

#### Введение в информационный поиск

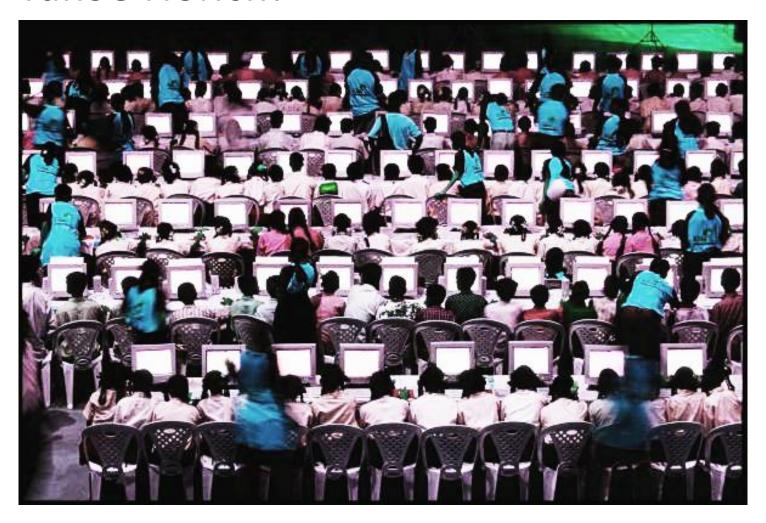
Сергукова Юлия, программист отдела инфраструктуры проекта Поиск@Mail.Ru



#### Что такое Поиск?



#### Что такое Поиск?





#### Что такое Поиск?





#### Поиск@Mail.Ru



рики-тики-тави

Найти

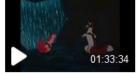
Интернет

Картинки

Видео Новости

#### Видео по запросу «рики-тики-тави»

3 тыс. результатов









×

#### Сборник Мультфильмов Рикки-Тикки-Тави - Рики-Тики-Тави

5 сентября 2013

Youtube

28 мая 2012

РИКИ. ТИКИ. **ТАВИ**. 1975 г. СССР

Ok.ru

7 сентября 2015

Рики -(1965)Ok.ru

15 июл:

Mail.ru

#### Рикки-Тикки-Тави — Youtube.com №

youtube.com/watch?v=ar-lv03EIXI

Мультики студии Союзмультфильм, 3 196 020 просмотров

Мультики: Рикки тикки тави. Советский мультфильм для детей по рассказу Р. Киплинга. Нажми лайк, если любишь советские мультфильмы. Добрые советски...

#### 👓 Сказка Рикки Тикки Тави - Редьярд Киплинг 🔮

deti-online.com/skazki/skazki-

Рикки Тикки Тави - сказка Киплинга. Рикки Тикки Тави читать онлайн или скачать в формате doc или pdf.







#### Рикки-Тикки-Тави

«Рикки-Тикки-Тави» — рассказ из «Книги джунглей» Редьярда Киплинга, а также имя главного героя этого рассказа. Википедия

Год первого издания: 1893 Авторы: Редьярд Киплинг

Сообщить об ошибке Википедия



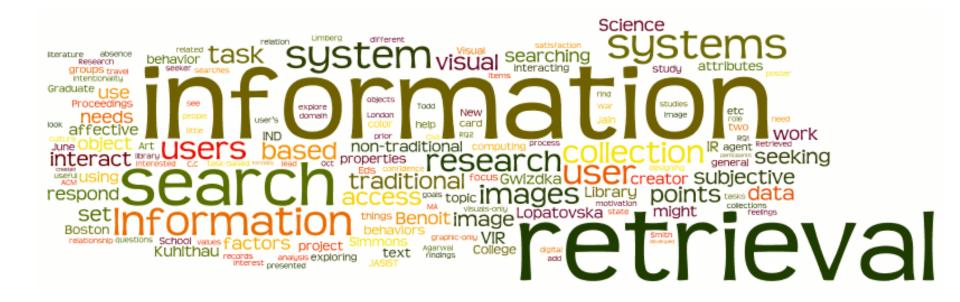
#### Поиск@Mail.Ru

- Текстовый поиск
- Картинки
- Видео
- Новости
- Ответы
- Карточка «быстрого ответа»

