

Добавляем поисковую строку в ваше приложение с Elasticsearch

Маленький google в вашем приложении

Шестаков Алексей

Инженер-Программист



Что будет в этом докладе

- Внутренний поиск
- История и Введение в Elasticsearch
- Примеры использования

Что будет в этом докладе

- Внутренний поиск
- История и Введение в Elasticsearch
- Примеры использования
- Теория
- Как сделать поиск умнее

Чего не будет в этом докладе

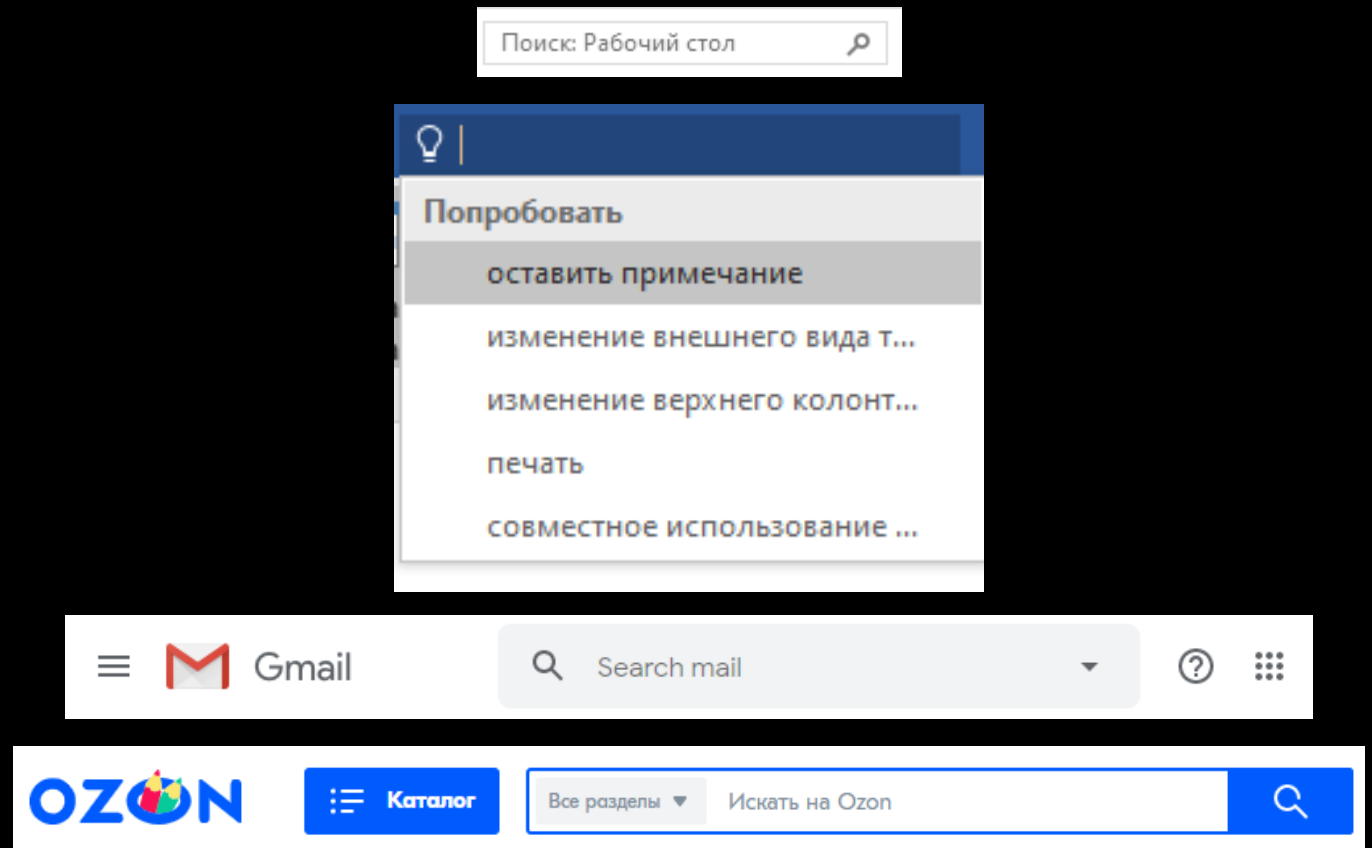
- Эксплуатация ES
- Использование ES для хранения логов
- ML (aka Машинное обучение) в поиске.

