Основные задачи NLP

Лектор: Алтухов Никита Александрович
Аналитик данных Сбербанк

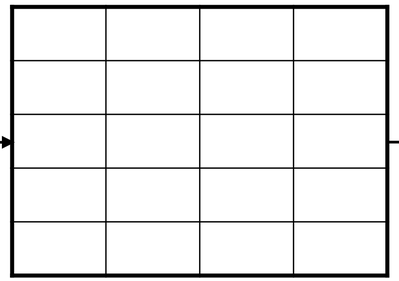
Что такое NLP?

Вход:

Естественный язык

-прив, как дела?
-норм, ахахах

Матрица



Модель

PLSA / LSTM / BERT / ELMO

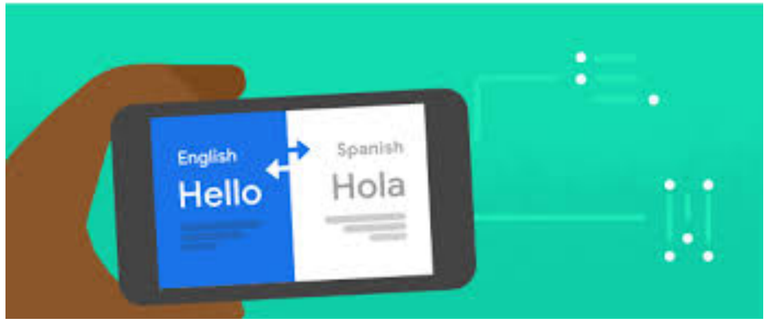
Выход:

Результат

Разговор друзей

Задачи

- Классификация
  - Тематическая классификация длинных текстов
  - Классификация коротких текстов (анализ тональности)
  - Классификация токенов
- Поиск
  - Поиск по запросу (поисковые машины)
  - Поиск похожих текстов (новости)
- Диалоговые системы (чат-боты)
- Машинный перевод (переводчики)
- Эксплоративный анализ (что тут вообще происходит?)



Формальное определение языка

- Язык - множество определенных цепочек символов из алфавита
- Цепочки строятся по определенным правилам
- Текст - одна цепочка
- Алфавит - множество символов, из которых строятся тексты

Особенности естественного языка

- Очень трудно описать формально все правила
- Флективность - изменчивость словоформ

Рам<u>а</u>, рамо<u>й</u>, раму, рам<u>е</u>

Бе<u>гу</u>, бе<u>жал</u>, побе<u>гу</u>

- Омонимия - изменение значения словоформы в зависимости от контекста

Мама <u>мыла</u> раму без <u>мыла</u>

Косил <u>косой</u> <u>косой</u> <u>косой</u>

- Вариативность порядка слов - зависимость смысла предложения от порядка слов в предложении

Мать обрадовала дочь

Дочь обрадовала мать

Токенизация

"Сказал он, не изменяя голоса и тоном, в котором из-за приличия и участия просвечивало равнодушие и даже насмешка"

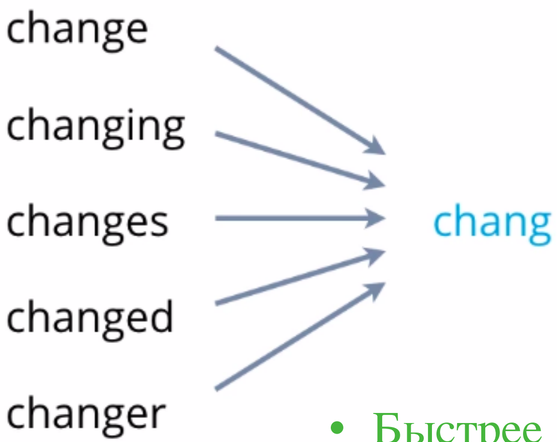
- С помощью регулярных выражений берем последовательности подряд идущих букв + приводим к нижнему регистру

['сказал', 'он', 'не', 'изменяя', 'голоса', 'и', 'тоном', 'в', 'котором', 'из', 'за', 'приличия', 'и', 'участия', 'просвечивало', 'равнодушие', 'и', 'даже', 'насмешка']

- Специализированные билиотеки (nltk, spacy, ...)

