Машинное обучение

Лекция 9

Категориальные признаки. Анализ текстов.

Михаил Гущин

mhushchyn@hse.ru

НИУ ВШЭ, 2024

Категориальные признаки

Задача

StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	One year	No	Mailed check	56.95	1889.50	No
No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

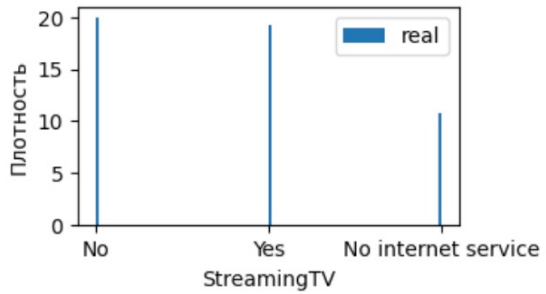
- Прогноз оттока клиентов в телекоме
- Задача классификации
- Есть категориальные признаки

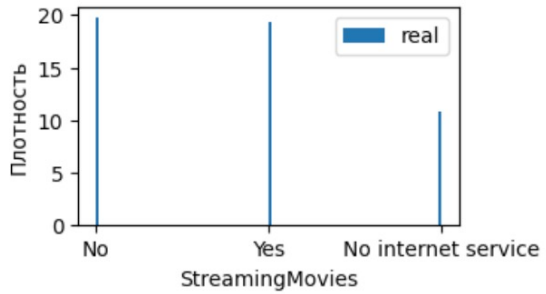
Данные: *https://www.kaggle.com/datasets/blastchar/telco-customer-churn*

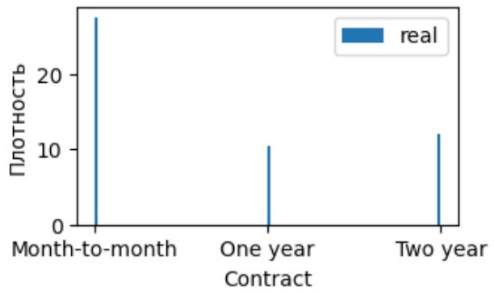
Решение

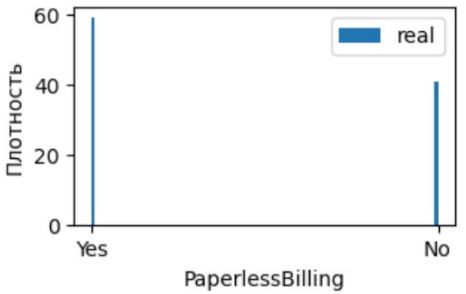
- Как бы вы решали эту задачу?
- Как предобработать данные?
- Что делать с нечисловыми признаками?
- Какие модели вы бы использовали?

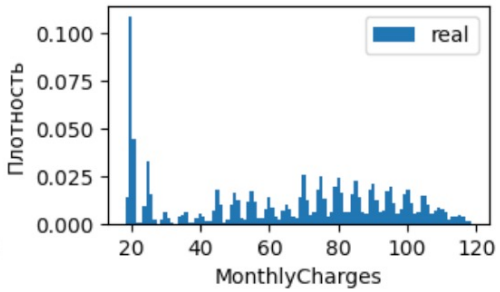
Распределение данных

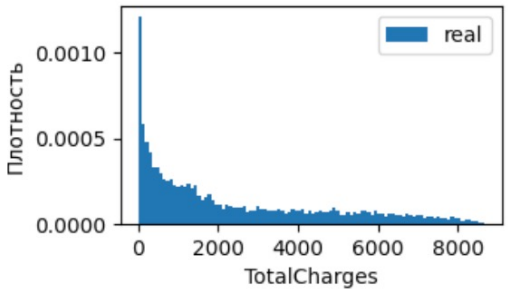


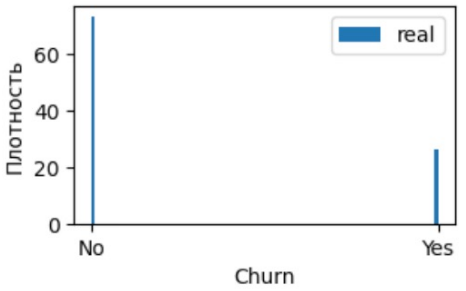












Данные: *https://www.kaggle.com/datasets/blastchar/telco-customer-churn*

Категориальные признаки

- Категориальные признаки (categorical features) – признаки, которые принимают дискретные значения. Количество значений ограничено.
- Примеры:
  - ['male', 'female']
  - ['France', 'USA', 'Canada', 'China', 'Israel', 'Japan', 'Russia']
  - [1, 5, 8, 3, 2, 7, 9]

