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 системы поиска шаблонов в данных портала GitHub

Научный руководитель:

профессор кафедры СП,

д.ф.-м.н., доцент

М.Л. Цымблер

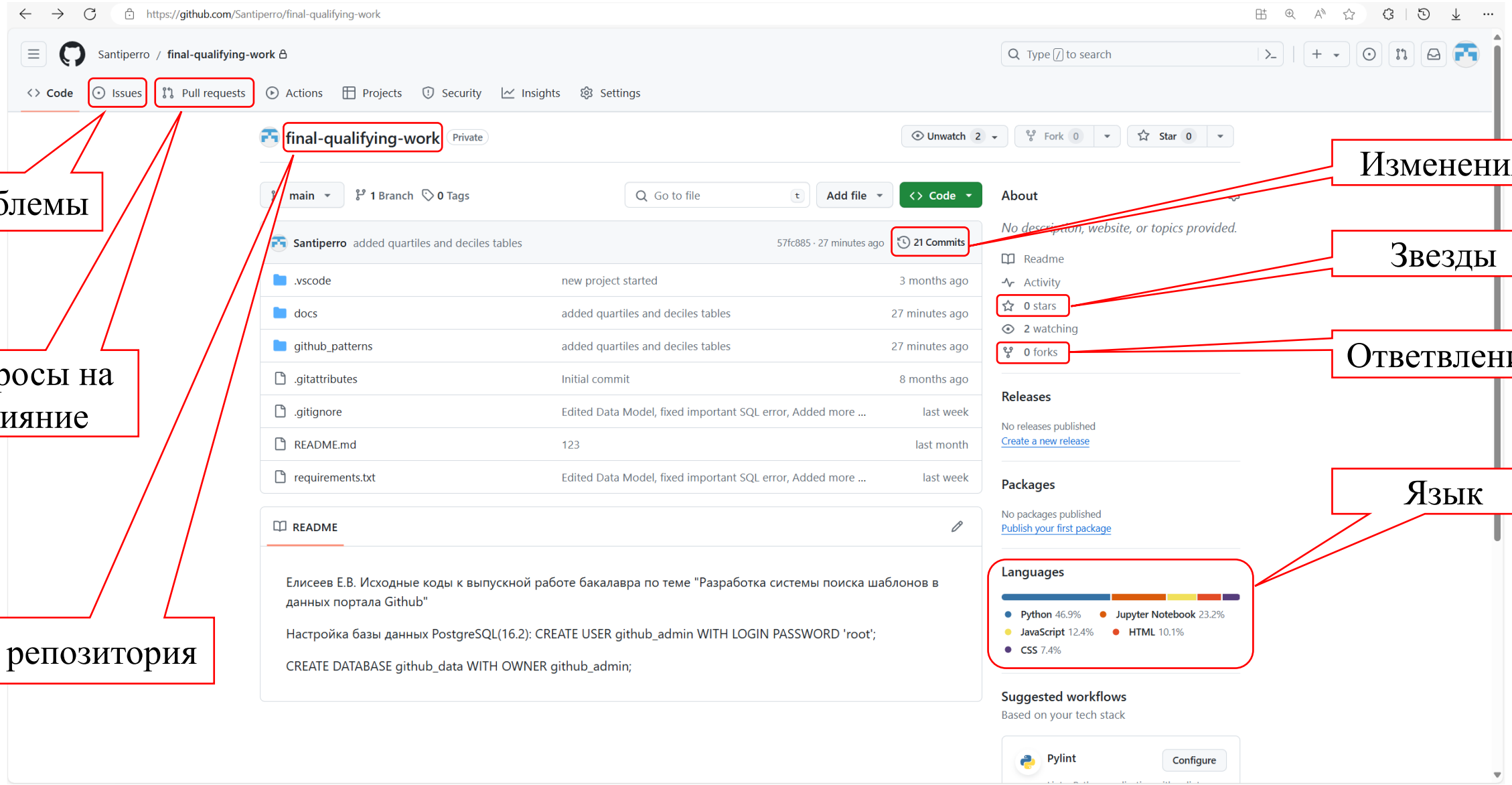
Автор работы:

студент группы КЭ-403

Е.В. Елисеев

Разработка системы поиска шаблонов в данных портала GitHub

АКТУАЛЬНОСТЬ



ЦЕЛИ И ЗАДАЧИ

Цель:
Разработка системы поиска шаблонов в данных портала GitHub.

Задачи:
1. Провести анализ и спецификацию предметной области.
2. Выполнить проектирование пользовательского интерфейса и архитектуры системы.
3. Реализовать систему поиска шаблонов.
4. Провести тестирование системы поиска шаблонов.
5. Провести эксперименты с применением разработанной системы

АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

Основные определения

Шаблон

$X \to Y$,

где $X \neq \emptyset, Y \neq \emptyset, X \cap Y = \emptyset$

Поддержка

$sup(X \to Y) = P(X \cup Y)$

Достоверность

$conf(X \to Y) = P(Y \backslash X)$

Шаблон $X \to Y$ устойчив,
если $sup(X \to Y) \geq minsup$
и $conf(X \to Y) \geq minconf$