Встроенный поиск

Зачем в вашем приложении встроенный поиск?

Встроенный поиск – Примеры

- Windows
- MS Office
- GMail
- Ozon
- IntelliJ Idea
- И ещё много-много примеров....



Встроенный поиск – Профит

- Единая точка навигации
- Быстрая навигация

Встроенный поиск – Когда

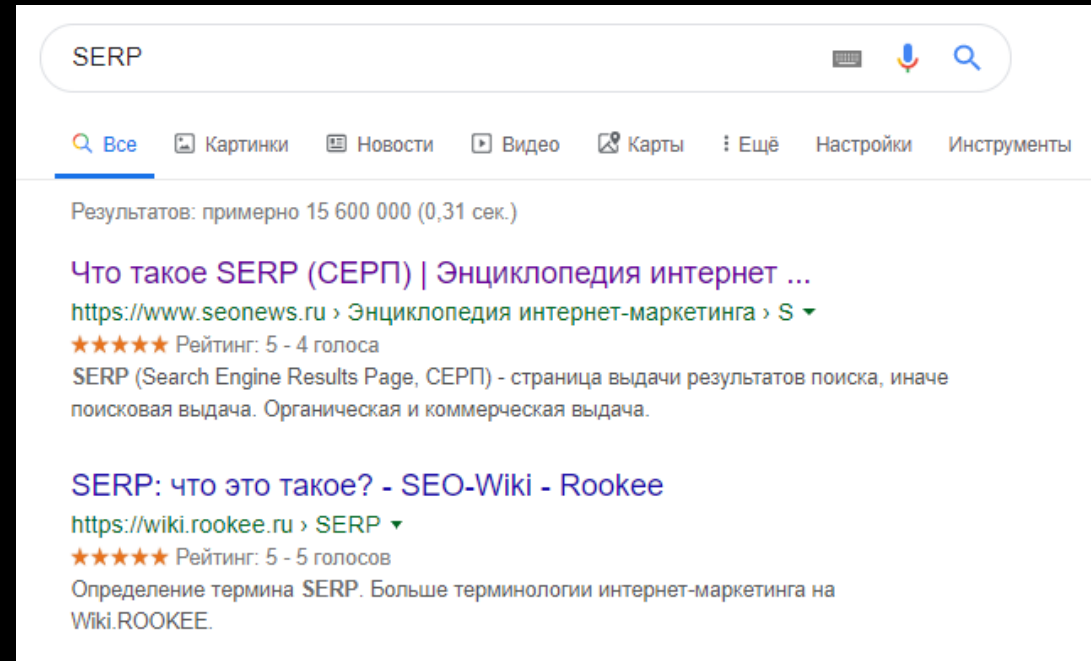
- Признаки того, что вам он нужен
 - Много контента
 - Много типов контента
 - Много разных функции (фичей)
 - Сложная навигация меню

ОСНОВНЫЕ ВИДЫ ПОИСКА

- SERP - Search Engine Result Page
- Поисковая подсказка

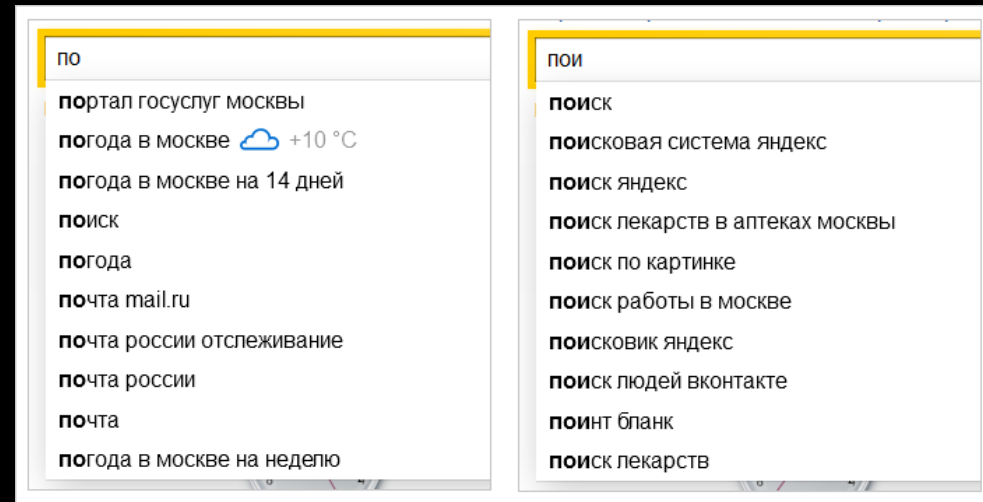
SERP – Search Engine Result Page

- По факту что у Google/Yandex
- Снippet – часть документа, важная в контексте запроса
- Это поиск по **целым словам(токены)**