#### а еще:

- Приложения
- Обсуждения
- и многое другое







### Что такое информационный поиск

Информационный поиск — 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 + Практика



## Информационный поиск

#### Наука

- математические модели
- вычислительная лингвистика
- статистика и теорвер



## 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 пользователя»



#### Цель

# «удовлетворение информационной потребности пользователя»

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



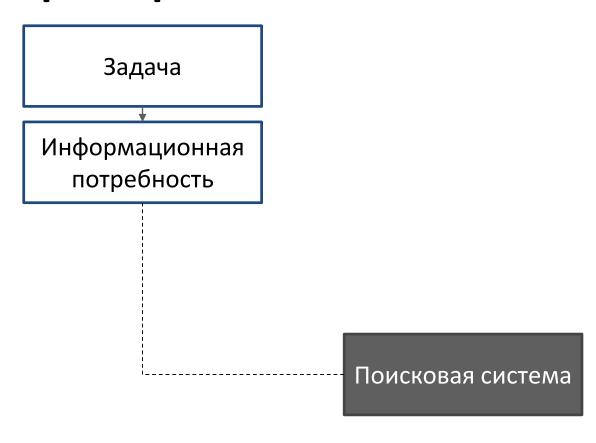
Задача

Поисковая система

Смерть («Плоский мир» Т.Пратчетта)







Смерть («Плоский мир» Т.Пратчетта)

Статья/обсуждение персонажа



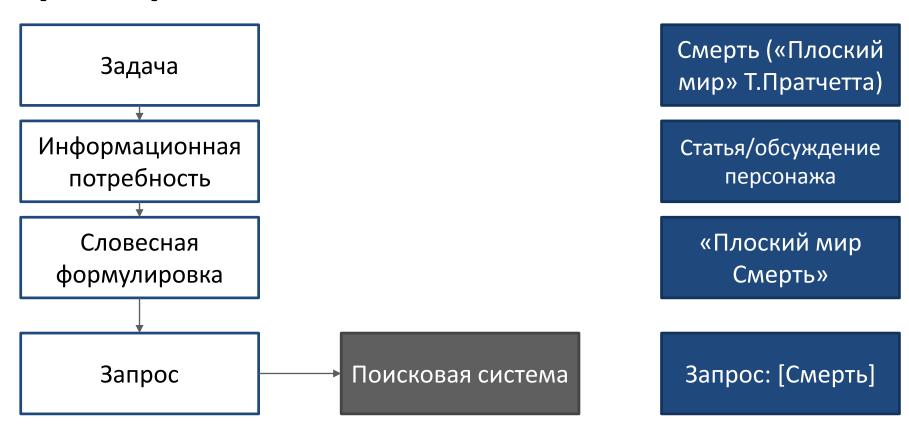


Смерть («Плоский мир» Т.Пратчетта)

Статья/обсуждение персонажа

«Плоский мир Смерть»











#### Что может пойти не так?



#### Что может пойти не так?

- Не поняли пользователя
- Нашли не то



#### Что может пойти не так?

- Не поняли пользователя
  - Синонимы
  - Опечатки
  - Несодержательный запрос
- Нашли не то
  - Омонимы
  - Нет знаем нужный документ (непопулярные запросы)
  - Пользователь не согласен с результатами