Пример поиска шаблонов

База транзакций

№	Изменения	Проблемы	Звезды
1	1	4	1
2	1	2	1
3	2	1	3
4	2	1	2
5	1	4	1

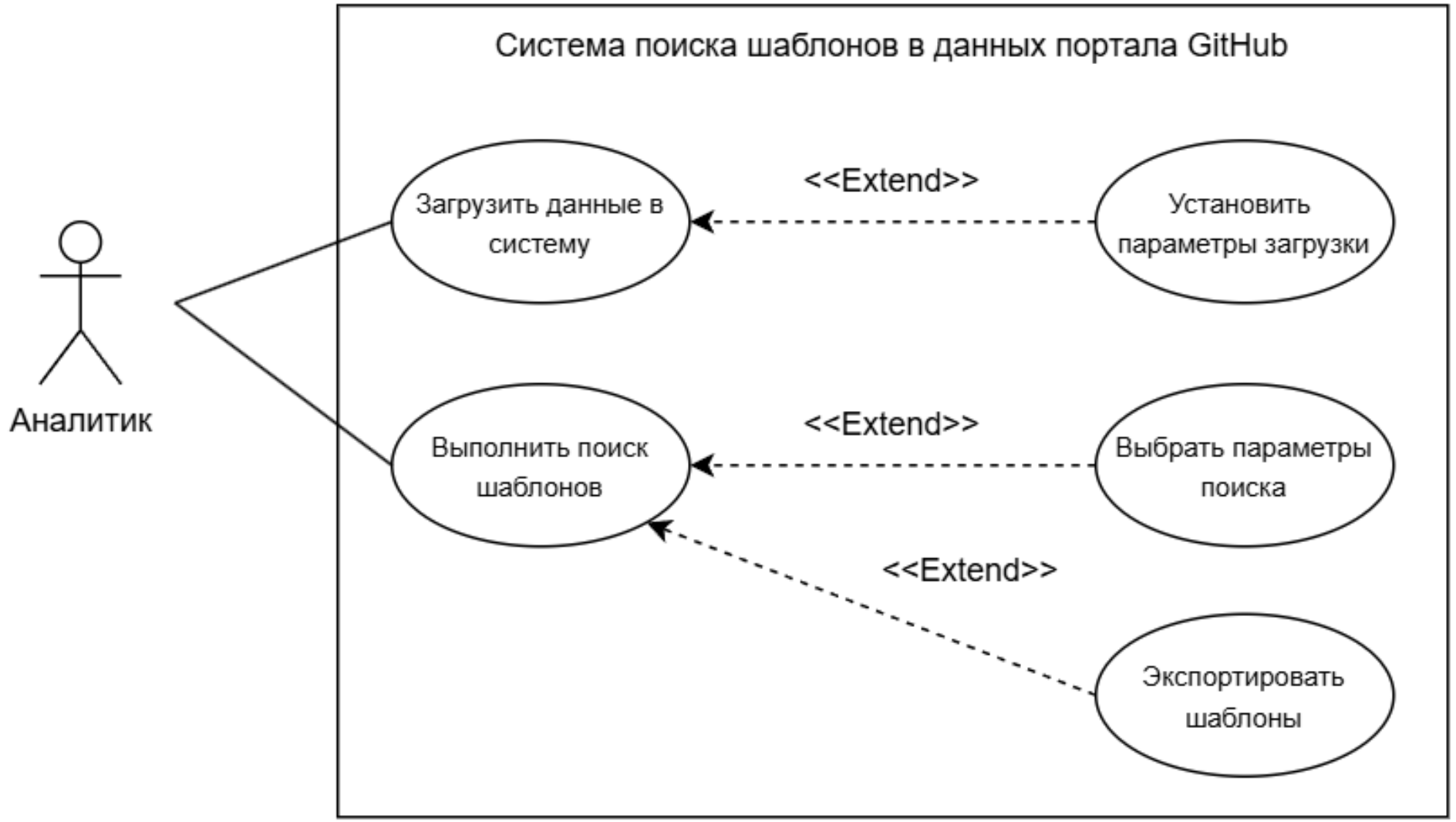
$minsup = 0.1$
$minconf = 0.5$

Устойчивый шаблон

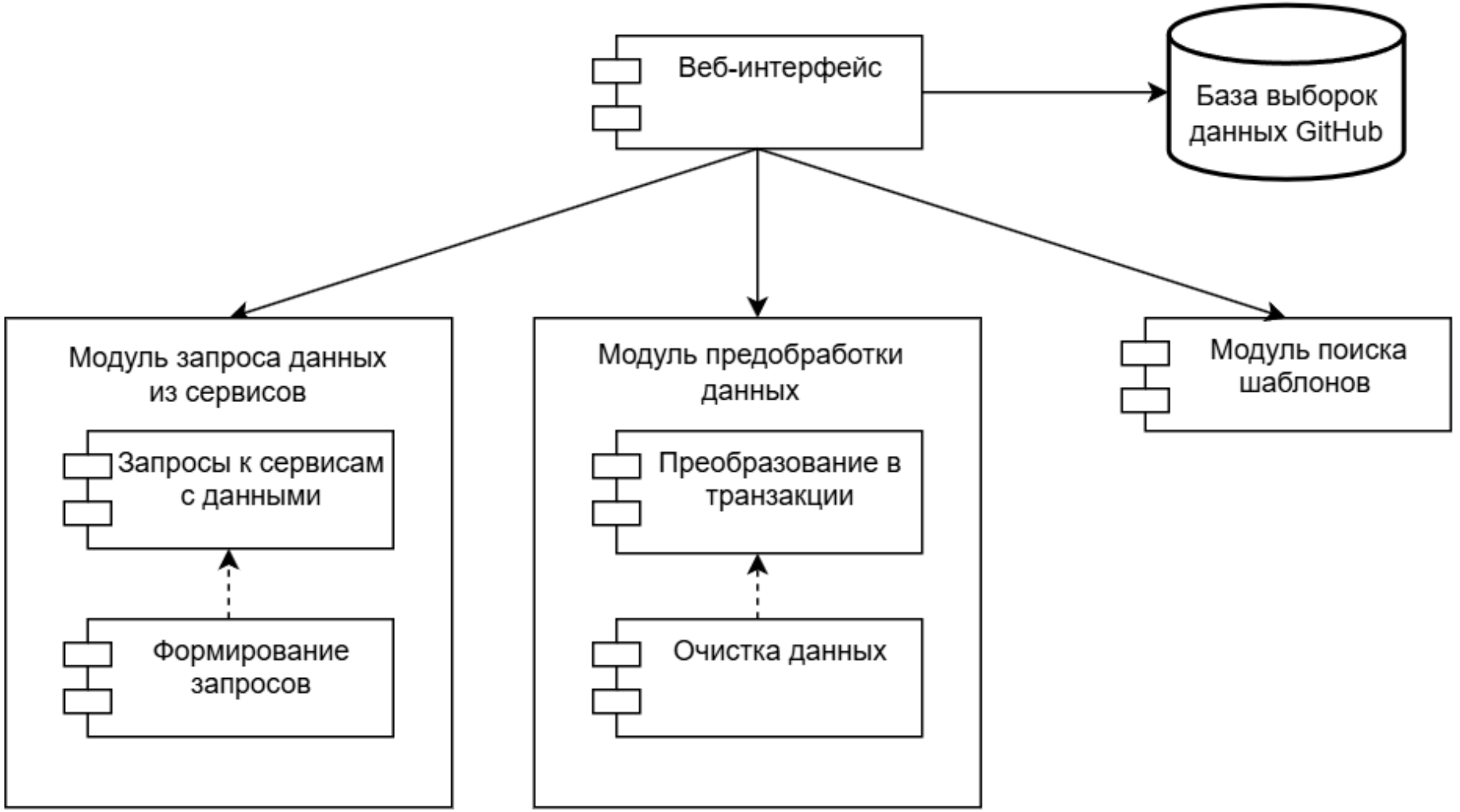
Изменения = 1, Проблемы = 4 → Звезды = 1

