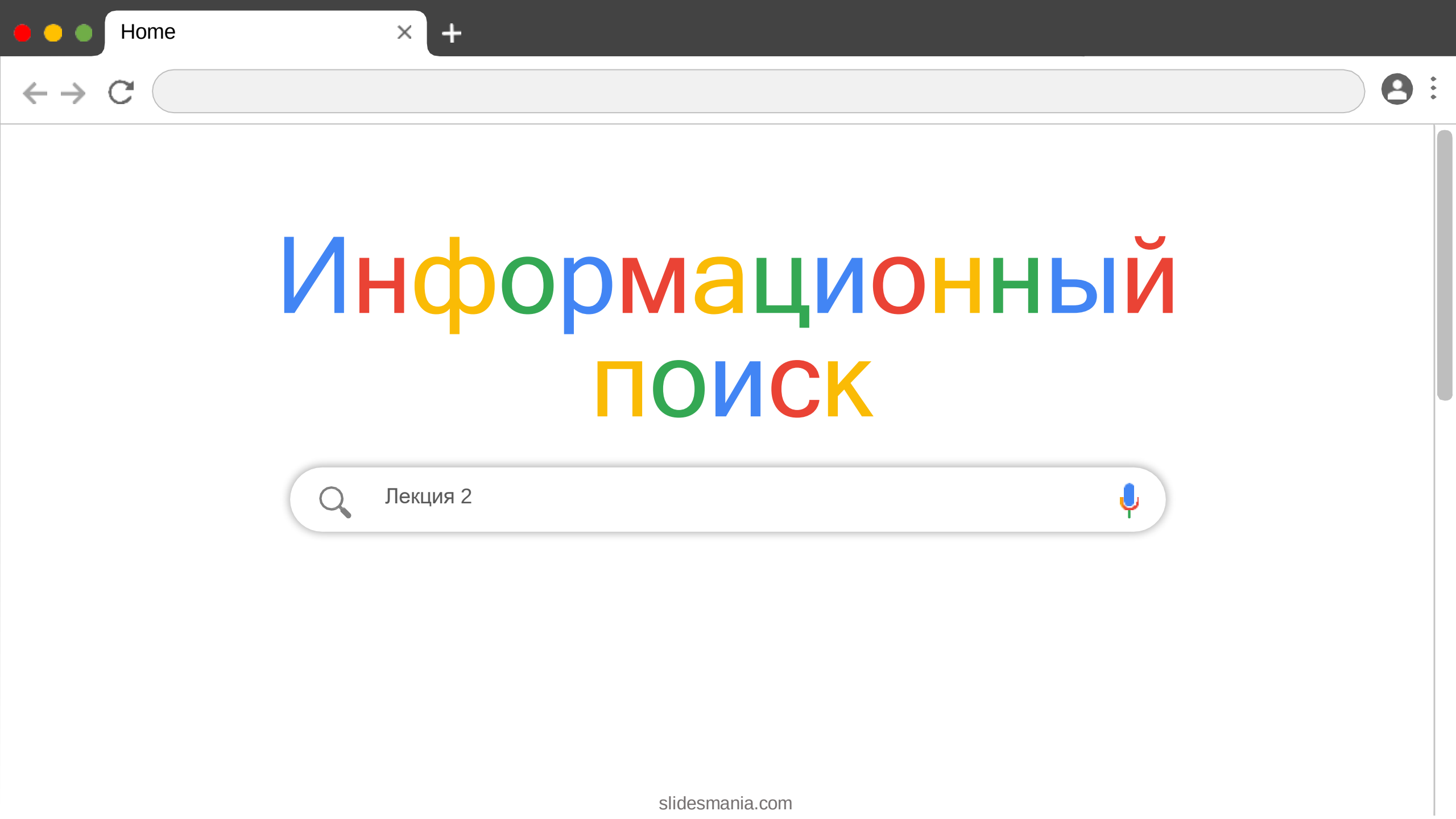
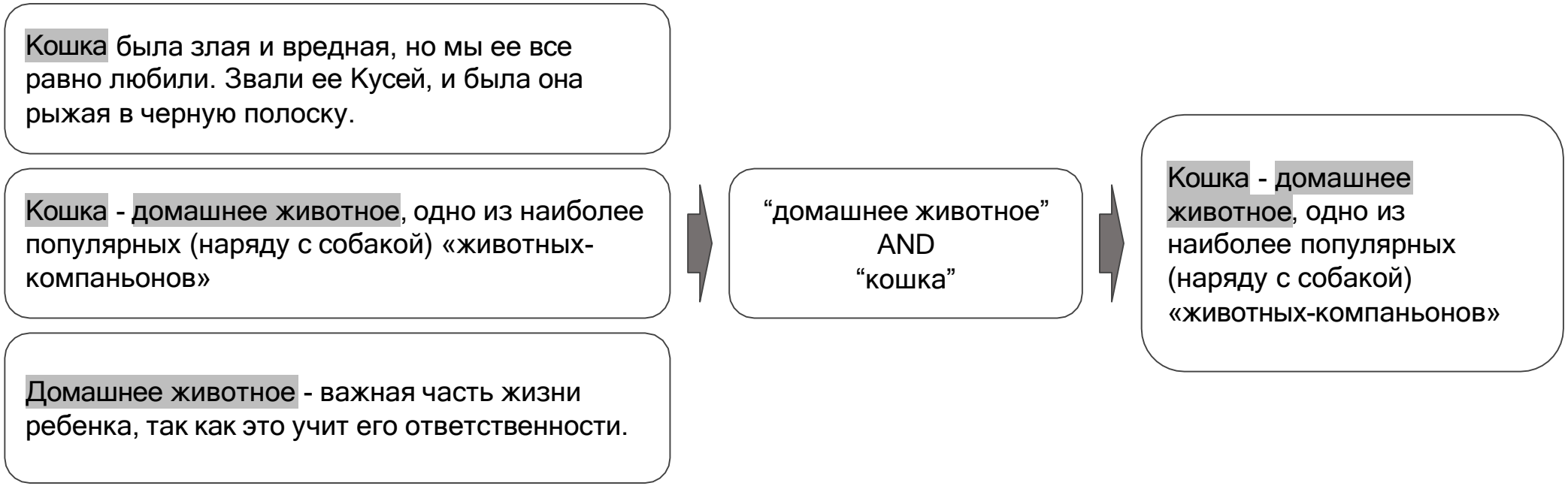


