



СИБИРСКИЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

SIBERIAN
FEDERAL
UNIVERSITY

Технологии автоматической обработки и анализа текстовой информации (NLP)

Аксаментова Елена, Аркадьева Ксения, Идрисова
Самира, Карасова Вероника
ГФ25-01Б



Технологии автоматической обработки и анализа текстовой информации (NLP) (Natural Language Processing) или обработка естественного языка – это методы и инструменты, позволяющие компьютерам понимать, интерпретировать и извлекать смысл из текстовых данных.

Прежде чем компьютер сможет «понять» и проанализировать текст, его необходимо преобразовать в машиночитаемый формат. Этот процесс лежит в основе всех NLP-технологий





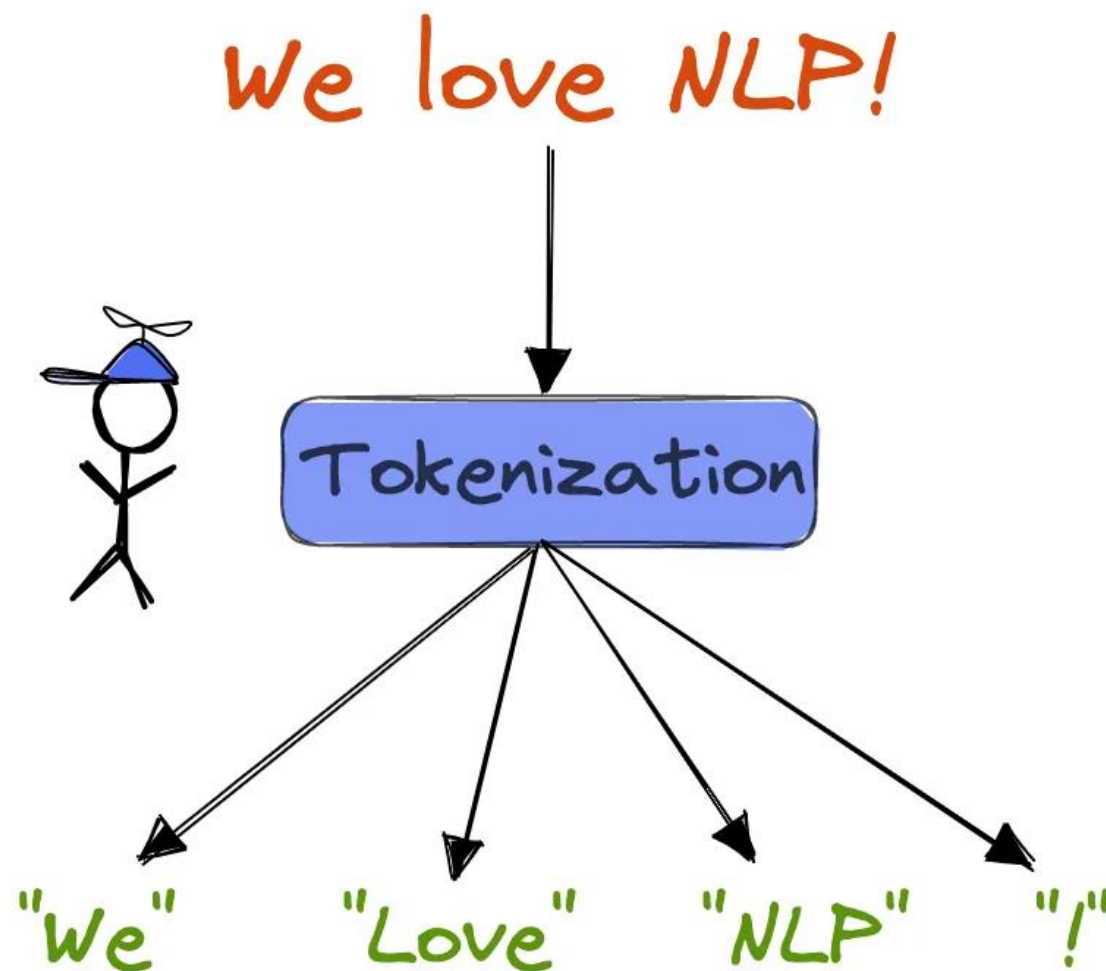
Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- Токенизация

Разбиение на отдельные слова
(токены)

Пример: «Книга лежит на столе» ->
[«Книга», «лежит», «на», «столу»]





