

Основы искусственного интеллекта

Лекция 19. Анализ текстов

2023/2024 учебный год

Доцент кафедры ИВЭ, Махно В.В.

©Создано при помощи <https://sberuniversity.ru/>



Обработка текстов

В качестве простого примера задачи анализа текстов можно привести классификацию текстов, например, определение темы документа или определение эмоционального окраса отзыва клиента.

Сама область обработки текстов по-английски называется **Natural Language processing (NLP)**. NLP охватывает все методы обработки текстов, включая, например, линейные модели для анализа текстов.

Однако большой прорыв в этой области произошел с приходом нейронных сетей, поэтому сегодня NLP во многом ассоциируется именно с нейросетевыми подходами.



Предобработка текстов

- **Очистка данных**

Данные, скачанные из социальных сетей, могут содержать ссылки, хэштеги, упоминания других пользователей (@username), а тексты, скачанные из интернета, — остатки разметки. Иногда также выполняют удаление пунктуации и приведение всех слов к нижнему регистру, однако это может быть полезно не во всех задачах. Например, в задаче определения эмоционального окраса отзыва использование большого количества заглавных букв может указывать на недовольство клиента.

- **Сборка словаря**

После очистки текстов выполняется сборка словаря: для каждого слова вычисляется количество раз, сколько оно встретилось в текстах, и словарь составляется из наиболее часто встречаемых слов, обычно 50 тысяч или 100 тысяч.

Редкие слова заменяются на слово «Пропуск»: если слово встретилась всего несколько раз во всем наборе данных, нейросеть не сможет хорошо выучить, что оно означает. Иногда, чтобы не терять информацию о редких словах, их делят на части, например, слово «эмбеддинг» можно разделить так: «эм», «бе», «д», «д», «инг». Такой подход называется **Byte Pair Encoding (BPE)**.

Предобработка текстов

Лето закончится через девятнадцать дней

2 6 5 7 5

Словарь:

я	1
лето	2
мы	3
через	4
дней	5
закончится	6
девятнадцать	7

Иногда, чтобы не терять информацию о редких словах, их делят на части, например, слово «эмбеддинг» можно разделить так: «эм», «бе», «д», «д», «инг». Такой подход называется **Byte Pair Encoding (BPE)**.

Векторные представления слов

Нейронные сети для анализа текстов обычно состоят из двух компонент: первая компонента обрабатывает лексику (распознает смысл слов, Embedding layer), а вторая обрабатывает структуру (извлекает зависимости из взаимного расположения слов).

Отличное



Замечательное



Мобильное



Приложение



Эмбеддинги

Эмбеddинг на практике

На практике человек не придумывает, какому смыслу соответствует каждое число в эмбеddинги — эмбеddинги слов компьютер учит сам.



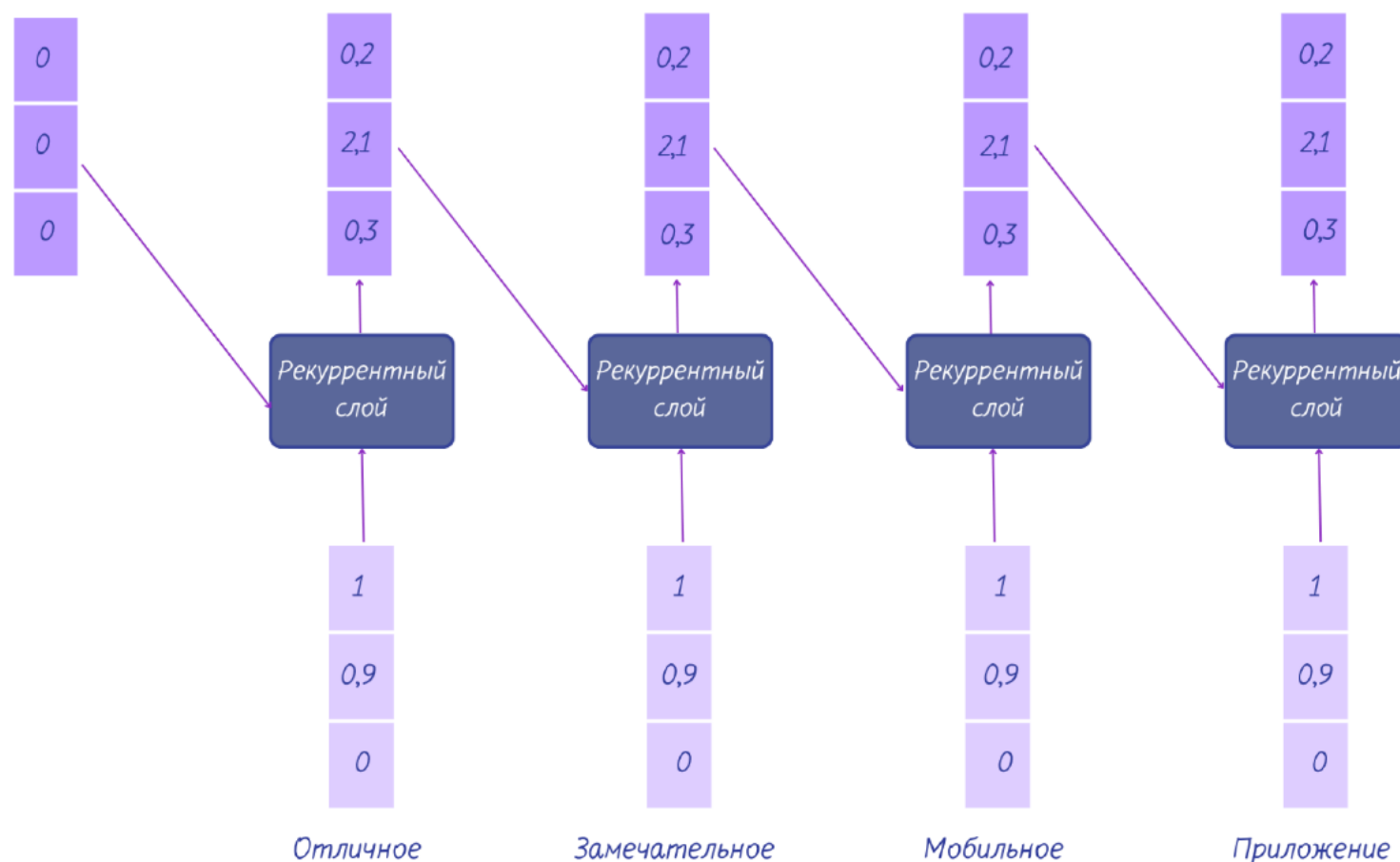
Пример визуализации для задачи анализа эмоционального окраса отзывов

