

Поисковые расширения

Владимир Гулин

Что такое поисковые расширения?

поиск@mail.ru айфон x q

Интернет Картинки Видео Приложения Новости Ответы

🍏 iPhone – Apple (RU)

apple.com/ru/iphone

Откройте для себя мир iPhone. Взгляните на iPhone 6s, iPhone 6 и iPhone SE. Зайдите на сайт Apple, чтобы изучить информацию, купить и получить поддержку.

поиск@mail.ru мгу x q

Интернет Картинки Видео Приложения Новости Ответы

🏛️ Московский государственный университет имени М.В. Ломоносова

msu.ru

Информация о факультетах, институтах, центрах, руководстве МГУ. Учеба: высшее, дополнительное, дистанционное образование, практическое обучение...

📍 Москва, мкр. Ленинские горы, д. 1 📞 +7 (495) 939-10-00

Об университете Поступающим

Образование Университетская жизнь

поиск@mail.ru гиппопотам x q

Интернет Картинки Видео Приложения Новости Ответы



Обыкновенный бегемот — Википедия

ru.wikipedia.org/wiki/...

Обыкновенный бегемот, или гиппопотам — млекопитающее из отряда парнокопытных, подотряда свинообразных (нежвачных), семейства бегемотовых, единственный современный вид рода Hippopotamus. Характерной особенностью бегемота является его полуводный образ жизни — большую часть времени он проводит...

Общая информация

Облик и строение

Название

Происхождение и систематика

Подвиды

Ареал и численность

Что такое поисковые расширения?

 × Найти ⋮

Интернет Соцсети beta Картинки Видео Новости Ответы

расшифровка→rasshifrovka

расшифровка→декодирование

расшифровка→дешифрование

расшифровка→интерпретация

расшифровка→истолкование

кндр → корейская народная
демократическая республика

кндр→северная корей

Простая схема работы ранжирования



Зачем нужны поисковые расширения

Расширения прежде всего влияют на:

- **Фильтрацию:**
 - находим новые релевантные документы и продвигаем их на следующие стадии поиска
- **Ранжирование**
 - с помощью расширений мы “изменяем” запрос, это влияет на факторы ранжирования
 - находим релевантные документы для редкочастотных запросов

