МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ Федеральное государственное автономное образовательное учреждение высшего образования «Южно-Уральский государственный университет (национальный исследовательский университет)» Высшая школа электроники и компьютерных наук Кафедра системного программирования

Разработка приложения для классификации литературных произведений по жанрам с помощью методов машинного обучения

Рецензент:
доцент кафедры ВМиИТ ЧелГУ,
к.ф.-м.н.
А.Ю. Маковецкий

Научный руководитель:
доцент кафедры СП, к.ф.-м.н.
Т.Ю. Маковецкая

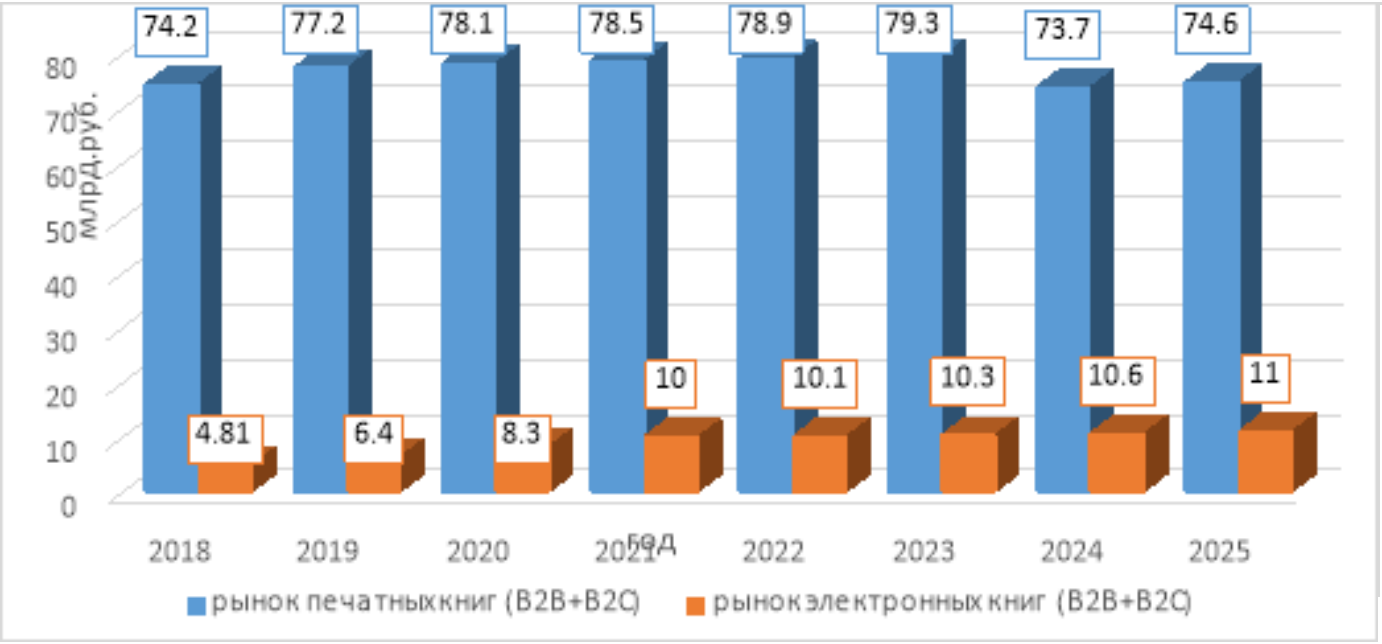
Автор:
студент группы КЭ-220
Д.С. Курочкин

Челябинск, 2024 г.