#### Связанные задачи

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





#### Примеры поисковых систем

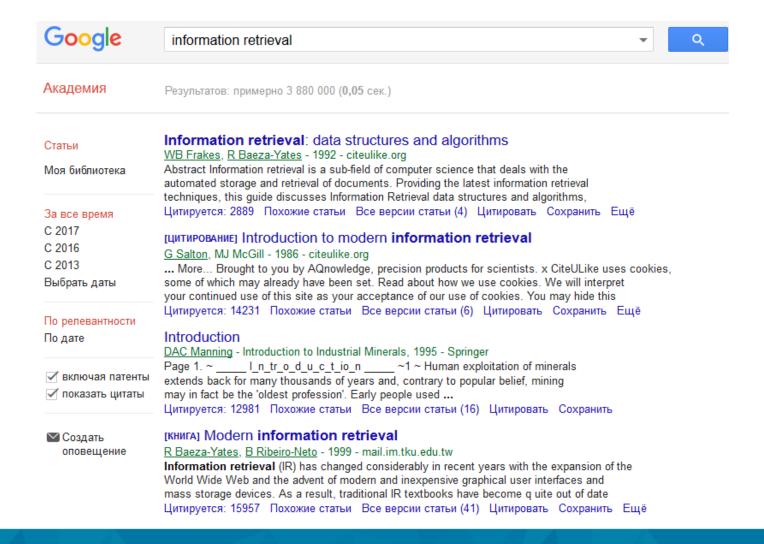


## Южная Корея - Naver



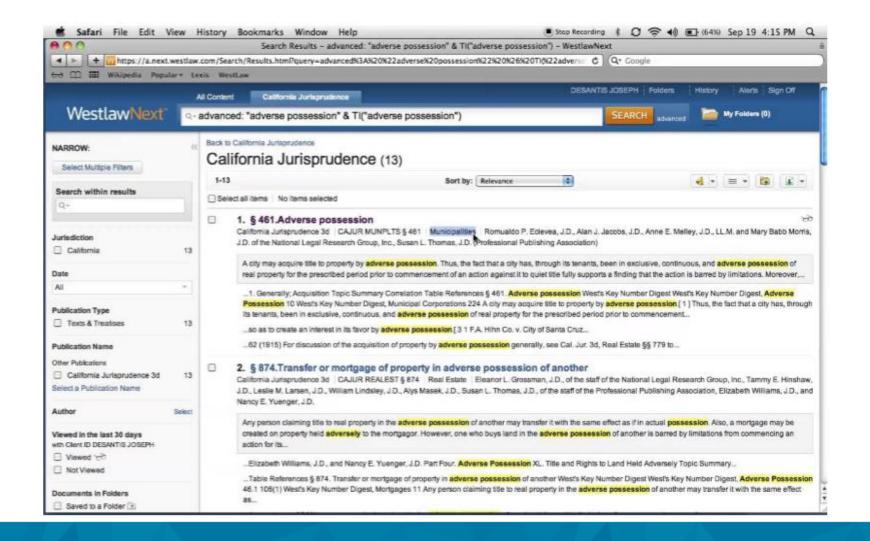


## **Google Scholar**



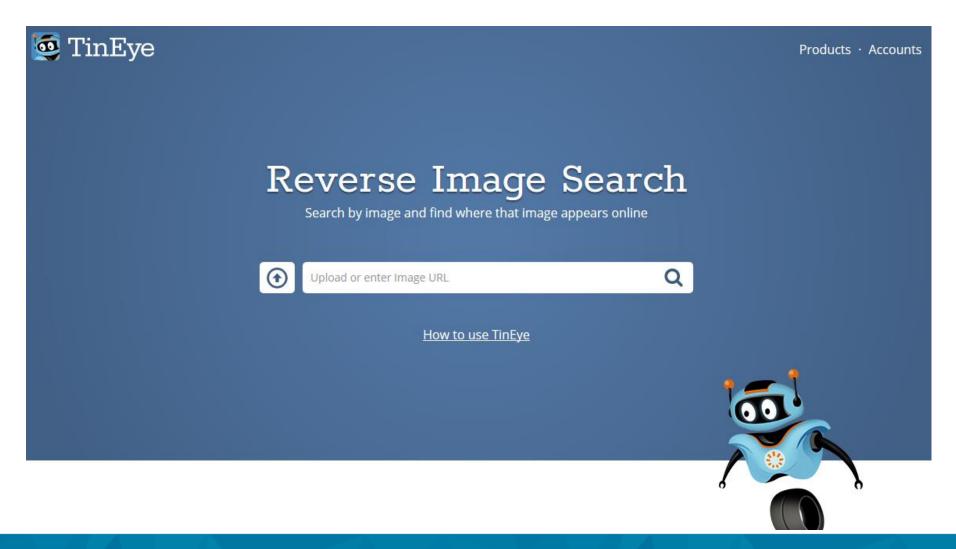