Поисковые подсказки

- Search-as-you-type
- Помогает сформировать поисковый запрос
- Переход в SERP или на сразу на результат
- Это **префиксный** поиск



Elastic Search - Введение

ES – Движок полнотекстового поиска

- Движки – специализированное ПО
 - Умеют искать документы по запросу
 - Умеют ранжировать результаты поиска

Почему реляционные БД не подходят ?

- Реляционные БД не подходят для поиска слов по многим таблицам
- Не умеют ранжировать
- Не умеют обрабатывать текст
- Некоторые БД включают поиск как доп. фичу
 - MSSQL и PostgreSQL
 - Функционал все равно далек от специального решения

Почему документно-ориентированные БД не подходят ?

- Не умеют ранжировать
- Не умеют обрабатывать текст

Что такое Elasticsearch?

- Реплицированный
- Распределенный
- Горизонтально масштабируемый

Что такое Elasticsearch?

- Реплицированный
- Распределенный
- Горизонтально масштабируемый
- Самодостаточный
- Быстрый
- движок поиска

Что такое Elasticsearch?

- REST сервер
- Как одиночный сервер
- Как кластер

Почему Я рассказываю про ES?

- Мы выбрали ES так-как:
 - Opensource
 - Самое популярное решение
 - Просто разворачивается

Конкуренты Elasticsearch

- Sphinx
- Solr
- Имеют свои плюсы и заслуживают отдельных докладов

Elasticsearch – История

История Elasticsearch



- Opensource
- Автор Shay Banon
- Начал проект Compass в 2004 г.,
- В 2010 г. как результат большого переписывания появился Elasticsearch
- В качестве основы используется **Lucene**



История Lucene



- Opensource
- Автор Douglass Cutting
- Начал проект в 1997 г.
- В 1999 г. выложил на SourceForge.net
- В 2001 г. вошел в состав проектов фонда Apache



Почему Elasticsearch > Lucene

- Lucene – Java библиотека
- Elasticsearch – REST сервер
- Elasticsearch добавляет поверх Lucene
 - Масштабируемость
 - Реплицируемость
 - Аналитические инструменты

Elasticsearch -

Начнем с простого примера

Индексируем

POST localhost:9200/information/person/1

```
{  
  "name" : "Paul",  
  "lastname" : "Smith",  
  "job_description" : "Business Analyst"  
}
```

ПОИСК

```
curl -X GET "localhost:9200/information/_search?pretty" -H 'Content-Type: application/json' -d'
```

```
{  
  "query" : {  
    "term" : { "user" : "kimchy" }  
  }  
}
```

Еще варианты:

GET localhost:9200/_search?q=kimchy

GET localhost:9200/_search?q=user:kimchy

Ищем по префиксу

```
curl -X GET "localhost:9200/_search?pretty" -H 'Content-Type: application/json' -d'
{
  "query": {
    "prefix": {
      "user": {
        "value": "ki"
      }
    }
  }
}
```

Теория чтобы искать умнее

Elasticsearch – не черная коробка

- Elasticsearch использует классическую теорию поиска
- Для того что бы качественно настроить поиск нужно понимать процесс обработки текста

Модель – Bag of Words

- Документом называется индексируемое текстовое поле
- Каждый документ режется на «слова» или «токены»
- Порядок не учитывается



Что такое «Токен»?

