 лекций + 3 домашки \rightarrow 1-й коллоквиум

$\frac{1}{2}$: 6 лекций + 3 домашки \rightarrow 2-й коллоквиум

Домашки из первой половины можно пересдавать после первого коллоквиума, но за половину оценки

Делайте домашки!

Для допуска к первому коллоквиуму нужно набрать пороговое значение баллов.

Не набрали – прощаемся с вами в середине семестра.

Оценки

За что можно получить баллы:

- домашние работы
- активная работа на лекция и семинарах
- сдача коллоквиума

Оценки:

50 – 74 == «3»

75 – 89 == «4»

90 – 100 == «5»

Домашние работы

- проверяются автоматически
- плагиат карается: 0 баллов и запрещена пересдача
- все работы: закодировать то, что рассказывали на лекции
- формат: каггл, тесты и т.д.
- **Дедлайн!!!!** Обычно 2 недели.

Если что-то непонятно:

в первую очередь спрашивайте преподавателя,
который выдал домашку/читал лекции.

Блог – ваш друг

<https://sphere.mail.ru/blog/view/40/>

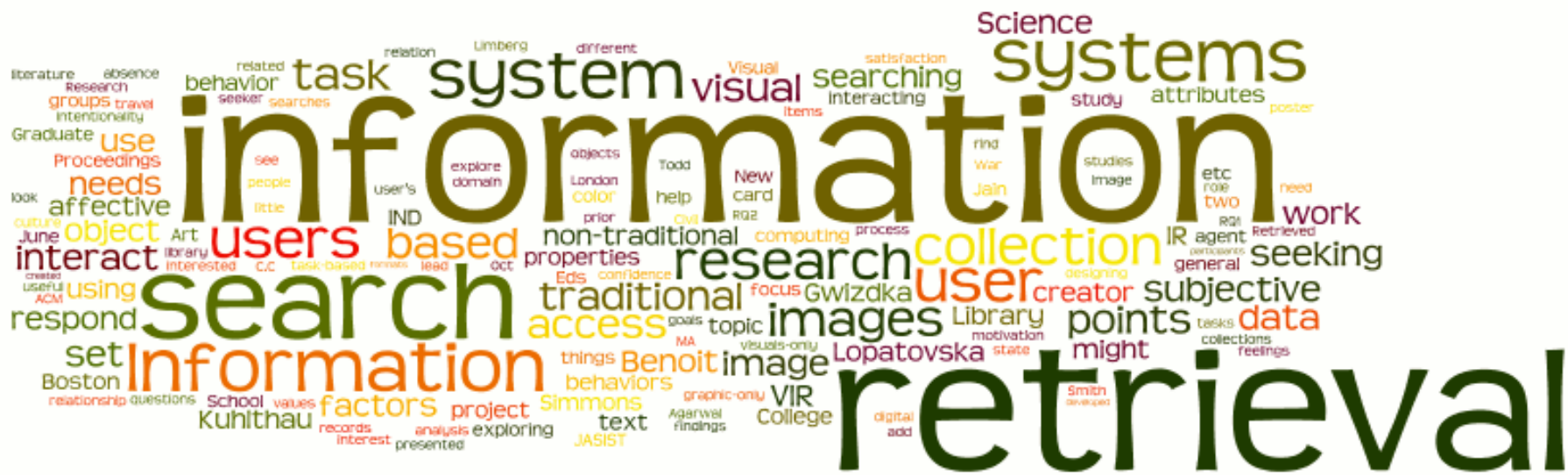
Зачем?

- Выкладываются лекции
- Перед семинаром даются указания
- Выкладываются домашки (и иногда результаты)
- Иногда бывают изменения в расписании

Лайки!



...а еще отзывы и посещаемость



Что такое информационный поиск

Информацио́нный по́иск — процесс поиска неструктурированной документальной информации, удовлетворяющей информационные потребности пользователя

Примеры информационных поисков

- **grep**
- **find**
- **SQL**
- **Maps.Me / Google Maps**
- **Запросы в багтрекере**
- **Поиск по микроблогам**
- **Поиск по картинкам**
- **Вэб-поиск**

Информационный поиск

Наука

- математические модели
- вычислительная лингвистика
- статистика и теорвер

Практика

- огромный объем данных
- KISS («keep it simple, stupid»)
- хаки практической реализации

Цель

«удовлетворение информационной потребности»

У пользователя есть потребность в получении определенной информации

Цель – найти **релевантную** его запросу информацию

Входные данные

- **корпус документов** – набор документов (текстов, изображений, новостных блоков и т.д.)
- **ссылочная информация** – PageRank
- **обратная связь** – запросы пользователей и клики в выдачу