['сказал', 'он', ',', 'не', 'изменяя', 'голоса', 'и', 'тоном', ',', 'в', 'котором', 'из-за', 'приличия', 'и', 'участия', 'просвечивало', 'равнодушие', 'и', 'даже', 'насмешка', '.',]

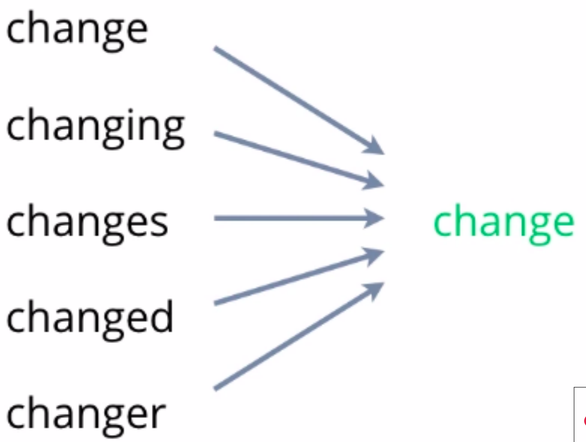
Стемминг и Лемматизация



- Быстрее
- Грубый

Проверь еду у кошек ->

Провер е у кош



- Медленнее
- Более точный

Проверь еду у кошек ->

Проверить еда у кошка

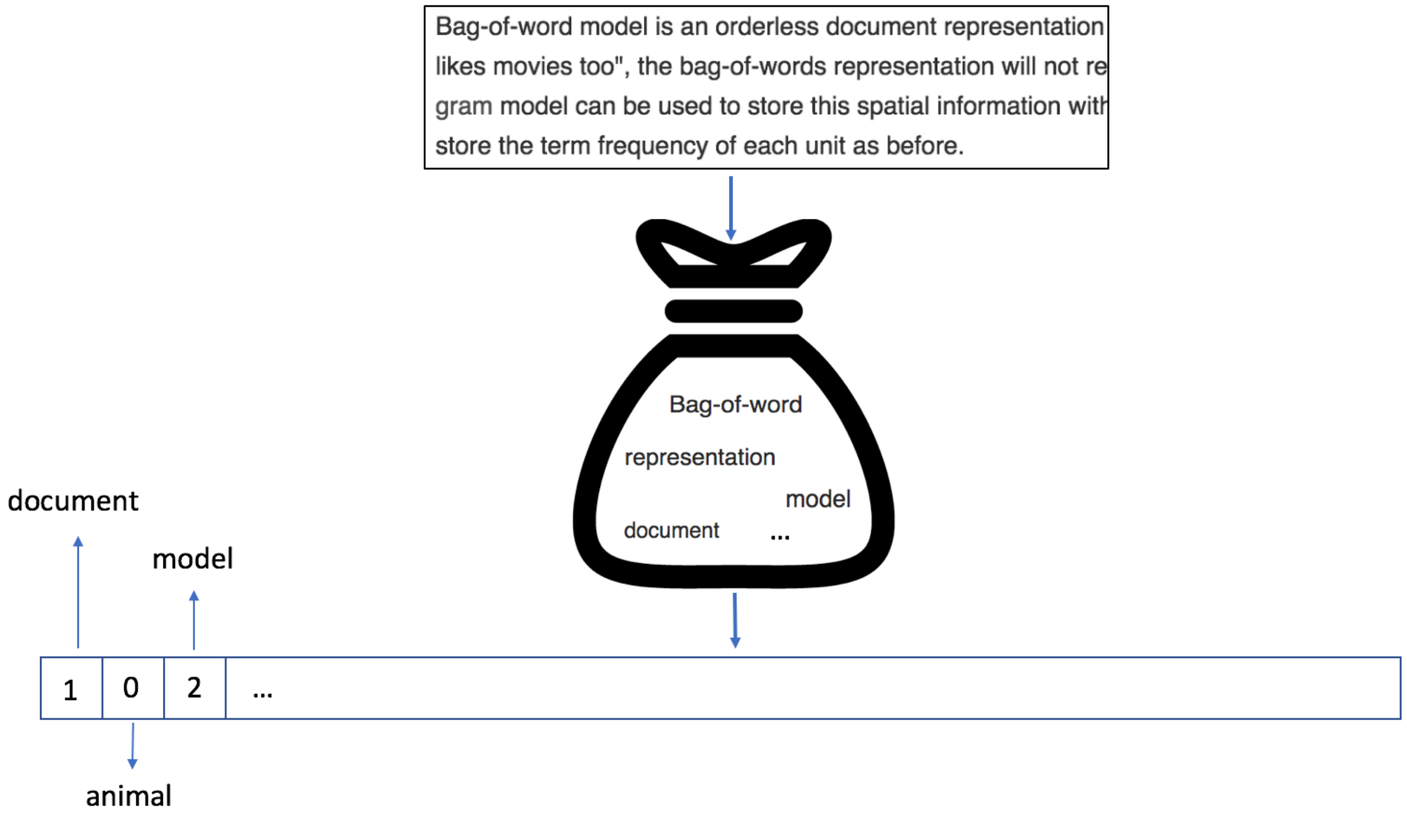
One Hot Encoding

Pet
Cat
Dog
Turtle
Fish
Cat

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

- Не учитываем частоту встречаемости разных слов
- Очень большая размерность вектора

