Home

Информационный поиск

Лекция 1. Индексирование

Telegram

GitHub

Ника Зыкова
2023/09/06

Важные вещи

Формула оценки:

О = 0.25 х дз_инфопоиск + 0.25 х проект_инфопоиск + 0.25 х дз_бд + 0.25 х проект_бд

По базам данных вам расскажет Янина на базах данных.

По инфопоиску:

Две домашние работы (предварительно). Первая про реализацию индекса руками, вторая - про добавление к этому эмбеддингов и CLI.

По проекту. Если вас устраивает 6, то можно просто защитить результат двух домашек. Если нет, то придется делать проект, требования будут чуть позже.

О чем курс?

1. Что такое информационный поиск, индексирование
2. Способ индексирования: Tf-Idf и BM25
3. Повторение программирования: numpy, gensim, flask, sklearn
4. Способ индексирования: эмбеддинги
5. Docker (возможно)
6. To be updated...
7. To be updated...
8. Защита проектов 1
9. Защита проектов 2

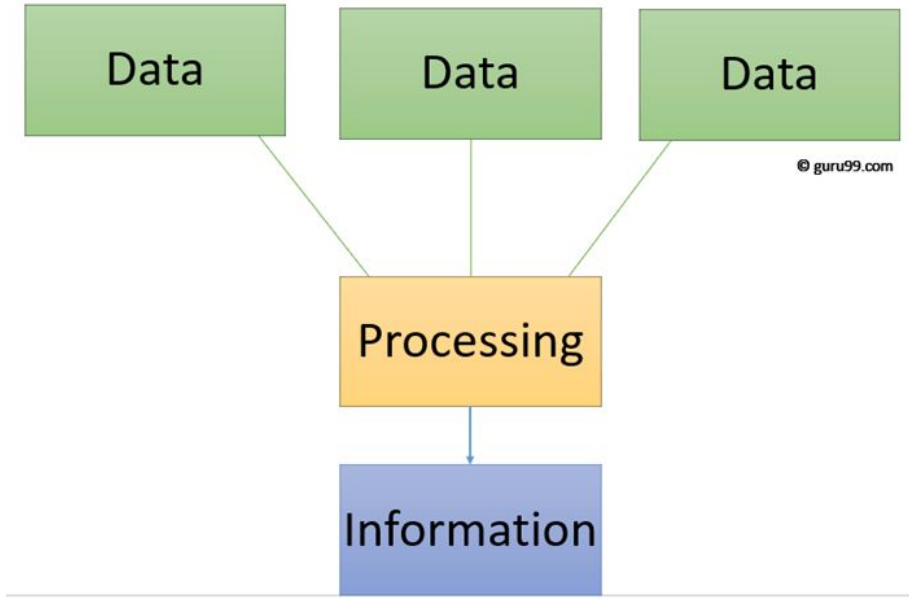
Что такое информация?

Есть три сущности: данные, информация и знание.

Данные - сырой набор объектов (картинок, текстов, аудио), содержащих информацию.

Информация - структурированный набор фактов (таблица, схема, осмысленная часть картинки или аудио).

Знание - абстрактное представление этой информации в голове человека.



Что такое информационный поиск?

В очень широком смысле - поиск объектов в массиве по условию.
Теоретически, многие задачи NLP можно сформулировать в подобном виде.

В чуть менее широком - выделение информации из неструктурированных данных.

В контексте NLP - извлечение информации из текста.

Типы поиска

По типу данных

- Текстовый
- По картинкам
- По аудио
- По метаинформации

По условию поиска

- Булев поиск
- Поиск по сходству
- Поиск по релевантности

Home | Recent Tabs | Current Page

You can type something here...