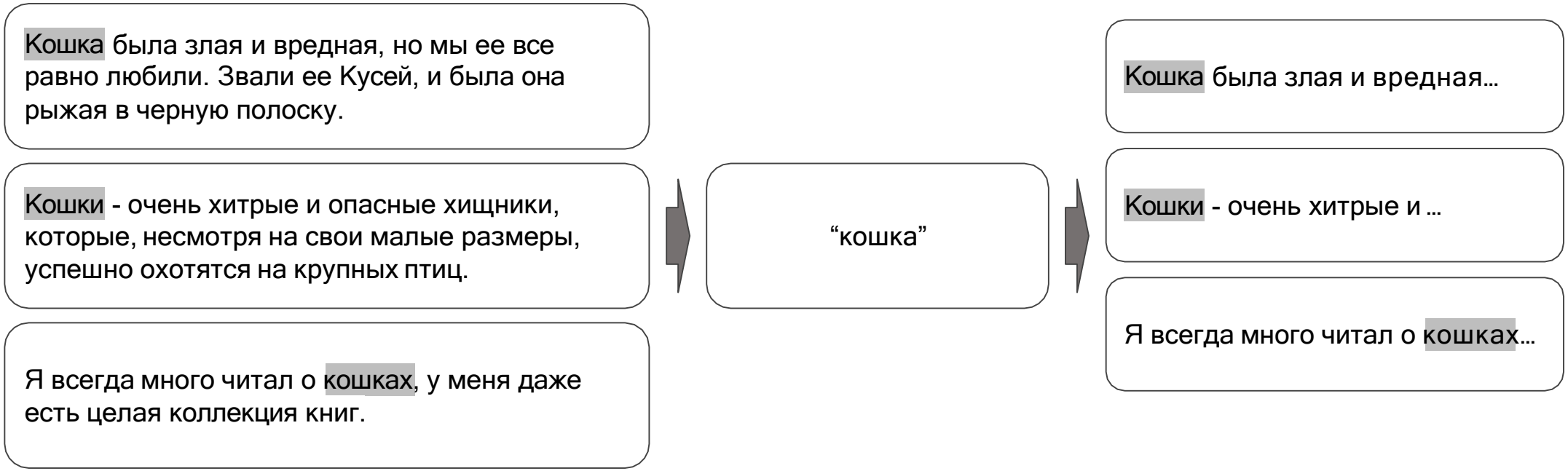
Булев поиск

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



Поиск по сходству (нечёткий поиск)

Надстройка над булевым поиском: допускаются отклонения от запроса.



Поиск по сходству (нечёткий поиск)

Зачем нужен поиск по сходству?

- чтобы могли находиться похожие слова (например, формы одного слова)
- но, конечно, для этого лучше использовать нормализацию текста, об этом поговорим чуть позже
- для исправления опечаток

Исправление опечаток

Реализация:

- исправление изолированного термина
  - расстояние редактирования
  - пересечение k-грамм
- исправление с учётом контекста
- фонетическое исправление

Расстояние редактирования

- чаще всего – расстояние Левенштейна

- расстояние Левенштейна между двумя строками символов s_1 и s_2 —
это минимальное количество операций редактирования, с помощью которых строку s_1 можно трансформировать в строку s_2

- операции редактирования, позволяющие это сделать, включают в себя следующие преобразования:
    1) вставка символа в строку
    2) удаление символа из строки
    3) замена символа в строке другим символом

- расстояние Дамерау-Левенштейна:
    + 4) транспозиция символов

Расстояние редактирования

- для слова с опечаткой ищем в словаре слово
с минимальным расстоянием редактирования до него

- если таких слов несколько, выбираем более распространенный вариант
(распространённость считаем либо по словарю, либо по запросам других пользователей)

Расстояние редактирования

- вариантов слов с указанным расстоянием редактирования может быть много, перебор может оказаться ресурсозатратным

- существуют разные эвристики: например, считается, что люди почти никогда не опечатываются в первой букве слова, поэтому такие варианты исправлений можно не учитывать

Пересечение k-грамм

Фактически k-граммный индекс используется для поиска терминов лексикона, содержащих большое количество k-грамм, общих с запросом. Идея заключается в том, что при разумной трактовке выражения "большое количество общих k-грамм" процесс поиска по существу сводится к однократному просмотру "словопозиций"[4] для k-грамм, входящих в запрос q.

