Российские интернет-технологии
2008

Исправление поисковых запросов в Яндексе

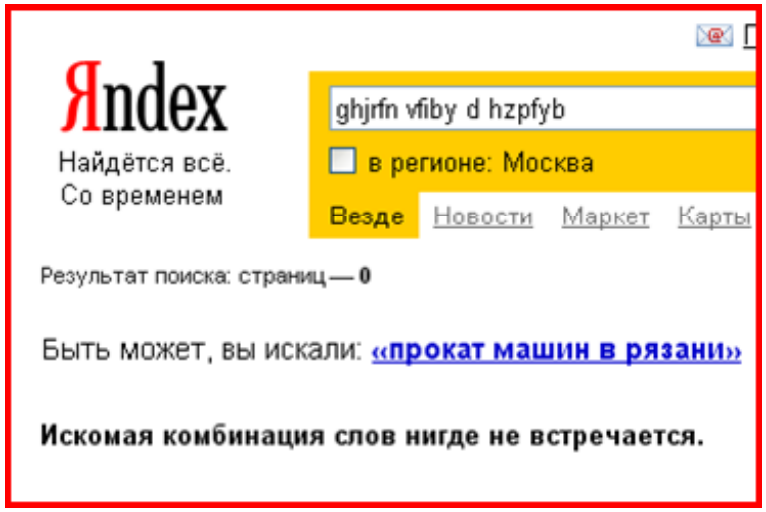
Вероятностная языковая модель

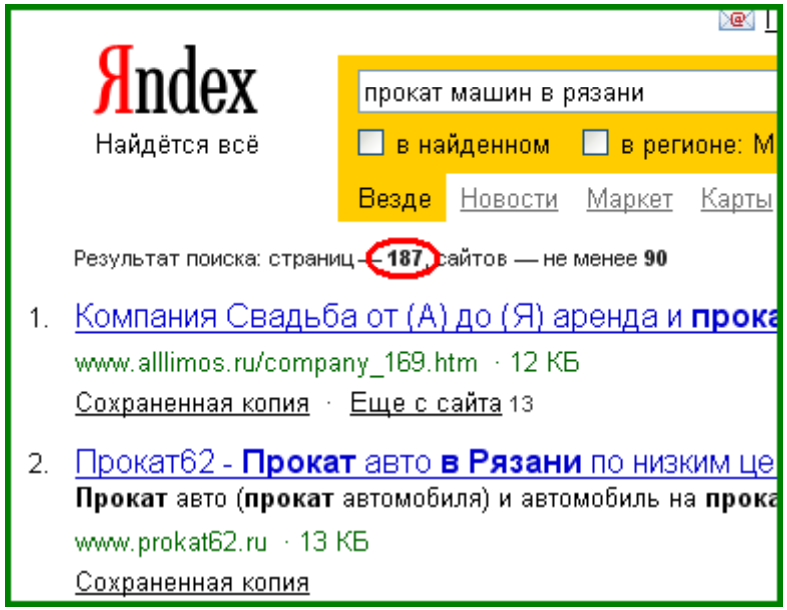
Ошибки в запросах

- …в **15%** запросов содержатся ошибки.

Ничего не найдено

Запрос: ghjrfn vfiby d hzpfyb
[прокат машин в рязани]