Пример

Задача

Попасть в клуб РАЙ

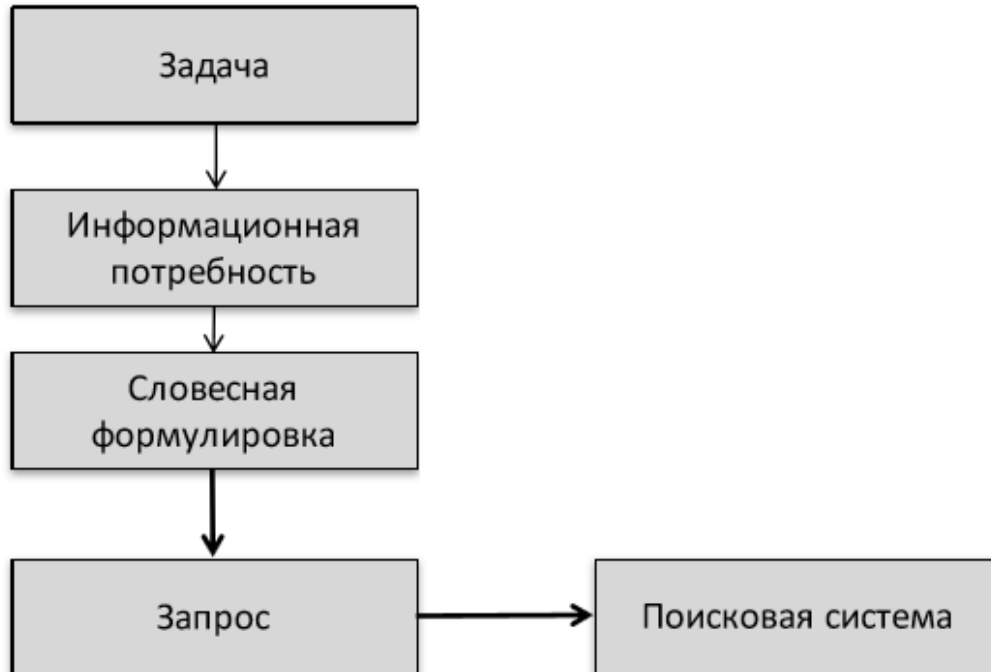
Пример



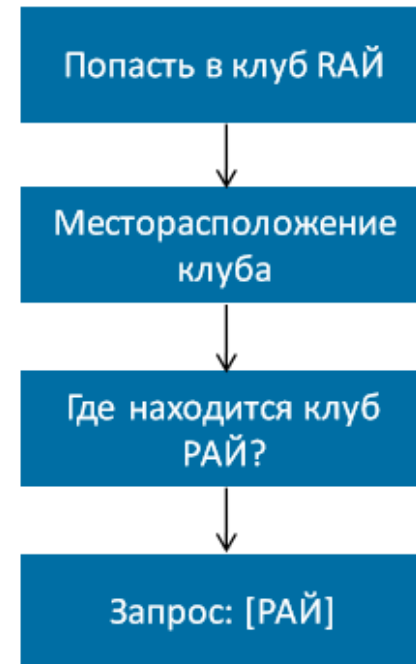
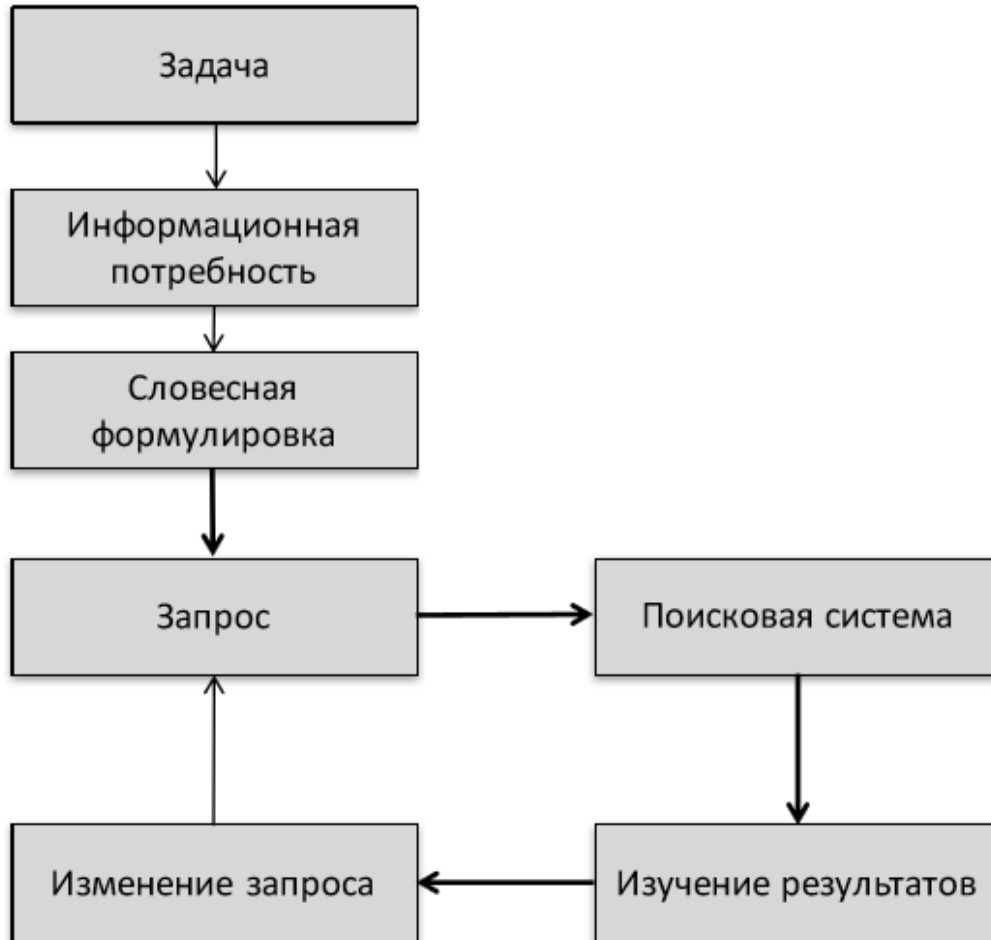
Пример



