Лекция 1
Индексирование

Дроздова Ксения
drozdova.xenia@gmail.com

Telegram

HSE Infosearch 19

1 member

https://t.me/infosearch19
наш общий чат вместо почты

@ksdrozdova можете писать мне лично

В чем фишка курса:

1. После курса вы будете четко представлять базовые концепты любого поисковика
2. Напишете свой поисковик
3. Создадите продакш проект

Постановка задачи

Дано: набор документов

Задача: отсортировать документы по релевантности к запросу

Технологии: индексирование, метрика близости запроса и документа

Список задач IR постоянно расширяется и теперь включает:

a) Классификацию документов
b) Фильтрацию документов
c) Кластеризацию документов
d) Проектирование архитектур поисковых систем и пользовательских интерфейсов
e) Извлечение информации, в частности аннотирования и реферирования документов
и др.

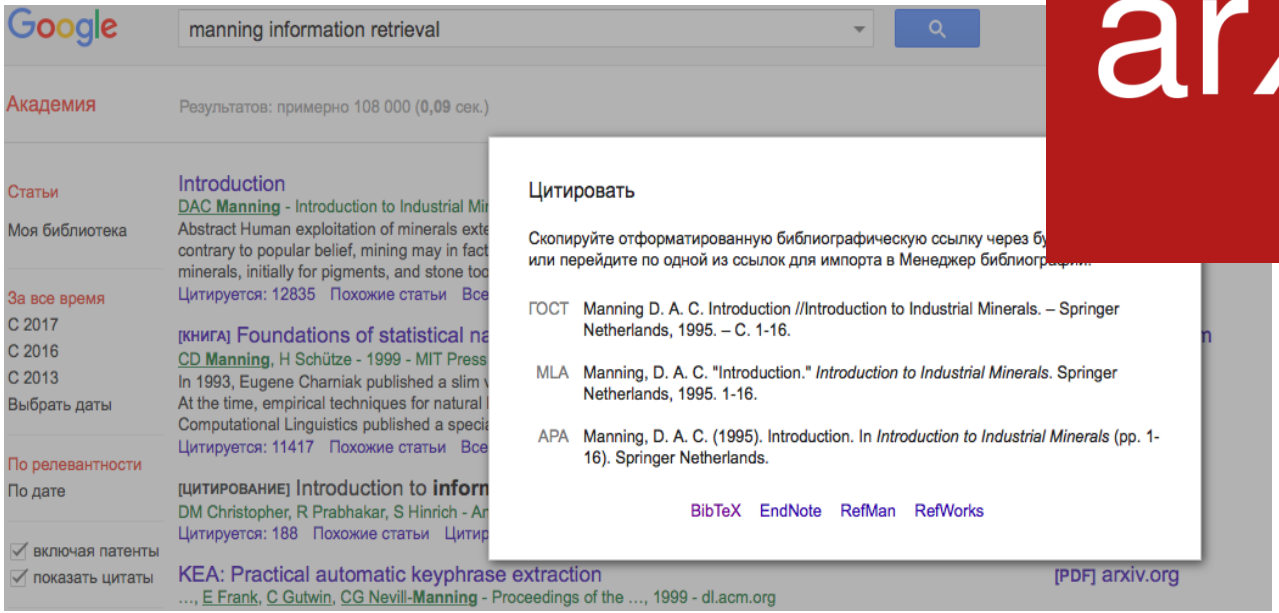
Также перед IR ставятся задачи по обработке естественного языка (морфологический анализ, разрешение лексической многозначности и тд)