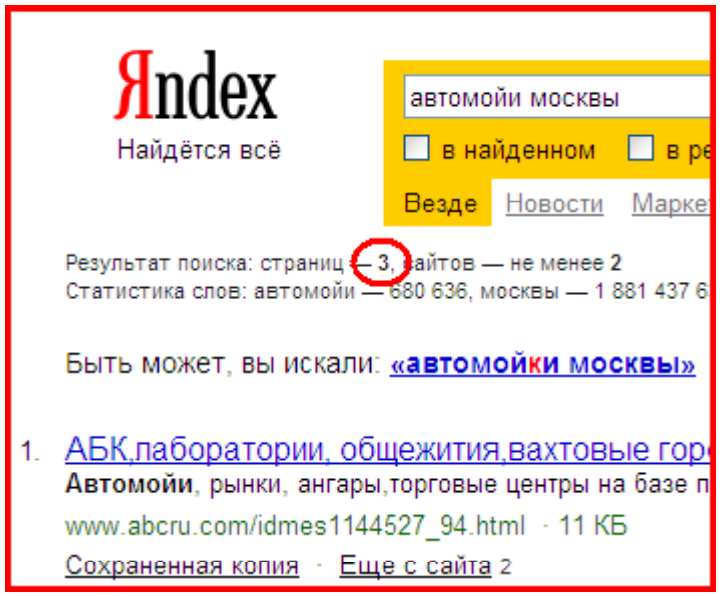


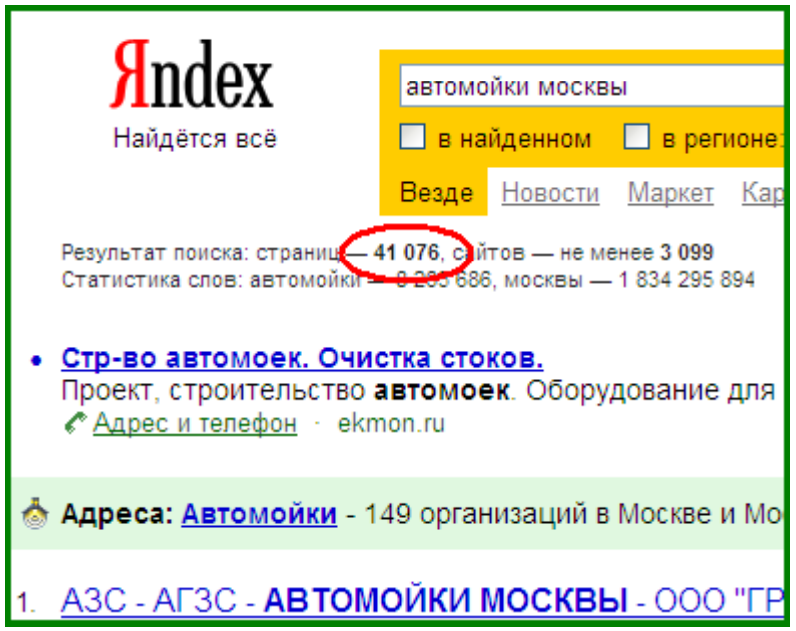


Что-то найдено…

Запрос: автомойи москвы

[автомойки москвы]