#### Westlaw



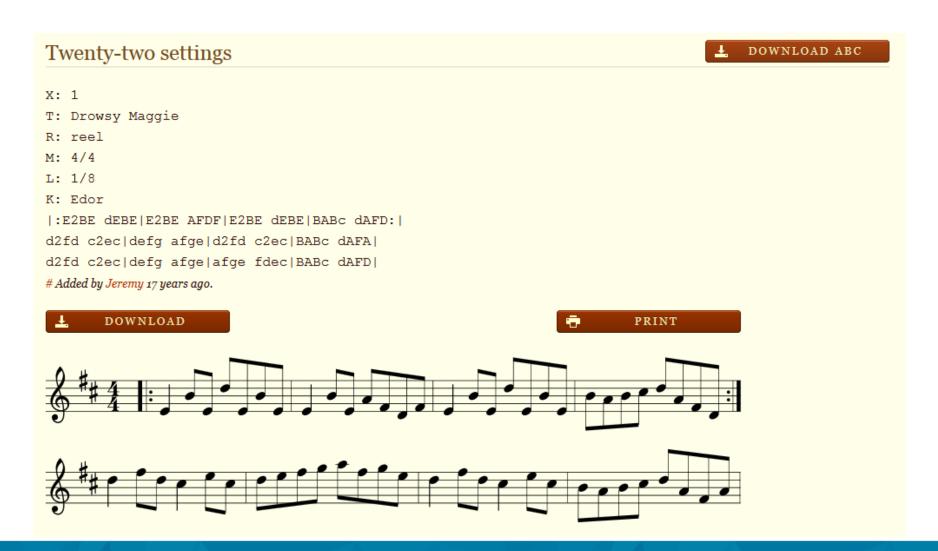


## He-текстовый поиск. TinEye



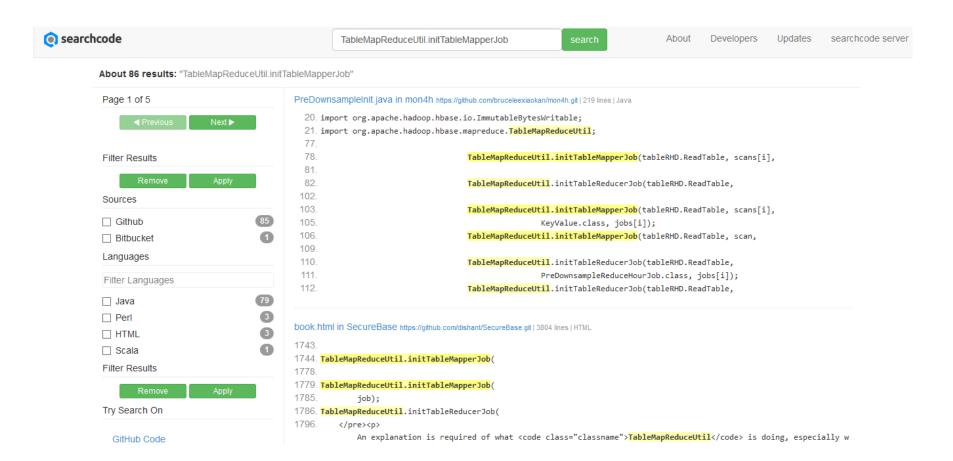


#### Не-текстовый поиск.



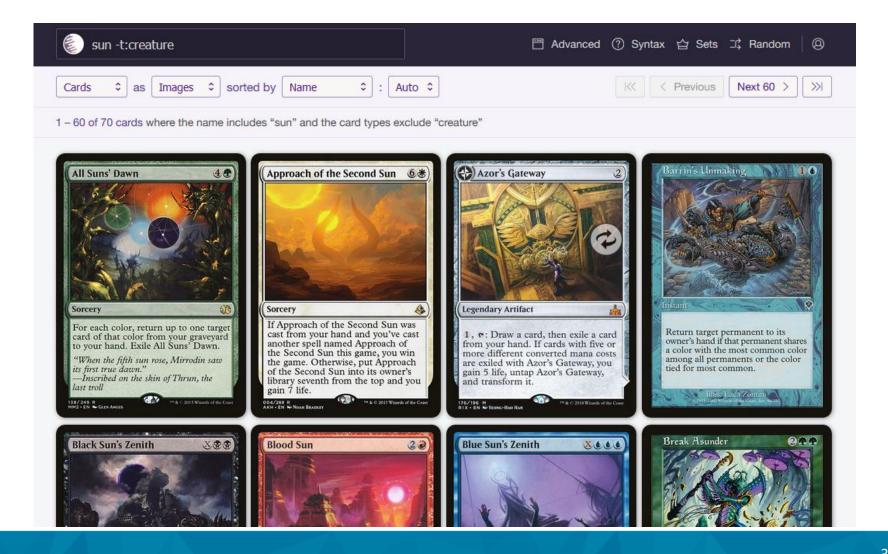


### Специализированный поиск.



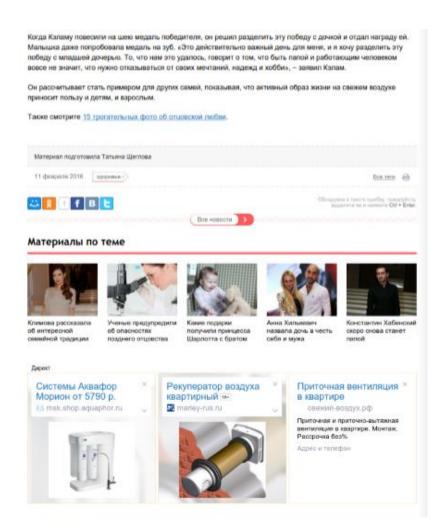


#### Специализированный поиск.





#### Контентные рекомендательные системы



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

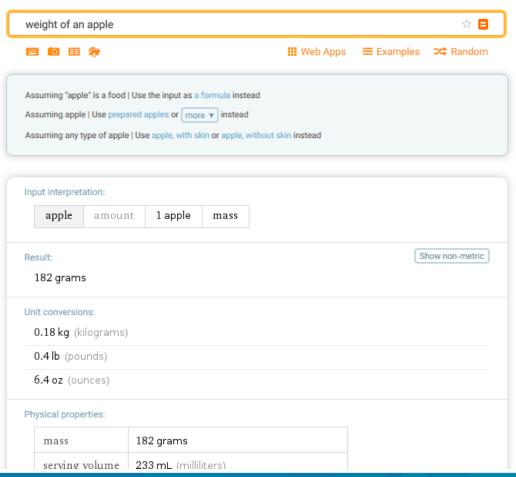
Twenty-five music analysts 'grade' 10,000 songs a month. It's a mountainous job, writes Rob Pegoraro, but the results will be serious business.