- «Токен» – минимальная единица поиска
- Обычно слово, но не всегда
- Поиск меньших частей «токена» возможен, но более затратен

Поисковый индекс

- ES Строит индекс почти как в книгах
- Запись в индексе:
- Связка токена – и номера документов, где слово встречается
- Еще называется «Инвертированным»

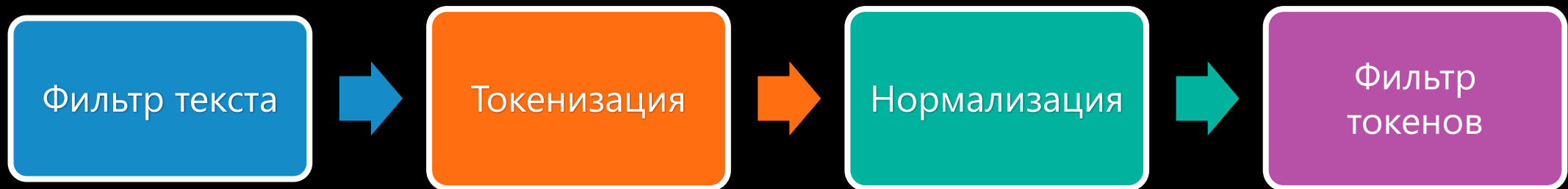
Index	Д	З
	Данные инструмента	Загрузка файлов помощи ... 142
	ввод в таблицу ... 150	Замена буферной батареи ... 475
	Индексация ... 156	Замена текстов ... 93
	Данные инструментов	Запись значений оцупывания в
	ввод в программу ... 149	таблицу нулевых точек ... 377
	вызов ... 161	Запись значений оцупывания в
	Дельта-значения ... 149	таблицу предустановок ... 378
	Движение по траектории	Захват текущей позиции ... 87, 183
	Декартовы координаты	Защита данных ... 124
	Круговая траектория с	И
	плавным переходом ... 190	Изменение скорости вращения
	Круговая траектория с	шпинделя ... 365
	указанием радиуса ... 188	Измерение заготовок ... 390
	Круговая траектория с	Измерение инструмента ... 153
	центром окружности	Изображение в 3 плоскостях ... 411
	СС ... 187	Имя программы: см. Управление
	Обзор ... 181	файлами, имя файла
	Прямая ... 182	Индексированные
	декартовы координаты	инструменты ... 156
	Полярные координаты	Индикация состояния ... 63
	Круговая траектория вокруг	дополнительная ... 65
	полюса СС ... 197	общая ... 63
	Круговая траектория с	Интерфейс передачи данных
	плавным переходом ... 197	Настройка ... 440
	Обзор ... 195	Разводка контактов ... 466
	Прямая ... 196	Индикаторы ... 474

Инвертированный индекс

Слово	Номера документов где слово встречается
все	2,4,7
всегда	2,7
делать	2
дело	3,3

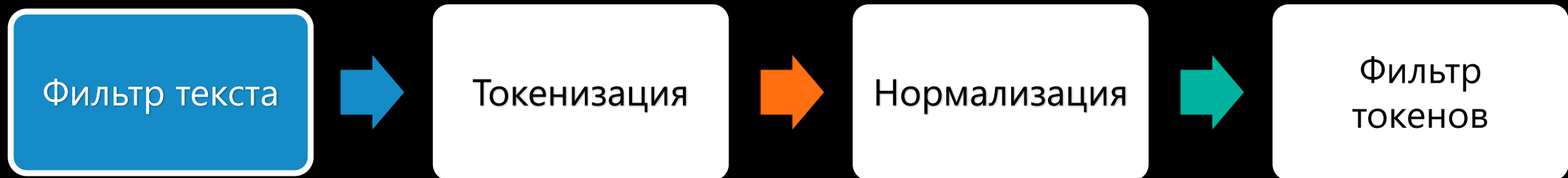
Процесс индексации

- Процесс построения индекса
- Управляя процессом индексации можно
 - Убрать лишнее из выдачи
 - Научить поиск находить больше



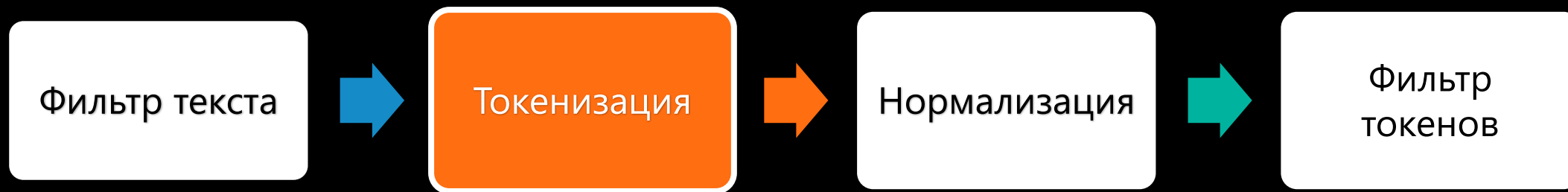
Фильтр текста

- Убираем из текста все что не должно искаться
- В терминах ES – «char_filter»
- Пример – «html_strip» - убирает HTML разметку