$sup = 0.4$
$conf = 1$

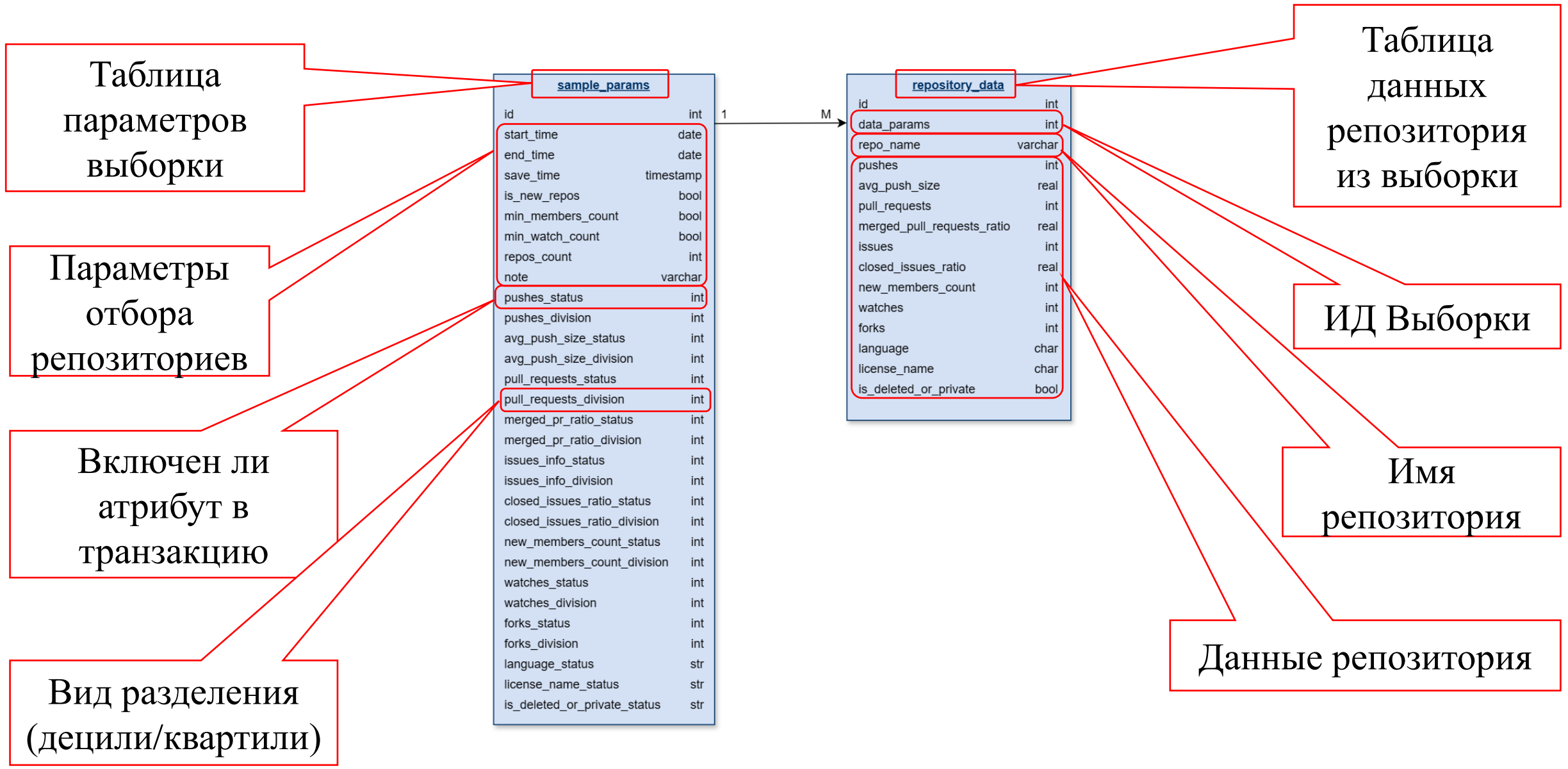
ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ



АРХИТЕКТУРА



СТРУКТУРА БАЗЫ ДАННЫХ



РЕАЛИЗАЦИЯ СИСТЕМЫ

Язык программирования: Python 3.11.7

Библиотеки:

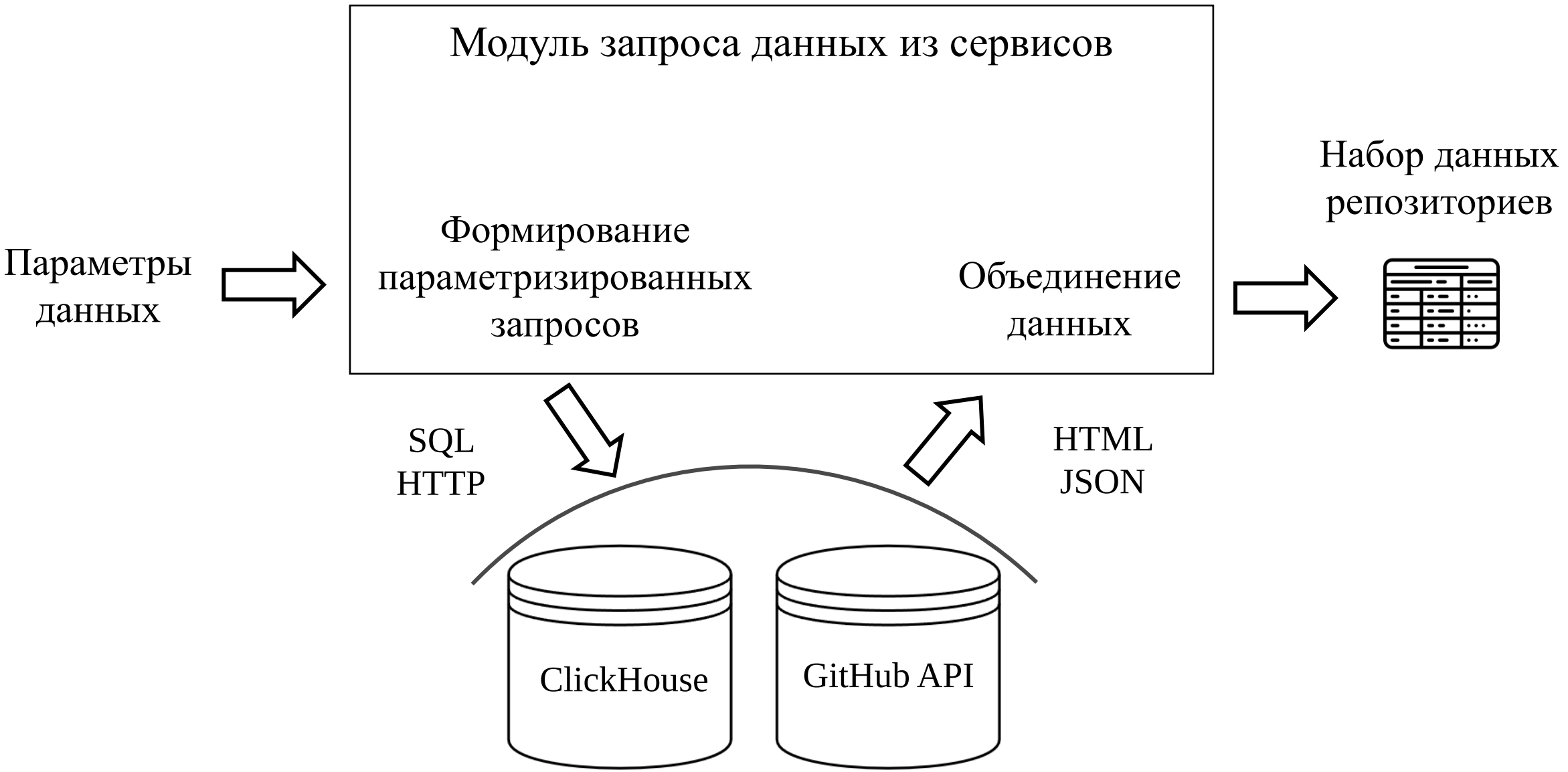
Веб-интерфейс	Django
Запрос данных из сервисов	Aiohttp Asyncio Selenium
Предобработка данных	Pandas Numpy
Поиск шаблонов	Mlxtend