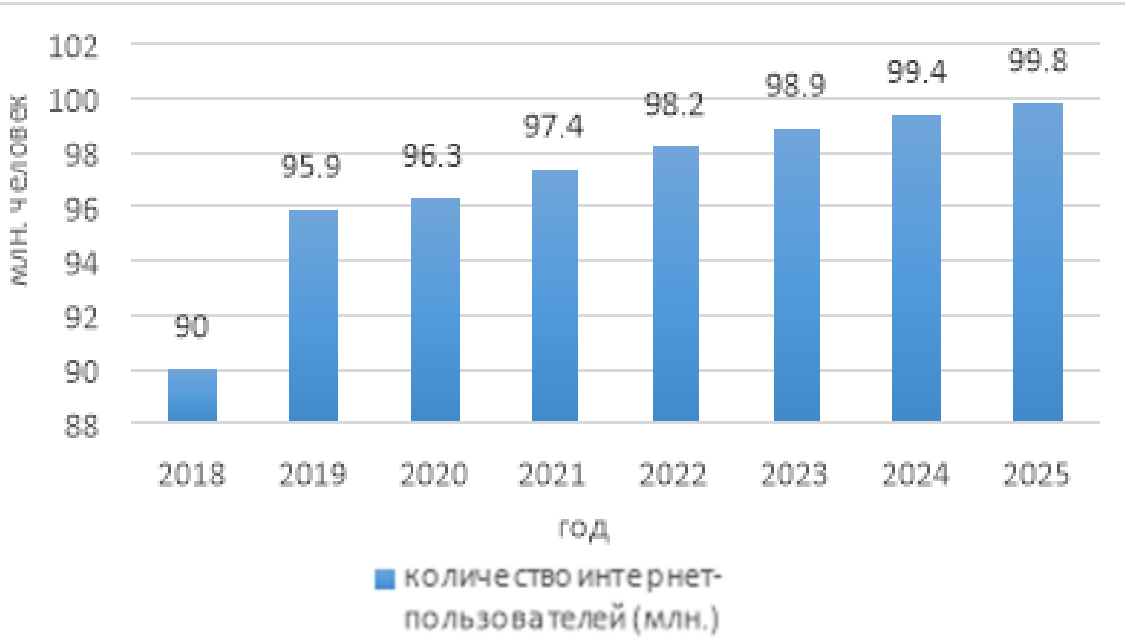
МИРОВОЙ РЫНОК ЭЛЕКТРОННЫХ КНИГ

РЫНОК ЭЛЕКТРОННЫХ КНИГ В РОССИИ

Прогноз развития рынка электронных книг, млрд руб.



Прогноз количества пользователей, млн. человек



САМИЗДАТ В РОССИИ

- Текущая доля самиздата в сегменте электронных книг – 35%
- Адаптация книжного сервиса «Литнет» для российского самиздата

ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель:

Разработка приложения для классификации литературных произведений по жанрам с помощью методов машинного обучения

Задачи:

1. Провести обзор аналогов и научной литературы
2. Осуществить сбор и предобработку данных для обучения
3. Разработать модель машинного обучения и оценить результаты работы
4. Реализовать приложение
5. Провести тестирование приложения

АНАЛИЗ СЕРВИСОВ ПО РАБОТЕ С ТЕКСТОМ

Функции	Amazon Comprehend	IBM Watson Natural Language Understanding	Natural Language AI	AI Azure
Классификация текстовых документов	+	+	+	–
Выделение сущностей	+	+	+	+
Поддержка русского языка	–	+/–	+	+/–
Наличие демоверсии	–	+	+	+/–
Выделение ключевых фраз	+	+	+	+
Анализ тональностей	+	+	+	+
Синтаксический анализ	–	+	+	+
Количество дополнительных возможностей (другие функции)	2	5	6	11

НАБОР ДАННЫХ

№	Жанр	Кол-во	№	Жанр	Кол-во
1	Бизнес	849	**9**	Психология/ мотивация	991
2	Боевики	917	**10**	Религия	611
3	Детективы	908	**11**	Романы	767
4	Дом/дача	470	**12**	Спорт/здоровье/ красота	995
5	Знания/навыки	1062	**13**	Ужасы/мистика	1019
6	История	948	**14**	Фантастика	802
7	Классика	472	**15**	Фэнтези	772
8	Поэзия	1012	**16**	Хобби/досуг	933
				Всего	**13 528**

- Формат электронных книг – pdf
- Обучающая выборка – 80%
- Тестовая выборка – 20%

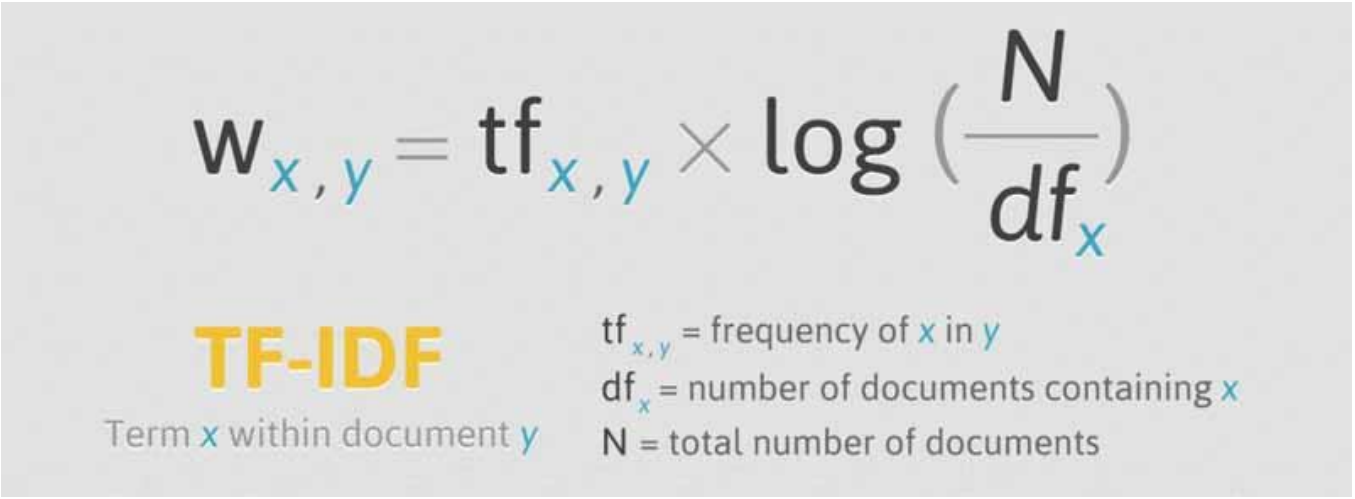
ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ

1. Удаление знаков препинания и сторонних символов
2. Приведение слов к нижнему регистру
3. Удаление стоп-слов

СРЕДСТВА РЕАЛИЗАЦИИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

1. Среда обучения:
   - ОС Linux, Intel(R) Xeon(R) CPU @ 2.20GHz, ОЗУ 12,7 ГБ
2. Язык и инструменты:
   - Язык программирования: Python
   - Редактор исходного кода: Google Collab
3. Библиотеки и фреймворки:
   - Библиотека для обучения моделей: sklearn
   - Библиотека для работы с естественным языком: NLTK
   - Библиотека для работы с pdf-файлами: PyPDF2

ПРЕОБРАЗОВАНИЕ ТЕКСТА В ЧИСЛОВОЙ ФОРМАТ



- Минимальное количество документов, содержащие токены – 50
- Максимальное количество документов, содержащие токены – 13 000

СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В БИБЛИОТЕКИ

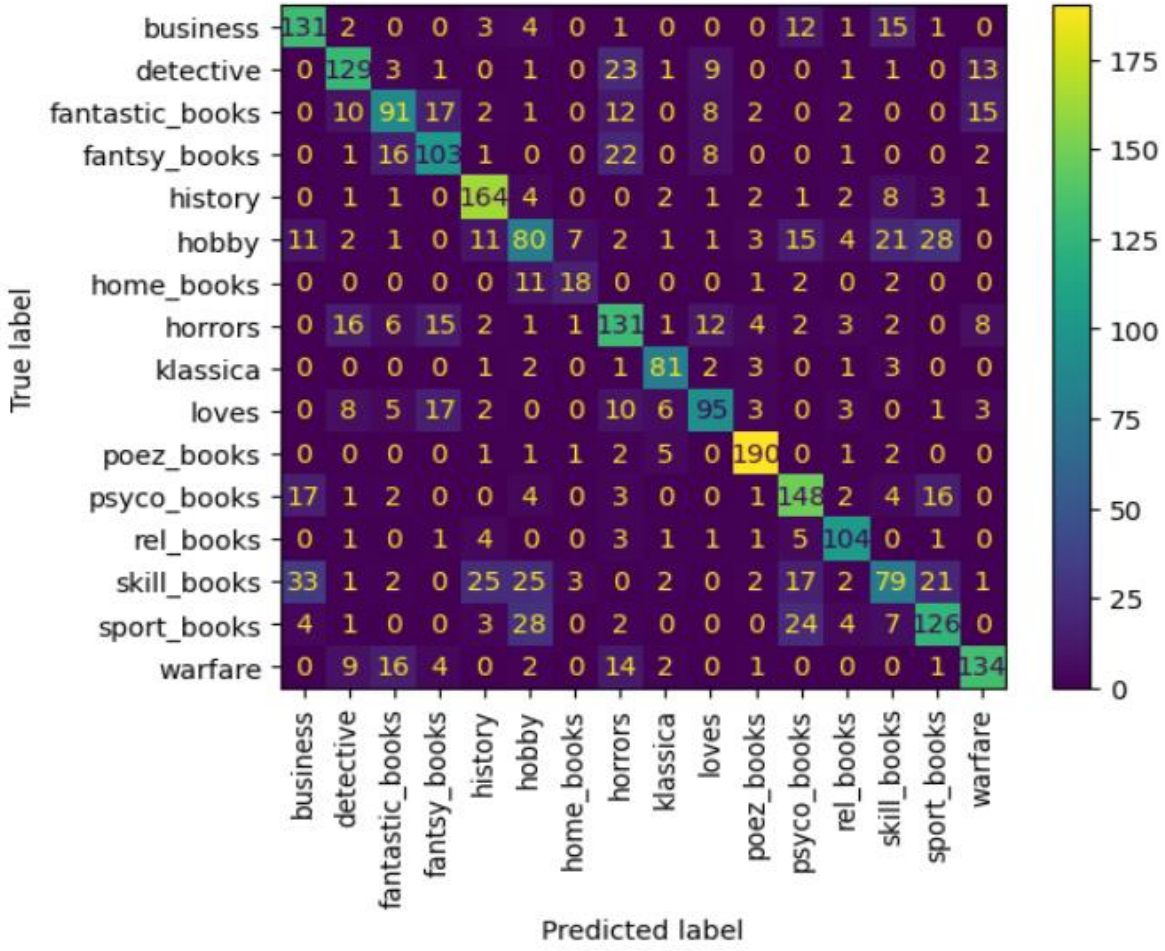
Методы	Accuracy	Macro avg (precision)	Macro avg (recall)
ComplementNB	0,6478	0,66	0,63
MultinomialNB	0,5877	0,62	0,53
SVC	0,6523	0,67	0,64
LogisticRegression	0,6776	0,69	0,66
SGDClassifier	**0,6818**	**0,68**	**0,68**
RidgeClassifier	0,6723	0,67	0,67
RandomForestClassifier	0,565	0,56	0,54

ОСНОВНЫЕ ПОКАЗАТЕЛИ ТЕСТИРОВАНИЯ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ МЕТОДА SGDClassifier

