Анализ репозиториев Git

HR

Первичный этап отбора откликов соискателей

- Соответствие используемых ЯП заявленным в резюме
- График работы над проектами
- Оценка вклада соискателя в проектах

Профиль соискателя - аккаунт на GitHub

Team Lead

Отслеживание эффективности работы сотрудников

- Действия сотрудника в проекте
- Нагрузка сотрудников по проекту
- Нагрузка конкретного сотрудника

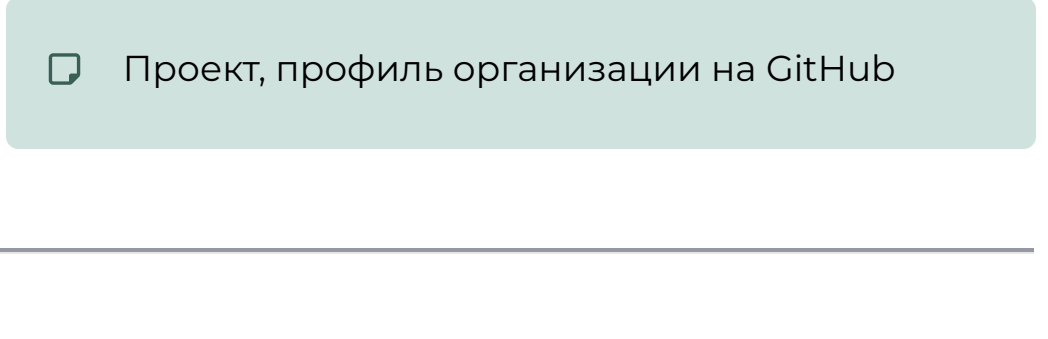
Проект, профиль организации на GitHub

Ожидаемые результаты

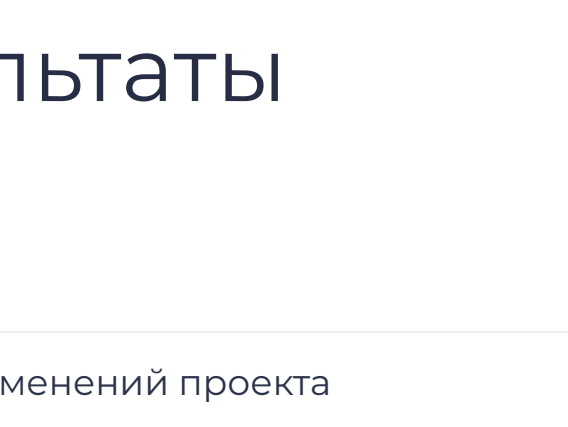
Этап визуализации

Использование языков программирования

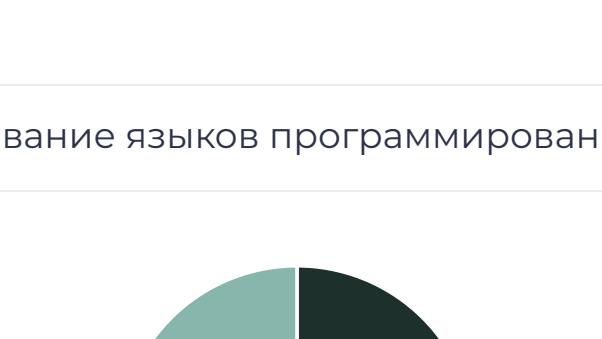
График изменений проекта



Типы изменений



Типы изменений по репозиториям



Статистика работы

Этапы разработки

ETL

Входные данные

Этап 1

- Имя пользователя на github.com
- Путь к директории, в которую будут склонированы репозитории пользователя

username

C:/Users/Admin/

Этап 2

- Ссылка для клонирования git-репозитория
- Путь к директории, в которую будет склонирован репозиторий

https://github.com/tidyverse/ggplot2.git

C:/Users/Admin/

Этап 3

- Путь к директории с локальным репозиторием

C:/Users/Admin/Repo

Этап 1 → Этап 2

1. MariUt005

Пользователь вводит имя пользователя на Github

2. **Ссылка**

Имя пользователя подставляется в ссылку GitHub API, после чего отправляется запрос

https://api.github.com/users/MariUt005/repos

3. **JSON**

В ответ на запрос приходит JSON-файл, содержащий информацию про все публичные репозитории пользователя на Github

```
{
  "id": 939957982,
  "node_id": "R_kgDOOAae3g",
  "name": "git-analysis",
  "full_name": "MariUt005/git-analysis",
  "private": false,
```