Верстка: HTML, CSS, JavaScript

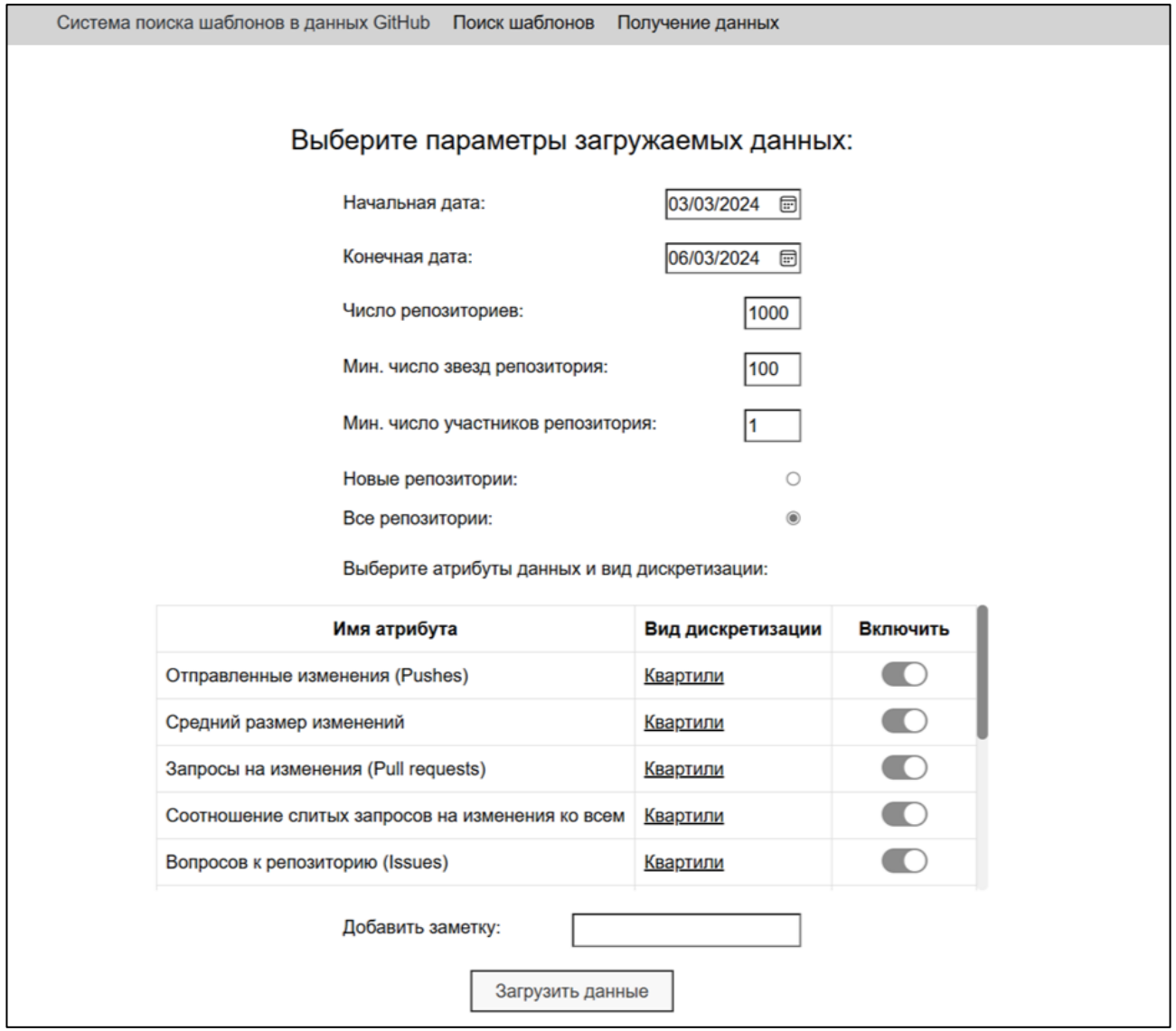
СУБД: PostgreSQL 16.2

Исходные коды:

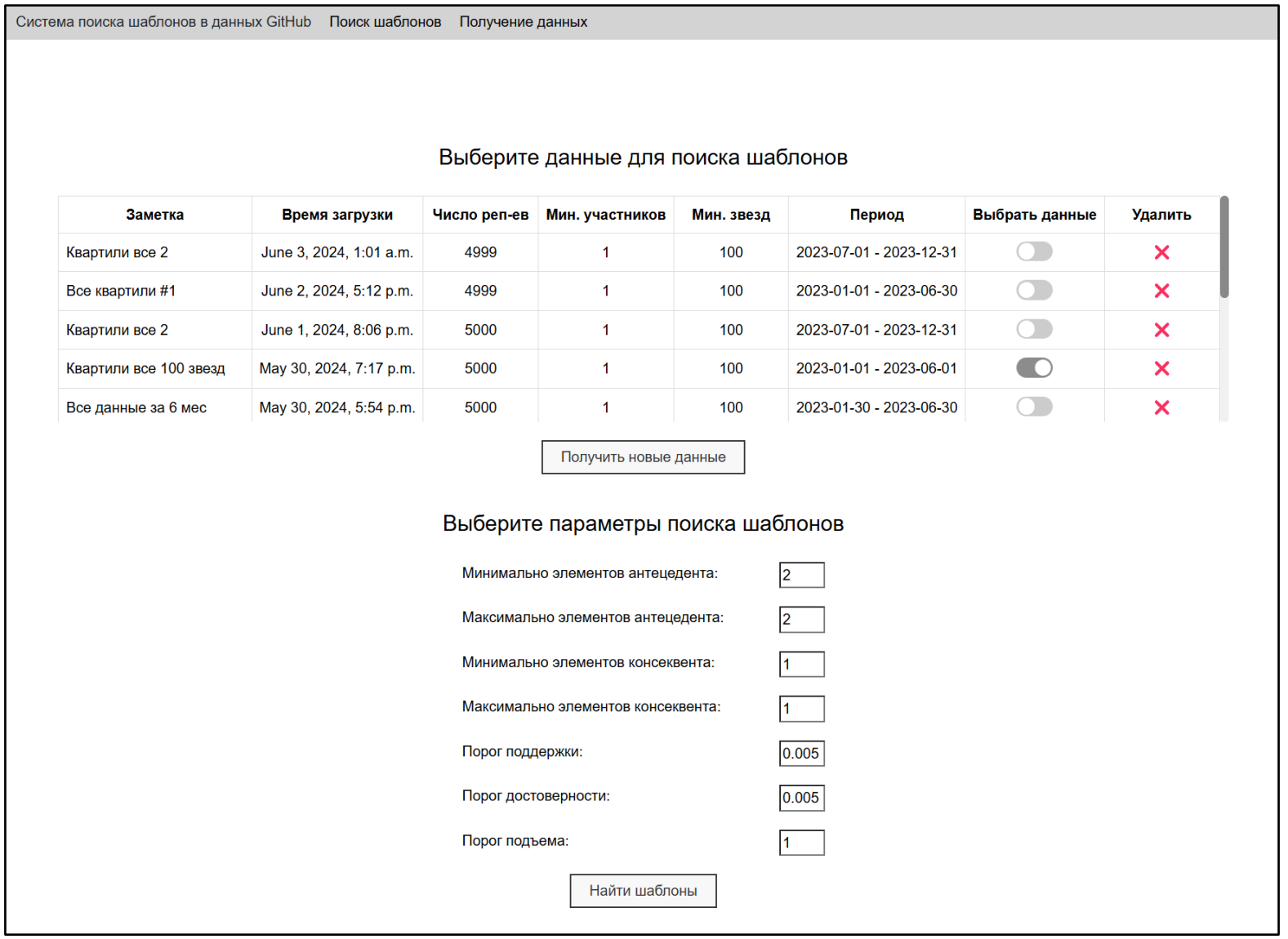
РЕАЛИЗАЦИЯ МОДУЛЯ ЗАПРОСА ДАННЫХ



ПОЛУЧЕНИЕ ДАННЫХ



ПОИСК ШАБЛОНОВ



РЕЗУЛЬТАТ ПОИСКА ШАБЛОНОВ

Найдено 838 шаблонов:

Антецедент: язык	Консеквент: лицензия	Поддержка	Достоверность	Подъем
(Язык Python)	(Лицензия MIT License)	0.0394	0.3512	1.7244
(Язык TypeScript)	(Лицензия MIT License)	0.0348	0.5859	2.8769
(Язык Python)	(Лицензия Apache License 2.0)	0.0226	0.2014	2.1379
(Язык JavaScript)	(Лицензия MIT License)	0.0224	0.4226	2.0754
(Язык Go)	(Лицензия MIT License)	0.0152	0.3762	1.8476
(Язык Go)	(Лицензия Apache License 2.0)	0.0146	0.3614	3.8356
(Язык Python)	(Лицензия Other)	0.0136	0.1212	1.9998
(Язык Python)	(Лицензия GNU General Public License v3.0)	0.0108	0.0963	1.9019
(Язык Java)	(Лицензия Apache License 2.0)	0.0096	0.378	4.0114
(Язык Rust)	(Лицензия Apache License 2.0)	0.0086	0.3583	3.8032

Скачать шаблоны

ТЕСТИРОВАНИЕ И ЭКСПЕРИМЕНТЫ

Выполнено 10 функциональных тестов

Выполнены вычислительные эксперименты:

1. Исследование зависимости времени поиска шаблонов от порога поддержки
2. Исследование зависимости времени поиска шаблонов от размера базы транзакций

Выполнен поиск шаблонов с целью исследования влияния языка программирования и лицензии на популярность репозитория

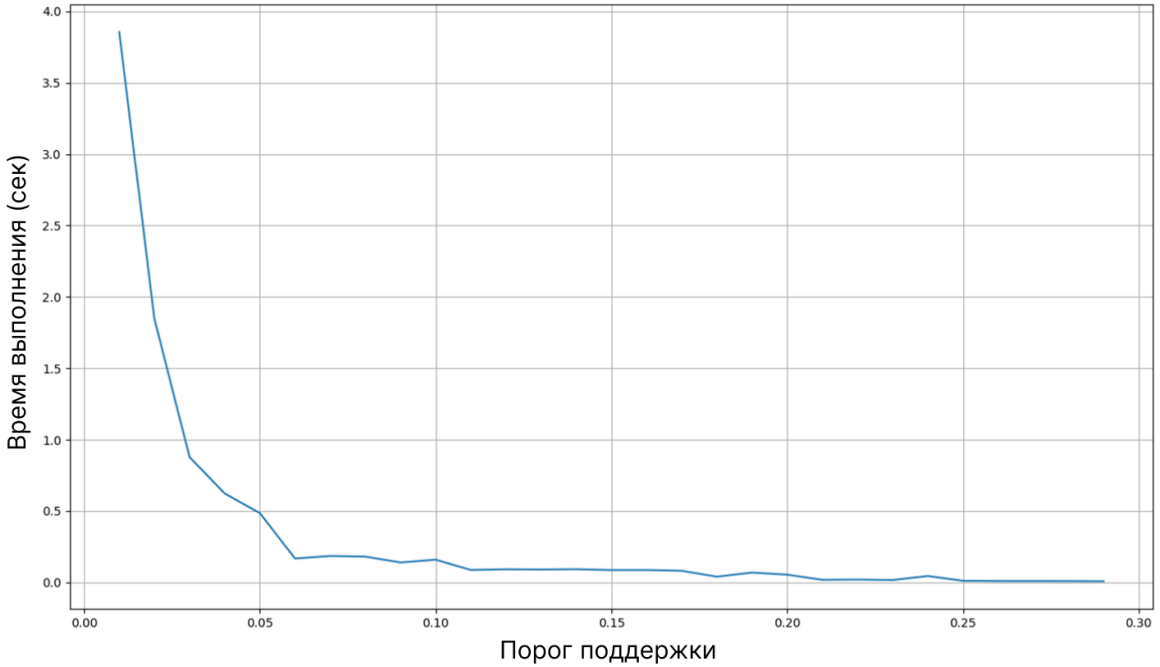
Наборы данных

№	Временной промежуток	Число репозиториев	Минимум звезд	supp	conf
1	01-01-2023 – 30-06-2023	5000	100	0.01 – 0.3	0.01
2	01-01-2023 – 30-06-2023	7000	100	0.01	0.01
3	01-01-2023 – 30-06-2023	10000	100	0.002	0.01

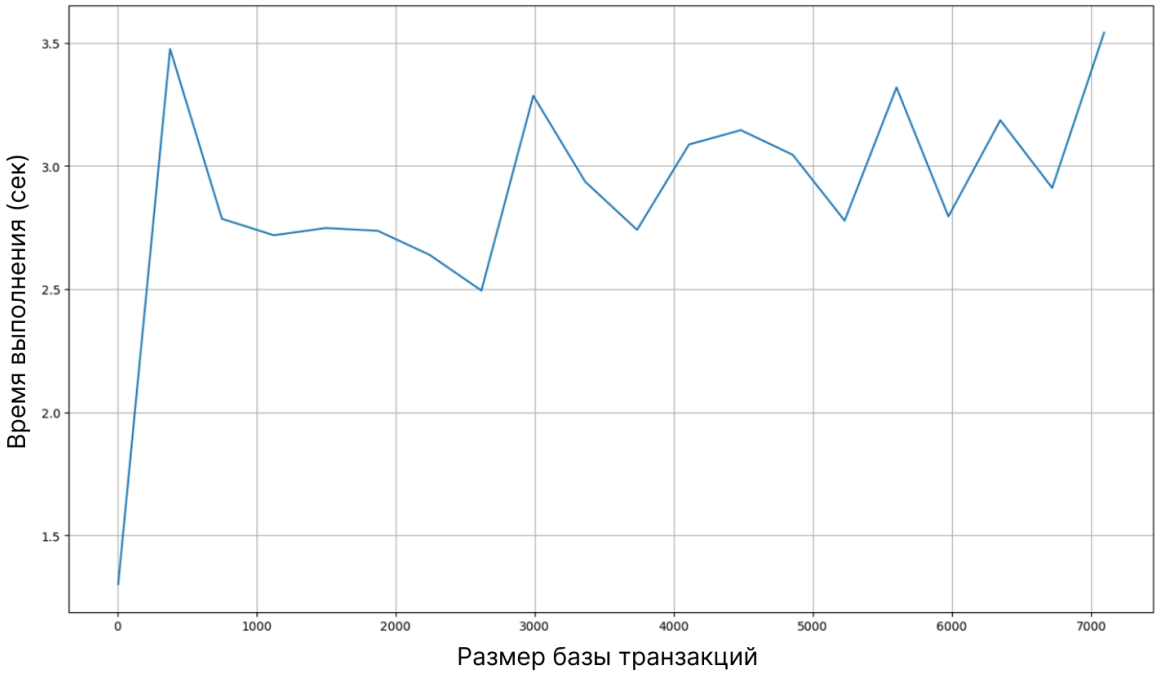
Характеристики системы

Компонент	Спецификация
ОС	Windows 11
Версия языка Python	3.11.7
Процессор	Intel Core i7-12700f
Оперативная память	32 Гб

ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ



Зависимость времени поиска шаблонов от порога поддержки



Зависимость времени поиска шаблонов от размера базы транзакций

ПОЛЕЗНЫЕ ШАБЛОНЫ

№	Если		То	Метрики	
	Язык	Лицензия	Популярность	supp	conf
1	Python	MIT	Звезды 1	0.0132	0.3350
2	TypeScript	MIT	Звезды 1	0.0116	0.3333
3	Go	Apache 2.0	Звезды 1	0.0070	0.4795
4	C++	MIT	Звезды 4	0.0030	0.3947
5	C#	MIT	Звезды 3	0.0026	0.3171
6	C	MIT	Звезды 4	0.0024	0.4444

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проведен анализ предметной области
2. Выполнено проектирование системы поиска шаблонов
3. Реализована система поиска шаблонов
4. Проведено тестирование системы поиска шаблонов
5. Проведен ряд экспериментов, которые оценили эффективность системы

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Шаблон

$X \to Y$,

где $X \neq \emptyset, Y \neq \emptyset, X \cap Y = \emptyset$

Шаблон $X \to Y$ устойчив,

если $sup(X \to Y) \geq minsup$

и $conf(X \to Y) \geq minconf$

Поддержка

$$supp(X \to Y) = \frac{|\{t \in D \mid (X \cup Y) \subseteq t\}|}{|D|}$$

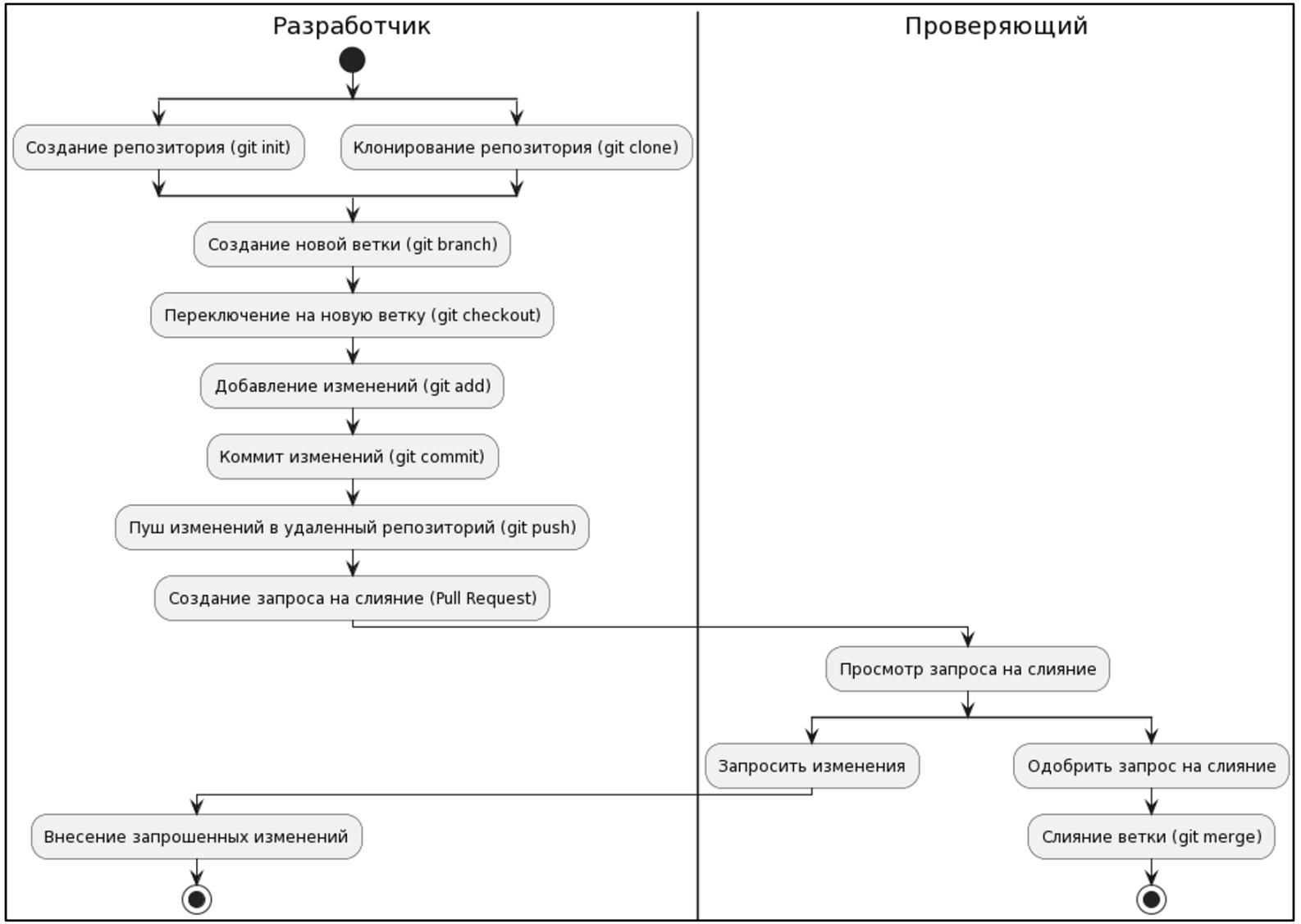
Достоверность

$$conf(X \to Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Подъем

$$lift(X \to Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$$

ПРИМЕР СЦЕНАРИЯ РАБОТЫ С GITHUB

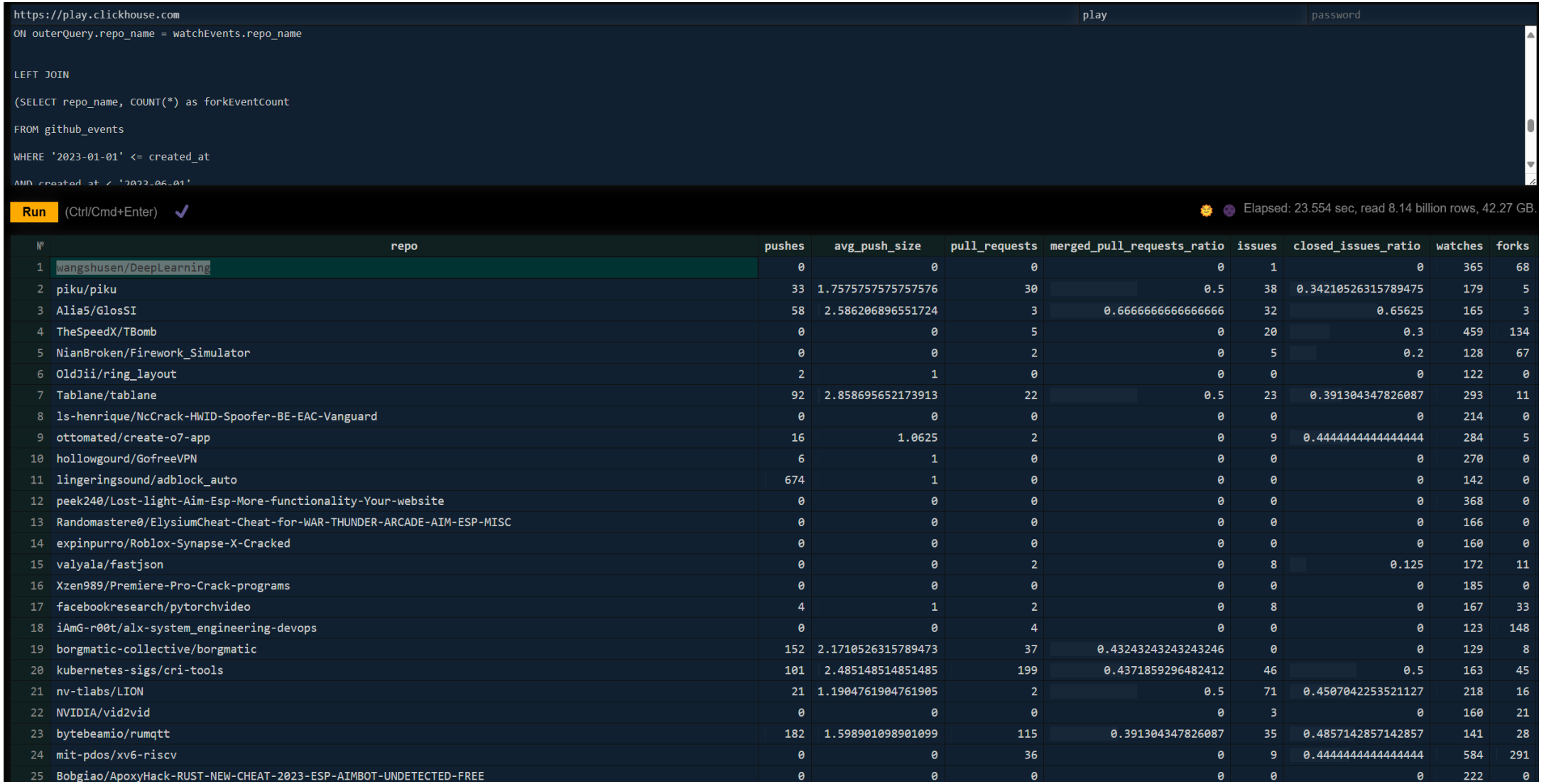


Разработка системы поиска шаблонов в данных портала GitHub

СЕМАНТИКА СОБЫТИЙ GITHUB

№	Команды	Семантика
1	init	Инициализация нового Git репозитория
2	add	Добавление файлов в индекс для последующего коммита
3	commit	Фиксация изменений в репозитории
4	push	Отправка изменений в удаленный репозиторий
5	merge	Слияние изменений из одной ветки в другую
6	branch	Создание новой ветки в репозитории
7	checkout	Переключение между ветками или коммитами в репозитории
