

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:

- а) Классификацию документов
 - б) Фильтрацию документов
 - с) Кластеризацию документов
 - д) Проектирование архитектур поисковых систем и пользовательских интерфейсов
 - е) 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

The screenshot shows a Google Scholar search interface. The search bar contains the text "manning information retrieval". Below the search bar, the results are displayed in a list format. The first result is titled "Introduction" by D.A.C. Manning, with a citation count of 12835. The second result is titled "Foundations of statistical natural language processing" by J. R. Brown and P. V. Brown, with a citation count of 11417. The third result is titled "Introduction to information retrieval" by D.M. Christopher, R. Prabhakar, S. Hinrich, and A. Kuhlmann, with a citation count of 188. The search results are filtered by "По релевантности" (By relevance). A sidebar on the left contains navigation links such as "Академия", "Статьи", "Моя библиотека", "За все время", "С 2017", "С 2016", "С 2013", "Выбрать даты", "По релевантности", and "По дате". At the bottom of the sidebar, there are checkboxes for "включая патенты" and "показать цитаты".

Google

manning information retrieval

Академия

Результатов: примерно 108 000 (0,09 сек.)

Статьи

Моя библиотека

За все время

С 2017

С 2016

С 2013

Выбрать даты

По релевантности

По дате

включая патенты

показать цитаты

Introduction

DAC Manning - Introduction to Industrial Minerals

Abstract Human exploitation of minerals extends far beyond the commonly held belief that mining is contrary to popular belief, mining may in fact be a natural part of human life, initially for pigments, and stone tools.

Цитируется: 12835 Похожие статьи Все

Foundations of statistical natural language processing

CD Manning, H Schütze - 1999 - MIT Press

In 1993, Eugene Charniak published a slim volume titled "Natural Language Processing: Theoretical Foundations". At the time, empirical techniques for natural language processing were limited. Computational Linguistics published a special issue on "Foundations of Statistical Natural Language Processing".

Цитируется: 11417 Похожие статьи Все

Introduction to information retrieval

DM Christopher, R Prabhakar, S Hinrich - Ar

Цитируется: 188 Похожие статьи Цитир

KEA: Practical automatic keyphrase extraction

..., E Frank, C Gutwin, CG Nevill-Manning - Proceedings of the ..., 1999 - dl.acm.org

ГОСТ Manning D. A. C. Introduction //Introduction to Industrial Minerals. – Springer Netherlands, 1995. – С. 1-16.

MLA Manning, D. A. C. "Introduction." Introduction to Industrial Minerals. Springer Netherlands, 1995. 1-16.

APA Manning, D. A. C. (1995). Introduction. In Introduction to Industrial Minerals (pp. 1-16). Springer Netherlands.

BibTeX EndNote RefMan RefWorks

[PDF] arxiv.org

arXiv.org

Примеры поисковых систем

Westlaw, Гарант

The screenshot displays the WestlawNext interface. At the top, the search bar contains the query "q- advanced: *adverse possession* & TI('adverse possession')". The results are filtered to "California Jurisprudence" and show 13 items. The left sidebar includes filters for Jurisdiction (California), Date (All), Publication Type (Texts & Treatises), and Publication Name (California Jurisprudence 3d). The main content area lists two results:

- 1. § 461. Adverse possession**
California Jurisprudence 3d CAJUR MUNPLTS § 461 Municipalities Romualdo P. Echevea, J.D., Alan J. Jacobs, J.D., Anne E. Melley, J.D., LL.M. and Mary Babb Morris, J.D. of the National Legal Research Group, Inc., Susan L. Thomas, J.D., Professional Publishing Association)
A city may acquire title to property by **adverse possession**. Thus, the fact that a city has, through its tenants, been in exclusive, continuous, and **adverse possession** of real property for the prescribed period prior to commencement of an action against it to quiet title fully supports a finding that the action is barred by limitations. Moreover, ...
...1. Generally: Acquisition Topic Summary Correlation Table References § 461. **Adverse possession** West's Key Number Digest West's Key Number Digest, **Adverse Possession** 10 West's Key Number Digest Municipal Corporations 224 A city may acquire title to property by **adverse possession** [1] Thus, the fact that a city has, through its tenants, been in exclusive, continuous, and **adverse possession** of real property for the prescribed period prior to commencement ...
...so as to create an interest in its favor by **adverse possession** [3 1 F.A. Hihn Co. v. City of Santa Cruz ...
...62 (1915) For discussion of the acquisition of property by **adverse possession** generally, see Cal. Jur. 3d, Real Estate §§ 779 to ...
- 2. § 874. Transfer or mortgage of property in adverse possession of another**
California Jurisprudence 3d CAJUR REALEST § 874 Real Estate Eleanor L. Grossman, J.D., of the staff of the National Legal Research Group, Inc., Tammy E. Hinshaw, J.D., Leslie M. Larsen, J.D., William Lindsay, J.D., Alys Masek, J.D., Susan L. Thomas, J.D., of the staff of the Professional Publishing Association, Elizabeth Williams, J.D., and Nancy E. Yuenger, J.D.
Any person claiming title to real property in the **adverse possession** of another may transfer it with the same effect as if in actual **possession**. Also, a mortgage may be created on property held **adversely** to the mortgagor. However, one who buys land in the **adverse possession** of another is barred by limitations from commencing an action for its ...
...Elizabeth Williams, J.D., and Nancy E. Yuenger, J.D. Part Four **Adverse Possession** XL. Title and Rights to Land Held Adversely Topic Summary ...
...Table References § 874. Transfer or mortgage of property in **adverse possession** of another West's Key Number Digest Municipal Corporations 224 A city may acquire title to property by **adverse possession** [1] Thus, the fact that a city has, through its tenants, been in exclusive, continuous, and **adverse possession** of real property for the prescribed period prior to commencement ...

yellow midi dress with long sleeves



Powered by  Twiggle



Pieces High Neck Midi Dress In Floral Print

\$76



PLUS-SIZE

TTYA BLACK Plus Midi Wrap Dress With Knot

\$198



ASOS PETITE

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 EXCLUSIVE

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

ASOS TALL WEDDING Rouché Midi Dress in

\$51.5

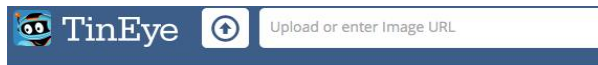


New Look One Shoulder Frill Sleeve Midi Dress

\$36

Примеры поисковых систем

Мультимедиа-поиск



JPEG, 600x764, 140.9 KB

8 results

Searched over **21.6 billion images** in 0.9 seconds.
for: lxH3ez2-DKg.jpg

Best match

Filter by domain/collection



JPEG, 500x637, 125.6 KB
[Compare Match](#)

miasmusings.tumblr.com

Filename: [tumblr_nrus3lkGAo1qez43mo1_500.jpg](#)

Found on: [miasmusings.tumblr.com/](#)
Page crawled on Oct 27, 2015

fancy-everything.tumblr.com

Filename: [tumblr_nrus3lkGAo1qez43mo1_500.jpg](#)

Found on: [fancy-everything.tumblr.com](#)
Page crawled on Jul 22, 2015

Яндекс
Поиск №1 в России*

Загруженная картинка

Найти

ПОИСК **КАРТИНКИ** ВИДЕО КАРТЫ РЫНОК НОВОСТИ ПЕРЕВОДЧИК ЕЩЕ

< Вернуться назад



Исходная картинка
600x764

Эта картинка в других размерах

Средние	Маленькие
760×430	300×300
605×769	300×300
600×764	236×300
600×764	150×150
600×764	136×136
500×600	100×100

Кажется, на картинке

[веснушки](#)

[макияж](#)

[люди с веснушками](#)

[splatter](#)

[beauty](#)

Похожие картинки



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.



<u>term</u>	<u>freq</u>	<u>documents</u>
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