4. **Ссылка на репозиторий**

У каждого репозитория в JSON указана ссылка для клонирования

"clone_url": "https://github.com/MariUt005/git-analysis.git"

5. **Список ссылок**

Все ссылки для клонирования каждого репозитория пользователя

```
 [1] "https://github.com/MariUt005/book_reader.git"
 [2] "https://github.com/MariUt005/dotfiles.git"
 [3] "https://github.com/MariUt005/Enlightsion.git"
 [4] "https://github.com/MariUt005/exam_mqwerty.git"
 [5] "https://github.com/MariUt005/git-analysis.git"
 [6] "https://github.com/MariUt005/learnGo.git"
 [7] "https://github.com/MariUt005/learnWeb.git"
 [8] "https://github.com/MariUt005/magma.git"
 [9] "https://github.com/MariUt005/MersOff.git"
[10] "https://github.com/MariUt005/mqwerty.git"
[11] "https://github.com/MariUt005/mqwerty_bot.git"
[12] "https://github.com/MariUt005/qq.git"
[13] "https://github.com/MariUt005/regex.git"
[14] "https://github.com/MariUt005/reports_i2z1.git"
[15] "https://github.com/MariUt005/Students_db.git"
[16] "https://github.com/MariUt005/tetris.git"
[17] "https://github.com/MariUt005/tic-tac-toe.git"
[18] "https://github.com/MariUt005/TMP.git"
```

Этап 2 → Этап 3

1. **Ссылка на репозиторий**

https://github.com/MariUt005/git-analysis.git

2. **Имя репозитория**

git-analysis

3. **Путь для клонирования**

C:/Users/Admin/git-analysis

4. **Клонирование репозитория**

git clone {repo_url} {repo_path}

Обновление репозитория

git -C {repo_path} pull {repo_url}

Этап 3. Краткая история коммитов

1. Новый ли репозиторий?

- Проверка, есть ли этот репозиторий в базе данных
- Если да, то получаем номер последнего коммита last_commit

2. **Получение истории коммитов**

git -C {repo_path} log --format="%H\t%P\t%an\t%ai\t%s" --all

git -C {repo_path} log {last_commit}..HEAD --format="%H\t%P\t%an\t%ai\t%s" --all

```
c1f3475f3348faa6e2c6115eeec01e11c0f10b38\t2d49d9c9dd7d77f91c906869548ae799938cc99d\tMaria\t2025-03-30 00:45:05 +0300\tChanges in files
2d49d9c9dd7d77f91c906869548ae799938cc99d\tfa3a78ff139856b542759417fadad2cd76b05461\tMaria\t2025-03-30 00:29:50 +0300\tChanges in files
fa3a78ff139856b542759417fadad2cd76b05461\tf1b13932a6e5dd9d8f3873514f53880f56547511\tMaria\t2025-03-26 22:38:01 +0300\tGit commit history
f1b13932a6e5dd9d8f3873514f53880f56547511\ta4fee3e2f8ebd87aeb31d86a962360166aea9b15\tMaria\t2025-03-26 22:36:42 +0300\tGit commit history
a4fee3e2f8ebd87aeb31d86a962360166aea9b15\tf418704449370c5ba16131037f6eed177f018247\tgigwrld\t2025-03-15 15:39:13 +0300\tUpdate README.md
f418704449370c5ba16131037f6eed177f018247\tde0e51208d93be7e1ff7af16f4404011d89bb5ac\tgigwrld\t2025-03-15 15:37:53 +0300\tAdd files via upload
de0e51208d93be7e1ff7af16f4404011d89bb5ac\te60d36cd53e97fb6447967e85f025709b6f55807\tMariUt005\t2025-03-01 13:58:00 +0300\tUpdate todo.md
e60d36cd53e97fb6447967e85f025709b6f55807\t934e9b638be6e8c0a4af353ea6b5cb18dc3f61d9\tMariUt005\t2025-03-01 13:32:54 +0300\tUpdate todo.md
934e9b638be6e8c0a4af353ea6b5cb18dc3f61d9\t73ae66ab58ba0f97c9388647a3c644ba6f108d6d\tMariUt005\t2025-03-01 13:22:49 +0300\tUpdate todo.md
73ae66ab58ba0f97c9388647a3c644ba6f108d6d\t0096aaee29bfe53ee147f53bcca54725390150b3\tMariUt005\t2025-03-01 13:17:49 +0300\tUpdate todo.md
0096aaee29bfe53ee147f53bcca54725390150b3\t18ae95e1b691b281af3830cda6aa2b87ea525d07\tMariUt005\t2025-03-01 13:14:55 +0300\tCreate todo.md
18ae95e1b691b281af3830cda6aa2b87ea525d07\te0640a396a3d1e2abf540a6772cb3bc7e38ba52e\tMariUt005\t2025-03-01 12:52:02 +0300\tUpdate README.md
e0640a396a3d1e2abf540a6772cb3bc7e38ba52e\t7d90b470783c143ba2cd36f33dcfe4acbd4048a9\tMariUt005\t2025-03-01 12:51:51 +0300\tUpdate README.md
7d90b470783c143ba2cd36f33dcfe4acbd4048a9\tda9aa921a63bebd4eed7c7966168a4768647683b\tArikova Kristina\t2025-03-01 11:42:24 +0300\tadd images
da9aa921a63bebd4eed7c7966168a4768647683b\t326efbce72dd45b9bdef01373e394c4dc4c473e7\tArikova Kristina\t2025-02-27 17:07:56 +0300\tthe main point
326efbce72dd45b9bdef01373e394c4dc4c473e7\t\tMaria\t2025-02-27 14:44:03 +0300\tInit commit
```