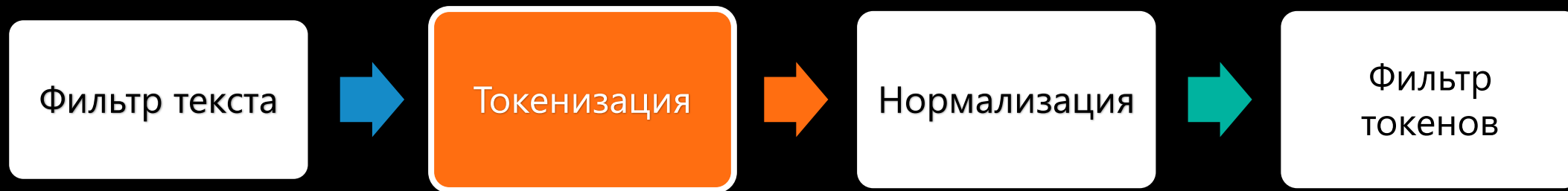
Токенизация

- Управляя токенизацией важно понять что мы хотим считать одним «словом»



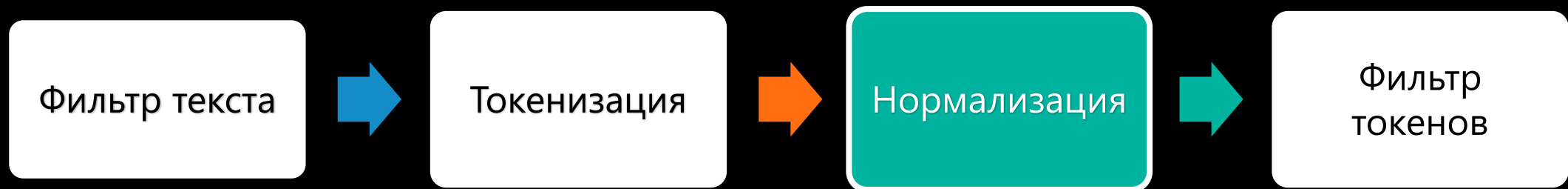
Токенизация

- Мобильные номера
 - + 7 (933) 34344322 → +7(933)34344322, (933)34344322, 34344322
- От выбора токенов зависит будет ли находиться тот или иной вариант



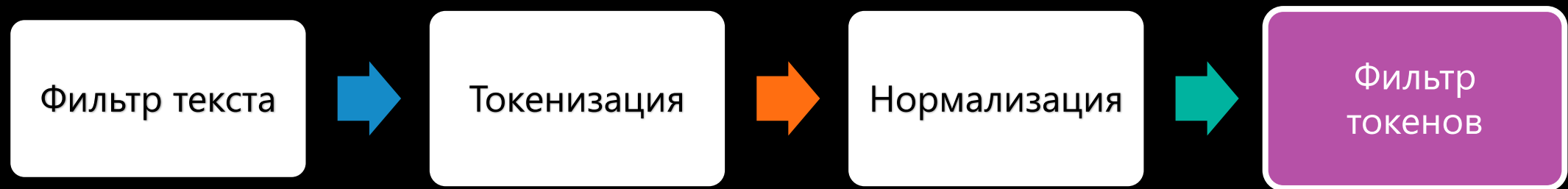
Нормализация

- Приводим схожие токены к одному
 - Хотим ли мы считать слова в разном падеже одним и тем же словом?
 - Мы теряем немного информации, но и будем находить больше



Фильтр токенов

- Не все токены нам интересны
 - Пример: предлоги



Делаем поиск умнее

Настраиваем маппинги

- Маппинг – аналог схемы в ES
- Что нам важно
- Тип поля
- Анализатор

Настраиваем маппинги

- Сущности в ES – иммутабельные
 - Настраивать можно только не индексируемые поля
 - Лучше настраивать при создании индексов

Настраиваем маппинги

```
curl -X PUT "localhost:9200/my-index?pretty" -H 'Content-Type: application/json' -d'
```

```
{  
  "mappings": {  
    "properties": {  
      "age": { "type": "integer" },  
      "email": { "type": "text" , "analyzer": "std_english" },  
      "name": { "type": "keyword" }  
    }  
  }  
}
```