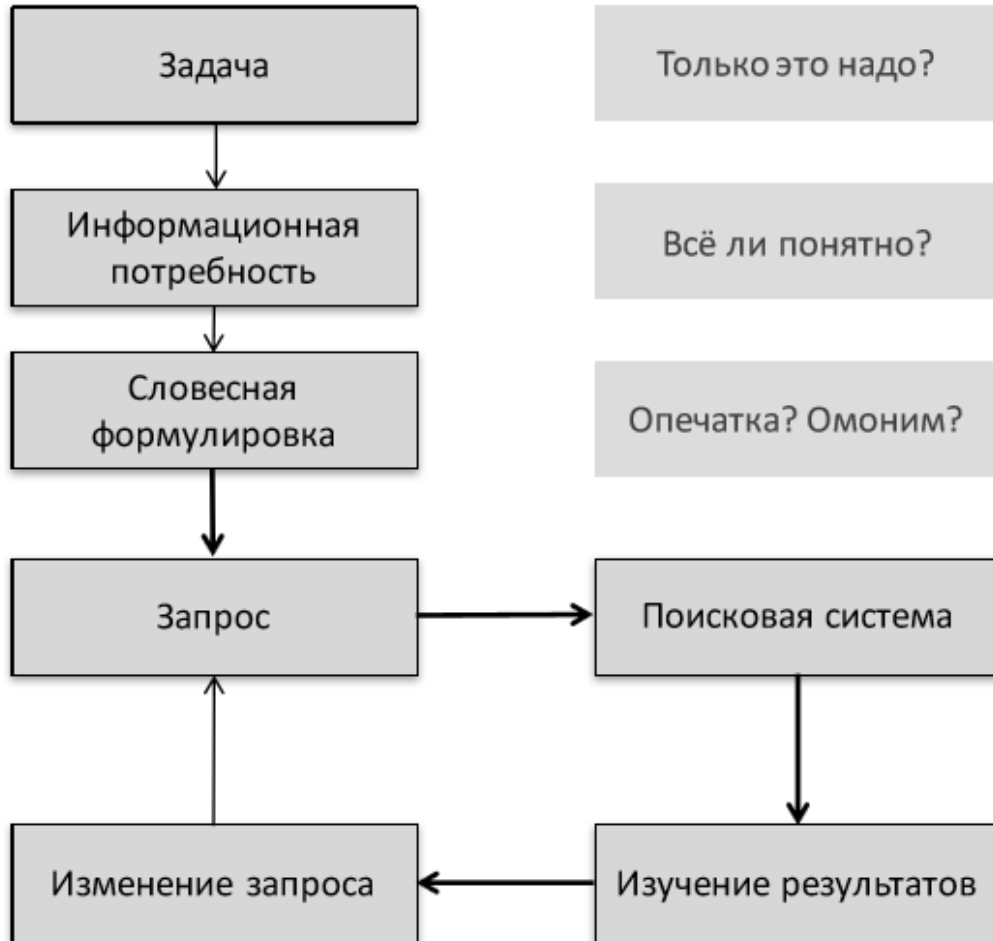
Пример



Пример



Пример



Сложности обработки запросов от людей

- лингвистика: что написано в тексте
- запросы: что хотел пользователь
- статистика
- машинное обучение
- огромные объемы данных
- большие нагрузки
- пользовательский интерфейс
- форматы данных

Связанные задачи

- рекомендательные системы
- машинный перевод
- извлечение мнений
- распознавание и синтез речи
- семантический анализ текста
- диалог с пользователем
- и т.д.





Примеры поисковых систем

Южная Корея - Naver



Google Scholar





Академия
Результатов: примерно 3 880 000 (0,05 сек.)

Статьи

Моя библиотека

За все время

С 2017

С 2016

С 2013

Выбрать даты

По релевантности

По дате

☒ включая патенты

☒ показать цитаты

☒ Создать оповещение

Information retrieval: data structures and algorithms
[WB Frakes](#), [R Baeza-Yates](#) - 1992 - citeulike.org
 Abstract Information retrieval is a sub-field of computer science that deals with the automated storage and retrieval of documents. Providing the latest information retrieval techniques, this guide discusses Information Retrieval data structures and algorithms,
 Цитируется: 2889 [Похожие статьи](#) [Все версии статьи \(4\)](#) [Цитировать](#) [Сохранить](#) [Ещё](#)