# Поиск по структурированной информации и неструктурированным запросам







#### **IBM Watson**



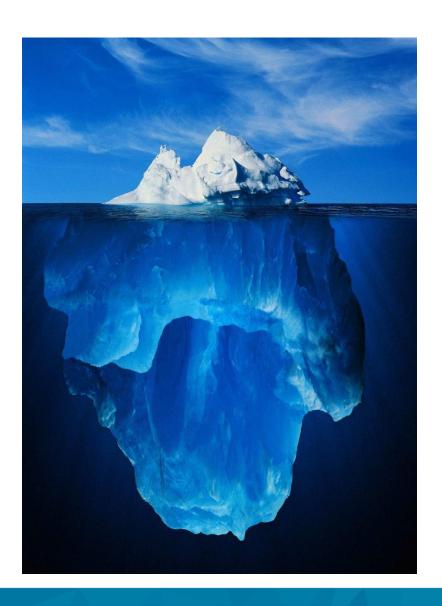


# Обзор курса





# Обзор курса

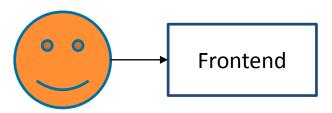








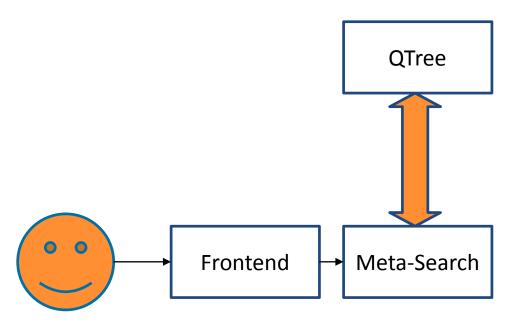




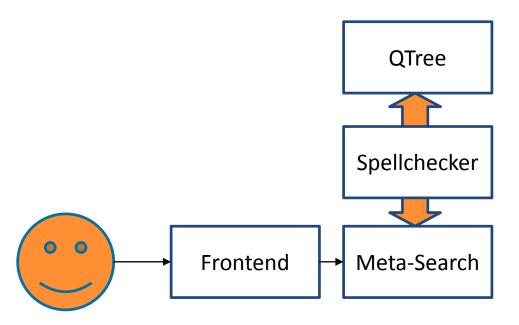