Например, на рис. 3.7 показана часть словопозиций для трех биграмм в запросе `bord`. Допустим, что мы желаем найти термины лексикона, содержащие по крайней мере две из этих трех биграмм. Однократное сканирование записей (как описано в главе 1) позволило бы перечислить все такие термины; в примере, показанном на рис. 3.7, перечислены термины `aboard`, `boardroom` и `border`.

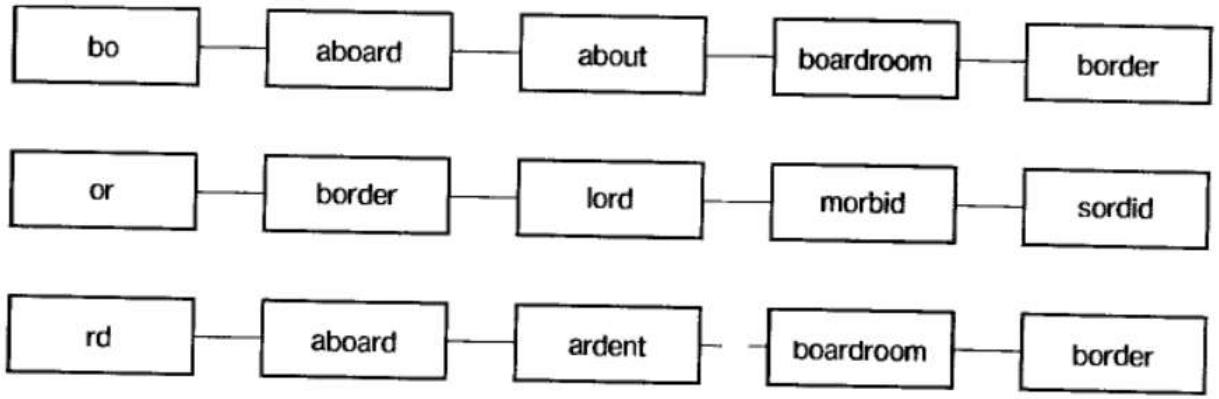


Рис. 3.7 Совпадение по крайней мере двух из трех биграмм в запросе `bord`

Исправление с учётом контекста

Допустим, опечатка в запросе привела к тому, что все слова в нём вроде бы существующие, но запрос получился бессмысленным. Например, «мальчик поел песню».
Если в ответ на фразу, подобную этой, будет возвращено лишь небольшое количество документов, то поисковая система может принять решение исправить этот запрос на «мальчик поет песню».
Простейший способ сделать это — перечислить исправления для каждого из трех терминов, даже если все термины запроса написаны правильно,
а затем попробовать произвести замену каждого термина фразы.
Более эффективный способ — перебирать только те комбинации возможных исправлений, которые чаще остальных встречаются в коллекции документов или в логах запросов.

Фонетическое исправление

- **Soundex** и другие похожие алгоритмы
- Основаны на фонетическом хешировании – слова с похожим звучаниям будут получать один и тот же хэш
- Для чего нужно фонетическое исправление?

Home | Recent Tabs | Current Page

You can type something here…

Нормализация текста

- приведение к единому регистру
- лемматизация / стемминг
- удаление стоп-слов (нужно ли?)

Нормализация текста

- приведение к единому регистру
- лемматизация / стемминг
- удаление стоп-слов (нужно ли?)

Какие задачи решает?

- нахождение разных форм слов из запроса
- сжатие словаря

Стемминг

Процесс приведения слов к их корневой форме путём удаления аффиксов, таких, как суффиксы и префиксы. Корневая форма не всегда может быть корректным словом в словаре, но она помогает стандартизировать варианты одного и того же слова.

Что лучше?

Лемматизация

Процесс приведения слов к их начальной форме (лемме), возвращает корректные слова из словаря.

Стоп-слова

Стоп-слова – **это слова**, которые часто встречаются в тексте,
но обычно не несут смысловой нагрузки.

- какие слова считать стоп-словами?
- какие слова могут считаться стоп-словами, скажем, для задачи частотного анализа,

но не будут таковыми для задач информационного поиска?

Home | Recent Tabs | Current Page

You can type something here…

Что почитать?

- Статья на Хабре про нечёткий поиск: https://habr.com/ru/articles/114997/
- Учебник «Введение в информационный поиск» (Маннинг, Рагхаван, Шютце)