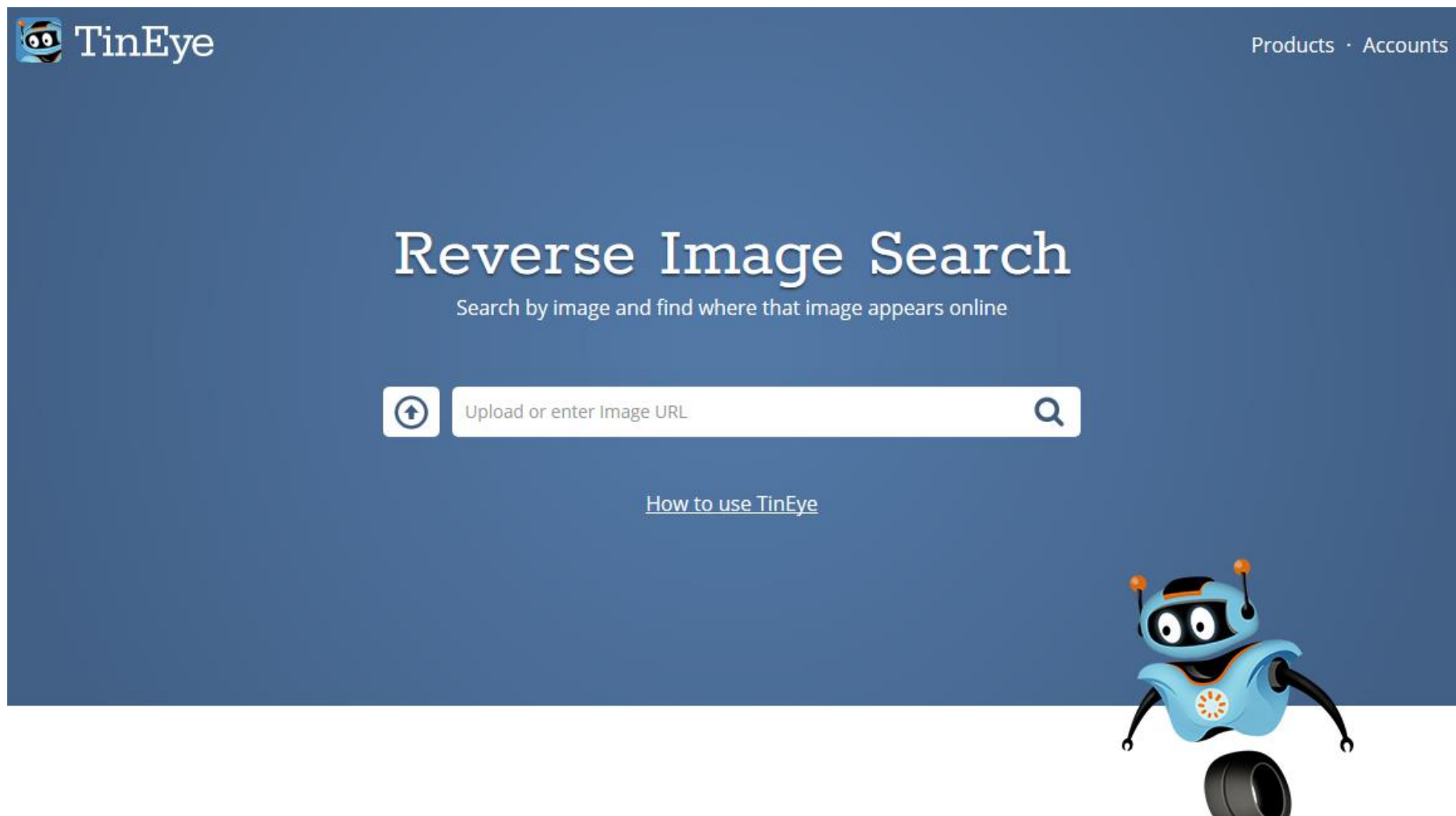
[ЦИТИРОВАНИЕ] Introduction to modern information retrieval
[G Salton](#), [MJ McGill](#) - 1986 - citeulike.org
 ... More... Brought to you by AQnowledge, precision products for scientists. x CiteULike uses cookies, some of which may already have been set. Read about how we use cookies. We will interpret your continued use of this site as your acceptance of our use of cookies. You may hide this
 Цитируется: 14231 [Похожие статьи](#) [Все версии статьи \(6\)](#) [Цитировать](#) [Сохранить](#) [Ещё](#)

Introduction
[DAC Manning](#) - Introduction to Industrial Minerals, 1995 - Springer
 Page 1. ~ ____ l_n_tr_o_d_u_c_t_i_o_n ____ ~1 ~ Human exploitation of minerals extends back for many thousands of years and, contrary to popular belief, mining may in fact be the 'oldest profession'. Early people used ...
 Цитируется: 12981 [Похожие статьи](#) [Все версии статьи \(16\)](#) [Цитировать](#) [Сохранить](#)

[книга] Modern information retrieval
[R Baeza-Yates](#), [B Ribeiro-Neto](#) - 1999 - mail.im.tku.edu.tw
Information retrieval (IR) has changed considerably in recent years with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out of date
 Цитируется: 15957 [Похожие статьи](#) [Все версии статьи \(41\)](#) [Цитировать](#) [Сохранить](#) [Ещё](#)



Не-текстовый поиск. TinEye



Контентные рекомендательные системы

Когда Каламу повесили на шею медаль победителя, он решил разделить эту победу с дочкой и отдал награду ей. Малышка даже попробовала медаль на зуб. «Это действительно важный день для меня, и я хочу разделить эту победу с младшей дочерью. То, что нам это удалось, говорит о том, что быть папой и работающим человеком вовсе не значит, что нужно отказываться от своих мечтаний, надежд и хобби», – заявил Калам.

Он рассчитывает стать примером для других семей, показывая, что активный образ жизни на свежем воздухе приносит пользу и детям, и взрослым.

Также смотрите [15 трогательных фото об отцовской любви](#).

Материал подготовила Татьяна Щеглова

11 февраля 2016

материал

Все темы



Обновления в статье: подписка, комментарии, выделите ее в закладки On • Email

Все новости

Материалы по теме



Климова рассказала об интересной семейной традиции



Ученые предупредили об опасностях позднего отцовства



Каким подарком получили принцесса Шарлотта с братом



Анна Хилькевич назвала дочь в честь себя и мужа



Константин Хабенский скоро снова станет папой

Директ

Системы Аквафор Морион от 5790 р.
telk.shop.aquaphor.ru



Рекуператор воздуха квартирный
mailey-rus.ru



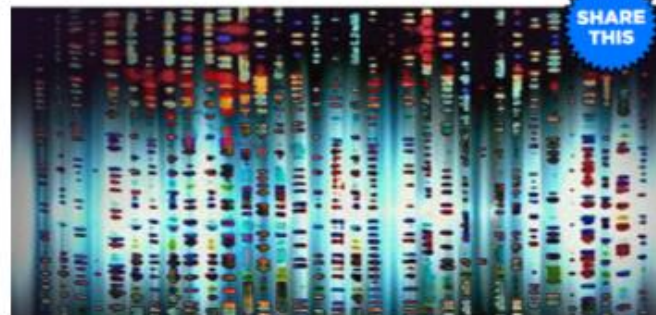
Приточная вентиляция в квартире
свежий-воздух.рф

Приточная и приточно-вытяжная вентиляция в квартире. Монтаж. Распродажа без%

Адрес и телефон

Pandora's "Music Genome Project" explores the cold hard facts of how we interact with music

Twenty-five music analysts "grade" 10,000 songs a month. It's a mountainous job, writes Rob Pegoraro, but the results will be serious business.