- Как распознать: каждому значению можем сопоставить целое число.
- Например:
  - ['Europe', 'Asia', 'Europe', 'South America'] -> [0, 1, 0, 2]

Кодирование признаков

Популярные алгоритмы кодирования категориальных признаков:

- Original Encoding (Label Encoding)
- One-Hot Encoding
- Binary Encoding
- Target Encoding

Original Encoding (Label Encoding)

- Каждой категории ставим в соответствие целое число.
- Например:
  - {A: 0, B: 1, C: 2}

- Заменяем категориальные значения признака этими числами.
- Например:
  - [A, C, B, C, A] -> [0, 2, 1, 2, 0]

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

One-Hot Encoding

- Для каждой категории создаем отдельную колонку в данных.
- Например:
  - Team_A, Team_B, Team_C

- Каждому значению категориального признака ставим 1 в его колонку, а в остальные колонки – 0.
- Например:
  - A -> [1, 0, 0]
  - B -> [0, 1, 0]
  - C -> [0, 0, 1]

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Binary Encoding

- Каждой категории ставим в соответствие натуральное число.
- Например:
  - {A: 1, B: 2, C: 3}

- Переводим натуральные числа в двоичную систему:
- Например:
  - 1 –> 001; 2 -> 010; 3 -> 011

- Каждому разряду двоичного числа делаем новую колонку в данных. Заменяем категории соответствующими 0 и 1 в колонках.
- Пример:
  - B -> [0, 1, 0]

Binary Encoding

Temperature
Hot
Cold
Very Hot
Warm
Hot
Warm
Warm
Hot
Hot
Cold

Order
1
2
3
4
1
4
4
1
1
2

Binary
001
010
011
100
001
100
100
001
001
010

Temperature_0	Temperature_1	Temperature_2
0	0	1
0	1	0
0	1	1
1	0	0
0	0	1
1	0	0
1	0	0
0	0	1
0	0	1
0	1	0

Источник: https://machinelearningmastery.ru/all-about-categorical-variable-encoding-305f3361fd02/

Target Encoding

workclass	target
State-gov	0
Self-emp-not-inc	1
Private	0
Private	0
Private	1

workclass	target mean
State-gov	0
Self-emp-not-inc	1
Private	1/3

workclass
0
1
1/3
1/3
1/3

- Для каждой категории считаем среднее значение таргета
- Заменяем значения категорий этими средними значениями

Алгоритм решения

- Пусть есть данные с категориальными признаками
- Кодируем все колонки с категориями в числовые признаки
- Получаем таблицу с числовыми признаками
- Используем любые модели классификации и регрессии для решения задачи

- Многие алгоритмы на решающих деревьях (CatBoost) умеют работать с категориями без кодирования ☺

Анализ текстов

Задача классификации текстов

"text": "Двое налетчиков совершили нападение на охранника банка \"ЦентрКредит\" в Алматы и завладели его оружием, сообщает пресс-служба ДВД Алматы.\n\"Сегодня, в 08.10 в Центр оперативного управления ДВД города Алматы поступило сообщение о нападении на охранника одного из отделений банка \"Центркредит\" в Алмалинском районе Алматы, в результате чего двое неустановленных лиц завладели его оружием\", — говорится в сообщении.\nПо неподтвержденным данным, чрезвычайное происшествие случилось в отделении, которое находится на улице Маметовой, между улицами Сейфуллина и Дзержинского.\nКак информирует ведомство, в настоящее время по городу объявлен спецплан \"Сирена\". Отрабатывается комплекс оперативных мероприятий, которые направлены на розыск и задержание преступников.\n",

"sentiment": "negative»

"text": "АСТАНА. КАЗИНФОРМ - Карим Масимов провел совещание по вопросам деятельности Актауского международного морского торгового порта. Read on the original site", "id": 2085,

"sentiment": "positive"