Примеры поисковых систем

WEB поисковики

Примеры поисковых систем

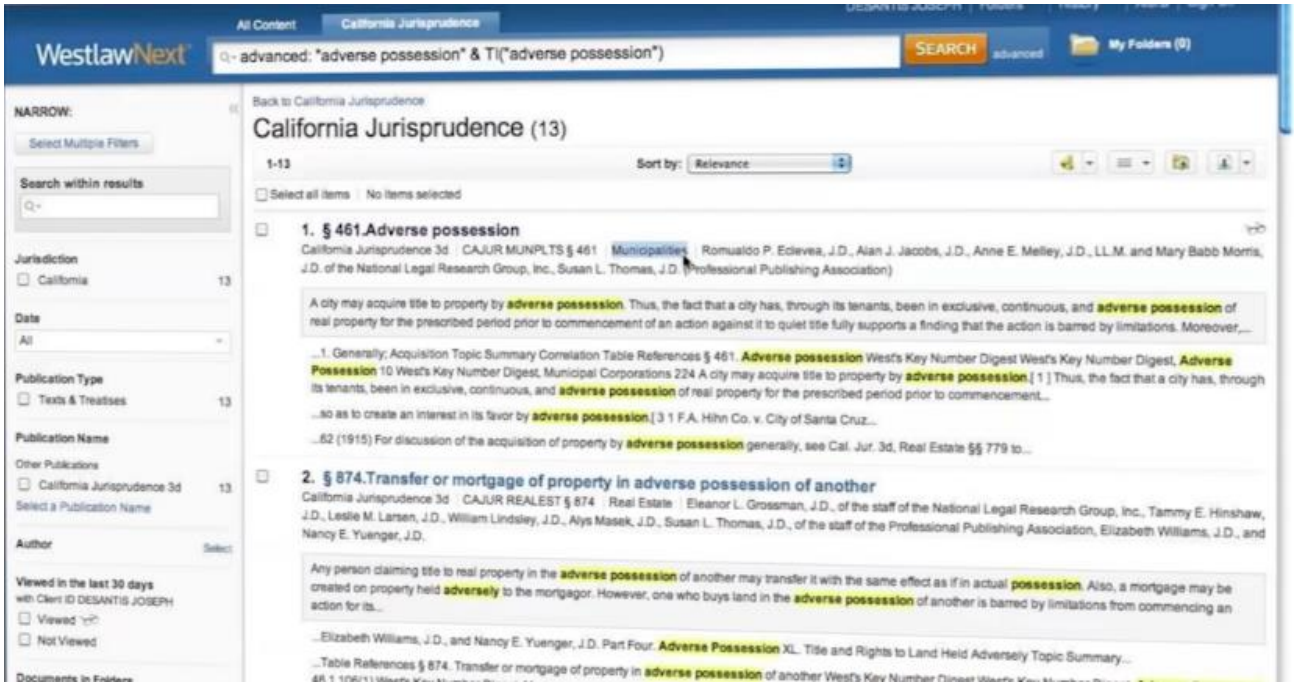
Google Scholar



arXiv.org

Примеры поисковых систем

Westlaw, Гарант



Twiggle

yellow midi dress with long sleeves

Powered by Twiggle

Pieces High Neck Midi Dress In Floral Print

$76

TTYA BLACK Plus Midi Wrap Dress With Knot

$198

ASOS PETITE Cold Shoulder Ruffle Tea Dress

$16

Monki Puff Sleeve Midi Sweat Dress

$14

ASOS Mini Bodycon Dress in Rib with Long Sleeves

$18

ASOS TALL Midi Dress with Pleated Skirt and Dip

$38

ASOS Long Sleeve Embroidered Sundress

$48

ASOS Pinny Lace Stripe Prom Midi Dress

$29.5

ASOS TALL WEDDING Rouched Midi Dress in

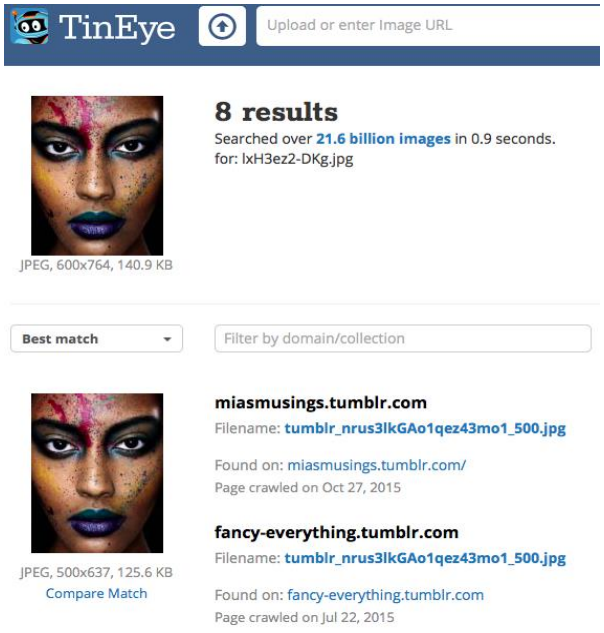
$51.5

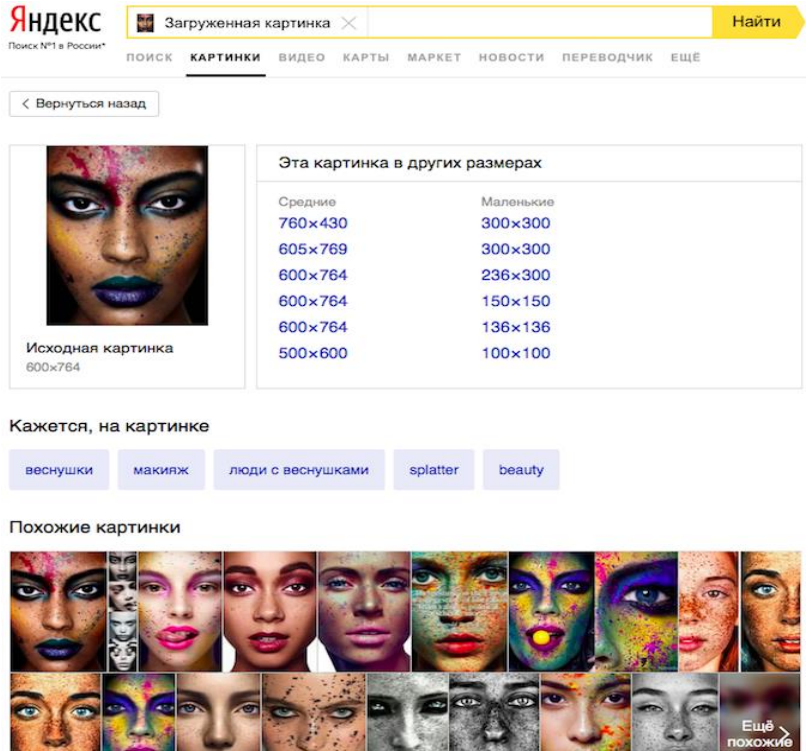
New Look One Shoulder Frill Sleeve Midi Dress

$36

Примеры поисковых систем

Мультимедиа-поиск





Булев поиск

Это самая простая структура данных в поиске.
Она основана на поиске с использованием всем знакомых логических операторов:

AND = Пересечение

OR = Объединение

NOT = Исключение

В этой модели документ или релевантен или нерелевантен запросу

Промежуточных состояний нет, логика TRUE / FALSE

Булев поиск

Плюсы:

- Простота
- Прозрачность результатов

Минусы:

- Неустойчив к опечаткам
- Не учитывает близость слов в запросе и документе
- Не учитывает ничего, кроме факта вхождения слова
- Не умеет ранжировать результаты поиска