Овалами были вручную выделены группы похожих слов, а в целом визуализация показывает, что у похожих по смыслу слов — похожие эмбеddинги.

Архитектуры нейросетей для анализа текстов

После слоя эмбедингов идет слой, отвечающий за обработку взаимного расположения слов в тексте — обычно используется рекуррентный слой или архитектура Трансформер.

Рекуррентная нейронная сеть





Задачи анализа и генерации текстов

Нейронные сети позволили решить множество интеллектуальных задач анализа текстов, которые не удавалось решить классическими методами. Эти задачи связаны с пониманием смысла или структуры текста и написанием новых текстов.

Более простыми считаются задачи, в которых текст выступает в качестве входных данных — такие задачи называют **дискриминативными**. Классические примеры — задача классификации текстов или задача регрессии.

Например, можно по тексту объявления о вакансии предсказывать заработную плату — такие алгоритмы используют на сайтах поиска вакансий

Разные примеры

Другой пример — задача модерации сообщений в социальных сетях или на сайтах объявлений: в ней нужно обнаруживать сообщения, написанные в грубой или оскорбительной форме. Например, Facebook и Instagram с 2016 года [используют](#) нейросети для модерации комментариев и рекомендации постов.

Отдельно выделяют задачи тегирования: в них нужно выполнить отдельное предсказание для каждого слова в тексте. К этой группе относятся задача определения частей речи для каждого слова и задача определения синтаксических ролей в предложении — такие алгоритмы помогают, в частности, выполнять автокоррекцию текста.

Еще один интересный пример — [задача выделения именованных сущностей](#). Она состоит в том, чтобы найти в тексте города, компании, персоналии и другие объекты реального мира. Их выделение помогает совершенствовать поиск по Интернету.

Противоположными дискриминативным считаются **генеративные задачи**: в них текст должна сгенерировать нейронная сеть, то есть текст — это выходные данные. Подробно генерация текстов разбирается в видеокейсе. Попробовать сгенерировать текст с помощью нейронной сети можно, используя сервис [Порфирьевич](#). А в [этом сервисе](#) можно загрузить фотографию, и нейросеть сгенерирует к ней подпись.

- a man holding a frisbee in his hands
- a couple of men standing next to each other
- a man holding a frisbee in a park
- a man holding a frisbee in his hand
- a man holding a frisbee in his right hand
- a couple of people that are standing in the grass
- a man and a woman standing next to each other
- a couple of men standing next to each other on a field
- a couple of men standing next to each other in a field
- a couple of men standing next to each other in a forest



Темы докладов на 20 декабря

1. Кейс "Компьютерное зрение« (распознавании изображений с помощью нейронных сетей, архитектура сверточной нейрости для решения задачи распознавания).
2. Кейс «Градиентный бустинг» (создание рекомендательной системы: свойства данных, методы коллаборативной фильтрации и вычисления с матрицами для создания признаков, обучении ансамбля, градиентного бустинга и тестирование получившихся рекомендаций на пользователях).
3. "Линейная классификация и тексты« (задача классификации тональности текста, особенности предподготовки текстовых данных: стемминге и токенизации; поиск признаков, решение задачи линейными моделями).
4. Различные виды сверток — [здесь](#) ([Матричные фильтры обработки изображений / Хабр \(habr.com\)](#)).

Онлайн-курс СберУниверситета

Генеративное искусство

Подробнее о курсе



Бесплатный курс от Сбера
по генеративному
искусству

https://courses.sberuniversity.ru/generative-art?utm_source=tg&utm_medium=organic&utm_campaign=courses&utm_content=gen_i&utm_term=01_09_2023