Поиск по тексту

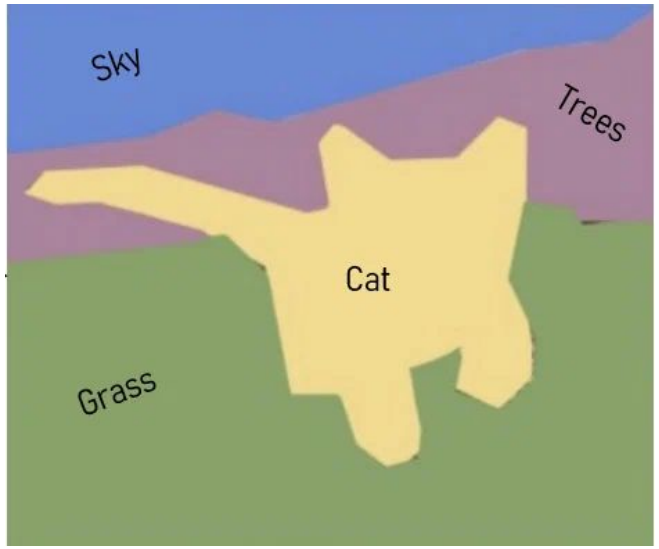
Чаще всего полнотекстовый поиск - поиск по содержанию документа.

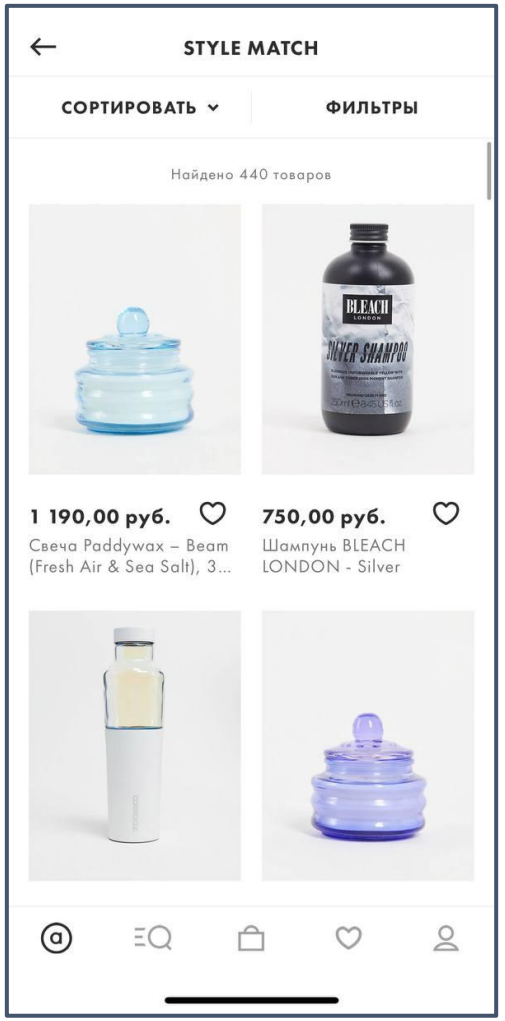
Home | Recent Tabs | Current Page

You can type something here...