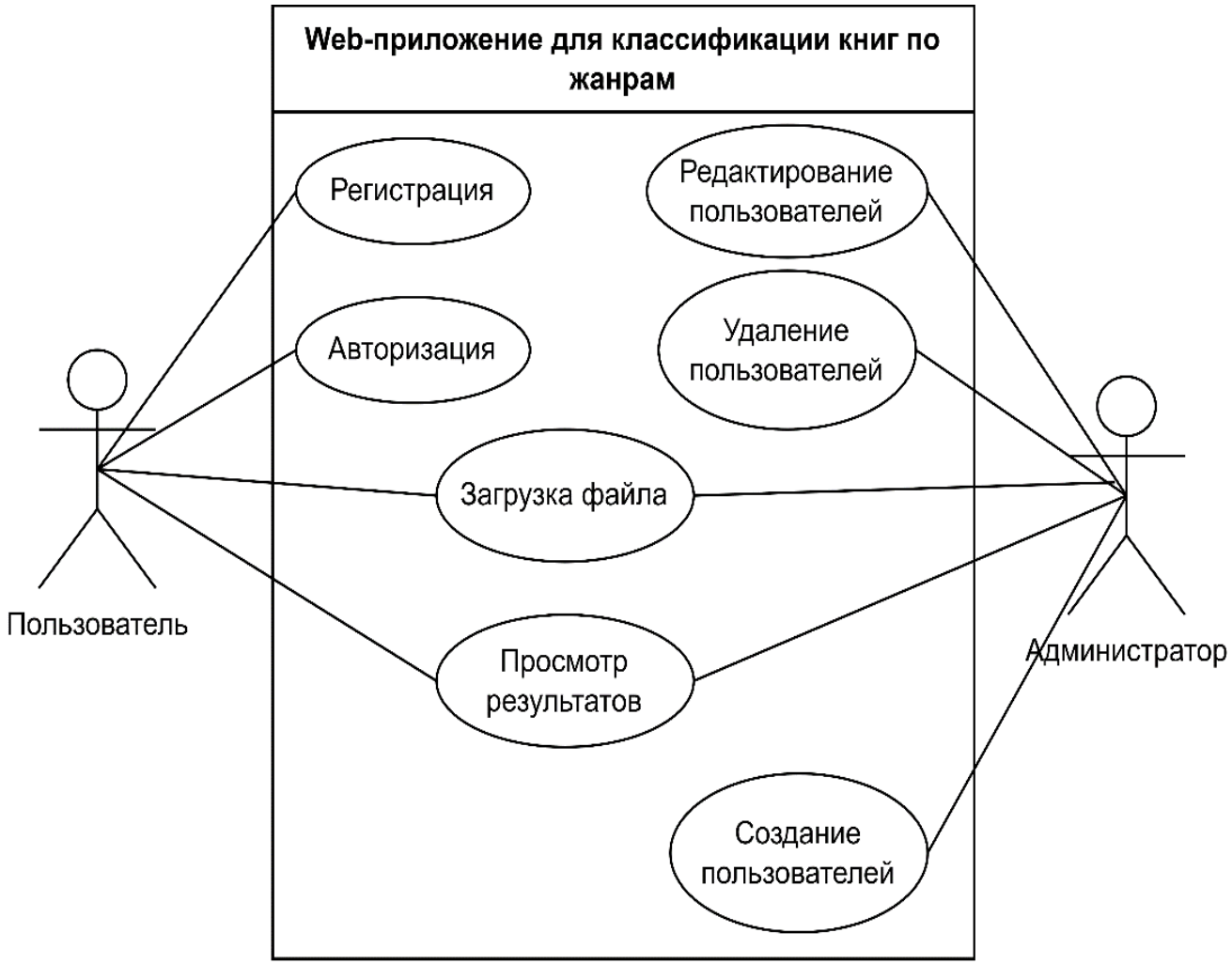
Отчет классификации

	precision	recall	f1-score	support
business	0.67	0.77	0.72	170
detective	0.71	0.71	0.71	182
fantastic_books	0.64	0.57	0.60	160
fantsy_books	0.65	0.67	0.66	154
history	0.75	0.86	0.80	190
hobby	0.49	0.43	0.46	187
home_books	0.60	0.53	0.56	34
horrors	0.58	0.64	0.61	204
klassica	0.79	0.86	0.83	94
loves	0.69	0.62	0.66	153
poez_books	0.89	0.94	0.91	203
psyco_books	0.65	0.75	0.70	198
rel_books	0.79	0.85	0.82	122
skill_books	0.55	0.37	0.44	213
sport_books	0.64	0.63	0.63	199
warfare	0.76	0.73	0.74	183
accuracy			0.68	2646
macro avg	0.68	0.68	0.68	2646
weighted avg	0.68	0.68	0.68	2646