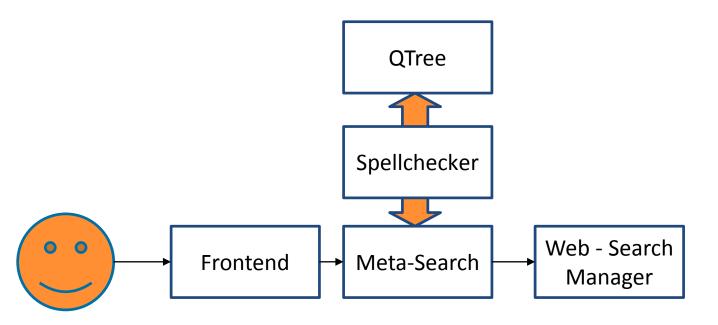




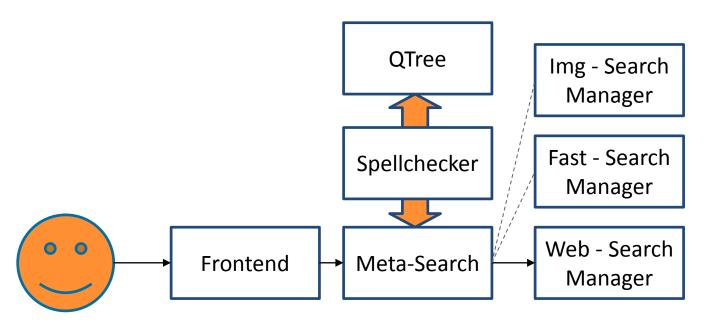






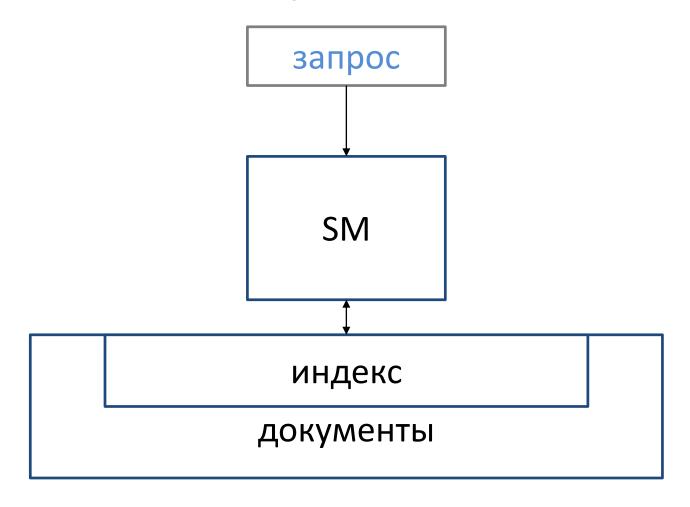






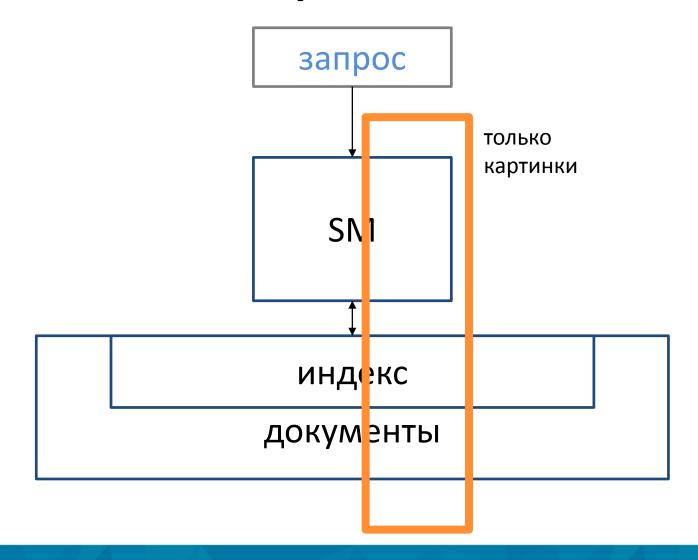


#### Немного о вертикалях

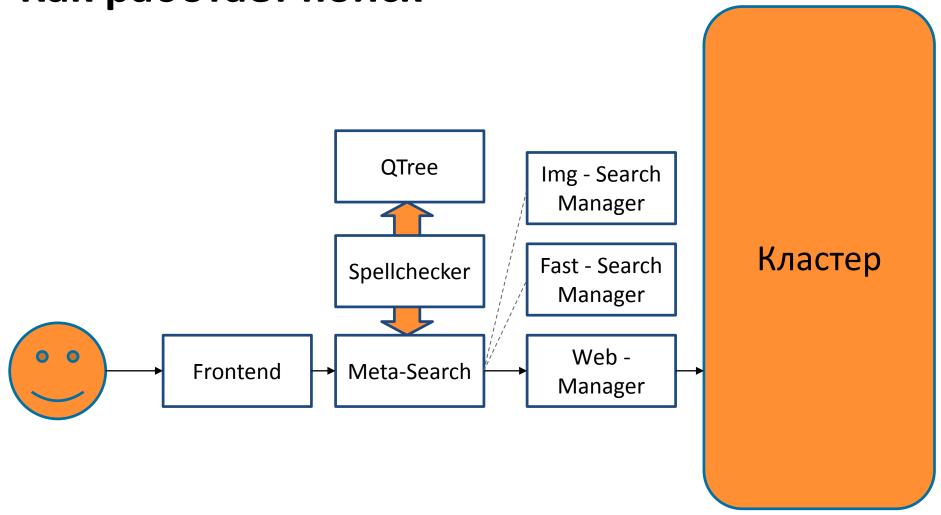




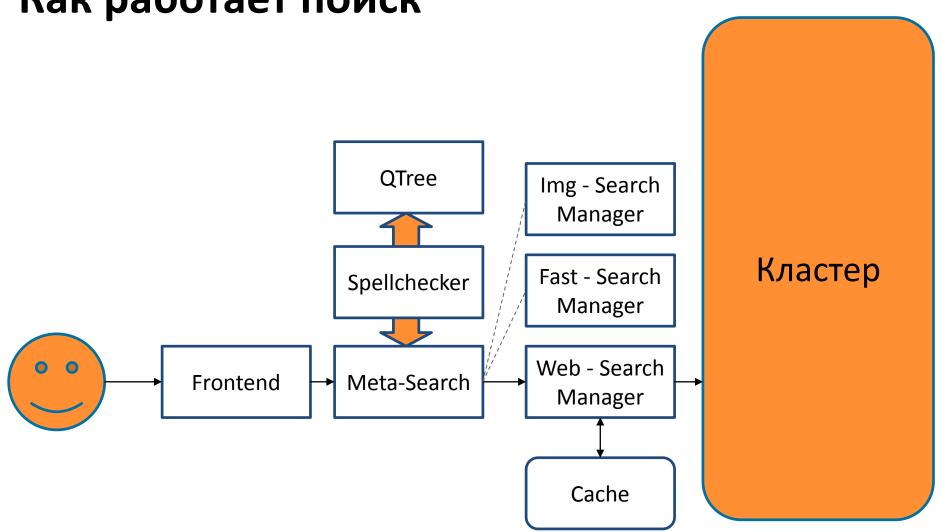
## Немного о вертикалях



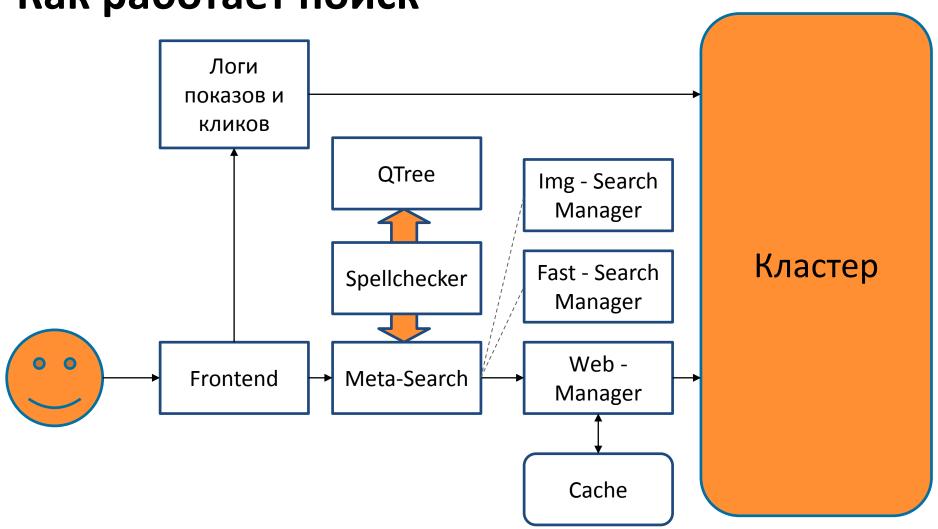