SHARE THIS

Поиск по структурированной информации и неструктурированным запросам



☆≡

📄 📷 📊 🔊 🌐 Web Apps ☰ Examples 🔀 Random

Assuming "apple" is a food | Use the input as a [formula](#) instead

Assuming apple | Use [prepared apples](#) or [more](#) instead

Assuming any type of apple | Use [apple, with skin](#) or [apple, without skin](#) instead

Input interpretation:

apple	amount	1 apple	mass
-------	--------	---------	------

Result:

182 grams

[Show non-metric](#)

Unit conversions:

0.18 kg (kilograms)

0.4 lb (pounds)

6.4 oz (ounces)

Physical properties:

mass	182 grams
serving volume	233 mL (milliliters)

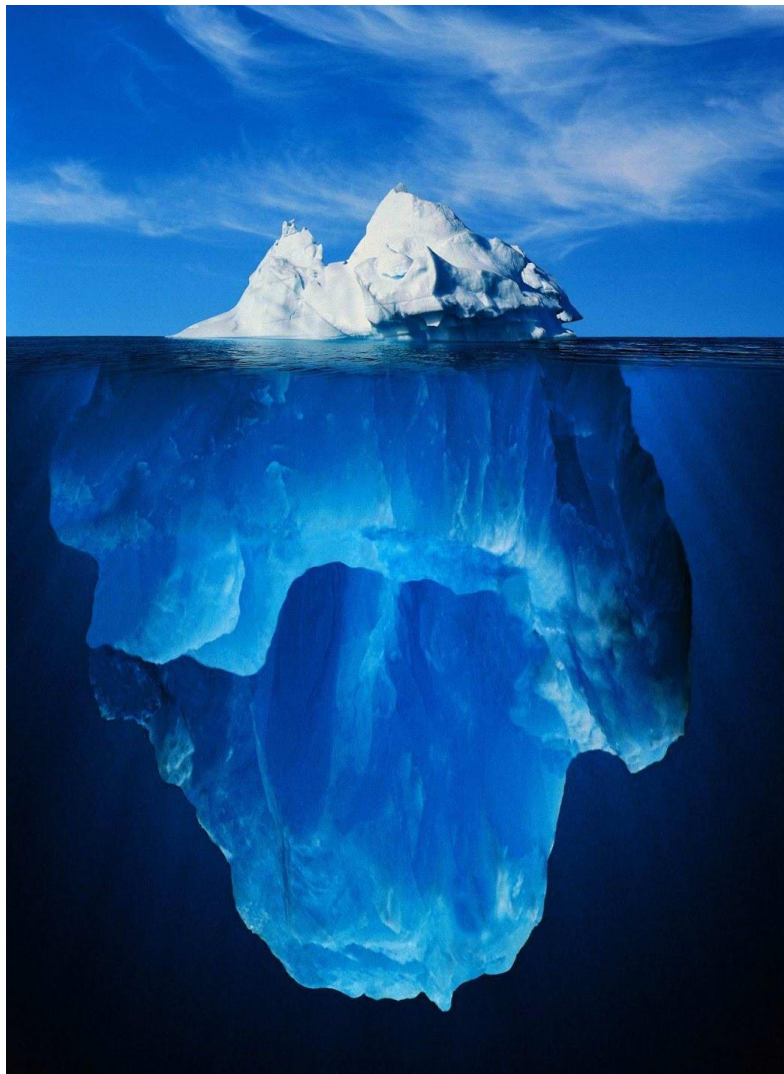