Решение

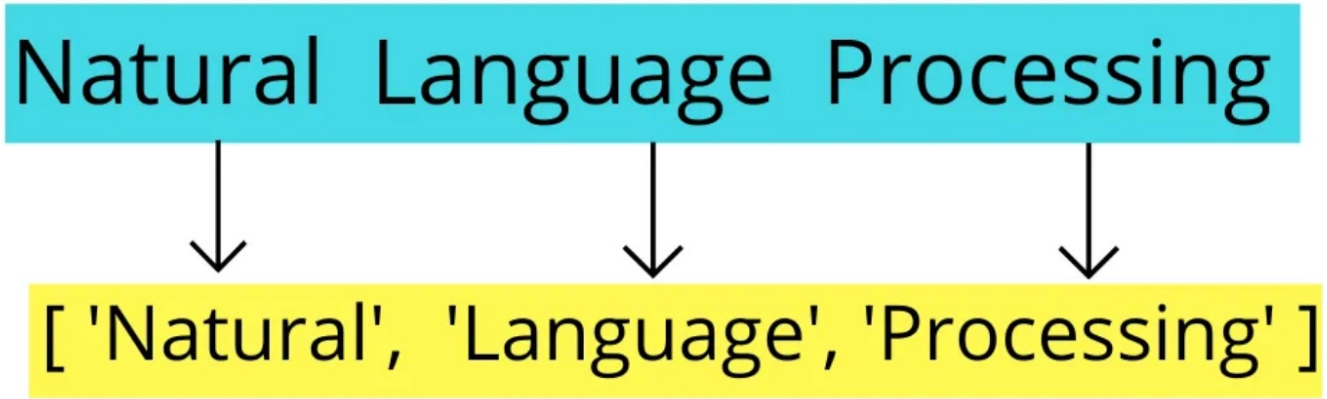
- Как бы вы решали эту задачу?
- Как предобработать текст?
- Что представить текст числовым вектором признаков?
- Какие модели вы бы использовали?

Основные этапы обработки текстов

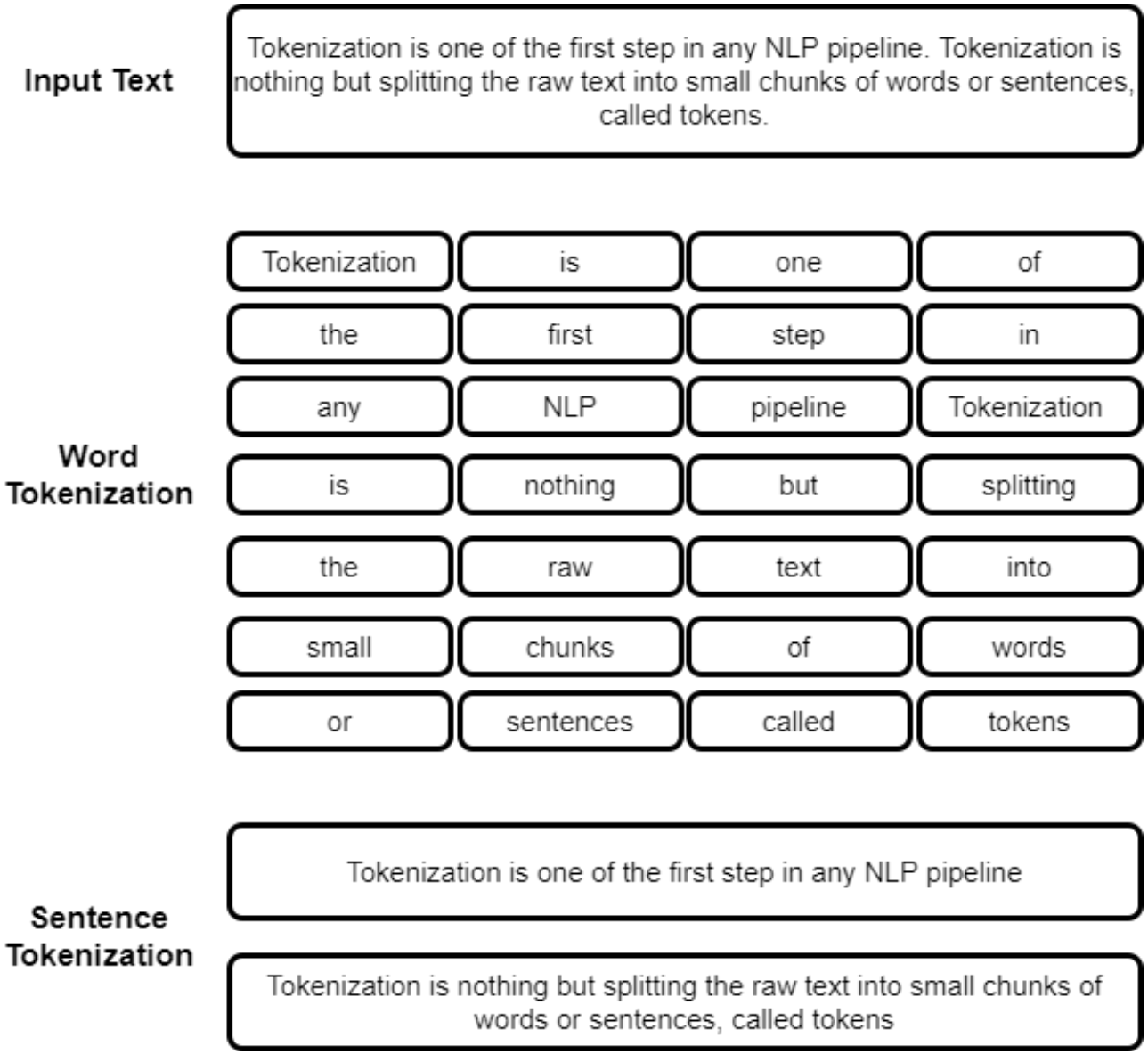
- Токенизация (tokenization)
- Нормализация текстов (text normalization)
  - Лемматизация (lemmatization)
  - Стемминг (Stemming)
- Мешок слов (Bag Of Words (BOW))
- TF-IDF (Term Frequency – Inversed Document Frequency)