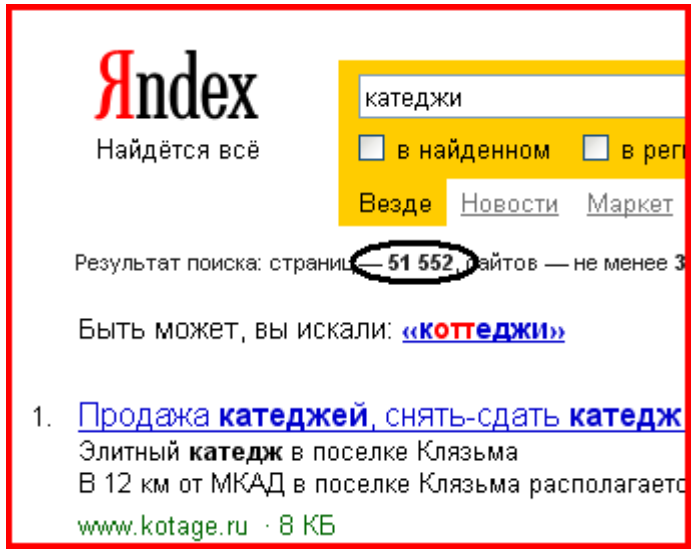


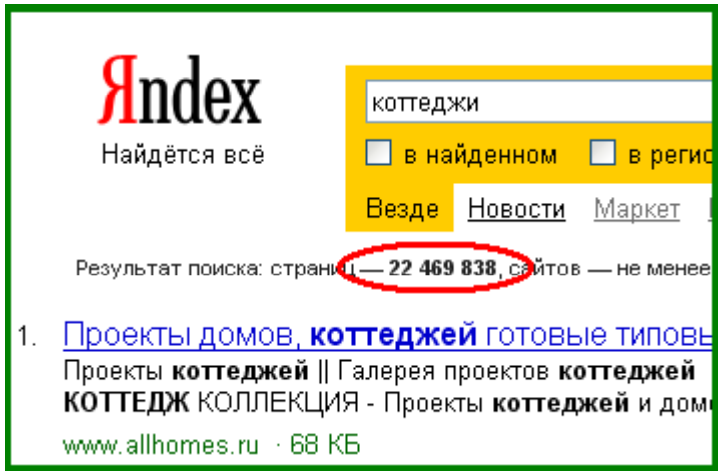


Найдено «то, что надо»?

Запрос: катеджи

[коттеджи]





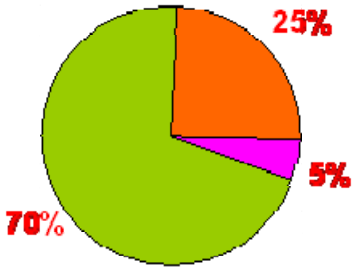
Источники ошибок в запросах

- **Случайные клавиатурные ошибки**

  приклбчение вместо *приключение*

- **Систематические когнитивные ошибки**
-
  - **Фонетические ошибки**
    игипет вместо *египет*
  - **Слитно-раздельное написание**
    фото ателье вместо *фотоателье*
  - **Заимствованные слова**
    фитнесс вместо *фитнес*
  - **Названия** фирм, брендов и т.п.
    лачетти вместо *лацетти*
  - **Марки** товаров
    Нокиа i_850_w вместо *Нокиа i850w*

Типы ошибок



- **Искажения в отдельных словах**
  - исчезновение буквы (*афавит* вместо *алфавит*)
  - вставка лишней буквы (*вврач* вместо *врач*)
  - замена буквы (*барабае* вместо *барабан*)
  - перестановка соседних букв (*притнер* вместо *принтер*)
- **Искажения в последовательности слов**
  - вставка пробела (*Вели кий Новгород* вместо *Великий Новгород*)
  - пропуск пробела (*КрасныйОктябрь* вместо *Красный Октябрь*)
- **Искажение смысла запроса**
  - контекстные ошибки (*вокруг меха* вместо *вокруг смеха*)
- **Латинский алфавит вместо русского**
  - транслитерация (*varezhka* вместо *варежка*)
- **Искаженная кодировка**
  - использование неправильной раскладки клавиатуры (*bnfkbz* вместо *италия*)

Что мы знаем об ошибках?

1. **Правильных слов больше**
   Частотность опечатки обычно на порядок меньше частоты правильного слова

2. **Ошибки повторяются**
   Повторяемость клавиатурных и фонетических ошибок очень высокая

3. **Ошибки зависят от контекста**
   Корректность слова определяется словарным окружением

Запросы – это наше все

Запросы это:

- информация о **частотности** слов

- информация о **сочетаемости** слов

- информация о **переформулировках** слов

•

Переформулировки запросов

- **Вот как исправляют запросы сами пользователи:**

райфаззен	-> райффайзен
ретеил	-> ритейл
колбассоф	-> колбасофф
крбина	-> корбина
тинидозол	-> тинидазол
скаэкспресс	-> скайэкспресс

- **Выбираем такие пары из пользовательских сессий и складываем их в словарь замен.**

Наш арсенал

- **Словари**
  - заменяем «плохие» слова на «хорошие»
- **Частоты слов**
  - заменяем слова на более частотные
- **Частоты ошибок**
  - заменяем с учетом частотности опечатки
- **Словарное окружение**
  - смотрим на сочетаемость со словом слева и справа
- **Пользовательские замены**
  - используем «знания» пользователей