Булев поиск

Булев поиск – классная вещь. Простая, но эффективная модель.

Когда мы ищем письма в своём электронной почте, то используем именно её.

Обратный индекс

Следующая по сложности структура хранения информации о документах коллекции – это обратный индекс

Прямой индекс: каждому документу соответствует список входящих в него слов

Обратный индекс: каждому слову соответствует списков документов, в которых оно встречается

Классический пример прямого и обратного индекса – содержание книги и предметный указатель в конце книги

Обратный индекс

1: Winter is coming.
2: Ours is the fury.
3: The choice is yours.

term	freq	documents
choice	1	3
coming	1	1
fury	1	2
is	3	1, 2, 3
ours	1	2
the	2	2, 3
winter	1	1
yours	1	3

Dictionary

Postings

Обратный индекс

В обратном индексе помимо id документов можно сохранить полезные параметры, которые могут быть использованы при ранжировании:

1. Частота встречаемости
2. TF-IDF
3. Местонахождение в документе (в заголовке или нет, в начале или в конце)
4. Автор документа, дата написания

Этапы создания поисковика, основанном на обратном индексе

1. Собрать базу для поиска
2. Сделать препроцессинг
    — лемматизировать или нет
    — удалять числа, пунктуацию, стоп-слова или нет
    — POS-тэггинг
    ...
3. Построить индекс
4. Определить метрику

Установка:
http://jupyter.org/install

Terminal —> jupyter notebook —> localhost