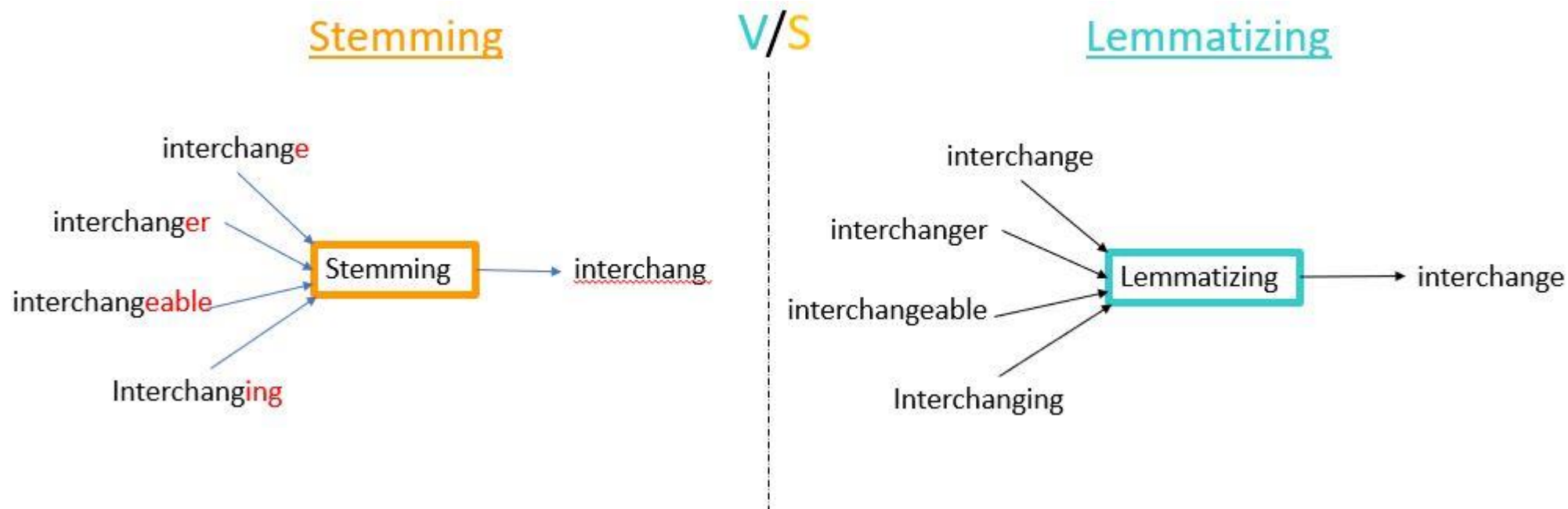
Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- Стемминг/Лемматизация

Приведение слов к корневой форме.
Стемминг удаляет окончания.
Лемматизация использует морфологический анализ

Пример: Стемминг: "бегущий" -> "бег",
Лемматизация: "бегущий" -> "бежать"





Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- Векторное представление текста

Преобразование текста в числовой вектор

1. Bag of words (BoW):

учитывает частоту слов, не учитывая порядок

Пример: «кот спит», «собака спит», -> [«кот»: 1, «спит»: 2, «собака»: 1]

	MARY	IS	HUNGRY	HAPPY	FOR	APPLES	NOT	JOHN	HE	
“Mary is hungry for apples.”	1	1	1	0	1	1	0	0	0	→ [1, 1, 1, 0, 1, 1, 0, 0, 0]
“John is happy he is not hungry for apples.”	0	2	1	1	1	1	1	1	1	→ [0, 2, 1, 1, 1, 1, 1, 1, 1]



Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- Векторное представление текста