3. **Преобразование в датасет**

- 1 строка = 1 запись
- В качестве разделителя используется табуляция

	commit	parent_commit	author	date	message
1	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	2d49d9c9dd7d77f91c906869548ae799938cc99d	Maria	2025-03-30 00:45:05 +0300	Changes in files
2	2d49d9c9dd7d77f91c906869548ae799938cc99d	fa3a78ff139856b542759417fadad2cd76b05461	Maria	2025-03-30 00:29:50 +0300	Changes in files
3	fa3a78ff139856b542759417fadad2cd76b05461	f1b13932a6e5dd9d8f3873514f53880f56547511	Maria	2025-03-26 22:38:01 +0300	Git commit history
4	f1b13932a6e5dd9d8f3873514f53880f56547511	a4fee3e2f8ebd87aeb31d86a962360166aea9b15	Maria	2025-03-26 22:36:42 +0300	Git commit history
5	a4fee3e2f8ebd87aeb31d86a962360166aea9b15	f418704449370c5ba16131037f6eed177f018247	gigwrld	2025-03-15 15:39:13 +0300	Update README.md
6	f418704449370c5ba16131037f6eed177f018247	de0e51208d93be7e1ff7af16f4404011d89bb5ac	gigwrld	2025-03-15 15:37:53 +0300	Add files via upload
7	de0e51208d93be7e1ff7af16f4404011d89bb5ac	e60d36cd53e97fb6447967e85f025709b6f55807	MariUt005	2025-03-01 13:58:00 +0300	Update todo.md
8	e60d36cd53e97fb6447967e85f025709b6f55807	934e9b638be6e8c0a4af353ea6b5cb18dc3f61d9	MariUt005	2025-03-01 13:32:54 +0300	Update todo.md
9	934e9b638be6e8c0a4af353ea6b5cb18dc3f61d9	73ae66ab58ba0f97c9388647a3c644ba6f108d6d	MariUt005	2025-03-01 13:22:49 +0300	Update todo.md
10	73ae66ab58ba0f97c9388647a3c644ba6f108d6d	0096aaee29bfe53ee147f53bcca54725390150b3	MariUt005	2025-03-01 13:17:49 +0300	Update todo.md
11	0096aaee29bfe53ee147f53bcca54725390150b3	18ae95e1b691b281af3830cda6aa2b87ea525d07	MariUt005	2025-03-01 13:14:55 +0300	Create todo.md
12	18ae95e1b691b281af3830cda6aa2b87ea525d07	e0640a396a3d1e2abf540a6772cb3bc7e38ba52e	MariUt005	2025-03-01 12:52:02 +0300	Update README.md
13	e0640a396a3d1e2abf540a6772cb3bc7e38ba52e	7d90b470783c143ba2cd36f33dcfe4acbd4048a9	MariUt005	2025-03-01 12:51:51 +0300	Update README.md
14	7d90b470783c143ba2cd36f33dcfe4acbd4048a9	da9aa921a63bebd4eed7c7966168a4768647683b	Arikova Kristina	2025-03-01 11:42:24 +0300	add images
15	da9aa921a63bebd4eed7c7966168a4768647683b	326efbce72dd45b9bdef01373e394c4dc4c473e7	Arikova Kristina	2025-02-27 17:07:56 +0300	the main point
16	326efbce72dd45b9bdef01373e394c4dc4c473e7		Maria	2025-02-27 14:44:03 +0300	Init commit