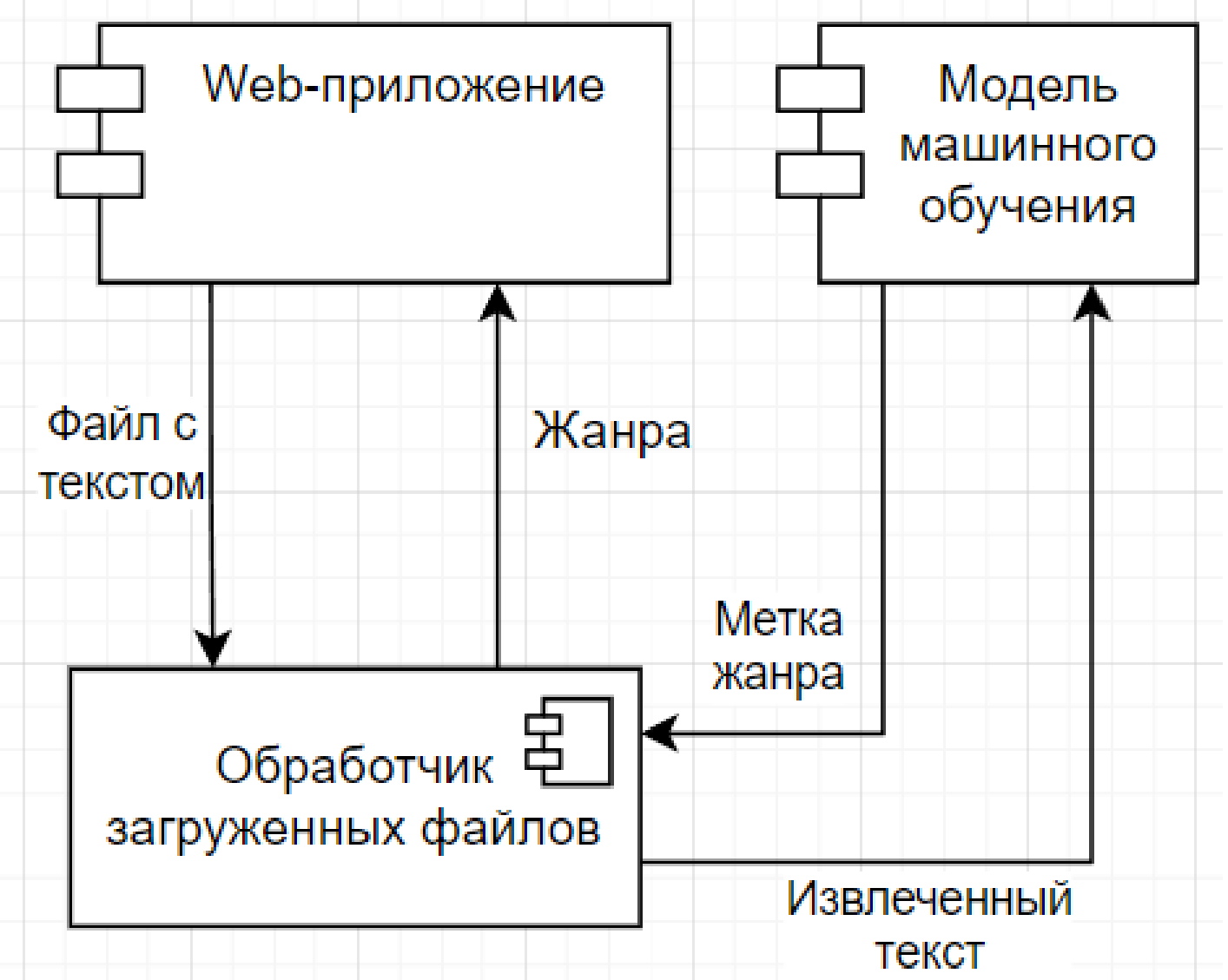
Матрица несоответствий



ФУНКЦИОНАЛЬНОСТЬ WEB-ПРИЛОЖЕНИЯ



АРХИТЕКТУРА СИСТЕМЫ



