



Факультет компьютерных
наук

Аналитика больших
данных

Москва
2025

Основы обработки естественного языка (NLP)



Курс “Основы NLP”

1. Введение в NLP и предобработка текста
2. Векторные представления слов и тематическое моделирование
3. Классификация текста
4. Последовательная разметка текста
5. Языковое моделирование и генерация текста
6. Seq2Seq-модели и механизм внимания
7. Передача знаний и предобученные языковые модели



- Базовая работа с текстом
- Векторизация текста
- Сверточные нейросети (CNN) и рекуррентные (RNN/LSTM)
- POS-теггинг, именованные сущности (NER)
- Seq2Seq-модели → энкодеры

Курс “Дизайн современных LLM”

1. Большие языковые модели и их архитектуры
2. Языковое моделирование и генерация текста с помощью LLM
3. Стратегии дообучения LLM под разные задачи
4. Работа с флагманскими моделями через API
5. Задачи RAG и Агенты
6. Ограничения LLM: галлюцинации, контекстное окно и другое
7. Бенчмарки и валидация





- Занятие: лекция + семинар
- Время: четверг в 18:10
- Репозиторий курса

HW	Задание	Выдача	Мягкий дедлайн	Жёсткий дедлайн
дз1	Предобработка текста и тематическое моделирование	чт 13 ноя	ср 26 ноя	ср 03 дек
дз2	Классификация текста и последовательная разметка	чт 27 ноя	вс 10 дек	вс 14 дек
дз3	Финал: Kaggle-соревнование	чт 04 дек	вт 16 дек	пт 19 дек

Итоговая оценка: $0.3 * \text{дз1} + 0.3 * \text{дз2} + 0.4 * \text{дз3}$

Политика дедлайнов: -0.1 балла в день, максимум -0.7 за одно ДЗ.



1. Введение в NLP и примеры применения
2. Основные задачи и сложности языка
3. NLP-пайpline: от текста к модели
4. Очистка и нормализация текста
 - Токенизация
 - Стоп-слова
 - Стемминг и лемматизация
5. Преобразование текста в признаки: Bag-of-Words и TF-IDF
6. Инструменты и библиотеки для NLP



Natural Language Processing

5 |

Information Retrieval

Doc A

Doc 1
Doc 2
Doc 3

Sentiment Analysis



Information Extraction



Machine Translation



Natural Language Processing

Question Answering



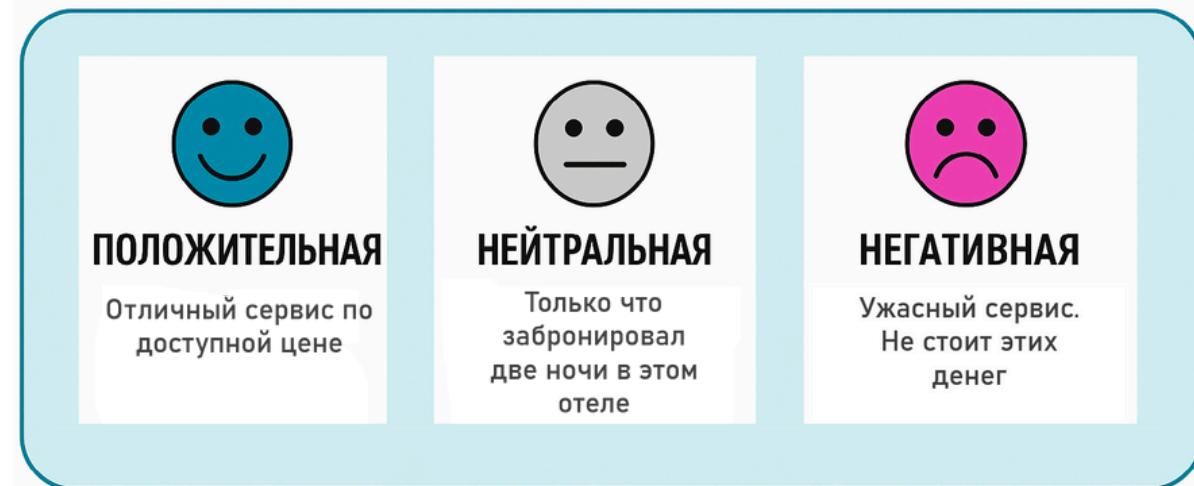
Human: When was Apollo sent to space?

