



Российские
интернет-технологии
2008

Исправление поисковых запросов в Яндексе

Вероятностная языковая модель



Российские
интернет-технологии
2008

Ошибки в запросах

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



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



Яndex
Найдётся всё.
Со временем

ghjrfn vfiby d hzpfyb

в регионе: Москва

Везде [Новости](#) [Маркет](#) [Карты](#)

Результат поиска: страниц — 0

Быть может, вы искали: [«прокат машин в рязани»](#)

Искомая комбинация слов нигде не встречается.



Яndex
Найдётся всё

прокат машин в рязани

в найденном в регионе: М

Везде [Новости](#) [Маркет](#) [Карты](#)

Результат поиска: страниц — **187** сайтов — не менее 90

1. [Компания Свадьба от \(А\) до \(Я\) аренда и прокат](#)
[www.allimos.ru/company_169.htm](#) · 12 КБ
[Сохраненная копия](#) · [Еще с сайта](#) 13
2. [Прокат62 - Прокат авто в Рязани по низким ценам](#)
Прокат авто (прокат автомобиля) и автомобиль на прокат
[www.prokat62.ru](#) · 13 КБ
[Сохраненная копия](#)





Запрос: **АВТОМОЙИ** МОСКВЫ
[автомойки москвы]



Яndex
Найдётся всё

автомойи москвы

в найденном в регионе

Везде [Новости](#) [Маркет](#)

Результат поиска: страниц — 3, сайтов — не менее 2
Статистика слов: автомойи — 680 636, москвы — 1 881 437 6

Быть может, вы искали: [«автомойки москвы»](#)

1. [АБК, лаборатории, общежития, вахтовые гор...](#)
Автомойи, рынки, ангары, торговые центры на базе п
[www.abcru.com/idmes1144527_94.html](#) · 11 КБ
[Сохраненная копия](#) · [Еще с сайта 2](#)



Яndex
Найдётся всё

автомойки москвы

в найденном в регионе

Везде [Новости](#) [Маркет](#) [Кар](#)

Результат поиска: страниц — 41 076, сайтов — не менее 3 099
Статистика слов: автомойки — 8 285 686, москвы — 1 834 295 894

- [Стр-во автомоек. Очистка стоков.](#)
Проект, строительство **автомоек**. Оборудование для
[Адрес и телефон](#) · [ekmon.ru](#)

Адреса: [Автомойки](#) - 149 организаций в Москве и Мо

1. [АЗС - АГЗС - АВТОМОЙКИ МОСКВЫ - ООО "ГР](#)





Запрос: **катеджи**
[коттеджи]



Yandex
Найдётся всё

катеджи
 в найденном в реги...

Везде Новости Маркет

Результат поиска: страниц **51 552**, сайтов — не менее 3

Быть может, вы искали: **«коттеджи»**

1. [Продажа катеджей, снять-сдать катедж](#)
Элитный **катедж** в поселке Клязьма
В 12 км от МКАД в поселке Клязьма располагаетс
www.kotage.ru · 8 КБ



Yandex
Найдётся всё

коттеджи
 в найденном в реги...

Везде Новости Маркет

Результат поиска: страниц **22 469 838**, сайтов — не менее

1. [Проекты домов, коттеджей готовые типовы](#)
Проекты **коттеджей** || Галерея проектов **коттеджей**
КОТТЕДЖ КОЛЛЕКЦИЯ - Проекты **коттеджей** и дом
www.allhomes.ru · 68 КБ





- **Случайные клавиатурные ошибки**

*прикл**б**чение* вместо *при**к**лючение*



- **Систематические когнитивные ошибки**

-

- **Фонетические ошибки**

***и**гипет* вместо *египет*

- **Слитно-раздельное написание**

*фото **а**телье* вместо *фото**а**телье*

- **Заимствованные слова**

*фит**н**ес**с*** вместо *фит**н**ес*

- **Названия фирм, брендов и т.п.**

*ла**ч**етти* вместо *ла**ц**етти*

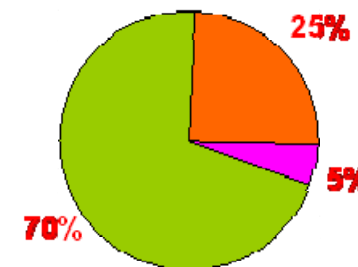
- **Марки товаров**

*Нокиа **і**_850_**w*** вместо *Нокиа **і**850**w***



■ 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*)