ОСНОВНОЙ ЦИКЛ РАБОТЫ

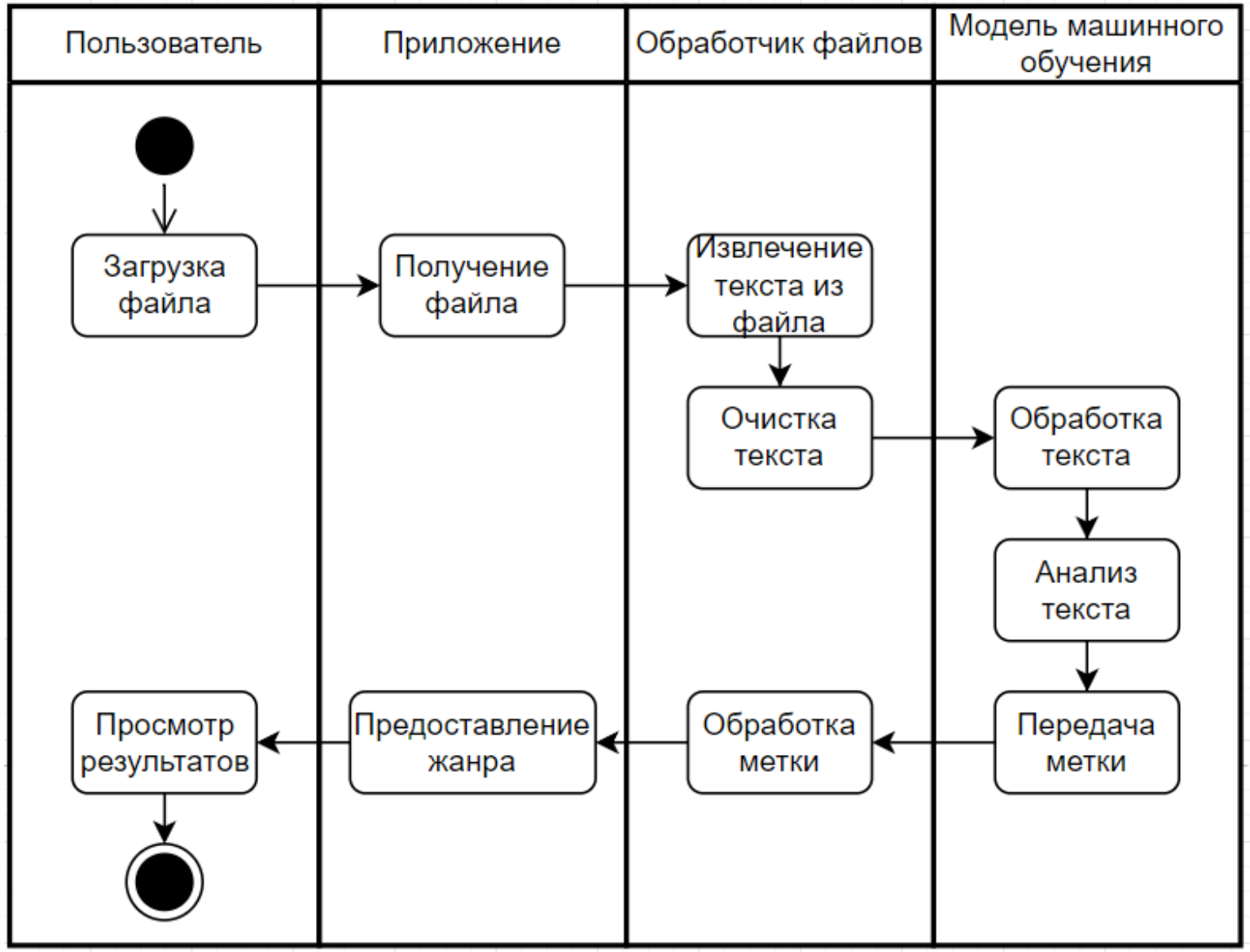
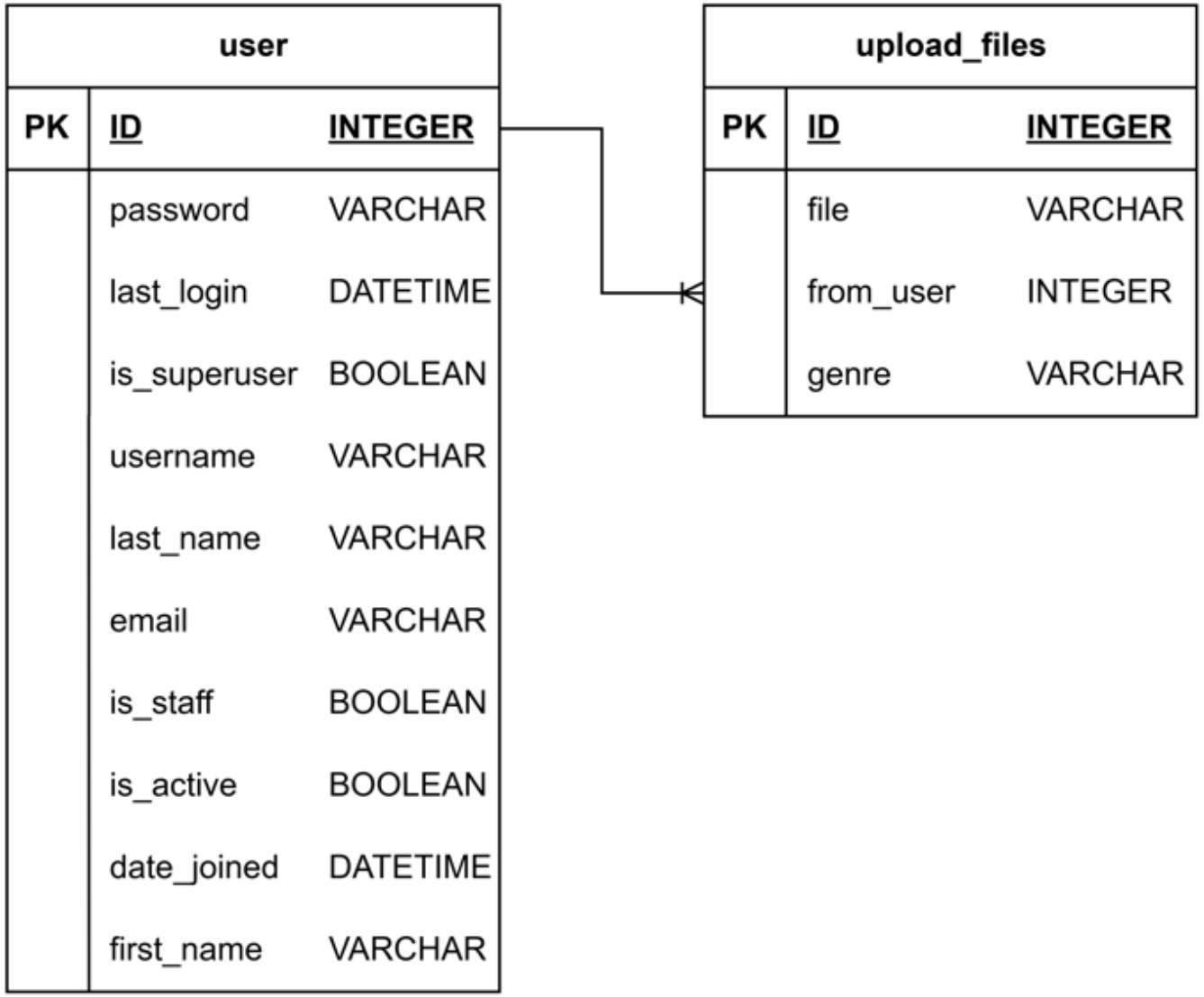