Machine: First flight -
AS-201,
February 26,
1966

1) классификация/кластеризация текстов.

текст → метка, несколько меток

- Фильтрация спама
- Анализ тональности
- Определение жанра
(кластеризация)
- Детекция AI-
сгенерированного
текста



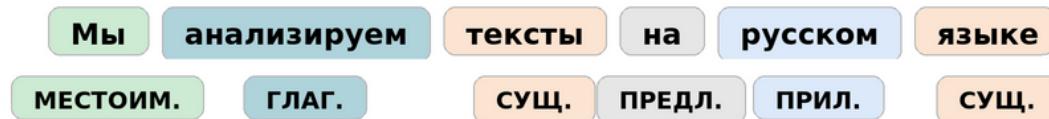
Основные задачи

2) классификация последовательности (sequence labeling)

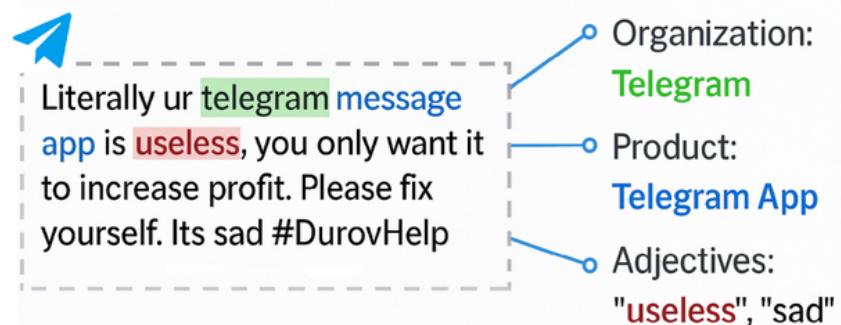
каждое слово в тексте → метка

- POS (part-of-speech) tagging

Разметка по частям речи (POS-tagging)



- NER – распознавание именованных сущностей





Основные задачи

8

3) sequence to sequence (seq2seq)

текст → текст

- Машинный перевод
- Вопросно-ответные системы
- Парафраз текста
- Суммаризация текста

The screenshot shows a machine translation interface. At the top, there are tabs for 'Text' and 'Documents'. Below that, language pairs are listed: ENGLISH - DETECTED, ENGLISH, GERMAN, ENGLISH, RUSSIAN. A double-headed arrow icon is between ENGLISH and GERMAN. The English input field contains the sentence 'How are you today?' and the German output field contains 'Wie geht es dir heute?'. There are also icons for microphone, speaker, and feedback at the bottom.

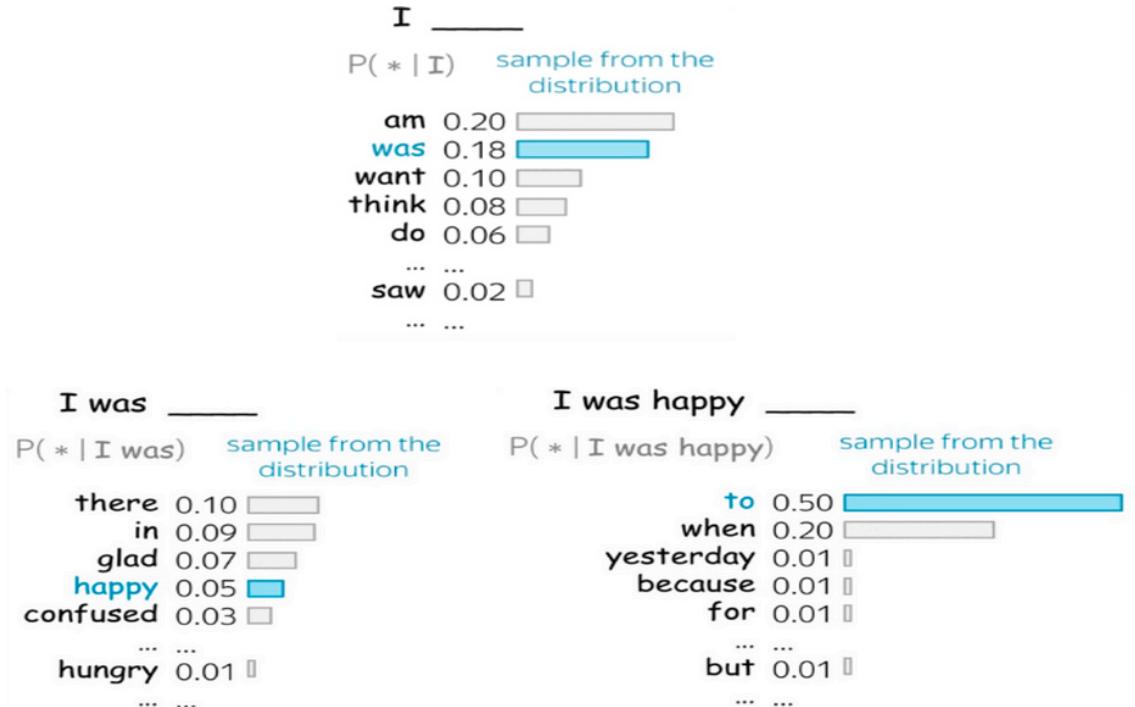
The screenshot shows a conversation with ChatGPT. The user asks 'Выбери число от 1 до 50' (Choose a number from 1 to 50). ChatGPT responds with '20'. The user then asks 'Мы не будем общаться и я не буду тобой пользоваться 20 дней' (We won't communicate and I won't use you for 20 days). ChatGPT replies 'Можно я выберу другое число?' (Can I choose another number?). The user then says 'Да' (Yes) and ChatGPT responds with '50'.