Преобразование текста в числовой вектор

2. TF-IDF:(Term Frequency-Inverse Document Frequency)

оценивает важность слова с учётом его частоты в документе
и редкости в коллекции

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

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

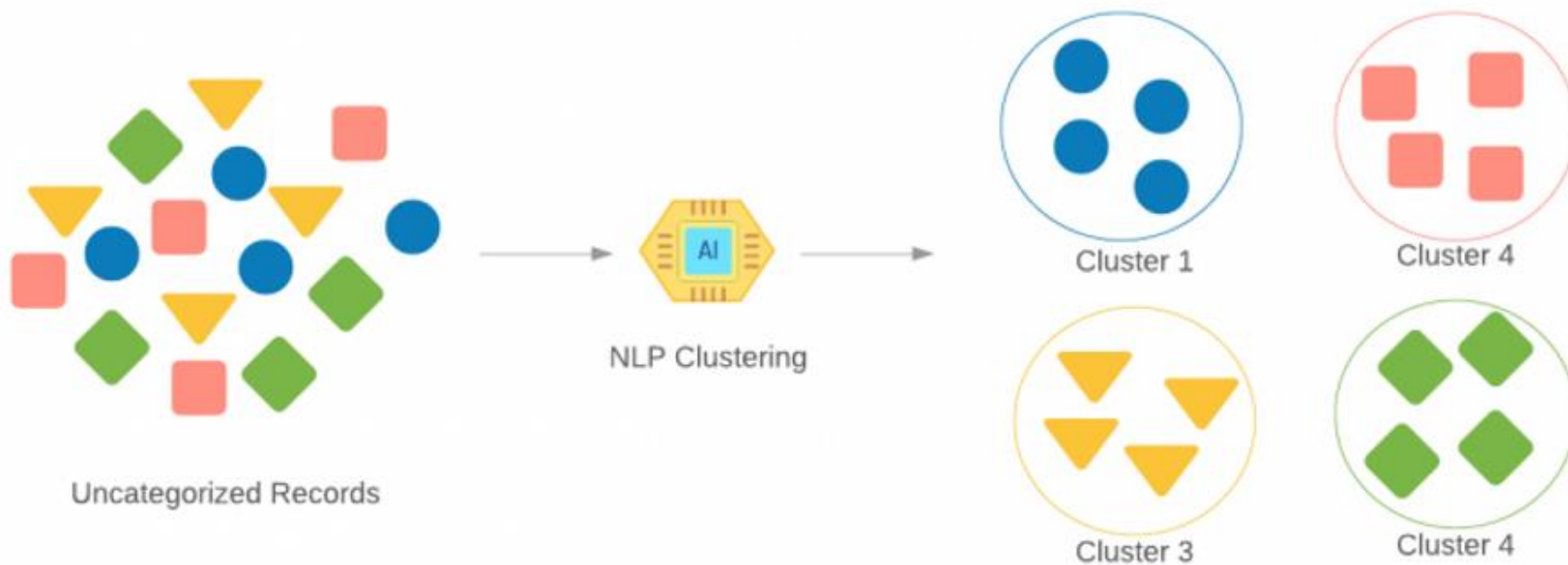


Традиционные методы

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Кластеризация**

Разбиение документов на тематические группы





Ключевые задачи и технологии

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Классификация текстов:** Отнесение текста к категории.
Пример: Определение, является ли электронное письмо спамом.
- **Извлечение именованных сущностей:** Выявление и классификация сущностей (люди, организации, места).
Пример: "Джон Смит работает в Google" -> Имя: Джон Смит, Организация: Google.
- **Определение тональности:** Оценка эмоциональной окраски текста: позитивная, негативная, нейтральная.
Пример: "Мне понравился этот фильм!" -> Позитивный.
- **Машинный перевод:** Автоматический перевод с одного языка на другой.
- **Суммаризация текста:** Создание краткого содержания.
- **Вопросно-ответные системы:** Поиск ответов на вопросы в тексте.
- **Диалоговые системы (чат-боты):** Поддержка диалога и выполнение команд.



Методы машинного обучения

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Наивный Байесовский классификатор** – простой и быстрый алгоритм классификации текстов, основанный на теореме Байеса. Этот метод широко используется для фильтрации спама, определения языка текста и категоризации новостей.
- **Метод опорных векторов (SVM)** – алгоритм для решения задач классификации и регрессии. SVM особенно эффективен при работе с большим количеством признаков, что делает его подходящим для анализа текстовых данных.
- **Решающие деревья и случайный лес (Random Forest)** – алгоритмы классификации и регрессии, основанные на построении иерархической структуры решений. Решающие деревья хорошо интерпретируемы и могут использоваться для извлечения правил из текстовых данных, в то время как случайный лес обеспечивает более высокую точность за счет ансамбля деревьев.

