



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

Евгений Борисов

# NLP частотный анализ

## Уровни сложности при автоматической обработке текстов

**Прагматика** - отдельные слова и устойчивые словосочетания

**Семантика** - смыслы последовательностей слов

**Синтаксис** - правила формирования последовательностей слов

**Морфология** - отдельные слова и устойчивые словосочетания

# 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

## **Общая схема системы обработки текстов**

1. подбор текстов для обучения модели
2. извлечение признаков из текста
3. обучение модели ML
4. тестирование результата

# NLP частотный анализ

## Общая схема системы обработки текстов

1. подбор текстов для обучения модели

**2. извлечение признаков из текста**

3. обучение модели ML

4. тестирование результата

## BoW (bag of words) - извлечение признаков из текста

1.определение языка

2.токенизация

3.очистка

4.составление словаря

5.частотный анализ текстов по словарю

# 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

## очистка текста

*способ очистки зависит от задачи*

- удаление лишних символов (знаки препинания и т. п.)
- удаление стоп-слов (предлоги и т.п.)
- преобразование чисел, интернет ссылок и т.п.
- лемматизация - приведение слов к нормальному виду
- стеминг - выделение основ слов
- ограничение по частоте (min, max)

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',  
'состоя',  
'очередн',  
'заседан',  
'дум',  
'последн',  
'перед',  
'каникул',  
'очередн',  
'заседан',  
'законодательн',  
'дум',  
'хабаровск',  
'кра',  
'состоя',

# NLP частотный анализ

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

из очищенного текста извлекаем словарь

```
[  
  '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  $V$

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

**Решение:** нормализованный 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