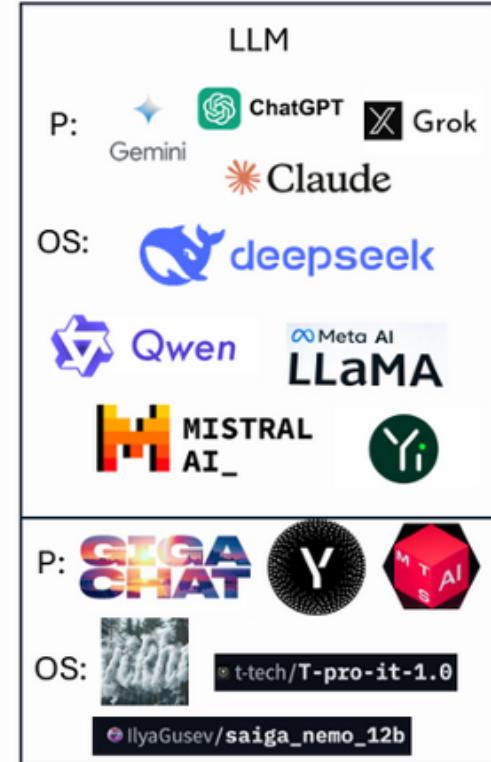
Языковое моделирование

9

задача курса “Дизайн современных LLM”