4. **Добавление названия репозитория**

	commit	parent_commit	author	date	message	repo
1	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	2d49d9c9dd7d77f91c906869548ae799938cc99d	Maria	2025-03-30 00:45:05 +0300	Changes in files	git-analysis
2	2d49d9c9dd7d77f91c906869548ae799938cc99d	fa3a78ff139856b542759417fadad2cd76b05461	Maria	2025-03-30 00:29:50 +0300	Changes in files	git-analysis
3	fa3a78ff139856b542759417fadad2cd76b05461	f1b13932a6e5dd9d8f3873514f53880f56547511	Maria	2025-03-26 22:38:01 +0300	Git commit history	git-analysis
4	f1b13932a6e5dd9d8f3873514f53880f56547511	a4fee3e2f8ebd87aeb31d86a962360166aea9b15	Maria	2025-03-26 22:36:42 +0300	Git commit history	git-analysis
5	a4fee3e2f8ebd87aeb31d86a962360166aea9b15	f418704449370c5ba16131037f6eed177f018247	gigwrld	2025-03-15 15:39:13 +0300	Update README.md	git-analysis
6	f418704449370c5ba16131037f6eed177f018247	de0e51208d93be7e1ff7af16f4404011d89bb5ac	gigwrld	2025-03-15 15:37:53 +0300	Add files via upload	git-analysis
7	de0e51208d93be7e1ff7af16f4404011d89bb5ac	e60d36cd53e97fb6447967e85f025709b6f55807	MariUt005	2025-03-01 13:58:00 +0300	Update todo.md	git-analysis
8	e60d36cd53e97fb6447967e85f025709b6f55807	934e9b638be6e8c0a4af353ea6b5cb18dc3f61d9	MariUt005	2025-03-01 13:32:54 +0300	Update todo.md	git-analysis
9	934e9b638be6e8c0a4af353ea6b5cb18dc3f61d9	73ae66ab58ba0f97c9388647a3c644ba6f108d6d	MariUt005	2025-03-01 13:22:49 +0300	Update todo.md	git-analysis
10	73ae66ab58ba0f97c9388647a3c644ba6f108d6d	0096aaee29bfe53ee147f53bcca54725390150b3	MariUt005	2025-03-01 13:17:49 +0300	Update todo.md	git-analysis
11	0096aaee29bfe53ee147f53bcca54725390150b3	18ae95e1b691b281af3830cda6aa2b87ea525d07	MariUt005	2025-03-01 13:14:55 +0300	Create todo.md	git-analysis
12	18ae95e1b691b281af3830cda6aa2b87ea525d07	e0640a396a3d1e2abf540a6772cb3bc7e38ba52e	MariUt005	2025-03-01 12:52:02 +0300	Update README.md	git-analysis
13	e0640a396a3d1e2abf540a6772cb3bc7e38ba52e	7d90b470783c143ba2cd36f33dcfe4acbd4048a9	MariUt005	2025-03-01 12:51:51 +0300	Update README.md	git-analysis
14	7d90b470783c143ba2cd36f33dcfe4acbd4048a9	da9aa921a63bebd4eed7c7966168a4768647683b	Arikova Kristina	2025-03-01 11:42:24 +0300	add images	git-analysis
15	da9aa921a63bebd4eed7c7966168a4768647683b	326efbce72dd45b9bdef01373e394c4dc4c473e7	Arikova Kristina	2025-02-27 17:07:56 +0300	the main point	git-analysis
16	326efbce72dd45b9bdef01373e394c4dc4c473e7		Maria	2025-02-27 14:44:03 +0300	Init commit	git-analysis

5. **Добавление датасета в базу данных**

Полученный датасет записывается в git_commit_history в DuckDB

Этап 3. Изменения в файлах

1. Новый ли репозиторий?

