# Кластеризация текстов и изображений

# Сегодня

Эльбрус буткемп

- как представить текст так, чтобы передать его в модель
- базовые алгоритмы обработки языка
- решение задачи кластеризации текста
- пример с класстеризацией изображений

# Предобработка: пословная токенизация

Эльбрус буткемп

Backgammon is one of the oldest known board games. Its history can be traced back nearly 5,000 years to archeological discoveries in the Middle East. It is a two player game where each player has fifteen checkers which move between twenty-four points according to the roll of two dice.

# Предобработка: пословная токенизация

Эльбрус буткемп

'Backgammon', 'is', 'one', 'of', 'the', 'oldest', 'known', 'board', 'games', '.', 'Its', 'history', 'can', 'be', 'traced', 'back', 'nearly', '5,000', 'years', 'to', 'archeological', 'discoveries', 'in', 'the', 'Middle', 'East', '.', 'It', 'is', 'a', 'two', 'player', 'game', 'where', 'each', 'player', 'has', 'fifteen', 'checkers', 'which', 'move', 'between', 'twenty-four', 'points', 'according', 'to', 'the', 'roll', 'of', 'two', 'dice', '.'

# Предобработка: стемминг

Эльбрус буткемп

**Стемминг** (stemming) – это грубый эвристический процесс, который отрезает «лишнее» от корня слов, часто это приводит к потере словообразовательных суффиксов.

- dog**s**, dog **'s**, dog**s'** $\longrightarrow$ dog
- хорош**ий**, хорош**ая** $\longrightarrow$ хорош

# Предобработка: лемматизация

Эльбрус буткемп

Лемматизация (lemmatization) – процесс, который использует словарь и морфологический анализ, чтобы в итоге привести слово к его канонической форме – лемме.

- drove $\rightarrow$ drive; seen $\rightarrow$ see
- хотеть, хочу, хотел $\rightarrow$ хотеть

# Предобработка: стопслова

Эльбрус буткемп

- 'и', 'в', 'во', 'не', 'что', 'он', 'на', 'я', 'с', 'со', 'как', 'а', 'то', 'все', 'чтоб', 'без', 'будто', 'впрочем', 'хорошо', 'свою', 'этой', 'перед', 'иногда', 'лучше', 'чуть', 'том', 'нельзя', 'такой', 'им', 'более', 'всегда', 'конечно', 'всю', 'между ...
- 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves'

Сильно зависят от языка, в разных библиотеках может отличаться

- до: **По асфальту мимо цемента, Избегая зевак под аплодисменты. Обитатели спальных аррондисманов**
- после: **асфальт мимо цемент избегать зевака аплодисменты обитатель спальный аррондисман**

# Мешок слов / Bag of words / BoW

Эльбрус буткемп

0. I like this movie, it's funny.
1. I hate hate this movie movie.
2. This was awesome! I like it.
3. Nice one. I love love love it.

| | awesome | funny | hate | it | like | love | movie | nice | one | this | was |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 |
| 2 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 1 | 1 | 1 | 0 |

# N-граммы / N-grams

Эльбрус буткемп

Можно использовать не отдельные слова, а сочетания из N слов. Это позволяет учитывать контекст, но увеличивает признаковое пространство

Оригинал:

> The office building is open today

- Юниграммы

> The, office, building, is, open, today

- Биграммы

> The office, office building, building is, is open, open today

- Триграммы...

# Tf-IDF

Эльбрус буткемп

$$W_{x,D} = \mathrm{tf}_{x,D} \times \log \frac{N}{\mathrm{df}_x}$$

- $x$ – слово
- $D$ – документ
- $N$ – общее число документов
- $\mathrm{df}_x$ – число документов, содержащих $x$
- $\mathrm{tf}_{x,D}$ – число появлений слова $x$ в $D$ / общее число слов в $D$

# Tf-IDF

Эльбрус буткемп

| Index | Document | tf-idf( `cat` ) |
|---|---|---|
| D1 | Simple example with Cat and Mouse | $W(cat, D_1) = 1/6 \times \log(4/3) = 0.06$ |
| D2 | Another simple example with dogs and cat | $W(cat, D_2) = 1/7 \times \log(4/3) = 0.04$ |
| D3 | Funny pigs | $W(cat, D_3) = 0/2 \times \log(4/3) = 0$ |
| D4 | Cat | $W(cat, D_4) = 1/1 \times \log(4/3) = 0.28$ |

# Итоги

Эльбрус буткемп

- обработка естесственного языка – развивающаяся область
- одна из основных задач – представить текст в виде чисел
  - классические подходы: bag of words, Tf-IDF
  - нейросетевые подходы: word2vec, fasttext - о них позже
  - современные подходы: о них позже
- обработанный текст – входные данные для модели

# Дополнительно

Эльбрус буткемп

- Чудесный мир Word Embeddings: какие они бывают и зачем нужны?
- A simple Word2vec tutorial