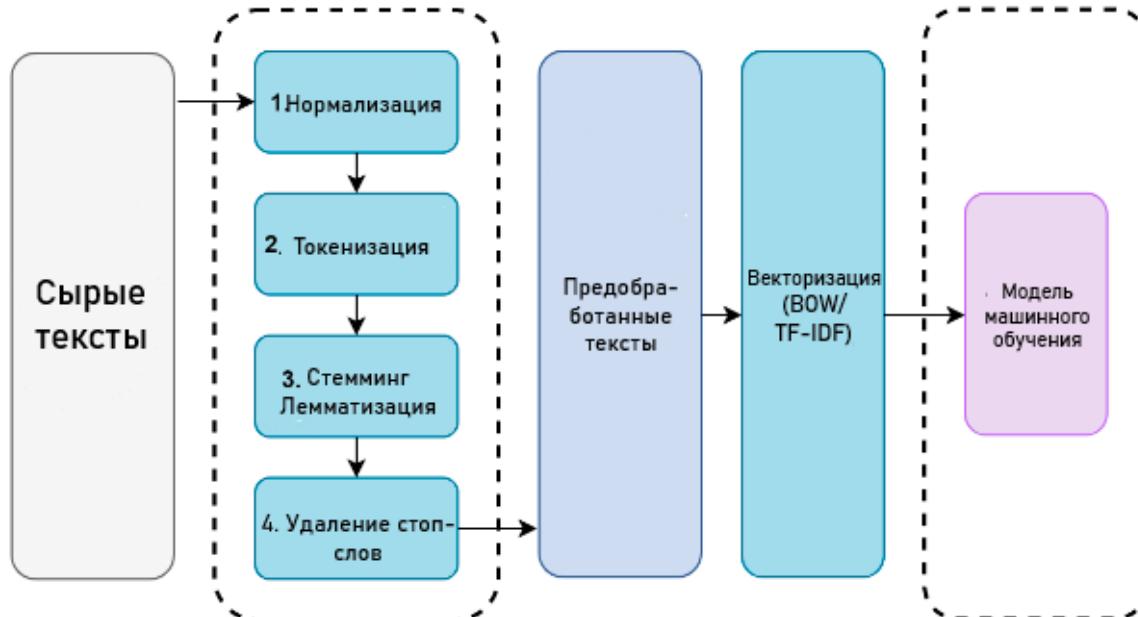
В Мире





NLP-пайплайн: от текста к модели

10





Очистка и нормализация текста

11



“I'll meet you at 3:00 PM, don't forget!!!”

“i'll meet you at 3:00 pm, don't forget!!!”

“i'll meet you at 3:00 pm dont forget”

“i'll meet you at pm dont forget”

“i will meet you at pm do nt forget”



Токенизация

12

Токенизация на уровне символов

```
["I", " ", "❤", " ", "N", "L", "P", "!", " ", "|", "t", "", "s", " ", "s", "o", " ", "m", "u", "c", "h", " ", "f", "u", "n", "!", " ", "😎"]
```

Токенизация на уровне слов

```
["I", "❤", "NLP", "!", "It's", "so", "much", "fun", "!", "😎"]
```

Токенизация по пробелам

```
["I", "❤", "NLP!", "It's", "so", "much", "fun!", "😎"]
```

Токенизация с учётом
знаков препинания

```
["I", "❤", "NLP", "!", "It", "", "s", "so", "much", "fun", "!", "😎"]
```

Токенизация через
кодирование пар байтов
Byte-Pair Encoding (BPE)

```
["I", "❤", "N", "LP", "!", "It", "", "s", "so", "much", "fun", "!", "😎"]
```

"much", "fun", "!", "😎"]

"I ❤ NLP! It's so much fun! 😎"



Byte-Pair Encoding

“newer never”

13

Алгоритм сжатия текста,
который используется в
токенизации для
уменьшения размера
словаря и обработки
редких слов.

1. Разбиение на символы
2. Подсчёт частоты пар
3. Слияние пар
4. Повторение

n e : 2 раза

e w : 1 раз

e r : 2 раза

n v : 1 раз

v e : 1 раз

n e w e r n e v e r

n e w e r n e v e r

["ne", "w", "e", "r", "ne", "v", "e", "r"].



“apple” -> ["app", "le"]

“unhappiness” -> ["un", "happiness"]

- Сжатия текста

Слияние частых символов позволяет эффективно уменьшить размер словаря, особенно для редких слов.

- Универсальности

Позволяет работать с редкими или неизвестными словами, разделяя их на более мелкие, часто встречающиеся компоненты.

- Улучшения обработки OOV (out-of-vocabulary)

Слова, которых нет в словаре, можно разобрать на более мелкие токены, что позволяет эффективно справляться с новыми или редкими словами.

Стемминг → стема
(основа)

Лемматизация → лемма
(словарная форма)

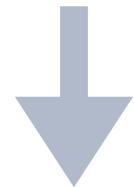
Оригинал:
“increased” (в смысле “увеличенный”)

После стемминга:
“increased” → “increas”

Слово	Стеммер →	Лемматизатор →
studies (англ)	studi	study
cars	car	car
better (англ)	better	good (условно)
бегущие (рус)	бегущ	бежать

! стемминг увеличивает recall (покрытие), объединяя больше форм, но может снизить precision, объединяя лишние

