Занятие 2

Предобработка

Катя Герасименко Зимняя олимпиадная школа МФТИ 04.01.2019

Как работать с текстом?

Как скормить машине текст?

- целиком:)
- посимвольно (в некоторых задачах работает хорошо)
- по словам

Что такое слово?

- кусок строки от пробела до пробела
 токенизация сплитом по пробелам дает нормальное качество, но чаще нужно лучше
- знаки препинания удалить, оставить?
- contractions и другие апострофы (don't, we're, Smith's)
- дефисы (Санкт-Петербург vs голубо-зеленый)
- пробелы (в течение, не работает (ср. некрасивый))
- точки конец предложения vs т. д.
- и многие, многие другие детали

Как это делается

- Большая и тяжелая система правил реализована в NLTK
- Машинное обучение спасение для беспробельных языков
- ➤ Если что-то специфическое иногда надо написать свой токенизатор

Регистр

Зависит от задачи.

Чаще всего – приводить все к нижнему.

Но, например, для именованных сущностей информацию о регистре надо обязательно сохранить.

Спеллингчек

Степень необходимости зависит от задачи.

В чат-боте и поиске лучше иметь (и вообще возможность нечеткого поиска — fuzzy search) В основе хорошего спеллингчека:

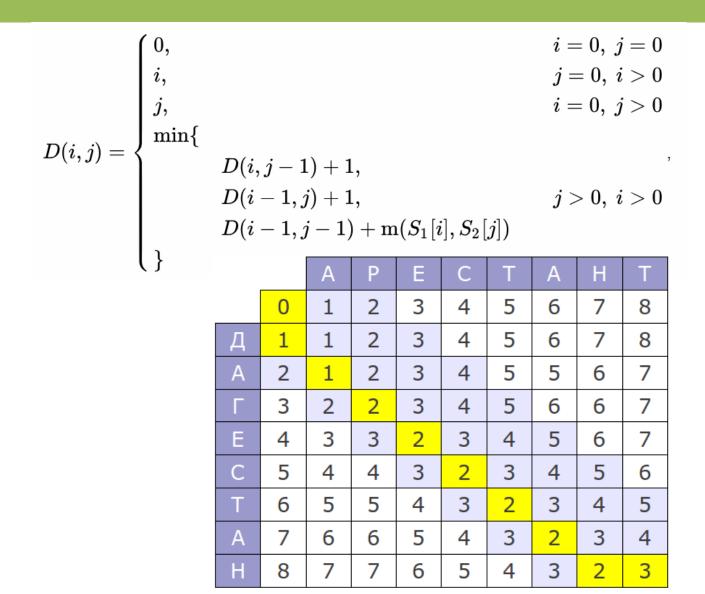
- модель для подбора кандидатов (расстояние Дамерау-Левенштейна, например)
- языковая модель для выбора среди кандидатов

Расстояние Левенштейна

Расстояние Левенштейна (одна из метрик редакционного расстояния (edit distance)) — минимальное количество операций вставки, удаления и замены, требующееся для того чтобы превратить одну строку в другую.

M	М	М	R		M	R	R
С	0	N	N		Ε	С	T
С	0	N	Е	Н	Е	Α	D

Расстояние Левенштейна



Расстояние Левенштейна

- > Можно давать больший вес замене
- Можно учитывать еще транспозицию (80% всех опечаток) – расстояние Дамерау-Левенштейна

Лемматизация

Русский язык – язык с богатой морфологией. *(с) каждая лекция о предобработке*

- Лемматизация для английского необязательно
- ➤ Лемматизация для русского а must (если у вас не очень большой корпус)

Связанная с этим проблема — снятие омонимии (стали, сорока)

Какие инструменты есть

pymorphy2:

- словарь + правила
- не смотрит на контекст, у слова одинаковые разборы и их порядок тоже одинаковый
- «скор» разбора считается по частотности парадигмы, не по словам
- много разных функций и хорошо выстроенная архитектура
- быстрый

Какие инструменты есть

mystem:

- префиксные деревья (trie) + статистика
- учитывает контекст при снятии омонимии
- есть обертка для питона, но не всегда удобно пользоваться + дольше по сравнению с запуском из командной строки
- mystem в целом чудовищно долго работает на Windows

Стоп-слова

Служебные и околослужебные слова не несут семантики и часто только создают шум.

Решение: сделать список таких слов, выкидывать их на этапе предобработки Список может быть разным в зависимости от задачи