Теория информационного поиска

Лекция 4. Обработка текста

Дмитрий Грановский

СПбГУ

14.04.2025

В предыдущей серии

1. Проблему мешка слов можно решить с помощью N-граммного индекса.
2. Или с помощью координатного индекса.
3. Терминам в разных частях документа (зонах) могут быть присвоены различные веса.
4. Иногда полезно хранить для документа дополнительные атрибуты.
5. Некоторые документы могут быть полезнее других независимо от запроса.

Проблема

Иногда (часто) хотим считать одним и тем же термином токены, которые по-разному написаны в запросе и в документе.

Обычно это называют *нормализацией*.

Проблема

Иногда (часто) хотим считать одним и тем же термином токены, которые по-разному написаны в запросе и в документе.

Обычно это называют *нормализацией*.

Включает:

- графику
- морфологию
- опечатки (потом)

Графика

Общий принцип: главное — делать единообразно при индексации и при поиске.

Графика

Общий принцип: главное — делать единообразно при индексации и при поиске.

1. Капитализация
2. Диакритика
3. Орфоварианты (регулярные)
4. Транслитерация (потом)

Капитализация

- обычно схлопываем в нижний регистр (*case-folding*)
- иногда есть омонимия: `[сибирское отделение ран]`, `[расписание от казань]`
  - не учитываем регистр $\longrightarrow$ теряем в точности
  - учитываем регистр $\longrightarrow$ теряем часть документов
- иногда зависит от языка (турецкий: `Iı İi`)

Диакритика

- проблема: диакритики в документах и (особенно) запросах употребляются непоследовательно
- в русском языке проблема незначительна (небо-нёбо)
- но в ряде языков пары слов чаще различаются только диакритикой (французский, турецкий, испанский)
  - срезаем диакритику $\longrightarrow$ не различаем омографы
  - не срезаем диакритику $\longrightarrow$ теряем часть документов

Диакритика

- проблема: диакритики в документах и (особенно) запросах употребляются непоследовательно
- в русском языке проблема незначительна (небо-нёбо)
- но в ряде языков пары слов чаще различаются только диакритикой (французский, турецкий, испанский)
  - срезаем диакритику $\longrightarrow$ не различаем омографы
  - не срезаем диакритику $\longrightarrow$ теряем часть документов

Hollink et al.

diacritic removal gains up to 23% (especially helpful for Finnish, French, and Swedish)

Орфоварианты

- тоннель? туннель?
- кэшбэк? кэшбек? кешбэк? кешбек?
- München / Muenchen
- это не опечатки
- простое решение: выберем один из вариантов, а все остальные будем приводить к нему (в индексе и в запросе)
- но бывают проблемы с омонимами:
  - [сколько граммов в *карате*]

Морфология

Задачи:

- понять, является ли токен словом
- нормализовать графику (см. выше)
- определить язык
- привести к нормальной форме
  - стемминг
  - лемматизация

Распознавание языка

Зачем:

- критерий фильтрации/ранжирования
- выбор морфоанализатора
  - а также модели опечаточника, расширений, поверхностного синтаксиса…
- выбор индекса (иногда)

Распознавание языка

Зачем:

- критерий фильтрации/ранжирования
- выбор морфоанализатора
  - а также модели опечаточника, расширений, поверхностного синтаксиса…
- выбор индекса (иногда)

Как:

- N-граммы символов
- регион запроса (по IP-адресу)
- настройки пользователя/браузера

Стемминг

- попытка отсечь всё лишнее от основы
- работает быстрее, не требует словаря
- самый известный алгоритм — стеммер Портера (1980), см. *Snowball*

Стемминг

- попытка отсечь всё лишнее от основы
- работает быстрее, не требует словаря
- самый известный алгоритм — стеммер Портера (1980), см. *Snowball*

> Stemming helped markedly for Finnish (30% improvement) and Spanish (10% improvement), but for most languages, including English, the gain from stemming was in the range 0-5%, and results from a lemmatizer were poorer still.

Сложные слова

- для русского языка в целом небольшая проблема
- но в других языках:
  - yliopistotutkintolautakunta
  - Kraftfahrzeughaftpflichtversicherung
- используется т.н. *compound splitting*
- отдельная, но связанная проблема — языки без пробелов

Сложные слова

- для русского языка в целом небольшая проблема
- но в других языках:
  - yliopistotutkintolautakunta
  - Kraftfahrzeughaftpflichtversicherung
- используется т.н. *compound splitting*
- отдельная, но связанная проблема — языки без пробелов

Compound splitting gained 25% for Swedish and 15% for German, but only 4% for Dutch.

Лемматизация

- для русского языка работает и важна
- в т.ч. нужна для сокращения словаря индекса
- при поиске важна скорость
- размер словаря может понадобиться ограничить
  - например, можно выбросить редкие формы
  - см. бессловарная морфология
- изменения в общем случае требуют переиндексации

Цитатный (фразовый) поиск

[техника безопасности]

Цитатный (фразовый) поиск

[техника безопасности]

- $d_1 =$ “техника безопасности”
- $d_2 =$ “безопасность техники”

Цитатный (фразовый) поиск

[техника безопасности]

- $d_1 =$ “техника безопасности”
- $d_2 =$ “безопасность техники”

Решение:

- в каком-то виде хранить в индексе грамматические признаки текстоформы
- сделать фактором ранжирования (например) долю их точных совпадений с запросом
- похожая идея — для капитализации и диакритики