Токенизация

- Токенизация (tokenization) – это разделение текста на отдельные токены (части). Примеры токенов:
  - Отдельные слова,
  - Последовательности из N слов (N-граммы),
  - Отдельные предложения.



Токенизация



Источник: https://www.kaggle.com/code/satishgunjal/tokenization-in-nlp

Лемматизация

- Лемматизация – это приведение слова в его нормальную (словарную) форму.

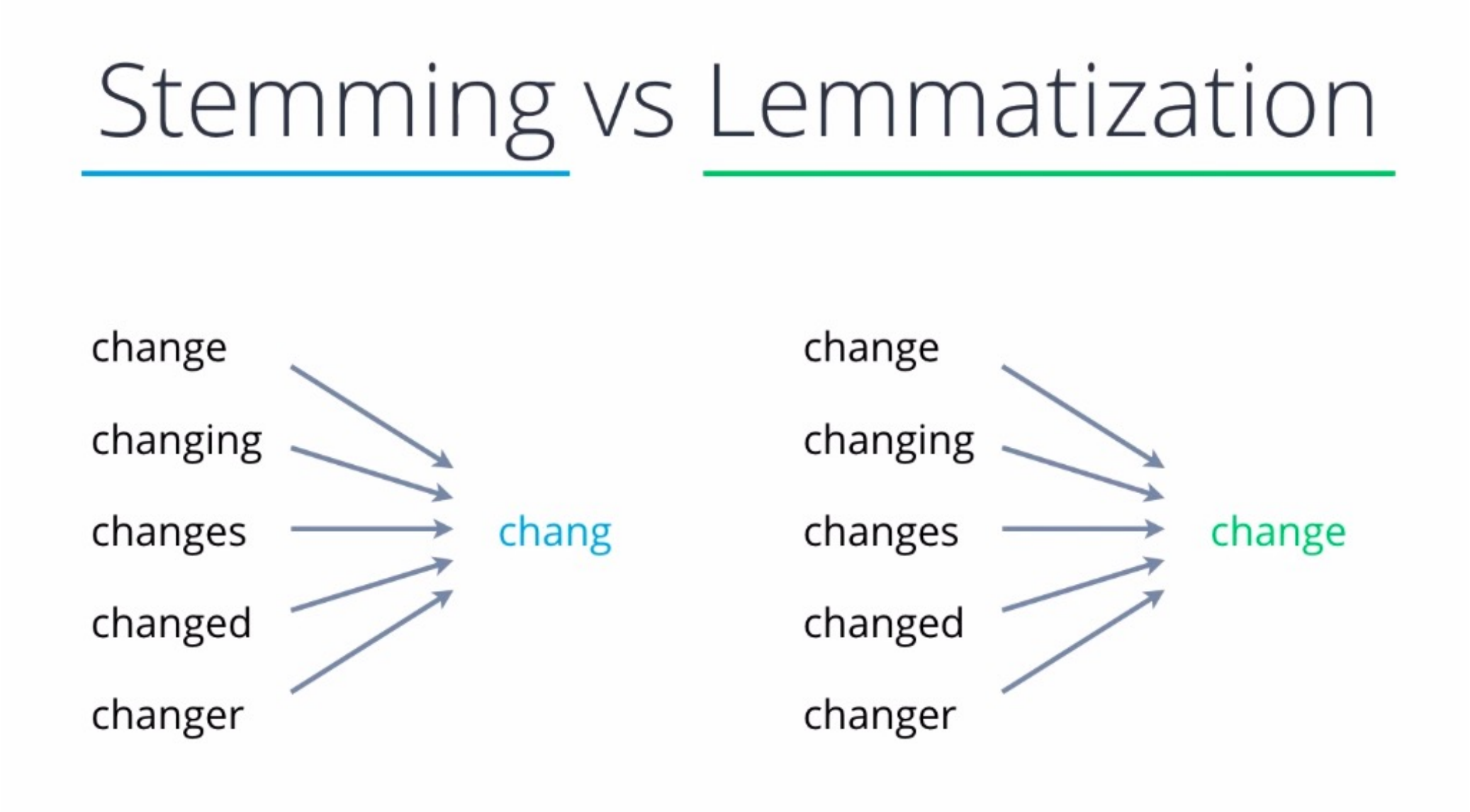
- Примеры:
  - для существительных — именительный падеж, единственное число;
  - для прилагательных — именительный падеж, единственное число, мужской род;
  - для глаголов, причастий, деепричастий — глагол в инфинитиве (неопределённой форме) несовершенного вида.