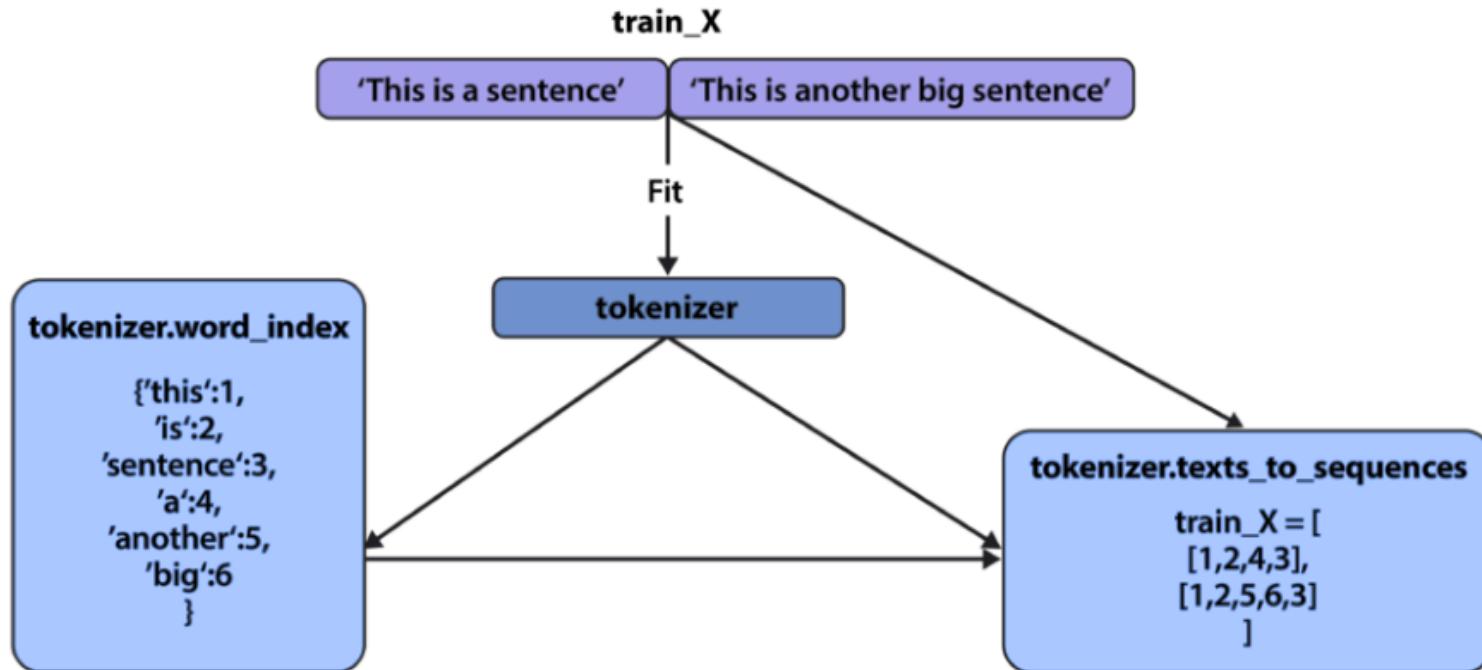
She is very happy to be going home.