IBM Watson



Обзор курса



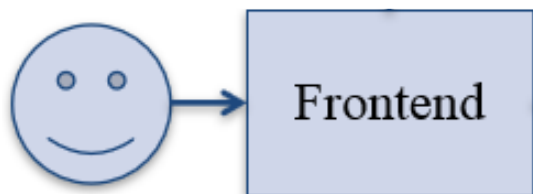
Обзор курса



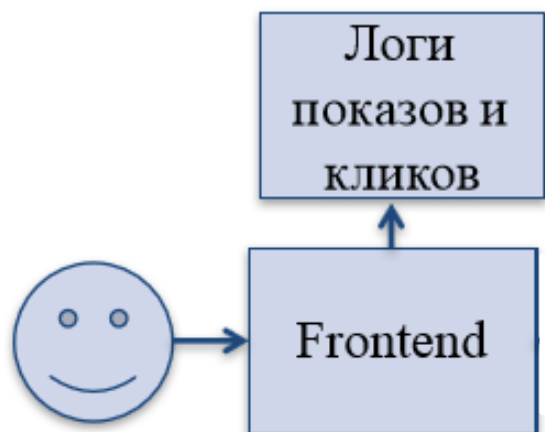
Как работает поиск



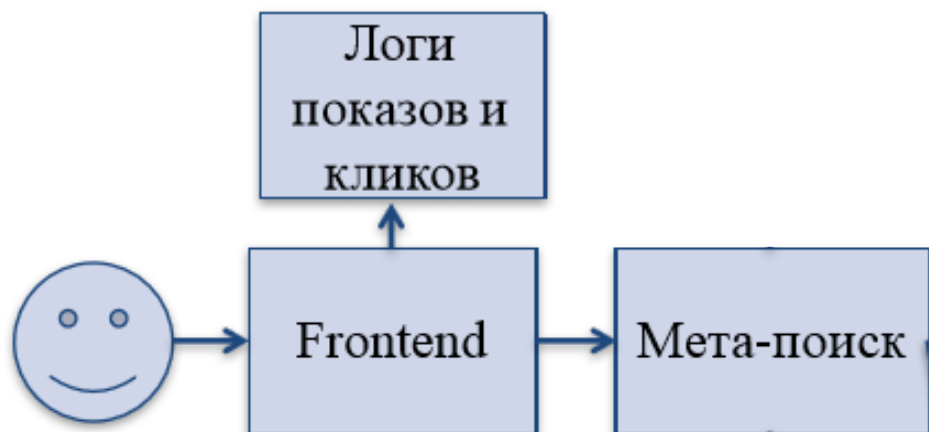
Как работает поиск



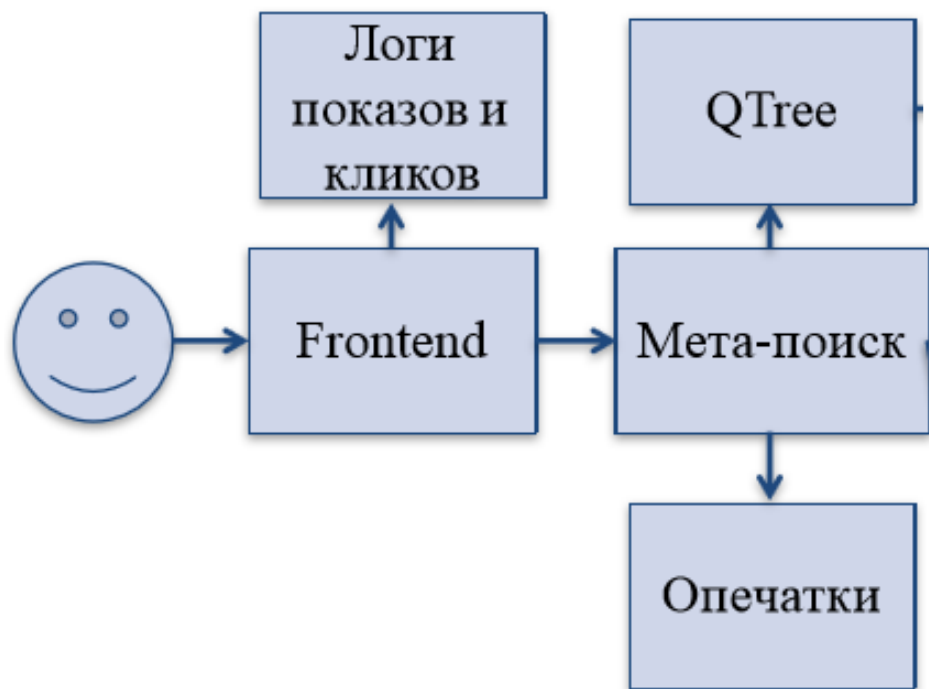
Как работает поиск



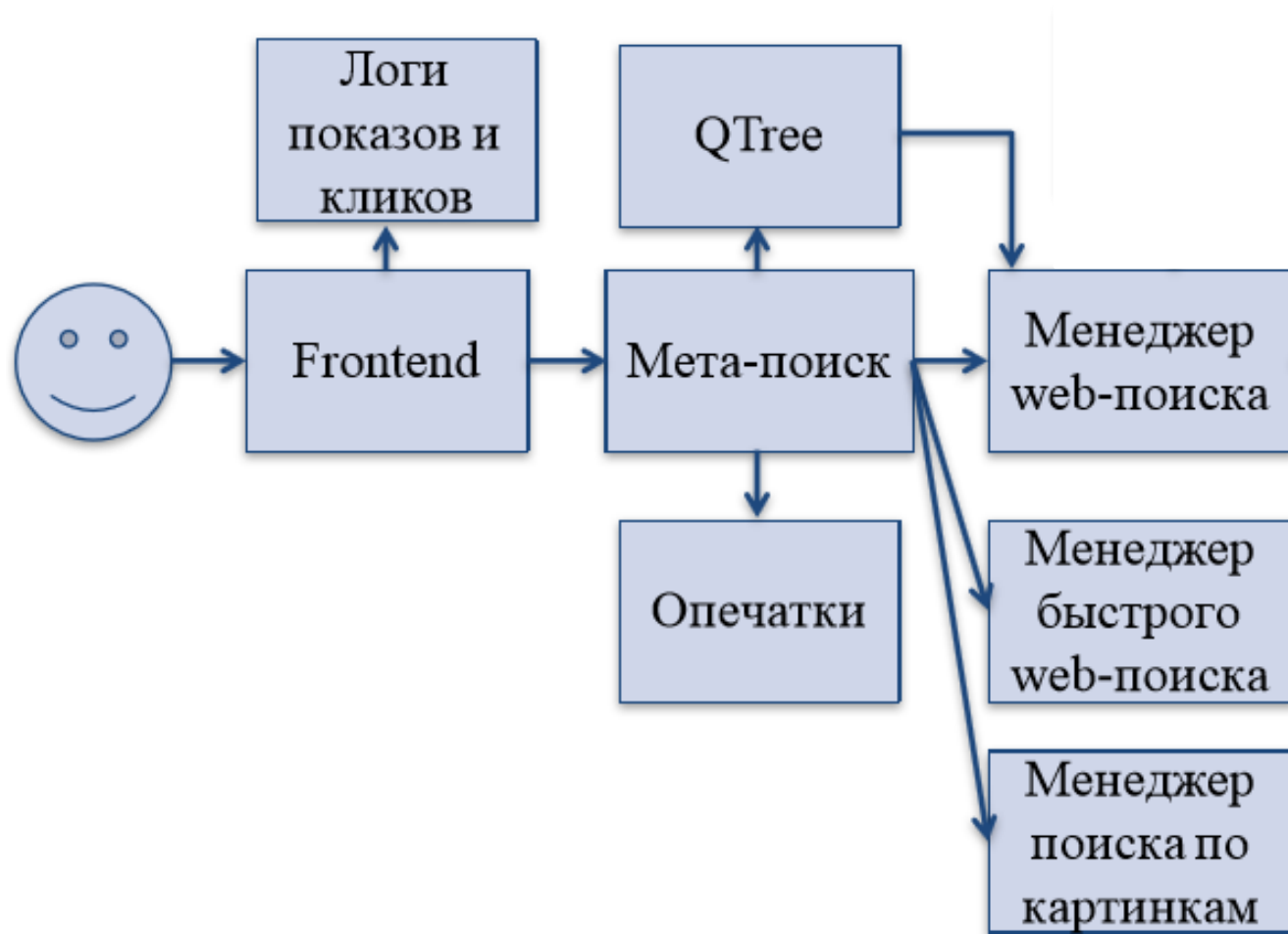
Как работает поиск



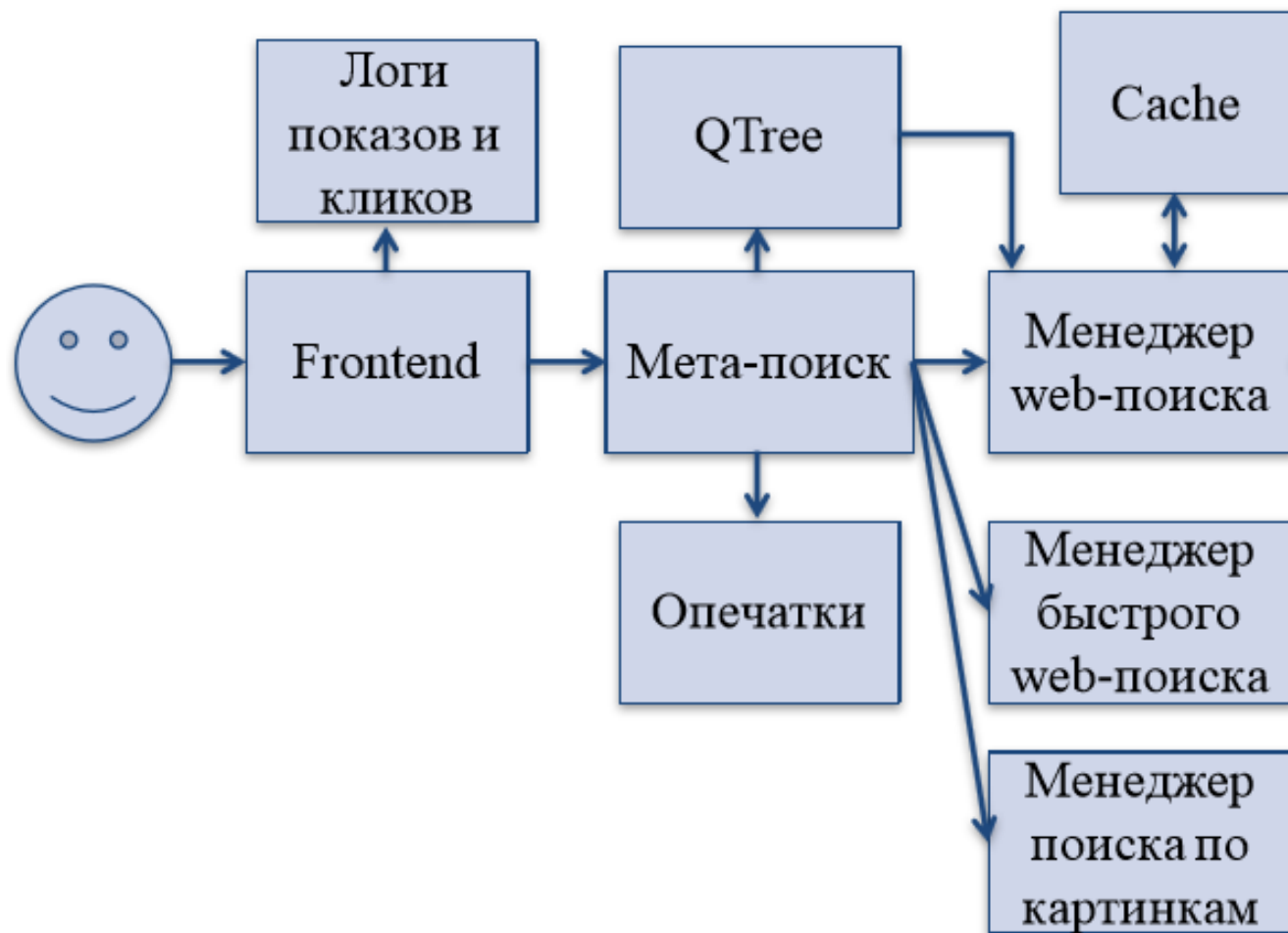
Как работает поиск



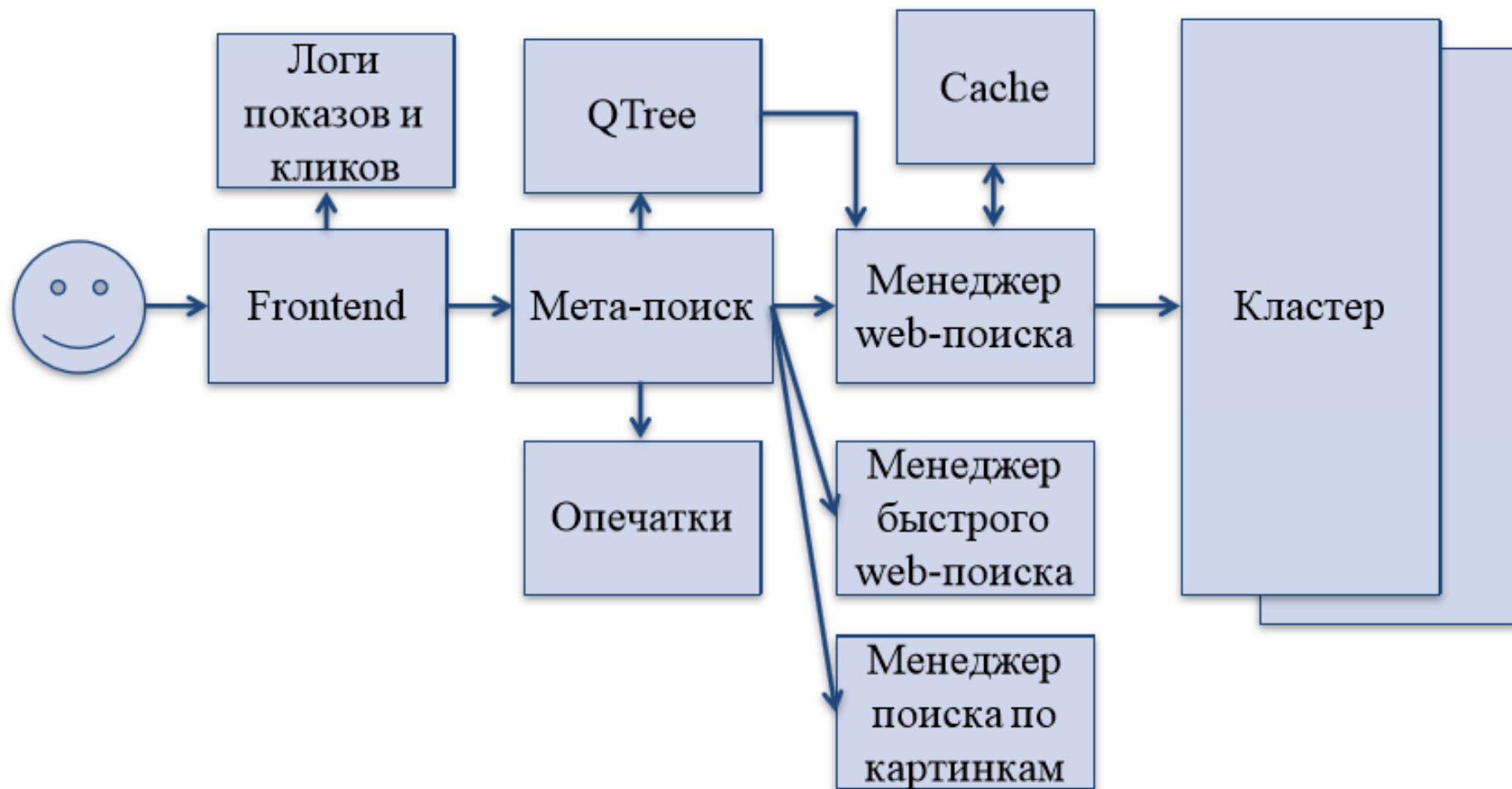
Как работает поиск



Как работает поиск



Как работает поиск



О чем мы будем говорить на лекциях

Откуда взять данные (документы)?

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. **Индексация** (лекция 4) и **оптимизация индекса** (лекция 5)