Поиск по картинке

В том числе задачи классификации и сегментации изображений



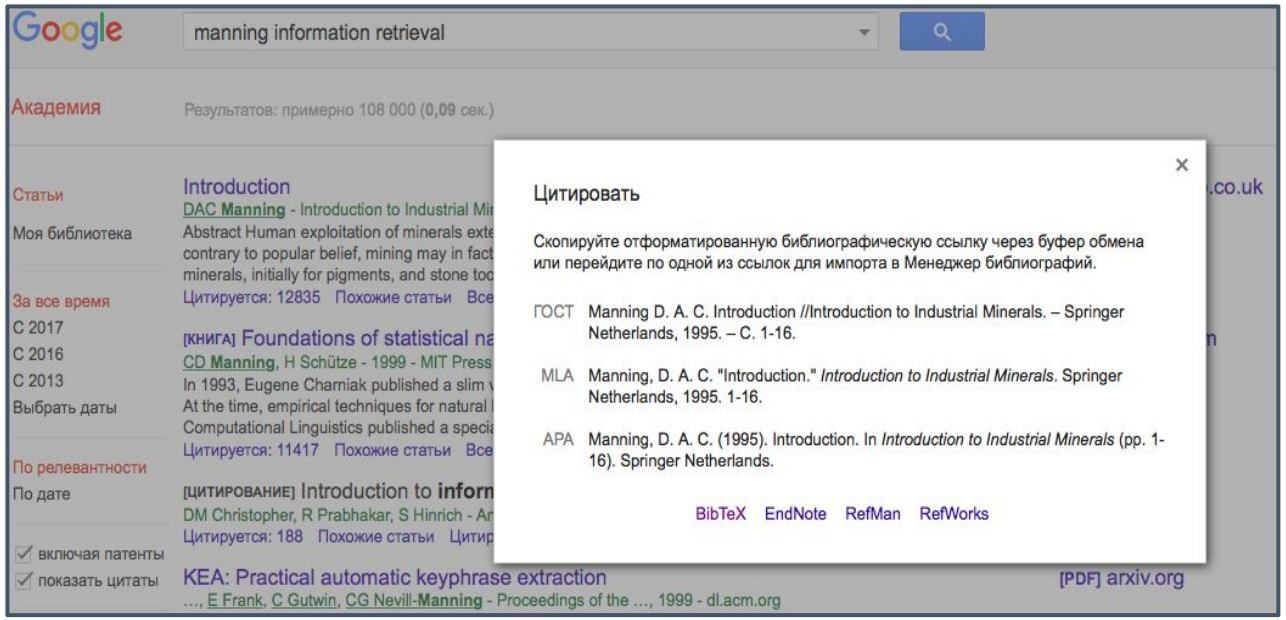


Home | Recent Tabs | Current Page

You can type something here...

Поиск по метаинформации

Фактически - поиск информации по информации, представленной в другом формате.



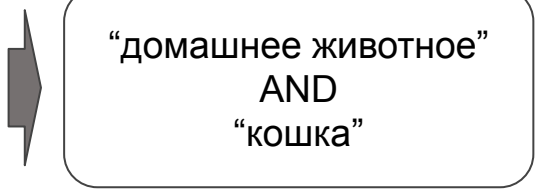
Булев поиск

Используются операторы алгебры-логики (or, and, xor). Есть только два варианта: объект либо подходит под запрос, либо нет.

Кошка была злая и вредная, но мы ее все равно любили. Звали ее Кусей, и была она рыжая в черную полоску.

Кошка - домашнее животное, одно из наиболее популярных (наряду с собакой) «животных-компаньонов»

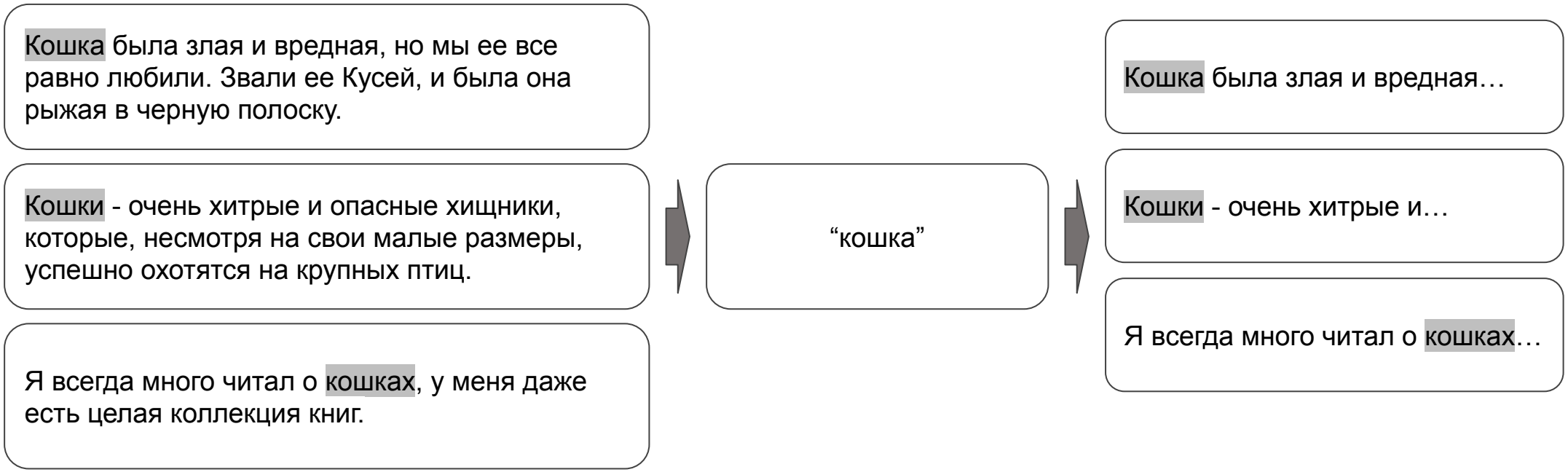
Домашнее животное - важная часть жизни ребенка, так как это учит его ответственности.