Стемминг

- Стемминг – это нахождение основы слова (стеммы), которая определяет его лексическое значение.

- Одна стемма соответствует разным формам слова
- Уменьшает число уникальных частей, на которые можем разделить исходный текст

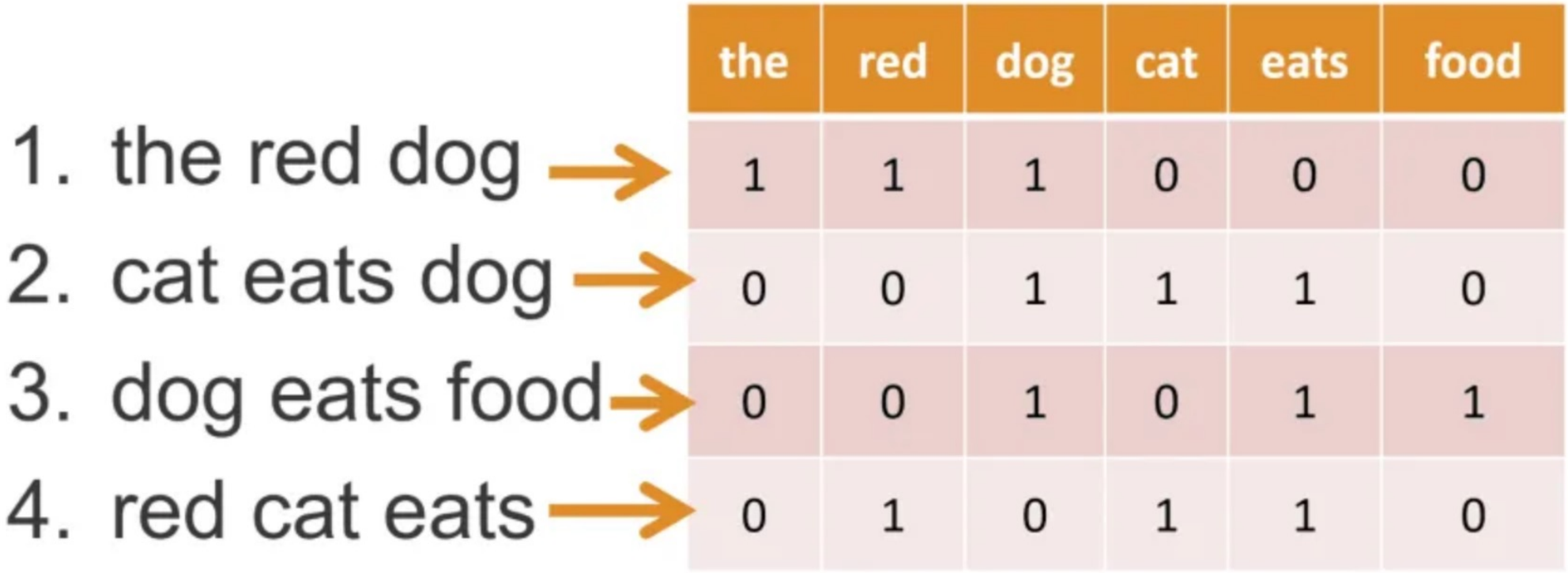
Пример


Stemming vs Lemmatization
change
changing
changes
changed
changer
chang
change
changing
changes
changed
changer
change

