Исправление опечаток в поисковых запросах

Александр Фонарев

newo@newo.su

Декабрь 2013

Содержание

Содержание

Где используется поиск и исправление опечаток?

1. в текстовых редакторах, браузерах и т.п.
2. в распознавании речи
3. при распознавании сканированного текста
4. ...
5. **в поисковых запросах**
6. ...

Опечаточная статистика

1. в среднем 3 слова в запросе
2. 12% запросов с опечатками
3. 80% полностью на русском
4. 84% ошибочных с одной ошибкой

Типы опечаток

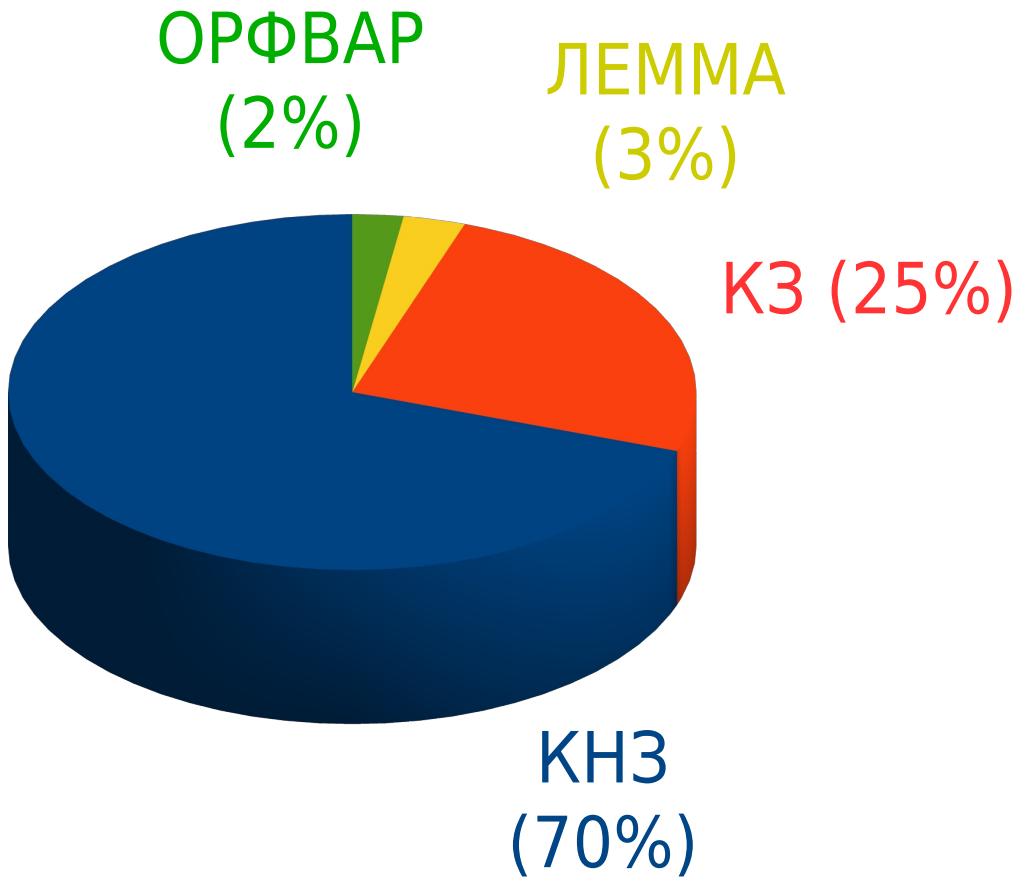
1. Ошибки с словах:
   1.1 Пропуск буквы (кросовки $\rightarrow$ кроссовки)
   1.2 Вставка буквы (фломастекр $\rightarrow$ фломастер)
   1.3 Замена буквы (эксперемент $\rightarrow$ эксперимент)
   1.4 Перестановка букв (пространтсво $\rightarrow$ пространство)
2. Склейка и разрезание:
   2.1 Пропуск пробела (купитьдиван $\rightarrow$ купить диван)
   2.2 Вставка пробела (пол года $\rightarrow$ полгода)
3. Раскладка клавиатуры (rfr cltkfnm cfqn $\rightarrow$ как сделать сайт)
4. Транслитерация (май нейм из саша $\rightarrow$ my name is sasha)

Опечаточная статистика

Контекстная статистика

1. 75% опечаток неконтекстные
2. 96% опечаток коротких слов контекстные

Контекстная статистика



Откуда брать данные?

1. Вручную размеченные пары *(запрос, исправление)* из поисковых логов. Таких данных мало, но они чистые.
2. Учет пользовательских кликов на подсказки. Таких данных много, но они шумные.
3. В качестве «правильных» текстов можно брать новостные статьи.
4. Можно почти рандомно делать ошибки
5. Можно посадить людей перепечатывать тексты, не давая им наживать на backspace и delete :)
6. Учет истории пользовательских исправлений в рамках сессии
7. ...