- Проверка, есть ли этот репозиторий в базе данных
- Если да, то получаем номер последнего коммита last_commit

не повторяется в программе, представлен для полноты понимания алгоритма

2. **Получение изменений в каждом файле для каждого коммита**

git -C {repo_path} log -p --unified=0 -w --ignore-blank-lines

git -C {repo_path} log {last_commit}..HEAD -p --unified=0 -w --ignore-blank-lines

```
commit c1f3475f3348faa6e2c6115eeec01e11c0f10b38 (HEAD -> master, origin/master, origin/HEAD)
Author: Maria <maria.ut.005@yandex.ru>
Date:   Sun Mar 30 00:45:05 2025 +0300

    Changes in files

diff --git a/ETL.R b/ETL.R
index 19bdfd0..481335a 100644
--- a/ETL.R
+++ b/ETL.R
@@ -47,0 +48,2 @@ df <- git_diff_df %>%
+  filter(!grepl('^deleted', lines)) %>%
+  filter(!grepl('^new', lines)) %>%
@@ -54 +56 @@ j = 1
-while (i < n) {
+while (i < 50000) {
@@ -86,0 +89,3 @@ while (i < n) {
+    if (!(grepl("^commit", df$lines[i]) | grepl("^---", df$lines[i]) | grepl("^\\+\\+\\+", df$lines[i]) | grepl("^@@", df$lines[i]))) {
+      i = i + 1
+    }
@@ -88 +92,0 @@ while (i < n) {
-  print(j)
@@ -90,0 +95 @@ while (i < n) {
+write_parquet(changesDf, paste(gsub("/", "_", strsplit(repo_url, "//")[[1]][2]),"_changes.parquet", sep=""))
```

3. **Запись информации в датасет**

	lines
1	commit c1f3475f3348faa6e2c6115eeec01e11c0f10b38
2	Author: Maria <maria.ut.005@yandex.ru>
3	Date: Sun Mar 30 00:45:05 2025 +0300
4	
5	Changes in files
6	
7	diff --git a/ETL.R b/ETL.R
8	index 19bdfd0..481335a 100644
9	--- a/ETL.R
10	+++ b/ETL.R
11	@@ -47,0 +48,2 @@ df <- git_diff_df %>%
12	+ filter(!grepl('^deleted', lines)) %>%
13	+ filter(!grepl('^new', lines)) %>%
14	@@ -54 +56 @@ j = 1
15	-while (i < n) {
16	+while (i < 50000) {
17	@@ -86,0 +89,3 @@ while (i < n) {

4. **Очистка от ненужных строк**

	lines
1	commit c1f3475f3348faa6e2c6115eeec01e11c0f10b38
2	--- a/ETL.R
3	+++ b/ETL.R
4	@@ -47,0 +48,2 @@ df <- git_diff_df %>%
5	+ filter(!grepl('^deleted', lines)) %>%
6	+ filter(!grepl('^new', lines)) %>%
7	@@ -54 +56 @@ j = 1
8	-while (i < n) {
9	+while (i < 50000) {
10	@@ -86,0 +89,3 @@ while (i < n) {

5. **Создание новых столбцов**

Для каждой строки определяется ее тип:

- хэш коммита
- путь к файлу до и после изменений;
- номер строки, на которой начался блок изменений и количество затронутых строк
- текст или код, который поменялся

Перед записью в новое поле данные очищаются от лишних символов и при необходимости дополнительно обрабатываются

Добавляются дополнительные поля для удобства при дальнейшей работе (например, segment_id)

*на скриншот не поместился столбец is_add, который указывает на характер изменений (+ или -)