8	clone	Клонирование удаленного репозитория на локальную машину

CLICKHOUSE PLAYGROUND



№	repo	pushes	avg_push_size	pull_requests	merged_pull_requests_ratio	issues	closed_issues_ratio	watches	forks
1	wangshusen/DeepLearning	0	0	0	0	1	0	365	68
2	piku/piku	33	1.7575757575757576	30	0.5	38	0.34210526315789475	179	5
3	Alia5/GlosSI	58	2.586206896551724	3	0.6666666666666666	32	0.65625	165	3
4	TheSpeedX/TBomb	0	0	5	0	20	0.3	459	134
5	NianBroken/Firework_Simulator	0	0	2	0	5	0.2	128	67
6	OldJii/ring_layout	2	1	0	0	0	0	122	0
7	Tablane/tablane	92	2.858695652173913	22	0.5	23	0.391304347826087	293	11
8	ls-henrique/NcCrack-HWID-Spoofer-BE-EAC-Vanguard	0	0	0	0	0	0	214	0
9	ottomated/create-o7-app	16	1.0625	2	0	9	0.4444444444444444	284	5
10	hollowgourd/GofreeVPN	6	1	0	0	0	0	270	0
11	lingeringsound/adblock_auto	674	1	0	0	0	0	142	0
12	peek240/Lost-light-Aim-Esp-More-functionality-Your-website	0	0	0	0	0	0	368	0
13	Randomastere0/ElysiumCheat-Cheat-for-WAR-THUNDER-ARCADE-AIM-ESP-MISC	0	0	0	0	0	0	166	0
14	expinpurro/Roblox-Synapse-X-Cracked	0	0	0	0	0	0	160	0
15	valyala/fastjson	0	0	2	0	8	0.125	172	11
16	Xzen989/Premiere-Pro-Crack-programs	0	0	0	0	0	0	185	0
17	facebookresearch/pytorchvideo	4	1	2	0	8	0	167	33
18	iAmG-r00t/alx-system_engineering-devops	0	0	4	0	0	0	123	148
19	borgmatic-collective/borgmatic	152	2.1710526315789473	37	0.43243243243243246	0	0	129	8
20	kubernetes-sigs/cri-tools	101	2.485148514851485	199	0.4371859296482412	46	0.5	163	45
21	nv-tlabs/LION	21	1.1904761904761905	2	0.5	71	0.4507042253521127	218	16
22	NVIDIA/vid2vid	0	0	0	0	3	0	160	21
23	bytebeamio/rumqtt	182	1.598901098901099	115	0.391304347826087	35	0.4857142857142857	141	28
24	mit-pdos/xv6-riscv	0	0	36	0	9	0.4444444444444444	584	291
25	Bobgiao/ApoxyHack-RUST-NEW-CHEAT-2023-ESP-AIMBOT-UNDETECTED-FREE	0	0	0	0	0	0	222	0