Мешок слов (Bag of Words)

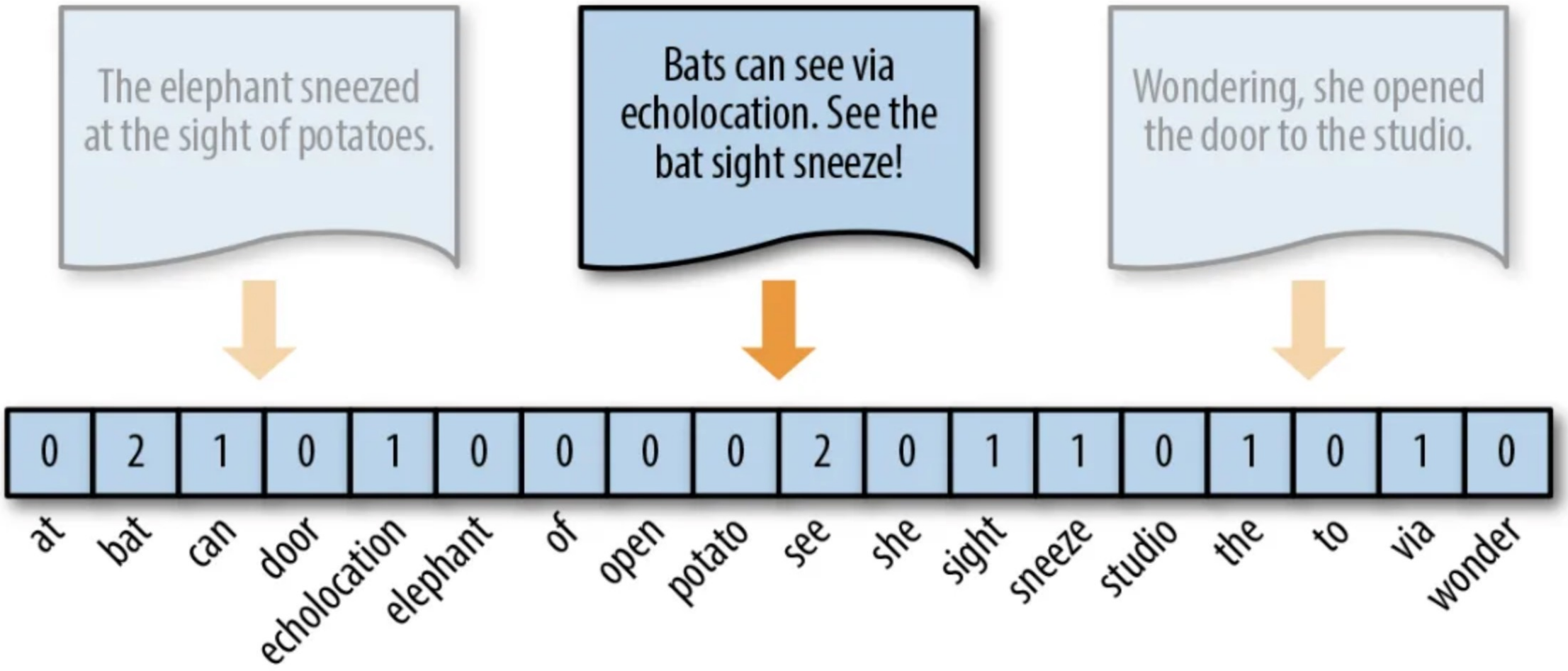
- Берем текст
- Делим его на слова (токенизация)
- Каждое слово приводим к его начальной форме (нормализация)
- Убираем лишние слова и символы (очень частые слова, знаки препинания)
- Оставшиеся слова образуют словарь слов
- Для каждого слова в словаре создаем свою колонку в данных
- Для каждого слова в тексте добавляем 1 в соответствующую колонку
- Получаем векторное представление текста

Мешок слов (Bag of Words)