	lines	commit_start	commit_id	commit	src_file	dst_file	start_del	count_del	start_add	count_add	metaline	segment_start	segment_id	src_code
1	commit c1f3475f3348faa6e2c6115eeec01e11c0f10...	TRUE	1	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	NA	NA	NA	NA	NA	NA	TRUE	FALSE	0	NA
2	--- a/ETL.R	FALSE	1	NA	a/ETL.R	NA	NA	NA	NA	NA	TRUE	FALSE	0	NA
3	+++ b/ETL.R	FALSE	1	NA	NA	b/ETL.R	NA	NA	NA	NA	TRUE	FALSE	0	NA
4	@@ -47,0 +48,2 @@ df <- git_diff_df %>%	FALSE	1	NA	NA	NA	47	0	48	2	TRUE	TRUE	1	NA
5	+ filter(!grepl('^deleted', lines)) %>%	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	1	filter(!grepl('^deleted', line
6	+ filter(!grepl('^new', lines)) %>%	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	1	filter(!grepl('^new', lines))
7	@@ -54 +56 @@ j = 1	FALSE	1	NA	NA	NA	54	1	56	1	TRUE	TRUE	2	NA
8	-while (i < n) {	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	2	while (i < n) {
9	+while (i < 50000) {	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	2	while (i < 50000) {
10	@@ -86,0 +89,3 @@ while (i < n) {	FALSE	1	NA	NA	NA	86	0	89	3	TRUE	TRUE	3	NA
11	+ if (!(grepl("^commit", df$lines[i]) | grepl("^---", ...	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	3	if (!(grepl("^commit", df$
12	+ i = i + 1	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	3	i = i + 1
13	+ }	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	3	}
14	@@ -88 +92,0 @@ while (i < n) {	FALSE	1	NA	NA	NA	88	1	92	0	TRUE	TRUE	4	NA
15	- print(j)	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	4	print(j)
16	@@ -90,0 +95 @@ while (i < n) {	FALSE	1	NA	NA	NA	90	0	95	1	TRUE	TRUE	5	NA
17	+write_parquet(changesDf, paste(gsub("/", "_", strs...	FALSE	1	NA	NA	NA	NA	NA	NA	NA	FALSE	FALSE	5	write_parquet(changesDf, pa

6. **Объединение кода**

Создается второй датасет, в котором весь измененный текст в рамках каждого блока объединяется в одну строку. При этом из нее удаляются все нечитаемые символы, в т.ч. пробелы, табуляции

	segment_id	is_add	src
1	1	TRUE	filter(!grepl('deleted',lines))%%filter(!grepl('new',lin...
2	2	FALSE	while(in){
3	2	TRUE	while(i50000){
4	3	TRUE	if(!(grepl("commit",dflines[i])grepl("---",dflines[i])gre...
5	4	FALSE	print(j)
6	5	TRUE	write_parquet(changesDf,paste(gsub("/","_",strsplit(...
7	6	FALSE	print(dir)
8	7	FALSE	print("pull")
9	8	FALSE	print("clone")
10	9	FALSE	renamefromgithub.com_tidyverse_ggplot2.git.parqu...
11	9	TRUE	git_diff_cmd-glue('git-C{repo_dir}log-p--unified0-w-...
12	10	TRUE	library(arrow)repo_url-"https://github.com/tidyverse...
13	11	FALSE	#gsub("/","\\\\",dir)
14	12	FALSE	dir_repo-getRepo("https://github.com/tidyverse/ggpl...
15	12	TRUE	repo_dir-getRepo(repo_url)git_log_cmd-glue('git-C{r...
16	13	TRUE	write_parquet(git_commit_history,paste(gsub("/","_"...

7. **Заполнение пропущенных значений**

Поля, где были пропущены значения, заполняются данными из поля выше

Удаляются все записи, где значение src_code было NA

Удаляются ненужные столбцы

	commit	src_file	dst_file	start_del	count_del	start_add	count_add	segment_id
1	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	47	0	48	2	1
2	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	47	0	48	2	1
3	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	54	1	56	1	2
4	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	54	1	56	1	2
5	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	86	0	89	3	3
6	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	86	0	89	3	3
7	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	86	0	89	3	3
8	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	88	1	92	0	4
9	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	90	0	95	1	5
10	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	15	1	14	0	6
11	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	18	1	16	0	7
12	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	21	1	18	0	8
13	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
14	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
15	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
16	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
17	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
18	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
19	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
20	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
21	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
22	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
23	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
24	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
25	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
26	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9