Мешок слов


Bag-of-word model is an orderless document representation
likes movies too", the bag-of-words representation will not re
gram model can be used to store this spatial information with
store the term frequency of each unit as before.
Bag-of-word
representation
model
document
...
document
model
1
0
2
...
animal

TF-IDF

TermFrequency InverseDocumentFrequency

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term *x* within document *y*

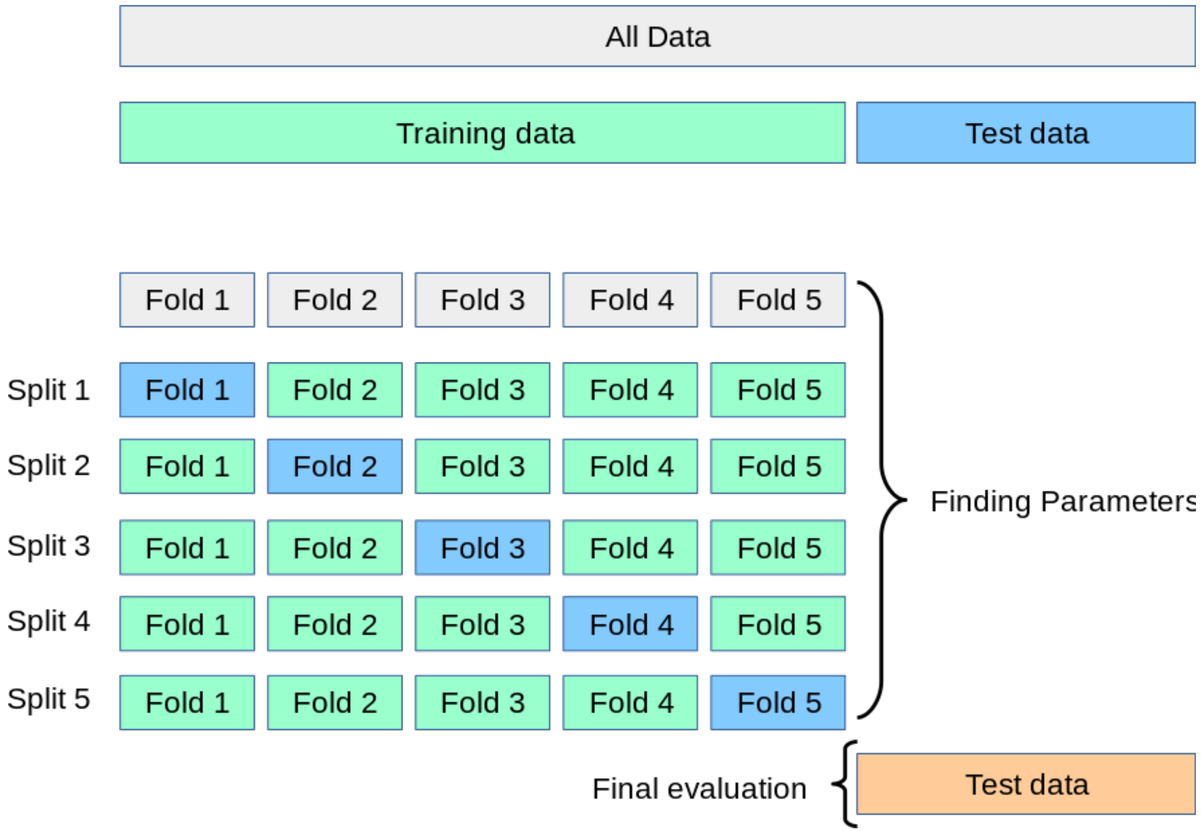
$tf_{x,y}$ = frequency of *x* in *y*

df_x = number of documents containing *x*

N = total number of documents

В чем огромный недостаток?

В чем огромный недостаток?



Строить словарь по всему тексту

||

утечка данных в таргет

Строить словарь по обучающей выборке

||

вероятность наткнуться на незнакомое слово на предикте

- **Использовать чужие большие словари**
- **Использовать в качестве токена не слово, а слог / букву**

Ссылки, источники

- Ссылка на Google Colab с кодом на Python: https://colab.research.google.com/github/nikitosl/spbu-nlp-2020/blob/master/text_preprocessing/war_and_peace.ipynb
- Очень хороший курс по NLP на русском: https://stepik.org/course/54098/syllabus
- Очень понятное введение в DL и NLP: https://dlcourse.ai