Содержание

Soundex

Используя фонетические правила, кодируем слова. Например:

1. удаляем гласные
2. m, n переходят в 5
3. d, t переходят в 3

Об алгоритме:

1. большое количество разновидностей этой идеи
2. зачастую работает лучше сложных и изощренных алгоритмов.
3. необходимы фонетические знания о языке

http://en.wikipedia.org/wiki/Soundex

Noisy channel

Пусть q — исходный запрос, c — исправление. Хочется:

$$c = \arg\max_c P(c|q) = \arg\max_c \frac{P(q|c)P(c)}{P(q)} = \arg\max_c P(q|c)P(c)$$

$$\approx \arg\max_c P(q|c)P^{\lambda}(c)$$

1. P(q|c) — модель ошибки (правдоподобие опечатки)
2. P(c) — языковая модель (априорная вероятность исправления)

Farooq Ahmad, Grzegorz Kondrak. Learning a Spelling Error Model from Search Query Logs, 2005

Языковая модель

1. Используем n-граммы
2. Сглаживаем модель

https://class.coursera.org/nlp/

Jurafsky D., James H. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech. 2000

Модель ошибки

Вспоминаем, что такое расстояние Левенштейна. Можно положить

$$-\log P(q|c) = dist(q, c).$$

А можно поступить умнее, обучив веса для каждого из переходов с помощью EM-алгоритма:

1. Инициализация: все нетождественные переходы малы и равновероятны, все тождественные велики и равновероятны (90%).
2. E-шаг: для каждого слова ищем близкие слова
3. M-шаг: пересчитываем вероятности

Можно поступать еще умнее и обучать трансфемы, такие как *ться*→*тся*. Получится трансфемная метрика.

Farooq Ahmad, Grzegorz Kondrak. Learning a Spelling Error Model from Search Query Logs. 2005

«Кластеризация» запросов и их опечаток

albert einstein	4834
albert einstien	525
albert einstine	149
albert einsten	27
albert einsteins	25
albert einstain	11
albert einstin	10
albert eintein	9
albeart einstein	6
aolbert einstein	6
alber einstein	4
albert einseint	3
albert einsteirn	3
albert einsterin	3
albert eintien	3
alberto einstein	3
albrecht einstein	3
alvert einstein	3

Misspelled query:	*anol scwartegger*
First iteration:	*arnold schwartnegger*
Second iteration:	*arnold schwarznegger*
Third iteration:	*arnold schwarzenegger*
Fourth iteration:	no further correction

Silviu Cucerzan, Eric Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. 2004

Некоторые детали

1. Исправления по парам соседних слов
2. Отдельный учет стоп слов вроде *and*
3. Учет словарных слов

Silviu Cucerzan, Eric Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. 2004

Содержание

Генерация кандидатов

Этот этап необходим для быстрой работы спеллчекера на реальном потоке запросов.

1. Генерация кандитатов исправления для каждого слова
2. Поиск k кратчайших путей в получившемся графе

Как работает:

1. Качественнее Soundex
2. Медленнее Soundex (на два-три порядка)
3. Для ускорения используется предподсчет частых комбинаций

Huizhong Duan, Bo-June (Paul) Hsu. Online Spelling Correction for Query Completion. WWW'11

Ранжирование кандидатов

1. Используем learning to rank, чтобы отсортивать кандидатов по надежности
2. Дальше будем смотреть только на верхний вариант
3. Возможно, понадобится отдельный классификатор надежности для верхнего варианта

Jianfeng Gao, Xiaolong Li, Daniel Micol, Chris Quirk, Xu Sun. A Large Scale Ranker-Based System for Search Query Spelling Correction. 2011

Какие признаки использовать?

Словарные признаки

1. Вес по буквенной языковой модели
2. Вес по словарной языковой модели
3. Длина слова
4. Присутствие в словарях
5. Вероятность быть именованой сущностью
6. Дистанция редактирования
7. Взаимный контекст

Запросные признаки

1. Результаты словарного классификатора
2. Вес по словарной языковой модели
3. Вес по буквенной языковой модели
4. Количество слов в запросе
5. Количество ошибок

Alexey Baytin, Irina Galinskaya, Marina Panina, Pavel Serdyukov. Speller Performance Prediction for Query Autocorrection. CIKM'13

Чего еще хочется?

1. Улучшать точность и полноту :)
2. Учитывать очень частые опечатки (аген**Т**ство)
3. Агглютинативные языки (турецкий, немецкий)
4. Учет результатов поиске по опечатке и исправлению
5. Персонализация исправлений
6. ...

Mu Li, Muhua Zhu, Yang Zhang, Ming Zhou. Exploring Distributional Similarity Based Models for Query Spelling Correction. 2006

Вопросы?