	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

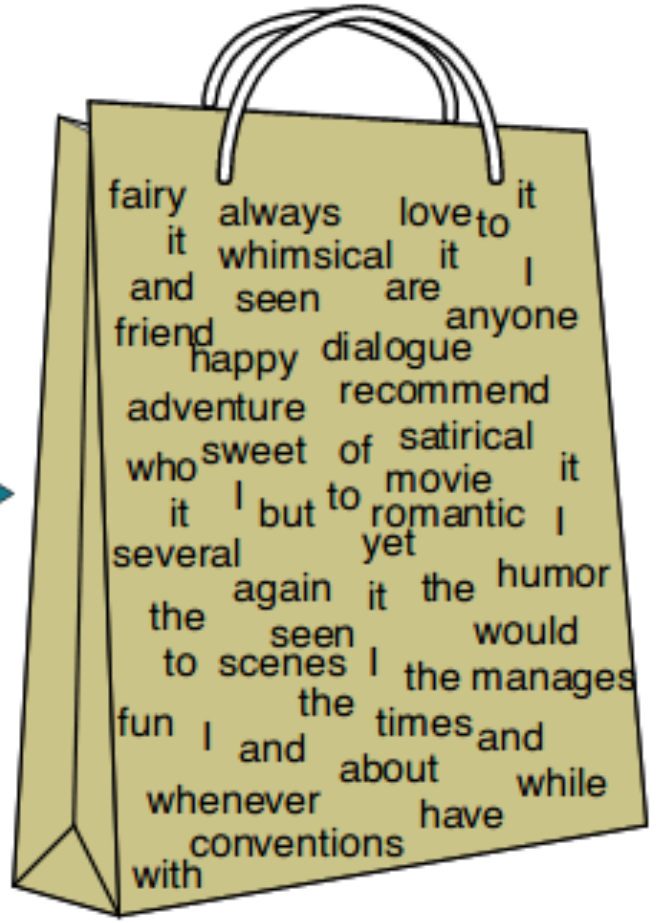
Мешок слов (Bag of Words)



Источник: https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-1-7ed0c7f3dfc5

Мешок слов (Bag of Words)

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Источник: https://dudeperf3ct.github.io/lstm/gru/nlp/2019/01/28/Force-of-LSTM-and-GRU/

TF-IDF

- TF-IDF (term frequency – inversed document frequency)

- Каждому слову (токену) в тексте ставим вес $w_{x,y}$:

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

  - $tf_{x,y}$ - частота слова (токена) x в тексте (документе) y
  - df_x - число текстов (документов), которые содержат слово x
  - N – общее число текстов (документов) в данных

TF-IDF

Term	Review 1	Review 2	Review 3	TF (Review 1)	TF (Review 2)	TF (Review 3)
This	1	1	1	1/7	1/8	1/6
movie	1	1	1	1/7	1/8	1/6
is	1	2	1	1/7	1/4	1/6
very	1	0	0	1/7	0	0
scary	1	1	0	1/7	1/8	0
and	1	1	1	1/7	1/8	1/6
long	1	0	0	1/7	0	0
not	0	1	0	0	1/8	0
slow	0	1	0	0	1/8	0
spooky	0	0	1	0	0	1/6
good	0	0	1	0	0	1/6

TF-IDF

Term	Review 1	Review 2	Review 3	IDF	TF-IDF (Review 1)	TF-IDF (Review 2)	TF-IDF (Review 3)
This	1	1	1	0.00	0.000	0.000	0.000
movie	1	1	1	0.00	0.000	0.000	0.000
is	1	2	1	0.00	0.000	0.000	0.000
very	1	0	0	0.48	0.068	0.000	0.000
scary	1	1	0	0.18	0.025	0.022	0.000
and	1	1	1	0.00	0.000	0.000	0.000
long	1	0	0	0.48	0.068	0.000	0.000
not	0	1	0	0.48	0.000	0.060	0.000
slow	0	1	0	0.48	0.000	0.060	0.000
spooky	0	0	1	0.48	0.000	0.000	0.080
good	0	0	1	0.48	0.000	0.000	0.080

Алгоритм решения

- Пусть есть набор текстов
- Делим каждый текст на токены и приводим слова к нормальной форме
- Используем Bag Of Words или TF-IDF, чтобы получить векторное представление каждого текста
- Используем эти вектора как вектора признаков
- Используем любые модели классификации и регрессии для анализа текстов