Настраиваем маппинги

- ES Автоматически настраивает поля и для текстовых полей выбирается анализатор "standard"
- Посмотреть настройки маппингов можно так:

```
curl -X GET "localhost:9200/my-index/_mapping?pretty"
```

Настраиваем анализаторы

- Pipeline обработки конкретного поля
- Создаем свой если стандартных мало

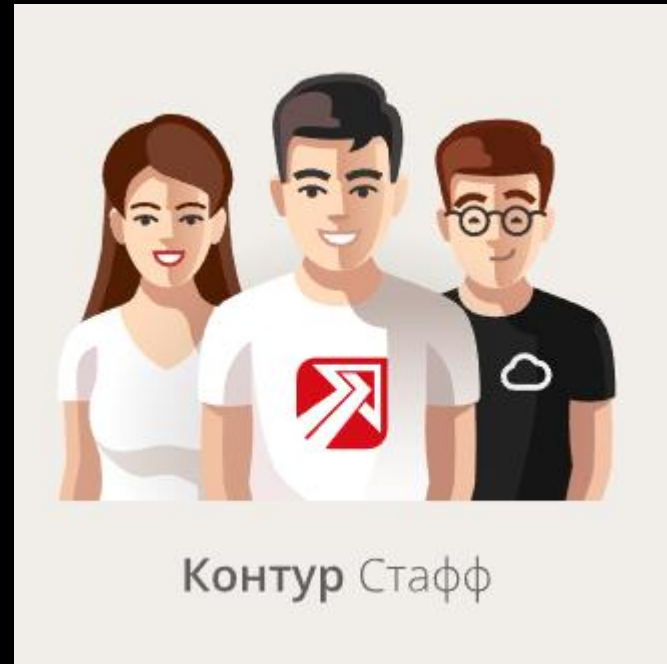
Настраиваем анализаторы

```
PUT my_index {  
  "settings": { "analysis": {  
    "analyzer": {  
      "my_custom_analyzer": { "type": "custom", "tokenizer": "standard",  
        "char_filter": ["html_strip" ],  
        "filter": [  
          "lowercase",  
] } } } } }
```