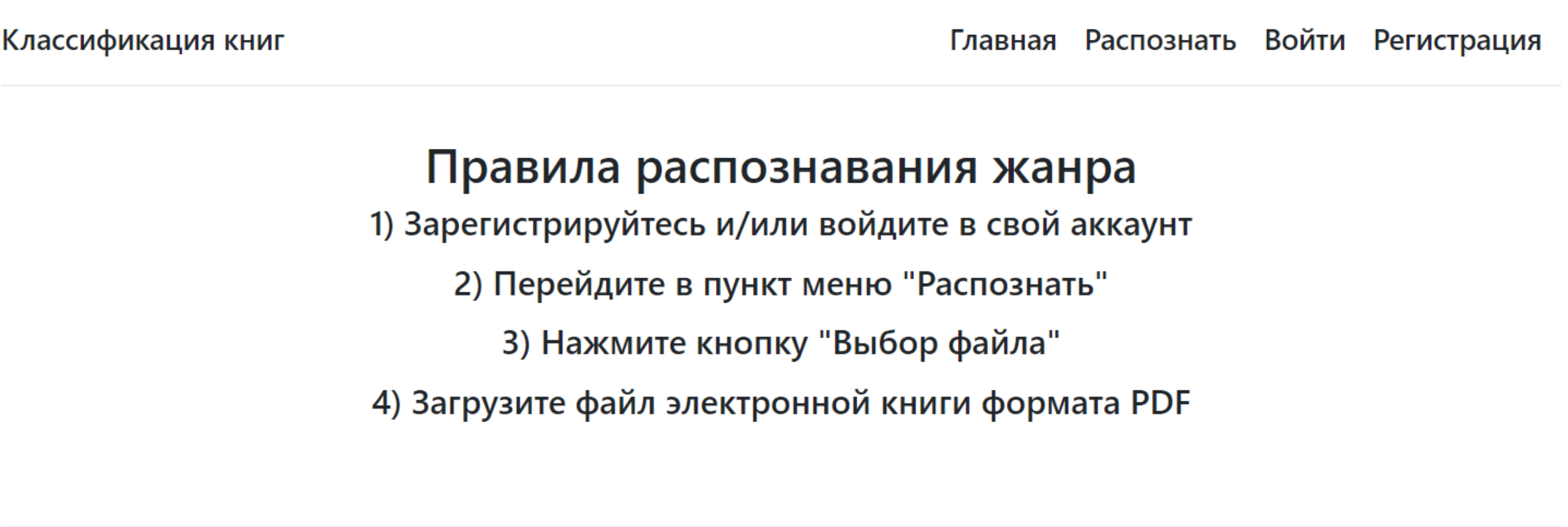
СХЕМА БАЗЫ ДАННЫХ



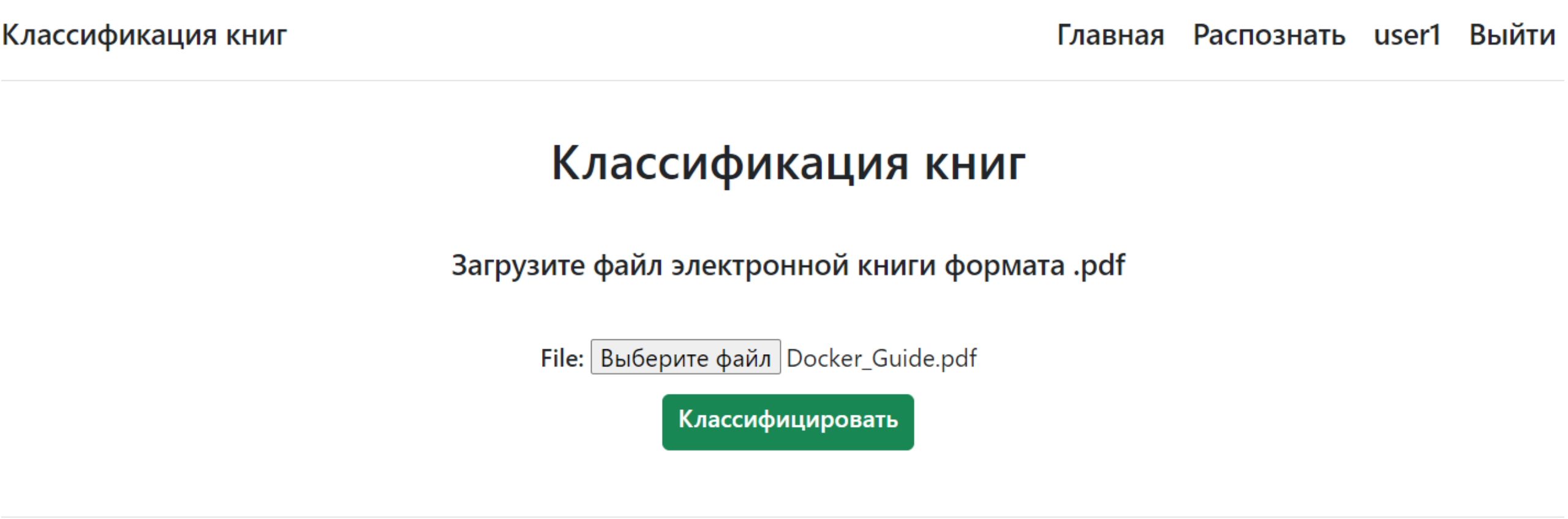
СРЕДСТВА РАЗРАБОТКИ WEB-ПРИЛОЖЕНИЯ

1. Библиотеки и фреймворки:
   - Фреймворк для создания web-приложений: Django
   - Библиотека для шаблонов HTML-страниц: Bootstrap
2. Дополнительные библиотеки:
   - Библиотека для работы с pdf-файлами: PyPDF2
3. Язык и инструменты:
   - Язык программирования: Python
   - Редактор исходного кода: PyCharm
   - Дополнительный редактор кода: Notepad++

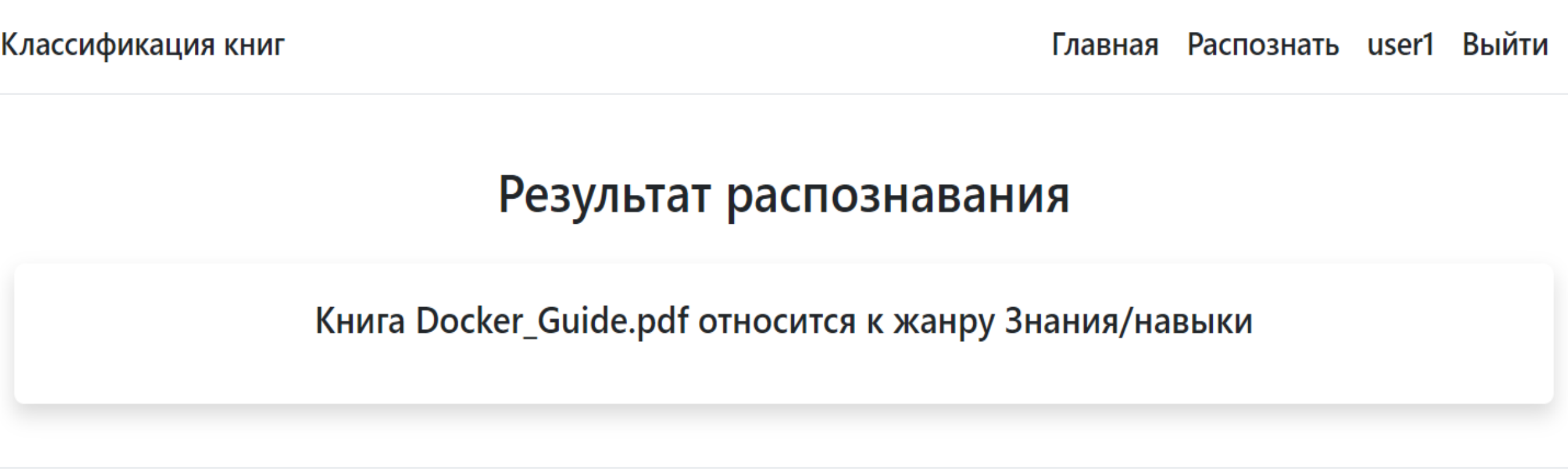
ГЛАВНАЯ СТРАНИЦА САЙТА



СТРАНИЦА ЗАГРУЗКИ ФАЙЛА



СТРАНИЦА С РЕЗУЛЬТАТОМ



ФУНКЦИОНАЛЬНОЕ ТЕСТИРОВАНИЕ

1. Сохранение файла пользователя
2. Корректное отображение результата классификации
3. Успешная авторизация
4. Предупреждение о неправильном вводе логина или пароля
5. Успешная регистрация
6. Успешный выход из аккаунта
7. …

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проведен обзор аналогов и научной литературы
2. Собран и обработан набор данных
3. Осуществлена разработка и тестирование модели машинного обучения
4. Спроектировано приложение
5. Реализовано web-приложение для классификации книг по жанрам
6. Выполнено тестирование приложения