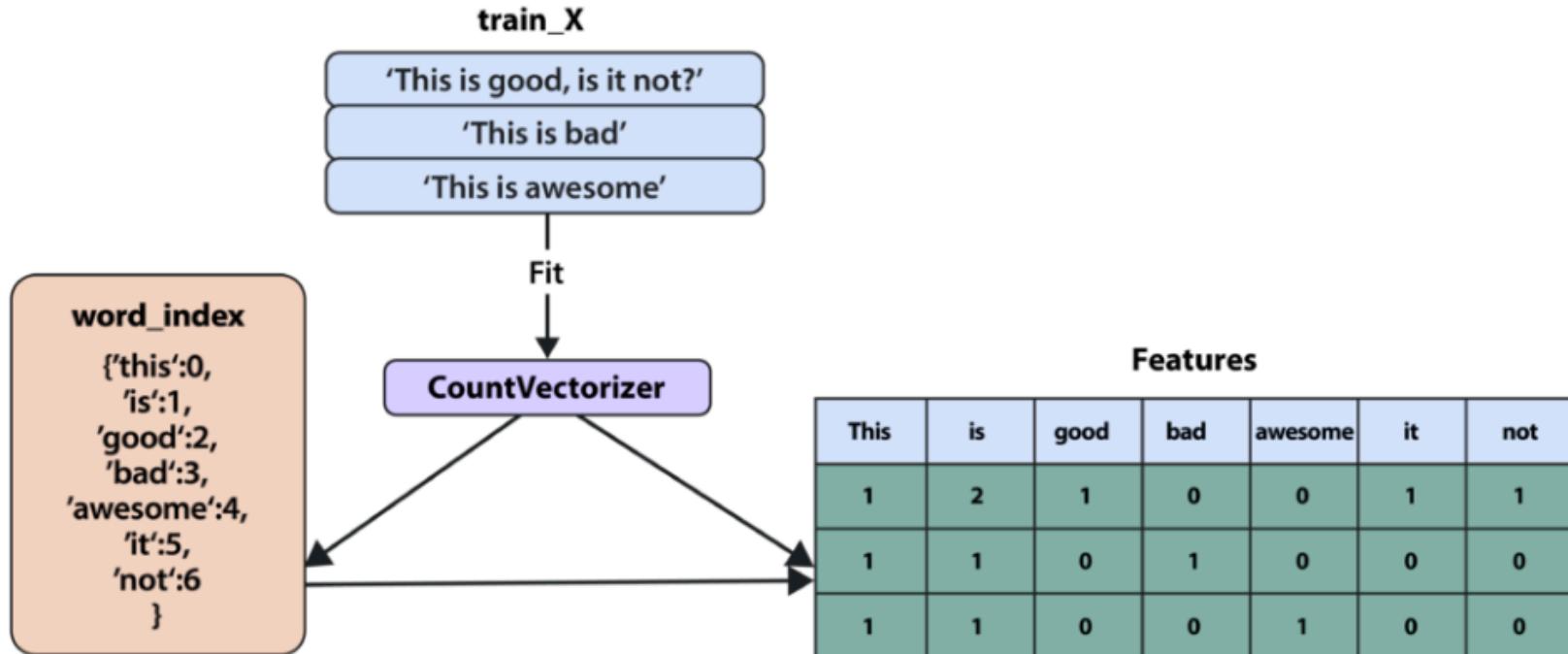


She happy going home.



Преобразование текста в признаки







Term Frequency – Inverse Document Frequency

19

Метод, где наибольший вес получают слова, которые часто встречаются в *одном конкретном документе* (высокий TF), но редко в *целом корпусе* (высокий IDF), что и выделяет их как уникальные и важные маркеры.

$$\text{TF}(t, d) = \frac{\text{число вхождений слова } t \text{ в документ } d}{\text{общее число слов в документе } d}$$

$$\text{IDF}(t, D) = \log \left(\frac{\text{общее число документов}}{\text{число документов, где встречается слово } t} \right)$$

Логарифмирование здесь проводится с целью уменьшить масштаб весов, ибо зачастую в корпусах присутствует очень много текстов.

В итоге каждому слову t из текста d теперь можно присвоить вес

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Интерпретируем:

- Если слово часто встречается в одном документе, но редко — в остальных, то оно важно для этого документа.
- Если слово встречается везде, например, такие как "и", "что", "это", то оно не помогает отличать один текст от другого — TF-IDF даст ему низкий вес.



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

TF-IDF

Термин x в документе y

$tf_{x,y}$ = частота x в документе y

df_x = число документов, содержащих x

N = общее число документов



TF-IDF всё еще - всегда

Место	Входные данные (Features)	Модель (Model)	Функция потерь (Loss)	Hitrate@3
11	app_name + shortDescription + full_description	cointegrated/rubert-tiny2	Стандартная	0.2846
10	app_name	FacebookAI/xlm-roberta-base	Стандартная	0.5576
9	app_name	FacebookAI/xlm-roberta-base	С весами для классов	0.7017
8	app_name + shortDescription	FacebookAI/xlm-roberta-base	Стандартная	0.7714
7	app_name + shortDescription	FacebookAI/xlm-roberta-base	С весами для классов	0.8409
6	app_name + shortDescription + full_description	TF-IDF + LSA + Logistic Regression	С весами для классов	0.8638
5	app_name + shortDescription + full_description	FacebookAI/xlm-roberta-base	Стандартная	0.8941
4	app_name + shortDescription + full_description	TF-IDF + ClassifierChain(LR)	С весами для классов	0.9092
3	app_name + shortDescription + full_description	FacebookAI/xlm-roberta-base	С весами для классов	0.9165
2	app_name + shortDescription + full_description	intfloat/multilingual-e5-base	С весами для классов	0.9196
1	app_name + shortDescription + full_description	TF-IDF + Logistic Regression	С весами для классов	0.9224



- Регулярные выражения (regex)
- NLTK (Natural Language Toolkit)
- spaCy

токенизацию, лемматизацию, POS-теггинг, NER

- scikit-learn
- CountVectorizer (BoW) и TfidfVectorizer