Кошка - домашнее животное, одно из наиболее популярных (наряду с собакой) «животных-компаньонов»

Поиск по сходству

Надстройка над булевым поиском: допускаются отклонения от запроса. Чаще всего оно ограничивается небольшим расстоянием Левенштейна.



Поиск по релевантности

Каждому документу присваивается число - мера его релевантности запросу

Примеры задач

А точнее проблемы, где необходимо использование методов инфопоиска

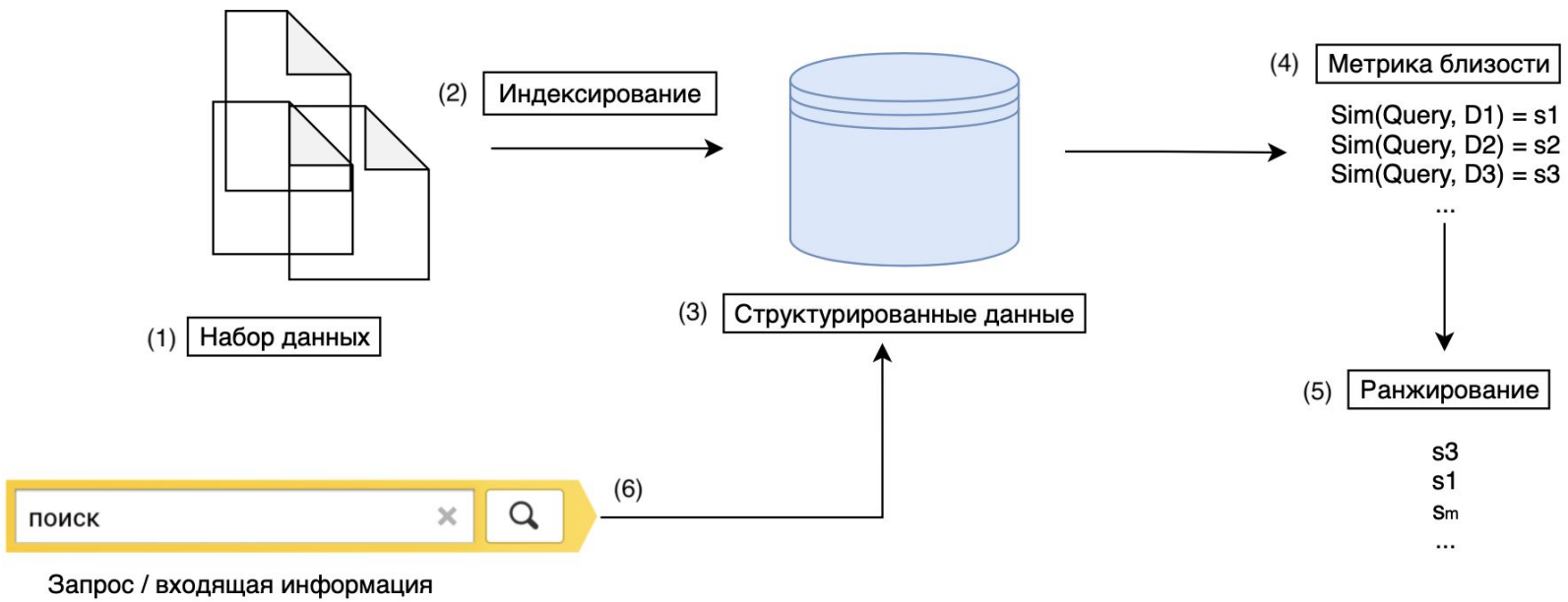
Ты шеф в большой компании. У тебя много разных отделов. Твоему стажеру надо узнать, как работает нечто, разработанное в другом отделе. Он не знает кому писать или боится спрашивать, но в итоге как-то находит источник информации.