Типы расширений. Синонимы

бедолага→бедняга

осваивать→изучать

беззвучно→бесшумно

оскорбительный→обидный

анальгетик→болеутоляющие средства

орган→органист

орган→организация

Типы расширений. Аббревиатуры

вмк→вычислительной математики и кибернетики

вет→ветеринарный

исполком→исполнительный комитет

начальник дивизии→начдив

Типы расширений. Вариативные написания

парадайс→парадайз

икея→икеа

дэдпул→дедпул

люксойл→люксоил

лютеранин→лютеранка

магазин шляп→шляпный магазин

маггл→маггловский

Типы расширений. Транслиты

lukoil→лукоил

пиканто→picanto

октавия→octavia

Типы расширений. Переводы

русский→russian

танцор→dancer

футбольный→football

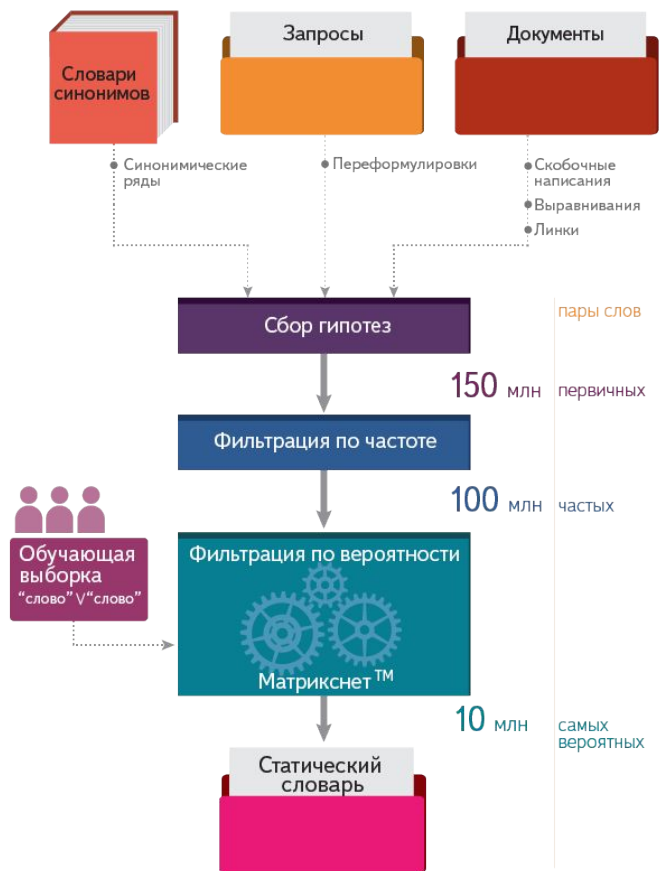
Слитное/раздельное написание

холодильникsharp→холодильник sharp

хотспот→хот спот

after noon→afternoon

Общая схема



Какие данные используются

- Запросы
- Клики/показы
- Переформулировки
- Тексты документов
- Снимпеты и заголовки
- N-grams
- Открытые базы синонимов
- Wikipedia, wiktionary
- WordNet
- Данные из соседних отделов

Какие данные используются

Запросы вместе с **документами**, на которые **кликали** несут основной сигнал:

- Что есть в тех документах, на которые они кликнули? Какие слова сделали запрос “похожим” на документ?

Запрос: “**танцующий миллионер**”

→

Заголовок: “Танцующий миллионер The **dancing millionaire**. Super Papa.”

- По двум разным запросам пользователи кликают на одни урлы. Что объединяет эти запросы? (ко-клики)

“**телеметрия** нижевартовск”→

“**телеметрические системы** для бурения”

Какие данные используются

N-grams:

- Текстовые
- Запросные
- Лемматизированные
- Скобочные/кавычечные

“ищейки с бродвея (bloodhounds of broadway, 1989)”

- Нейроязыковые модели

Какие данные используются

Таблицы машинного перевода:

genetically modified products → генетически измененная продукция

genetically modified products → генетически модифицированные продукты

Можно замкнуть:

измененная продукция → модифицированные продукты

Какие данные используются

Переформулировки:

физическая **рекреация**→палата для **восстановления** сил

Майнеры: аббревиатуры

- из скобочных нграмм:

Антиблокировочная система (АБС, ABS; *нем. Antiblockiersystem, англ. Anti-lock braking system*)

- из табличек с парами текстов:

“принцип работы **абс** на автомобиле”→

“устройство **антиблокировочной системы** тормозов”

Майнеры: википедия

Редиректы:

Дрифт

Материал из Википедии — свободной энциклопедии
(перенаправлено с «Управляемый занос»)

Текущая версия страниц

Ссылки:

Референдум

Материал из Википедии — свободной энциклопедии

Текущая версия страниц пока не проверялась опытными!

Референдум (лат. *referendum*) — форма непосредственного волеизъявления граждан,

Майнеры: Wiktionary

“Строитель”

Семантические свойства [править]

Значение [править]

1. специалист по строительству, то есть по возведению и реконструкции зданий и сооружений ♦ Не указан пример употребления
2. **высок.** человек, который создает что-либо ♦ Это были **строители** нового общества.
3. профессия, связанная со строительством ♦ Не указан пример употребления (см. **рекомендации**).

Синонимы [править]

1. -
2. созидатель

Антонимы [править]

1. -
2. разрушитель

Гиперонимы [править]

1. специалист
- 2.
3. профессия

Гипонимы [править]

1. монтажник, каменщик, штукатурщик
2. демиург

Майнеры: splitjoin

- Склеиваем слова в таблице с биграммami
- Смотрим часто ли встречаются слова в таблице с униграммами
- Фильтруем по частотам

Майнеры

По таблице с парами текстов смотрим как часто слова появляются в одних и тех же контекстах:

“как поступить в **мгу**” ->

“как поступить в **московский государственный университет**”

Или в походих контекстах:

“как поступить в **мгу**” ->

“поступление в **московский государственный университет**”

Майнеры: аббревиатуры

Ищем аббревиатуры внутри пары текстов

“**московский областной суд** как добраться” →

“**мособлсуд** адрес”

Запоминаем статистику встречаемости частей аббревиатуры:

мос → московский

обл→областной

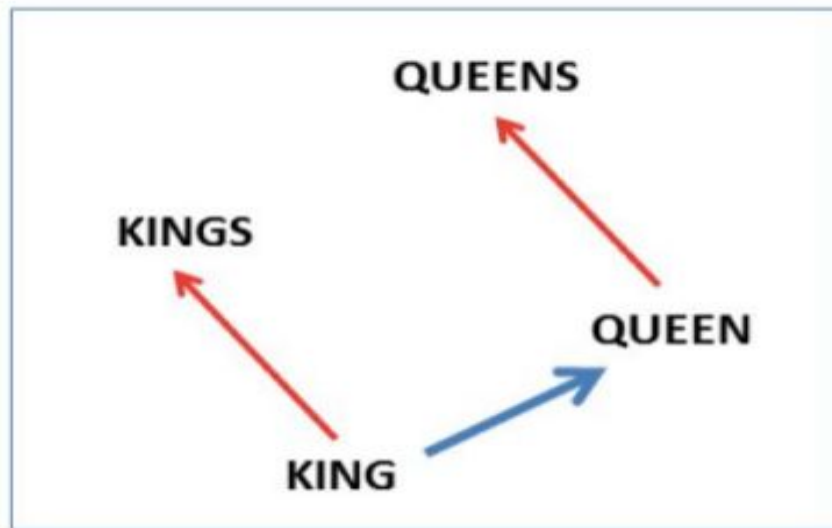
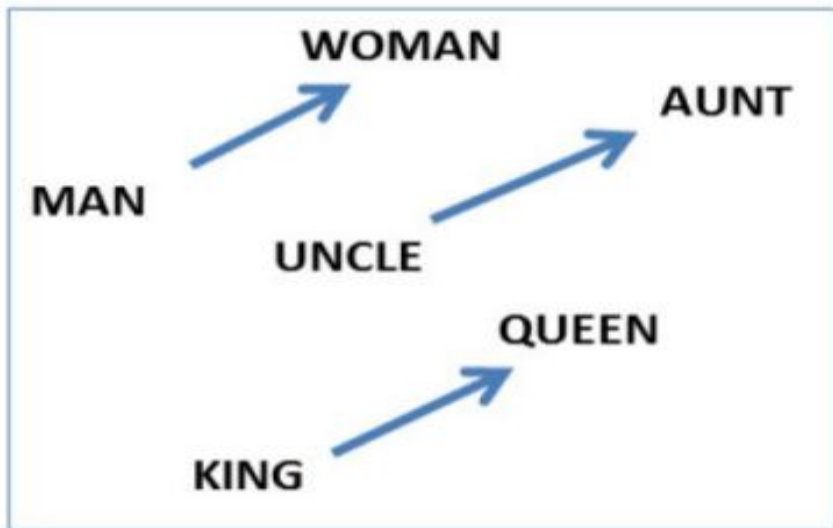
суд→суд

Генерируем из этих кусочков правдоподобные аббревиатуры:

нижоблсуд→нижегородский областной суд

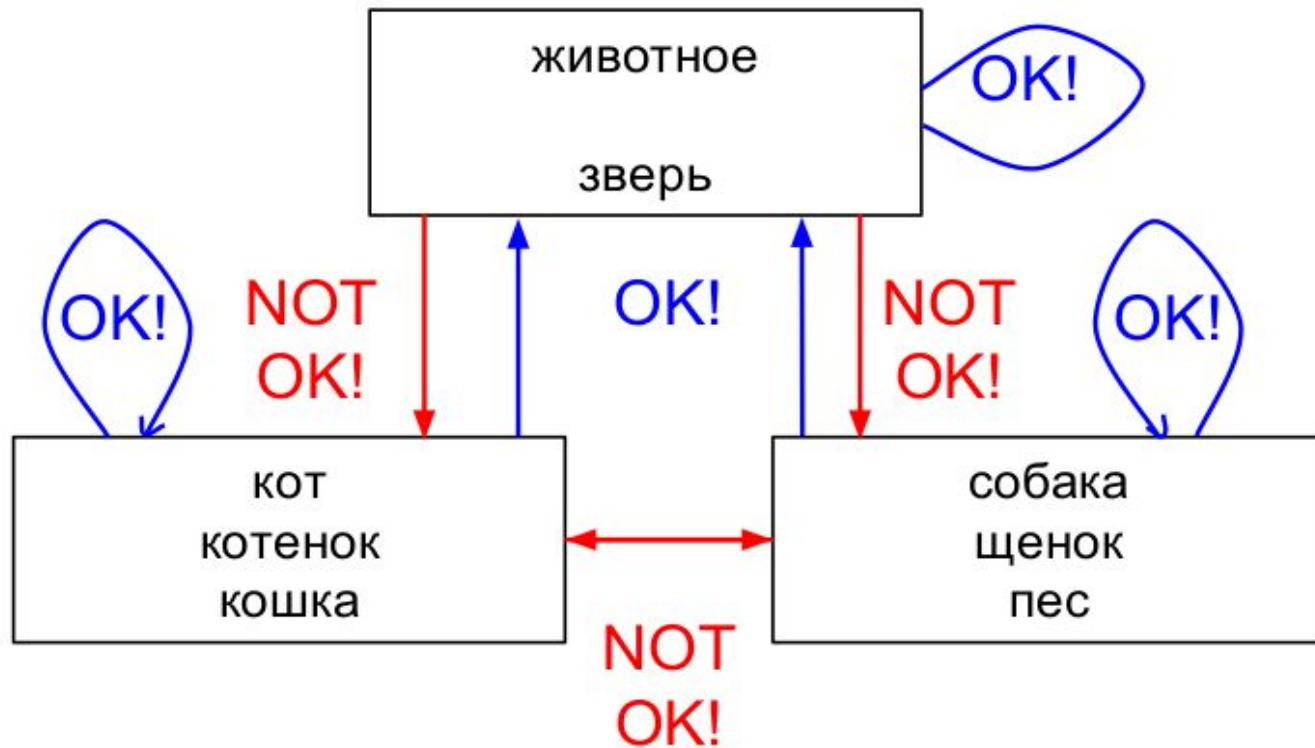
Майнеры: word2vec

- Возьмем много текстов/запросов и обучим word2vec
- Возьмем в качестве гипотез близкие по косинусному расстоянию



Проблема майнинга синонимов

Контекстно-близкие слова **далеко не всегда** хорошие расширения



Проблема майнинга синонимов

А еще:

- Месяцы/дни недели
- Марки автомобилей/телефонов
- Цвета
- Страны
- Имена
- ...

Контекстно похожи, но расширения из них не очень.

При этом синонимы должны встречаться в похожих контекстах.

Фильтрация

План:

- Посчитаем факторы для собранных гипотез
- Обучим классификатор
- Отберем лучшие расширения

Факторы

Около сотни факторов:

- Меры близости между строками,
- Лингвистические факторы
- Статистические (вхождения левых и правых частей в разные таблицы с данными)
- Факторы, пришедшие из майнеров

Факторы: лингвистические факторы

Отдельно для левой и правой частей смотрит на лингвистические признаки:

- Части речи
- Является ли леммой
- На каком языке
- Является ли числом
- Тип расширения
- ...

Факторы: edit distance

Машинно обученный edit distance минимизирует расстояние

Сделать разные стоимости операций в зависимости от букв и контекста.

Хотим минимизировать расстояния для опечаточных примеров:

рассчитать - расчитать

пщщпду - google

патечками - аптечками

любертсы - люберцы

клиенрты - клиенты

овуц - овец

Факторы: фонетические алгоритмы

Насколько схоже звучат левая и правая части гипотезы?

- SoundEx, Metaphone
- Левенштейн на транскрипциях

BP	1
FV	2
CKS	3
GJ	4
QXZ	5
DT	6
L	7
MN	8
R	9

Факторы: embeddings

- Представляем части гипотез в виде вектора
- Считаем близость векторов

символьные нграммы, word2vec, DSSM, ...

Факторы: ngrams

Построим языковую модель по собранным нграммам.

Посчитаем числовые факторы на основе этой модели

Факторы: статистические

- Во сколько запросов/текстов входит часть гипотезы и отдельные слова
- Сколько раз запрашивалась левая часть?
- ...

Факторы: PMI

Возьмем табличку с парой текстов (например, запрос -> заголовок)

Найдем вхождения левой части в запрос,

правой - в заголовок

Посчитаем PMI по этим вхождениям.

Факторы: из майнеров

- Сколько майнеров генерируют гипотезу/левую/правую часть
- Майнерозависимые факторы

Оценка

- Асессоры
- Клики

Качество факторов

- Считаем качество без фактора
- Считаем качество с добавлением фактора
- Смотрим насколько стало лучше

Проблема: сложно размечать

- улыбаться→ослабляться
- hidden markov model→hmm
- хиаб→hiab
- hike→пешеходный туризм

Контекстная модель

Собранные словари содержат расширения абстрактные “хорошие” расширения

Но уместность расширения очень зависит от контекста:

- “пм поступление” (пм→**прикладная математика**)
- “купить пм” (пм→**пистолет макарова**)

Контекстная модель пытается снять такую неопределенность



Качество

Считаем как сильно мы улучшили:

- Фильтрацию
- Ранжирования

В терминах метрик качества поиска