Разработка системы поиска шаблонов в данных портала GitHub

GITHUB API

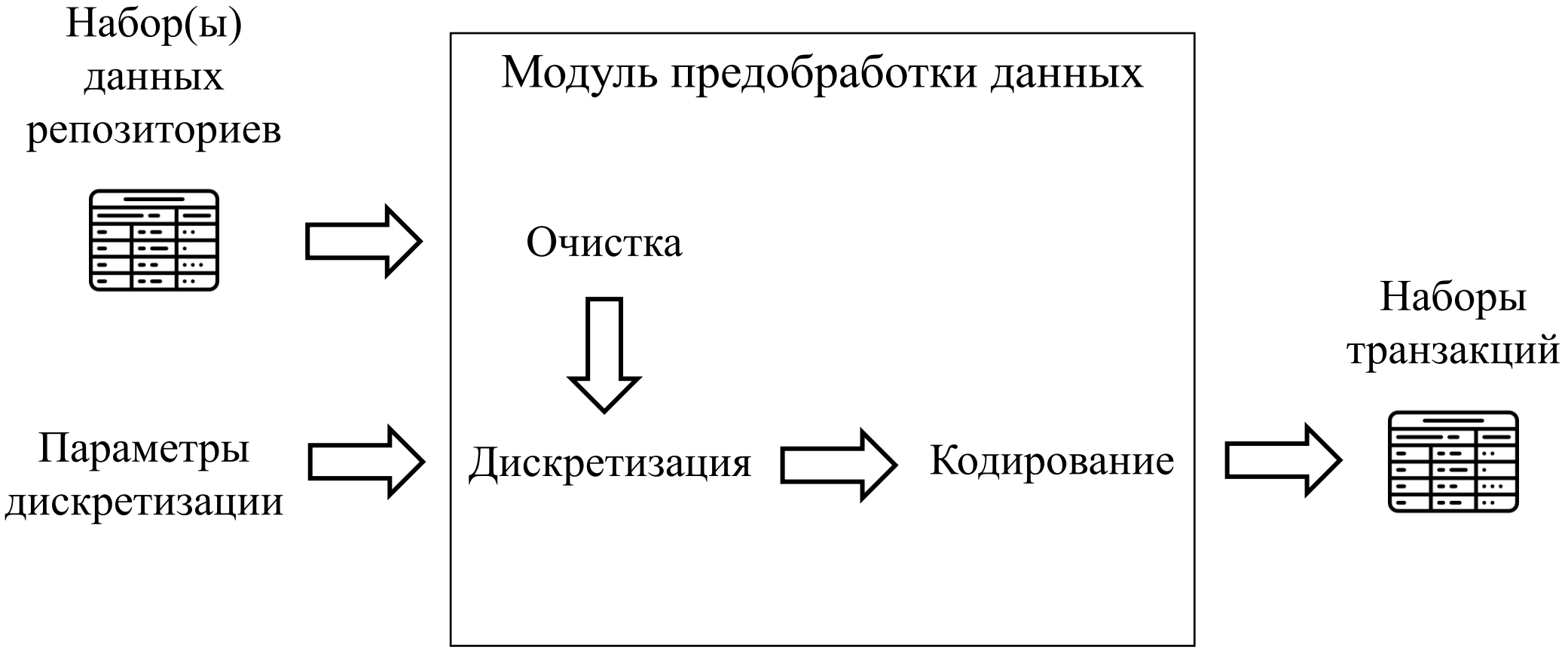

```
{
    "id": 180447794,
    "node_id": "MDEwOlJlcG9zaXRvcnkxODA0NDc3OTQ=",
    "name": "DeepLearning",
    "full_name": "wangshusen/DeepLearning",
    "private": false,
    "owner": {
        "login": "wangshusen",
        "id": 12660689,
        "node_id": "MDQ6VXNlcjEyNjYwNjg5",
        "avatar_url": "https://avatars.githubusercontent.com/u/12660689?v=4",
        "gravatar_id": "",
        "url": "https://api.github.com/users/wangshusen",
        "html_url": "https://github.com/wangshusen",
        "followers_url": "https://api.github.com/users/wangshusen/followers",
        "following_url": "https://api.github.com/users/wangshusen/following{/other_user}",
        "gists_url": "https://api.github.com/users/wangshusen/gists{/gist_id}",
        "starred_url": "https://api.github.com/users/wangshusen/starred{/owner}{/repo}",
        "subscriptions_url": "https://api.github.com/users/wangshusen/subscriptions",
        "organizations_url": "https://api.github.com/users/wangshusen/orgs",
        "repos_url": "https://api.github.com/users/wangshusen/repos",
        "events_url": "https://api.github.com/users/wangshusen/events{/privacy}",
        "received_events_url": "https://api.github.com/users/wangshusen/received_events",
        "type": "User",
        "site_admin": false
    },
    "html_url": "https://github.com/wangshusen/DeepLearning",
    "description": null,
    "fork": false,
    "url": "https://api.github.com/repos/wangshusen/DeepLearning",
    "forks_url": "https://api.github.com/repos/wangshusen/DeepLearning/forks",
    "keys_url": "https://api.github.com/repos/wangshusen/DeepLearning/keys{/key_id}",
    "collaborators_url": "https://api.github.com/repos/wangshusen/DeepLearning/collaborators{/collaborator}",
    "teams_url": "https://api.github.com/repos/wangshusen/DeepLearning/teams",
    "hooks_url": "https://api.github.com/repos/wangshusen/DeepLearning/hooks",
    "issue_events_url": "https://api.github.com/repos/wangshusen/DeepLearning/issues/events{/number}",
    "events_url": "https://api.github.com/repos/wangshusen/DeepLearning/events",
    "assignees_url": "https://api.github.com/repos/wangshusen/DeepLearning/assignees{/user}",
    "branches_url": "https://api.github.com/repos/wangshusen/DeepLearning/branches{/branch}",
    "tags_url": "https://api.github.com/repos/wangshusen/DeepLearning/tags",
    "blobs_url": "https://api.github.com/repos/wangshusen/DeepLearning/git/blobs{/sha}",
    "git_tags_url": "https://api.github.com/repos/wangshusen/DeepLearning/git/tags{/sha}",
    "git_refs_url": "https://api.github.com/repos/wangshusen/DeepLearning/git/refs{/sha}",
    "trees_url": "https://api.github.com/repos/wangshusen/DeepLearning/git/trees{/sha}",
    "statuses_url": "https://api.github.com/repos/wangshusen/DeepLearning/statuses/{sha}",
    "languages_url": "https://api.github.com/repos/wangshusen/DeepLearning/languages",
    "stargazers_url": "https://api.github.com/repos/wangshusen/DeepLearning/stargazers",
    "contributors_url": "https://api.github.com/repos/wangshusen/DeepLearning/contributors",
    "subscribers_url": "https://api.github.com/repos/wangshusen/DeepLearning/subscribers",
    "subscription_url": "https://api.github.com/repos/wangshusen/DeepLearning/subscription",
    "commits_url": "https://api.github.com/repos/wangshusen/DeepLearning/commits{/sha}",
```


Разработка системы поиска шаблонов в данных портала GitHub

ДАННЫЕ ДЛЯ ПОИСКА ШАБЛОНОВ

Атрибут	Сервис	Тип данных	Семантика
pushes	Clickhouse	Целое Число	Количество отправленных изменений (pushes) в репозиторий
avg_push_size	Clickhouse	Вещественное число	Средний размер отправленных изменений (pushes)
pull_requests	Clickhouse	Целое Число	Количество запросов на внесение изменений (pull requests)
merged_pull_requests_ratio	Clickhouse	Вещественное число	Отношение успешно принятых запросов на внесение изменений (pull requests) к общему числу таких запросов
issues	Clickhouse	Целое Число	Количество созданных вопросов или проблем (issues)
new_members_count	Clickhouse	Целое Число	Количество новых участников репозитория
closed_issues_ratio	Clickhouse	Вещественное число	Отношение успешно закрытых вопросов или проблем (issues) к общему числу таких вопросов
watches	Clickhouse	Целое Число	Количество пользователей, отслеживающих обновления репозитория (поставивших звезду)
forks	Clickhouse	Целое Число	Количество ветвлений (forks) репозитория
language	GitHub API	Строка	Язык программирования, на котором написан код в репозитории
license_name	GitHub API	Строка	Название лицензии, под которой распространяется код в репозитории
is_deleted_or_private	GitHub API	Булево	Индикатор, указывающий, удален репозиторий или является ли он приватным

РЕАЛИЗАЦИЯ МОДУЛЯ ПРЕДОБРАБОТКИ ДАННЫХ



ФУНКЦИОНАЛЬНОЕ ТЕСТИРОВАНИЕ

№	Название теста	Шаги	Ожидаемый результат	Тест пройден?
1.	Поиск шаблонов при некорректных параметрах.	На странице поиска шаблонов нужно выполнить следующие шаги. 1. Задать отрицательные или вещественные значения в полях «Минимально элементов антецедента», «Максимально элементов антецедента», «Минимально элементов консеквента», «Максимально элементов консеквента». 2. Задать отрицательное число или число больше 1 в поле «Порог поддержки» или «Порог достоверности». 3. Задать отрицательное число в поле «Порог подъема». 4. Нажать на кнопку «Найти шаблоны».	При задаче каждого некорректного параметра будет показываться уведомление и выделяться поле некорректного параметра. Поиск шаблонов не будет выполнен.	Да