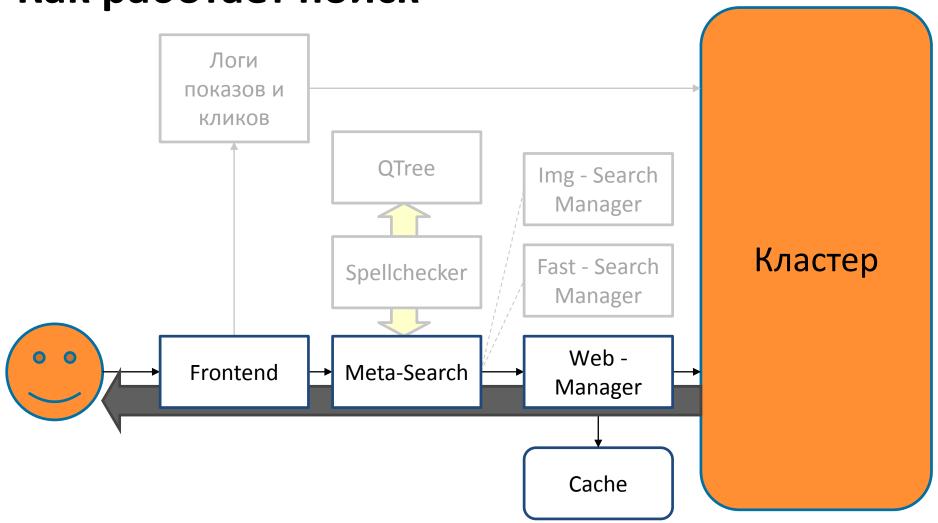














#### О курсе

# Основан на Introduction to information retrieval (Stenford)

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

Читаем более 10 лет

Основной упор – на практическое применение

Делайте домашки!



## Спасибо за внимание!