Ты РЖД. У тебя есть колл-центр. Его задача отвечать на вопросы клиентов РЖД. Ты знаешь, что 60% вопросов повторяются из раза в раз. Использовать для этого человеческие ресурсы - дорого и малоэффективно.

Home | Recent Tabs | Current Page

You can type something here...

Последовательность действий



Общая постановка задачи

Дано

Набор объектов = база данных (1)

- Набор корпоративных документов
- Набор типичных вопросов и ответов на них
- Набор продаваемых товаров

Задача

Пришел новый объект - запрос (6)

- Описание сервиса, к которому ищем документацию
- Новый вопрос от юзера
- Фото штанов, которые нужно найти среди товаров

Надо найти самый подходящий к нему объект из базы

Шаг 1. Индексируем данные (2)

Что значит “индексируем”?

Ищем, обрабатываем и сохраняем данные таким образом, чтобы потом по ним было удобно искать.

Индекси́рование, совершаемое поисковой машиной, — процесс сбора, сортировки и хранения данных с целью обеспечить быстрый и точный поиск информации (то же самое на языке Википедии)

Шаг 2. Сохраняем индекс

Что такое индекс (3)?

В результате индексирования получаются структурированные данные или индекс.

Именно к этим данным - индексу, мы обращаемся во время поиска. Исходные данные, из которых он был получен, можно не использовать.

Почему? Смотри определение индекса.

Шаг 3. Выбираем метрику близости (4)

Что такое метрика?

Любая метрика (функция) близости, подходящая для измерения схожести тех объектов, с которыми мы работаем.

Это может быть:

- Сумма и среднее отклонений
- Или квадратов отклонений
- Косинусная близость
- И т.д.

Шаг 4. Ранжируем результаты (5)

Что значит “ранжируем”?

Сортируем в соответствии со значением метрики. На первом месте должен оказаться самый релевантный объект к запросу.

Ранжи́рование — сортировка сайтов в поисковой выдаче, применяемая в поисковых системах (и снова мнение википедии).

Прямой индекс

Есть корпус, состоящий из нескольких текстов:

doc_1 = Буря мглою небо кроет

doc_2 = Вихри снежные крутя

doc_3 = То, как зверь, она завоет

doc_4 = То заплачет, как дитя

Прямой индекс ставит каждому документу в соответствие слова, содержащиеся в нем. Например, в виде списка

Документ	Списко слов
doc_1	буря, кроет, мглою, небо
doc_2	вихри, крутя, снежные
doc_3	завоет, зверь, как, она, то
doc_4	дитя, заплачет, как, то

Обратный индекс

В обратном индексе каждому слову ставится в соответствие набор документов, где оно встречается. Может быть представлен (как и прямой индекс):

В виде словаря:

```
{
        "буря": [
                "doc_1"
        ],
        "то": [
                "doc_3",
                "doc_4"
        ],
        "как": [
                "doc_3",
                "doc_4"
        ],
  ...
}
```

В виде Document-Term матрицы:

	буря	мглою	небо	кроет	вихри	снежные	крутя	...
doc_1	1	1	1	1	0	0	0	
doc_2	0	0	0	0	1	1	1	
doc_3	0	0	0	0	0	0	0	
doc_4	0	0	0	0	0	0	0	