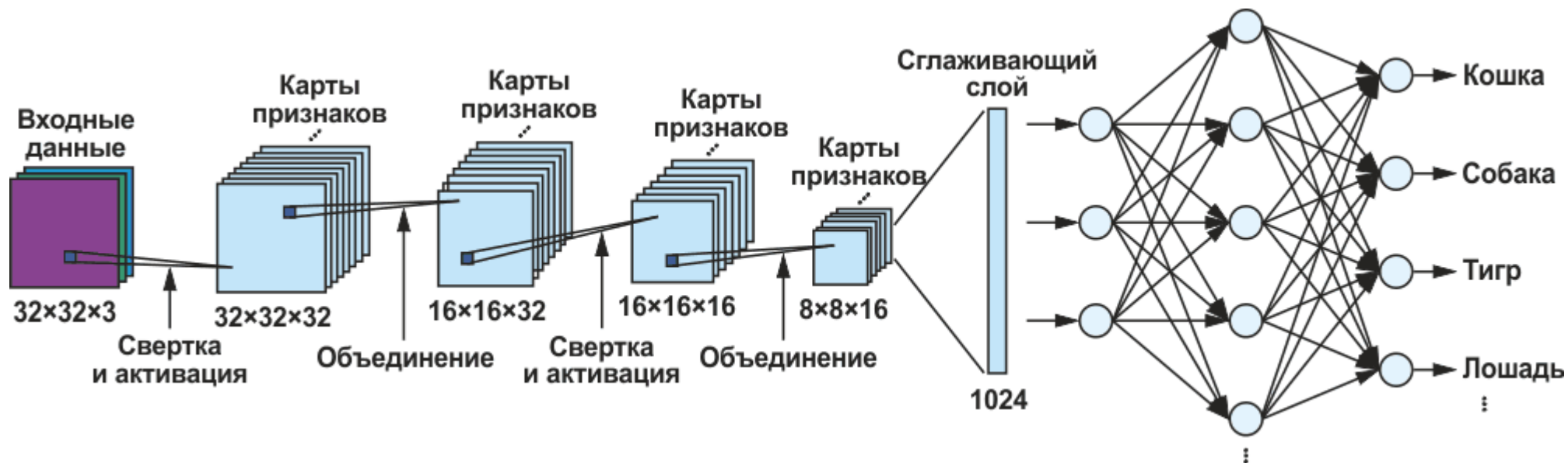


Методы глубокого обучения

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Сверточные нейронные сети (CNN)**

анализируют пиксели, которые находятся близко друг к другу, и ищут закономерности — сначала простые (линии, углы, пятна), а затем всё более сложные (глаза, уши, очертания объектов)



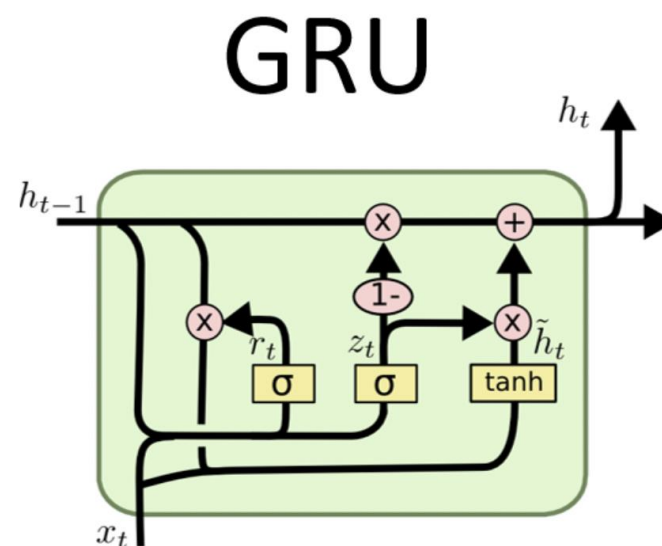
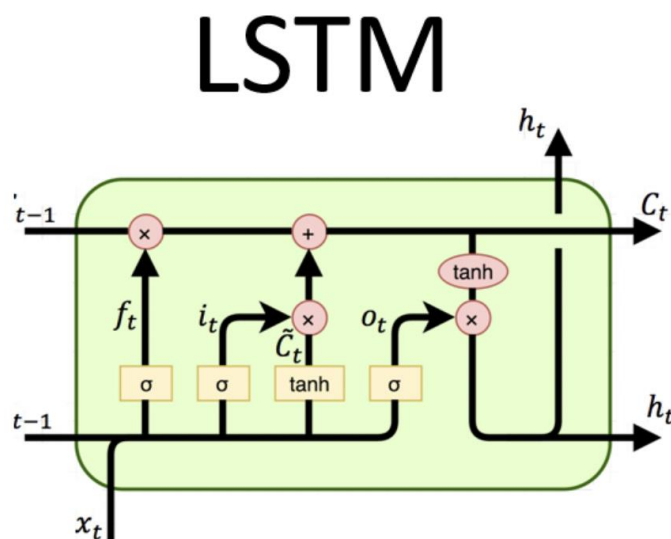


Методы глубокого обучения

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- Рекуррентные нейронные сети (RNN) и LSTM

способны улавливать контекст и долгосрочные зависимости в тексте, что делает их подходящими для задач генерации текста, машинного перевода и распознавания речи





Методы глубокого обучения

Технологии автоматической обработки
и анализа текстовой информации (NLP)

- **Трансформеры (BERT) и модели на основе BERT**

BERT изучает контекст в обоих направлениях одновременно. Это позволяет более точно определять значение слов и взаимосвязи с другими словами. Используют механизм внимания для обработки больших объемов текста и достижения высокой точности в различных задачах.