Пример использования – Контур.Стафф

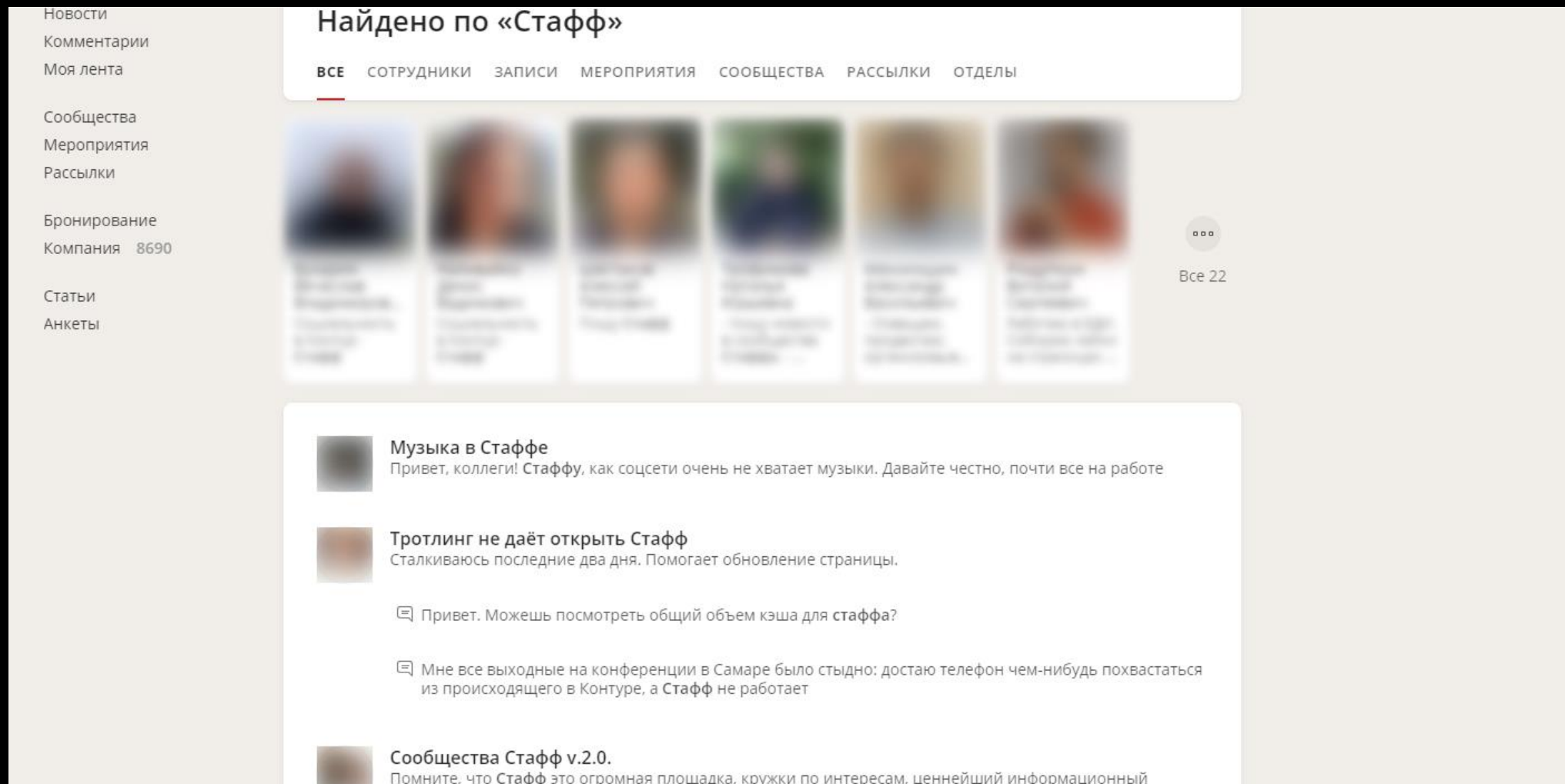
Контур.Стафф

- Справочник сотрудников
- Социальная сеть
- Использует ES для Поиска
- Много разного контента
- ~4к обращений к SERP за день



Контур.Стафф

- Использует продвинутые фишки ES
 - Скрипты ранжирования для управления результатом
 - Вложенные объекты
 - Релевантностью полей в запросе
- Настроены фильтры управления видимости
- Агрегирование



Поиск в Стафф

Это и есть SERP Стафф

Шестаков Алексей

Найдено по «Шестаков Алексей»

ВСЕ СОТРУДНИКИ ЗАПИСИ



Шестаков Алексей Петрович

Программист

Отдел развития информационных ресурсов (ОРИР)