Исправление запросов

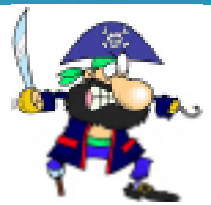
- **1. Интерпретация запроса**
  - Определение алфавита и кодировки
  - Разбиение запроса на слова
- **2. Поиск альтернативных вариантов**
  - Варианты словарных замен
  - Выбор лучшего кандидата

Рассмотрим запрос **скчать кодык игре ыефдлук**

- **3. Выбор оптимального исправления**

скчать | кодык | игре | ыефдлук

Восстановление алфавита и кодировки



Восстановление кодировки

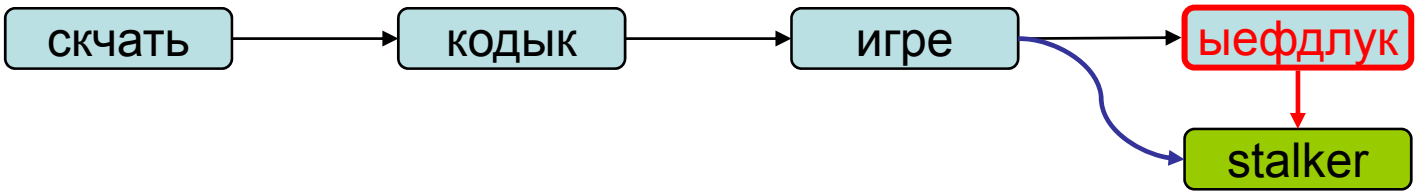
латиница -> кирилица (cytu -> *снег)*

кирилица -> латиница (ешьу -> *time)*

Восстановление алфавита

латинский –> русский (*mashina -> машина*)

- ыефдлук -> stalker



Сегментация запроса

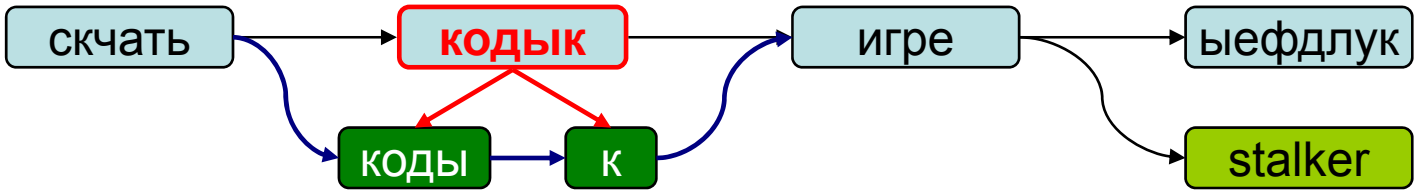
- Используем динамическое программирование (**алгоритм Витерби**).
- Вероятность каждой последовательности определяем по формуле:
  - $P = P_1 * P_2 * \ldots * P_N$

P_i – условная вероятность: слишком много вычислений…

Упрощаем: используем **модель однословных сочетаний**.

- **P(*альфа банк*) = P(*альфа*) * P(*банк | альфа*) ~ P(*альфа*) * P(*банк*)**

- Выбор лучшего разбиения: **модель двусловных сочетаний**.



Варианты словарных замен

- Для каждого слова составляем список возможных замен.
- **Метод простого перебора для многобуквенных ошибок не годится…**
- Используем метод **SOUNDEX**:

1. Убираем из слова гласные (***вода*** -> вд)
2. Оглушаем согласные (**вд** -> фт)
3. Производим удаления и перестановки (**фт** –> т,ф,тф)

Вода превратилась в [**фт,т,ф,тф**]…

Такие же саундексы имеют слова:

видео вход два ведь …

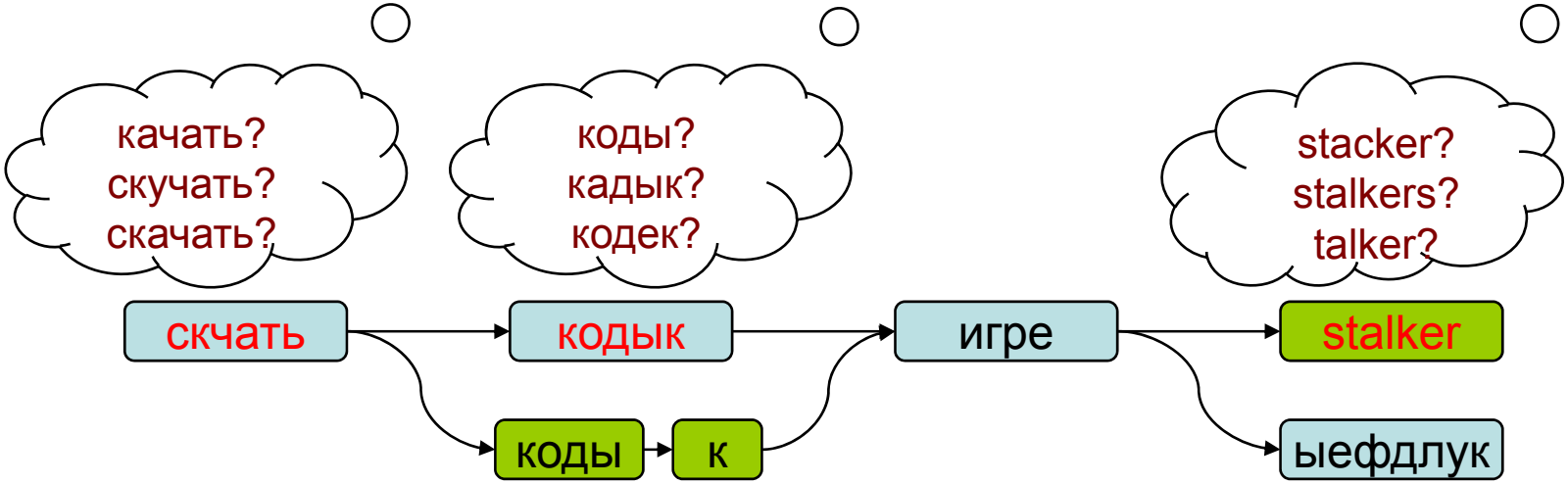
Мера близости

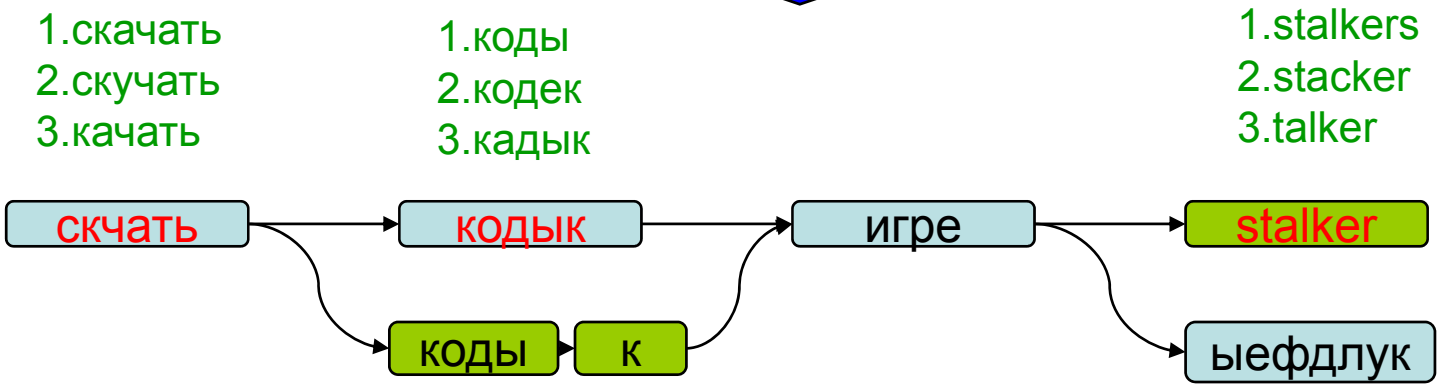
- ***скчать:*** *ска**а**чать* или *ску**у**чать?*
-
- Вероятность опечатки *скачть* [*скача**а**ть*] :
-
  P(*скачать* | *скчать*) = P(*скачать*) * $P_{ошибки}$(*скачаа**ть* : *скчать*)**
-
  $P_{ошибки}$(*скачаа**ть* : *скчать*) = ???**

