# Лекция 14: Автоматическая обработка текстов на естественном языке. Метод частотного анализа.

Евгений Борисов

обработка текстов на естественом языке, natural language processing (NLP) с помощью <u>частотного анализа</u>

сортировка по заданным темам определение авторства поиск похожих текстов

текст должен содержать слова в достаточном количестве

### система обработки текстов

подбор текстов для обучения извлечение признаков из текста обучение модели ML тестирование результата

### извлечение признаков из текста

токенизация

очистка

составление словаря

частотный анализ текстов по словарю

(bag of words, BoW)

### извлечение признаков из текста токенизация

разбиения текста на отдельные слова и/или словосочетания

n-gram - последовательность из n слов

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

### извлечение признаков из текста <u>очистка</u>

зависит от задачи

лемматизация - приведение слов к нормальному виду <u>или</u> стеминг - выделение основ слов

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.

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',
```

### извлечение признаков из текста составление словаря

из очищенного текста извлекаем словарь

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

### извлечение признаков из текста <u>частотный анализ текстов по словарю</u>

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

### 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 х делятся на общее число слов в тексте t.

$$TF(t,V) = x(t,V) / size(t)$$

извлечение признаков из текста частотный анализ текстов по словарю

Удалять часто употребляемые слова или нет?

### 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

### извлечение признаков из текста <u>частотный анализ текстов по словарю</u>

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

$$IDF(v) = log size(T) / size(T(v))$$

количество текстов Т разделить на количество текстов Т содержащих слово v

$$\mathsf{TF}\mathsf{-}\mathsf{IDF}(\mathsf{t},\mathsf{T},\mathsf{v}) = \mathsf{TF}(\mathsf{t},\mathsf{v}) * \mathsf{IDF}(\mathsf{v},\mathsf{T})$$

### 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

#### практическое применение

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

#### 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

#### 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)

### Литература

git clone <a href="https://github.com/mechanoid5/ml\_lectorium.git">https://github.com/mechanoid5/ml\_lectorium.git</a>

Евгений Борисов Автоматизированная обработка текстов на естественном языке, с использованием инструментов языка <a href="Python http://mechanoid.kiev.ua/ml-text-proc.html">Python http://mechanoid.kiev.ua/ml-text-proc.html</a>

Sebastian Raschka Python Machine Learning - Packt Publishing Ltd, 2015



Вопросы?

# NLP частотный анализ : практика





sklearn.datasets UCI Repository kaggle www.ruscorpora.ru



#### задание

- применить TF-IDF для кластеризации
- применить HashingVectorizer для кластеризации
- применить методы на других наборах текстов