



# **Автоматическая обработка текстов на естественном языке. Метод частотного анализа.**

Евгений Борисов

# NLP частотный анализ

## **о языке и задача автоматической его обработки**

обработка текстов на естественном языке (ЕЯ)

natural language processing (NLP)

NLU / natural language understanding

NLG / natural language generation

SP / speech processing (recognition/generation)

# NLP частотный анализ

## **обработка текстов на естественном языке**

NLP/ NLU natural language understanding

- natural entity recognition - распознавание именованных сущностей
- classification intent - классификация намерений
- sentiment analysis - оценка тона

# NLP частотный анализ

## метод частотного анализа

Какие задачи можно решать?

сортировка по заданным темам

определение авторства

определение тона текста

ПОИСК ПОХОЖИХ ТЕКСТОВ

текст должен содержать слова в достаточном количестве

# NLP частотный анализ

## схема системы обработки текстов

подбор текстов для обучения

извлечение признаков из текста

обучение модели ML

тестирование результата

# NLP частотный анализ

## извлечение признаков из текста

определение языка

токенизация

очистка

составление словаря

частотный анализ текстов по словарю

( bag of words, BoW)

# NLP частотный анализ

**извлечение признаков из текста**

**токенизация**

разбиения текста на отдельные слова  
и/или словосочетания

*10кг, АИ-97, к.ф.м.н.*

n-gram - последовательность из n слов

Законодательная дума Хабаровского края (duma.khv.ru)

[ 'Законодательная', 'дума', 'Хабаровского', 'края', '(duma.khv.ru)' ]

# NLP частотный анализ

**извлечение признаков из текста**

**очистка**

способ очистки зависит от задачи



# NLP частотный анализ

**извлечение признаков из текста**

**очистка**

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

# NLP частотный анализ

**извлечение признаков из текста**

## очистка

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

удаление лишних символов (знаки препинания и т.п.)  
«смайлики» - отдельное слово

# NLP частотный анализ

**извлечение признаков из текста**

## очистка

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

удаление лишних символов (знаки препинания и т.п.)  
«смайлики» - отдельное слово

преобразование чисел, интернет ссылок и т.п.

# NLP частотный анализ

**извлечение признаков из текста**

## очистка

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

удаление лишних символов (знаки препинания и т.п.)  
«смайлики» - отдельное слово

преобразование чисел, интернет ссылок и т.п.

лемматизация - приведение слов к нормальному виду

или

стеминг - выделение основ слов

# NLP частотный анализ

## извлечение признаков из текста

### очистка

способ очистки зависит от задачи

удаление стоп-слов (предлоги и т.п.)

удаление лишних символов (знаки препинания и т.п.)  
«смайлики» - отдельное слово

преобразование чисел, интернет ссылок и т.п.

лемматизация - приведение слов к нормальному виду

или

стеминг - выделение основ слов

Законодательная дума Хабаровского края (duma.khv.ru) Состоялось очередное заседание Думы На последнем перед каникулами очередном заседании Законодательной Думы Хабаровского края, состоявшемся 28

```
['законодательн',  
'дум',  
'хабаровск',  
'кра',  
'url',  
'состоя',  
'очередн',  
'заседан',  
'дум',  
'последн',  
'перед',  
'каникул',  
'очередн',  
'заседан',  
'законодательн',  
'дум',  
'хабаровск',  
'кра',  
'состоя',  
'digit',
```

# NLP частотный анализ

**извлечение признаков из текста**  
**составление словаря**

из очищенного текста извлекаем словарь

```
[  
  'digit',  
  'url',  
  'администрац',  
  'большинств',  
  'бурн',  
  'бюджетн',  
  'верхнебуреинск',  
  'власт',  
  'возьмет',  
  'войдет',  
  'вопрос',  
  'врем',  
  'втор',  
  'вызва',  
  'год',  
  ...  
]
```

# NLP частотный анализ

**извлечение признаков из текста**

**частотный анализ текстов по словарю**

простой частотный анализ  
считаем в тексте  $t$  количество повторов  $x_i$   
каждого слова  $v_i$  из словаря  $V$

текст должен содержать слова в достаточном количестве

# NLP частотный анализ

**извлечение признаков из текста**

## **частотный анализ текстов по словарю**

простой частотный анализ  
считаем в тексте  $t$  количество повторов  $x_i$   
каждого слова  $v_i$  из словаря  $V$

значения  $x$  зависят от размера текста  $t$ ,  
чем больше текст тем больше повторов

нормализованны частотный анализ (TF, term frequency)  
значения частоты  $x$  делятся на общее число слов в тексте  $t$ .

$$TF(t, V) = x(t, V) / \text{size}(t)$$



# NLP частотный анализ

**извлечение признаков из текста**  
**частотный анализ текстов по словарю**

Удалять часто употребляемые слова или нет?

# NLP частотный анализ

**извлечение признаков из текста**  
**частотный анализ текстов по словарю**

Удалять часто употребляемые слова или нет?

TF-IDF - компромиссный вариант формирования вектор-признаков.

не выбрасывает часто употребляемые слова из словаря  
но уменьшает их вес в вектор-признаке

# NLP частотный анализ

## **извлечение признаков из текста** **частотный анализ текстов по словарю**

Удалять часто употребляемые слова или нет?

TF-IDF - компромиссный вариант формирования вектор-признаков.

не выбрасывает часто употребляемые слова из словаря  
но уменьшает их вес в вектор-признаке

коэффициент обратной частоты (IDF, inverse document frequency)  
чем чаще встречается слово тем меньше значение его IDF

$$IDF(v) = \log \text{size}(T) / \text{size}(T(v))$$

количество текстов  $T$   
разделить на  
количество текстов  $T$  содержащих слово  $v$

$$TF-IDF(t, T, v) = TF(t, v) * IDF(v, T)$$

# NLP частотный анализ

**извлечение признаков из текста**  
**частотный анализ текстов по словарю**

хэш-векторизация

заменяем слова на их хэш ограниченной длины

сокращаем размер словаря  
и число признаков

экономия ресурсов для больших датасетов

# NLP частотный анализ

## практическое применение

**сортировка по заданным темам** - классификация  
собираем и размечаем тексты  
чистим текст  
применяем частотный анализ  
обучаем классификатор  
тестируем

# NLP частотный анализ

## практическое применение

### **сортировка по заданным темам** - классификация

собираем и размечаем тексты

чистим текст

применяем частотный анализ

обучаем классификатор

тестируем

### **определение авторства** - классификация

собираем и размечаем тексты

чистим текст (частота употребления предлогов - важный признак)

применяем частотный анализ

обучаем классификатор

тестируем

# NLP частотный анализ

## практическое применение

### **сортировка по заданным темам** - классификация

собираем и размечаем тексты

чистим текст

применяем частотный анализ

обучаем классификатор

тестируем

### **определение авторства** - классификация

собираем и размечаем тексты

чистим текст (частота употребления предлогов - важный признак)

применяем частотный анализ

обучаем классификатор

тестируем

### **поиск похожих текстов** - кластеризация

собираем тексты

чистим текст

применяем частотный анализ

выполняем кластеризацию (размечаем тексты)

# NLP частотный анализ

## Литература

`git clone https://github.com/mechanoid5/ml\_nlp.git`

К.В. Воронцов Вероятностные тематические модели коллекций текстовых документов.

Евгений Борисов Автоматизированная обработка текстов на естественном языке, с использованием инструментов языка Python  
<http://mechanoid.su/ml-text-proc.html>

Sebastian Raschka Python Machine Learning - Packt Publishing Ltd, 2015