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
- 4. Поиск дубликатов (лекции 6-7)**

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
4. Поиск дубликатов (лекции 6-7)

* коллоквиум! *

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
4. Поиск дубликатов (лекции 6-7)
5. **Анτισпам и антипорн** (лекции 8 и 9)

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
4. Поиск дубликатов (лекции 6-7)
5. Антиспам и антипорн (лекции 8 и 9)
6. **Микроразметка** (лекция 10)

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
4. Поиск дубликатов (лекции 6-7)
5. Антиспам и антипорн (лекции 8 и 9)
6. Микроразметка (лекция 10)
- 7. С니ппеты** (лекция 11)

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
4. Поиск дубликатов (лекции 6-7)
5. Антиспам и антипорн (лекции 8 и 9)
6. Микроразметка (лекция 10)
7. С니ппеты (лекция 11)
8. **Исправление опечаток** (лекция 12)

О чем мы будем говорить на лекциях

1. Поисковый спайдер (лекция 2)
2. Планировщик поискового спайдера (лекция 3)
3. Индексация (лекция 4) и оптимизация индекса (лекция 5)
4. Поиск дубликатов (лекции 6-7)
5. Антиспам и антипорн (лекции 8 и 9)
6. Микроразметка (лекция 10)
7. С니ппеты (лекция 11)
8. Исправление опечаток (лекция 12)
9. **Саджесты и переформулировки** (лекция 13)

Спасибо за внимание!

**Не забудьте отметить на сайте
и оставить отзыв**