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



Российские
интернет-технологии
2008

Запросы – это наше все



Запросы это:

- информация о **частотности** слов
- информация о **сочетаемости** слов
- информация о **переформулировках** слов





- Вот как исправляют запросы сами пользователи:



райфаззен -> райфрайзен

ретеил -> ритейл

колбассоф -> колбасофф

крбина -> корбина

ТИНИДОЗОЛ -> ТИНИДАЗОЛ

скаэкспресс -> скайэкспресс

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



- **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



- **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**

скчать

КОДЫК

игре

ыефдлук

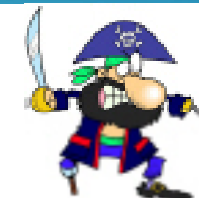
исправления



Восстановление кодировки

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

кирилица -> латиница (*ешью* -> *time*)

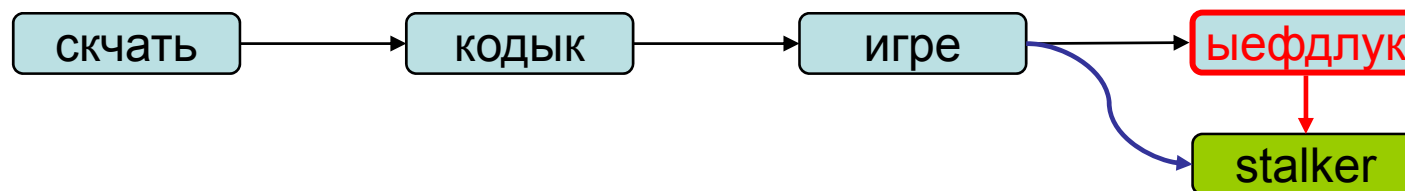


Восстановление алфавита

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



ыефдлук -> stalker





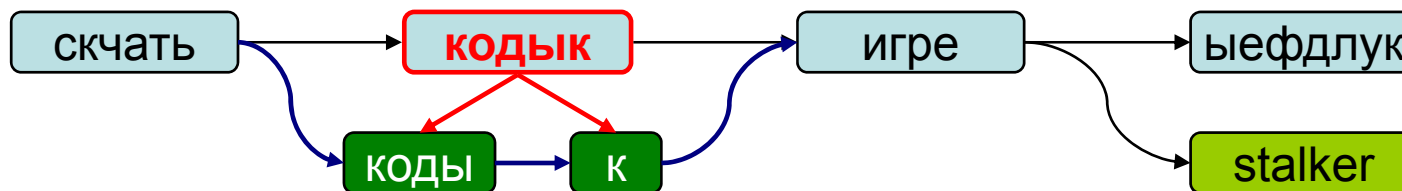
- Используем динамическое программирование (алгоритм Витерби).
- Вероятность каждой последовательности определяем по формуле:

$$P = P_1 * P_2 * \dots * P_N$$

P_i – условная вероятность: слишком много вычислений...

Упрощаем: используем модель однословных сочетаний.

- $P(\text{альфа банк}) = P(\text{альфа}) * P(\text{банк} | \text{альфа}) \sim P(\text{альфа}) * P(\text{банк})$
- Выбор лучшего разбиения: модель двусловных сочетаний.





- 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:

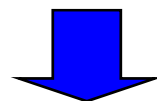
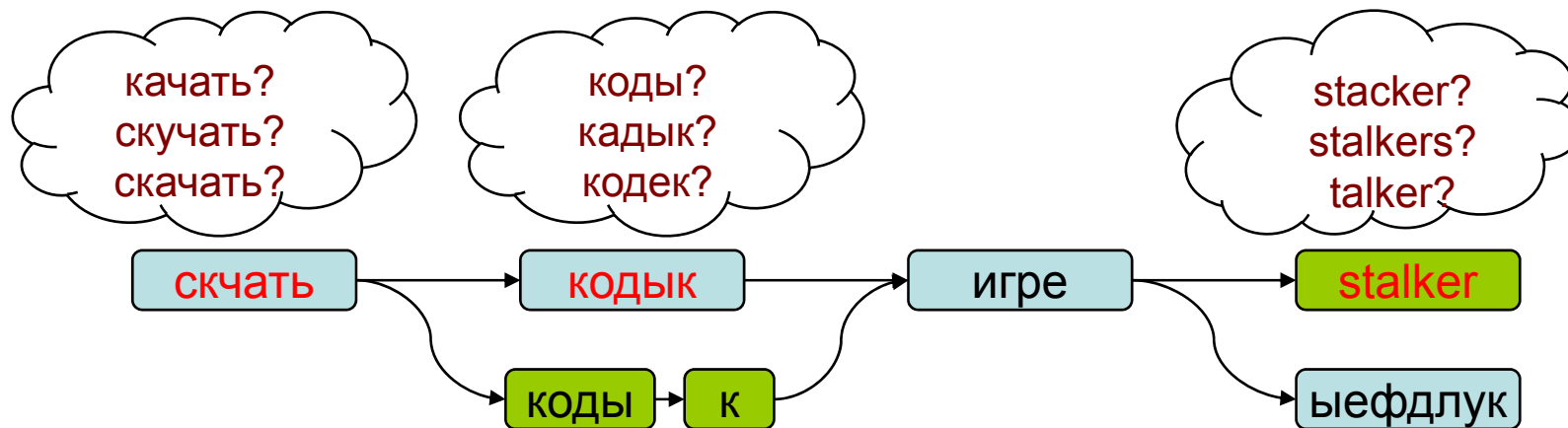
видео вход два ведь ...



- *скчатъ*: *скачатъ* или *скучатъ*?
-
- Вероятность опечатки *скачтъ* [*скачатъ*]
:
- $P(\text{скачатъ} \mid \text{скчатъ}) = P(\text{скачатъ}) * P_{\text{ошибки}}(\text{скачатъ} : \text{скчатъ})$
-
- $P_{\text{ошибки}}(\text{скачатъ} : \text{скчатъ}) = ???$
- Упрощаем:
- $P_{\text{ошибки}}(\text{скачатъ} : \text{скчатъ}) \sim P_{\text{ошибки}}(:\text{а})$



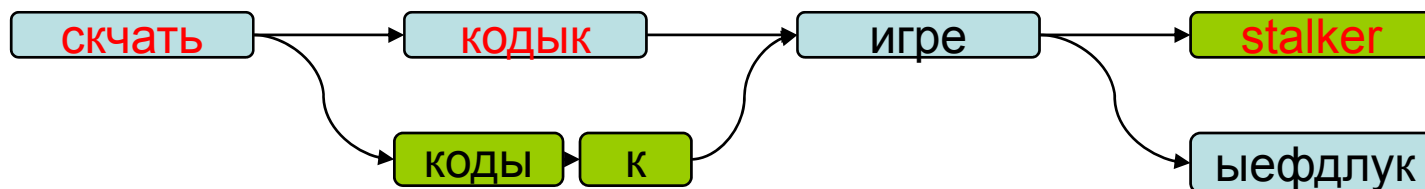
Сортируем кандидатов



1.скачать
2.скучать
3.качать

1.коды
2.кодек
3.кадык

1.stalkers
2.stacker
3.talker





- Вероятность каждого варианта исправления:
-
- $P(w_1, w_2, \dots, w_N) = \prod_{i=1, N} (P(w_i | w_1 w_2 \dots w_{i-1})) * \prod_{k=1, K} (P_{ok})$
- где
- $P(w_i | w_1 w_2 \dots w_{i-1})$ – 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

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

- применяем:

скачать

коды

$P(w_2$ к $* P(w$

игре

stalker



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



- | | |
|------------------|----------------------------------|
| белявский | -> ми лявский |
| олбас | -> к олбас |
| брендмауэры | -> бр андмауэры |
| термису | -> термин у |
| трассологическая | -> граф ологическая |
| любочка | -> юбочка |
| берег у моря | -> бере гу моря |
| ВХОД или ВЫХОД | -> ВХО дили ВЫ ХОД |



Российские
интернет-технологии
2008

Фильтруем подсказку

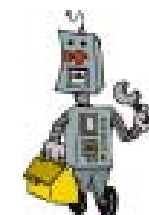
Убираем **ненужные** подсказки
олбас -> колбас

Исправляем **неправильные** подсказки
термису -> тирамису

Убираем из подсказки **нецензурные** слова
~~XX~~ -> XXX

Источник информации о качестве подсказки:
статистика кликабельности.





Яndex

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](#)

Яndex

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 и](#)



- 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