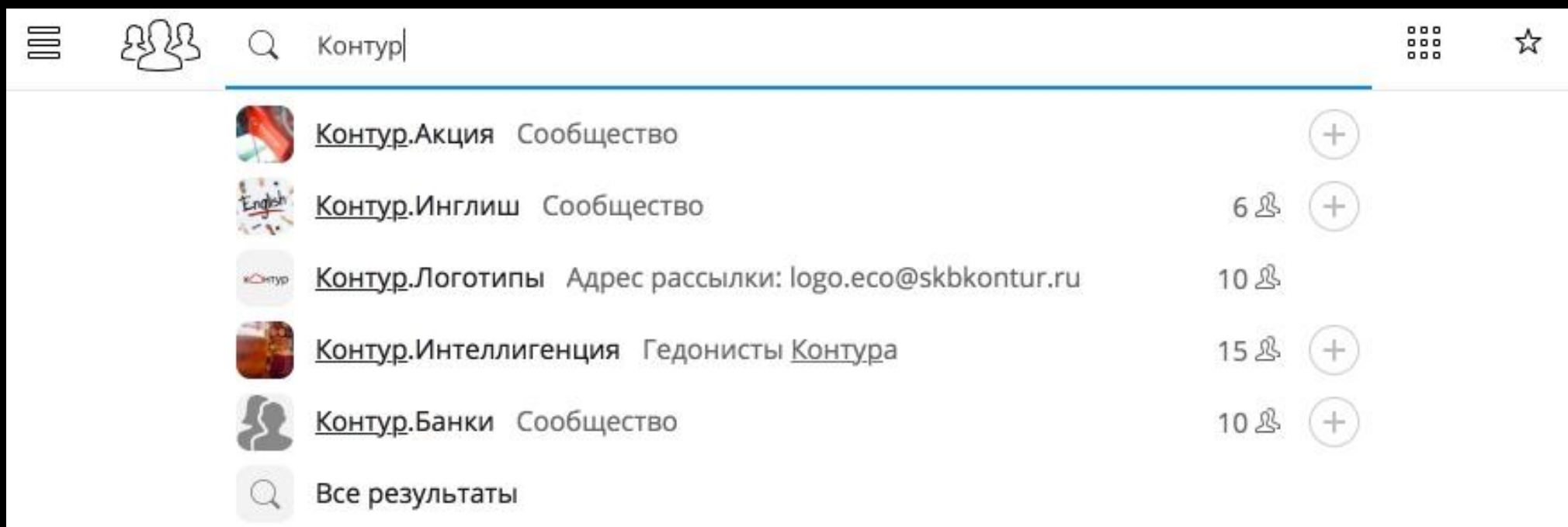
java.ural.Meetup @3

на Хабре: С докладами выступят: — Алексей Шестаков, — Владимир Лиля — Григорий Кошелев

Java Community • Григорий Кошелев • 9 сентября

Поиск в Стафф

Это и есть SERP Стафф



Поисковая подсказка в Стафф

Участники и руководители

УЧАСТНИКИ 78

РУКОВОДИТЕЛИ 1

Шестаков

Шестаков Алексей
Программист

Пригласить

Текст приглашения можно задать в [настройках сообщества](#)

Это тоже подсказки

Популярные темы

#марафонзнаний

#контур_интроверт

#впередкзнаниям

#вызов2019

#трекдня

#деньhr

#неоспоримые2года

#неоспоримые

#алло

#еслинетактотак

Популярные теги за последнее время

Что дальше

Литература

- Хорошее введение в теорию информационного поиска
- Идеально если вы хотите глубже понимать ES либо написать свой аналог
- Новой редакции нет с 2008 г.
- Устарела в ML части.

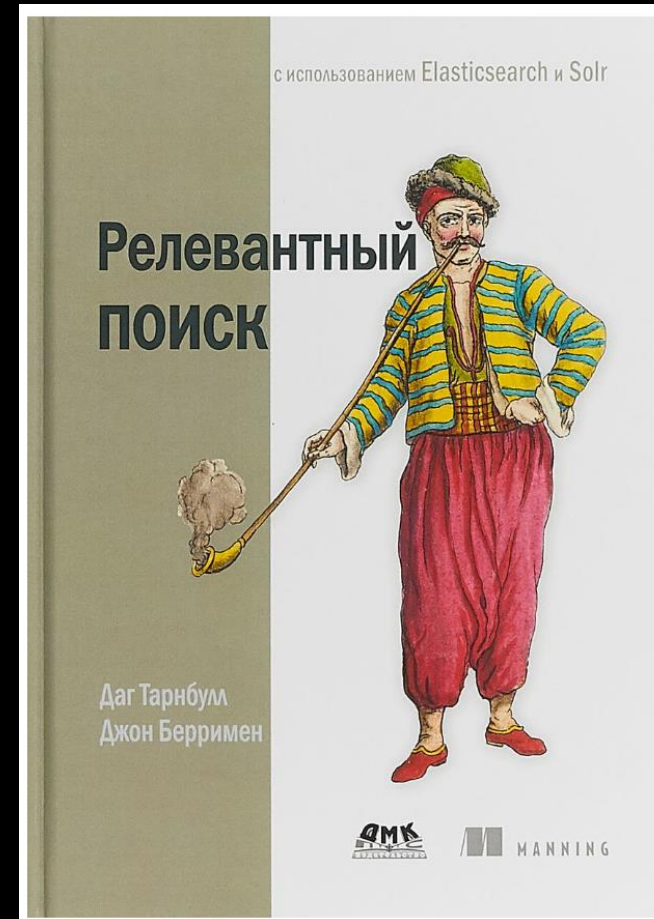




<https://www.ozon.ru/context/detail/id/5497130/>

Литература

- Конкретные рецепты
- Практика использования ES
- Ориентация на ручное управление ранжированием без ML





<https://www.ozon.ru/context/detail/id/144631193/>

Вопросы?



СКБ Контур

Алексей Шестаков

Инженер-Программист

shestakovap@skbkontur.ru

kontur.ru