- Упрощаем:

-
  $P_{ошибки}$(*скачаа**ть* : *скчать*) ~ $P_{ошибки}$(:a)**

Сортируем кандидатов





Выбор оптимального исправления

- Вероятность каждого варианта исправления:
- $P(w_1,w_2,..,w_N) = \Pi_{i=1,N} (P(w_i|w_1w_{2...}w_{i-1}) * \Pi_{k=1,K} (P_{ok})$
- где
- $P(w_i|w_1w_{2...}w_{i-1})$ – условная вероятность слова w_i
- P_{ok} – вероятность **k**-ой ошибки
- **Опять условная вероятность! Слишком много вычислений…**

- Упрощаем: используем модель **двусловных сочетаний**

- Для запроса из 3-х слов вместо

$$P(w_1) * P(w_2|w_1) * P(w_3|w_1w_2)$$

- применяем:

скачать коды $P(w_2$ к $* P(w$ игре stalker

Российские интернет-технологии 2008

Удачные подсказки

одеяло стебаное	–> одеяло стеганое
лодки катра	–> лодки катера
квадратный мэтр	–> квадратный метр
выборы мера	–> выборы мэра
желательная резинка	–> жевательная резинка
грибница фараона	–> гробница фараона
вышел с ухой из воды	–> вышел сухой из воды

Было:	Стало:
гадостное настроение	радостное настроение

Неудачные подсказки

белявский	-> милявский
олбас	-> колбас
брендмауэры	-> брандмауэры
термису	-> термину
трассологическая	-> графологическая
любочка	-> юбочка
берег у моря	-> берегу моря
вход или выход	-> входили выход

Фильтруем подсказку

Убираем **ненужные** подсказки

олбас -> колбас

Исправляем **неправильные** подсказки

термису -> тирамису

Убираем из подсказки **нецензурные** слова

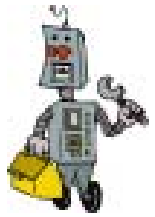
xx -> xxx

Источник информации о качестве подсказки:

статистика кликабельности.

Российские интернет-технологии 2008

Автоматическое исправление запросов



Yndex

Найдётся всё

прокат машин в рязани

☐ в найденном ☐ в регионе: Москва

Везде | Новости | Маркет | Карты | Словари | Блоги | Карт

Результат поиска: страниц — **128**, сайтов — не менее **64**
Статистика слов: прокат — 43 100 180, машин — 849 231 523, в — 25 944 559 503, ряз
30 359 042.

Запрос исправлен, так как по исходному ничего не найдено.

1. Компания Свадьба от (А) до (Я) аренда и **прокат машин в Р**

Yndex

Найдётся всё

коттеджи

☐ в найденном ☐ в регионе: Москва

Везде | Новости | Маркет | Карты | Словари | Блоги | Картинки

Результат поиска: страниц — **21 553 676**, сайтов — не менее **111 488**, в каталоге — **473**
Статистика слов: коттеджи — 62 973 117.

Запрос исправлен на «коттеджи». Показать результаты для «катеджи»?

1. Проекты домов, **коттеджей** готовые типовые проекты **коттеджей**

Немного статистики

- **Находим ошибки в 10% запросов**
- **Точность исправления 75%**

- **Используем словарные базы:**

Список двусловий	**29M** сочетаний
Словарь	**2.7M** слов
Пользовательские замены	**190K** замен
Индекс кликабельности	**320K** замен

- **«Обслуживаем» службы Яндекса:**
- Поиск Блоги Новости Карты Маркет Картинки
- **Нагрузка кластера исправления**

Алексей Байтин

адрес: 111033, Россия, Москва,
ул. Самокатная д.1, стр. 21.
телефон: +7 (495) 739-00-00
факс: +7 (495) 739-70-70

эл. почта: baytin@yandex-team.ru