ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ» МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОНИКИ И МАТЕМАТИКИ

**Программа для обнаружения спам-сообщений**

Руководство разработчика к заданию по дисциплине «Проектный семинар «Python в науке о данных» студентов образовательной программы «Информатика и вычислительная техника» учебной группы БИВ248:

* Погоняев Сергей Сергеевич, [sspogoniaev@edu.hse.ru](mailto:sspogoniaev@edu.hse.ru), 89234102924
* Шишелякин Матвей Романович, [mrshisheliakin@edu.hse.ru](mailto:mrshisheliakin@edu.hse.ru), 89142355181

Москва 2025

# **Оглавление**

[**Оглавление** 2](#_Toc200291599)

[**Назначение программы. Использование приложения** 3](#_Toc200291600)

[**Архитектура приложения** 4](#_Toc200291601)

[**Структура приложения** 4](#_Toc200291602)

[**Описание функций модулей и функций.** 5](#_Toc200291603)

[**Используемые библиотеки** 6](#_Toc200291604)

[**Визуальная часть приложения.** 6](#_Toc200291605)

# **Назначение программы. Использование приложения**

Приложение предназначено для классификации пользовательского SMS-сообщения в группы «Спам» и «Не спам», для создания графических и текстовых отчетов по датасету, на основе которого программа делает предположение, а также по обучению модели на данном датасете.

Приложение разворачивается на локальном сервере (порт 8000). Включает в себя 4 страницы:

1. Главная страница;
2. Настройки;
3. Страница добавления записи в датасет;
4. Страница загрузки отчетов.

Основной функционал.

* Классификация сообщения в группы «Спам» и «Не спам»;
* Добавление новых записей с сообщениями и маркерами для них в датасет, на основе которого обучена модель;
* Переобучение модели (если в датасет добавляются новые записи);
* Создание текстовых отчетов (отчет по обучению модели, отчет по датасету – формат .txt);
* Создание графического отчета (отчет по датасету – формат .jpg);
* Настройка пользовательского интерфейса.

Базовые требования для нормального функционирования программы:

* ОС Windows 10;
* 2 ГБ ОЗУ
* 15 МБ памяти на жестком диске

Приложение было написано на языке программирования Python. Для создания шаблонов их отображения использовался язык разметки HTML и язык программирования Java Script.

# **Архитектура приложения**

## **Структура приложения**

python\_project/

runner.bat

work/

Data/

SMSSpamCollection - датасет с сообщениями

Graphics/

graghic.png

library/

create\_reports.py - модуль для создания отчетов

database\_manipulation.py - работа с датасетом

model.py - модуль для работы с моделью

spam\_model.pkl

vectorizer.pkl

notes/

Руководство\_пользователя\_Python.docx

Руководство\_разработчика\_Python.docx

Output/

model\_report.txt

text\_report.txt

scripts/

main.py - основной файл

environment.yml

static/

templates/

## **Описание функций модулей и функций.**

Список модулей:

* main.py;
* model.py;
* database\_manipulation.py
* create\_reports.py

main.py.

Основой файл, включающий в себя функционал из других модулей, а также endpoint`ы для работы приложения. В файле также происходит настройка параметров для сервера и его запуск. Используемые методы запроса:

* GET – для отображения и рендеринга страниц;
* POST – для выполнения функциональной части программы.

model.py.

Модуль для работы с моделью (используется модель RandomForest):

* save\_f(filename, object\_to\_save) – функция для сохранения объекта в файл .pkl;
* prepare\_df(df\_name) – функция для чтения и подготовки датафрейма к обучению на нем модели;
* make\_model(df) – функция для создания и непосредственного обучения модели;
* reteach\_model() – функция для переобучения модели с последующим созданием всех отчетов;
* load\_model(model\_path, vectorizer\_path) – функция для загрузки модели для ее использования в анализе SMS-сообщения.

database\_manipulation.py.

В файле представлена функция add\_message(\_message, \_type, df\_path) для добавления новой записи в датасет.

create\_reports.py.

Модуль для работы с отчетами (создание и запись в файлы):

* create\_text\_report(report\_path, df\_path) – считывает датасет, создает на его основе отчет о датасете и записывает его в text\_report.txt;
* create\_graphic\_report(report\_path, df\_path) – считывает датасет, создает на его основе графический отчет по датасету и записывает его в graphic.png;
* create\_model\_report(conf\_matrix, report, accuracy, unhandled\_messages, report\_path) – получает на вход данные о результатах обучения модели и записывает их в model\_report.txt.

## **Используемые библиотеки**

|  |  |
| --- | --- |
| Библиотека | Версия |
| pandas | 2.2.3 |
| sklearn | 1.0.0 |
| matplotlib | 3.7.0 |
| zipfile | -стандартная библиотека |
| pickle | -стандартная библиотека |
| pathlib | -стандартная библиотека |
| fasapi | 0.68.0 |
| os | -стандартная библиотека |
| jinja2 | 3.0.0 |

Python 3.11, pip 21.0

## **Визуальная часть приложения.**

Для визуального отображения приложения используются шаблоны, расположенные в /templates и стили со скриптами к ним в /static. В приложении использован шаблонизатор Jinja2.

templates/

addinfo.html - шаблон страницы для добавления записи

index.html - шаблон главной страницы

reports.html - шаблон страницы для загрузки отчетов

settings.html - шаблон страницы с настройками приложения

static/

style.css - файл со стилями

settings.js - файл со скриптом для настройки интерфейса