8. **Удаление дубликатов**

Удаляются все дублирующиеся записи

	commit	src_file	dst_file	start_del	count_del	start_add	count_add	segment_id
1	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	47	0	48	2	1
3	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	54	1	56	1	2
5	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	86	0	89	3	3
8	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	88	1	92	0	4
9	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	90	0	95	1	5
10	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	15	1	14	0	6
11	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	18	1	16	0	7
12	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	21	1	18	0	8
13	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	9
71	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	2	0	3	3	10
74	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	11	1	13	0	11
75	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	25	3	27	11	12
89	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	28	0	39	2	13
91	f1b13932a6e5dd9d8f3873514f53880f56547511	/dev/null	b/ETL.R	0	0	1	28	14
20	a4fee3e2f8ebd87aeb31d86a962360166aea9b15	a/README.md	b/README.md	13	1	13	1	15
22	f418704449370c5ba16131037f6eed177f018247	a/README.md	b/README.md	13	1	13	1	15

9. **Объединение данных**

Данные из двух датасетов (см. шаг 6 и 8) объединяются в один датасет через segment_id

	commit	src_file	dst_file	start_del	count_del	start_add	count_add	src	is_add	repo
1	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	47	0	48	2	filter(!grepl('deleted',lines))%%filter(!grepl('new',lin...	TRUE	git-analysis
2	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	54	1	56	1	while(in){	FALSE	git-analysis
3	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	54	1	56	1	while(i50000){	TRUE	git-analysis
4	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	86	0	89	3	if(!(grepl("commit",dflines[i])grepl("---",dflines[i])gre...	TRUE	git-analysis
5	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	88	1	92	0	print(j)	FALSE	git-analysis
6	c1f3475f3348faa6e2c6115eeec01e11c0f10b38	a/ETL.R	b/ETL.R	90	0	95	1	write_parquet(changesDf,paste(gsub("/","_",strsplit(...	TRUE	git-analysis
7	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	15	1	14	0	print(dir)	FALSE	git-analysis
8	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	18	1	16	0	print("pull")	FALSE	git-analysis
9	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	21	1	18	0	print("clone")	FALSE	git-analysis
10	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	renamefromgithub.com_tidyverse_ggplot2.git.parqu...	FALSE	git-analysis
11	2d49d9c9dd7d77f91c906869548ae799938cc99d	a/ETL.R	b/ETL.R	39	0	37	54	git_diff_cmd-glue('git-C{repo_dir}log-p--unified0-w-...	TRUE	git-analysis
12	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	2	0	3	3	library(arrow)repo_url-"https://github.com/tidyverse...	TRUE	git-analysis
13	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	11	1	13	0	#gsub("/","\\\\",dir)	FALSE	git-analysis
14	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	25	3	27	11	dir_repo-getRepo("https://github.com/tidyverse/ggpl...	FALSE	git-analysis
15	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	25	3	27	11	repo_dir-getRepo(repo_url)git_log_cmd-glue('git-C{r...	TRUE	git-analysis
16	fa3a78ff139856b542759417fadad2cd76b05461	a/ETL.R	b/ETL.R	28	0	39	2	write_parquet(git_commit_history,paste(gsub("/","_"...	TRUE	git-analysis
17	f1b13932a6e5dd9d8f3873514f53880f56547511	/dev/null	b/ETL.R	0	0	1	28	library(glue)#Длявызовасистемныхкомандlibrary(...	TRUE	git-analysis
18	a4fee3e2f8ebd87aeb31d86a962360166aea9b15	a/README.md	b/README.md	13	1	13	1	imgsrc"images/План%20работы-02.png"data-fig-ali...	FALSE	git-analysis

10. **Добавление датасета в базу данных**

Полученный датасет записывается в git_diff в DuckDB

Этап 3. Сохранение пути к репозиторию

1. **Создание датасета**

	repo	path
1	git-analysis	D:/Творчество/git-analysis/MariUt005/git-analysis

2. **Добавление датасета в базу данных**

Полученный датасет записывается в repo_path в DuckDB

	repo	path
1	book_reader	D:/Творчество/git-analysis/MariUt005/book_reader
2	dotfiles	D:/Творчество/git-analysis/MariUt005/dotfiles
3	Enlightsion	D:/Творчество/git-analysis/MariUt005/Enlightsion
4	exam_mqwerty	D:/Творчество/git-analysis/MariUt005/exam_mqwerty
5	git-analysis	D:/Творчество/git-analysis/MariUt005/git-analysis
6	learnGo	D:/Творчество/git-analysis/MariUt005/learnGo
7	learnWeb	D:/Творчество/git-analysis/MariUt005/learnWeb
8	magma	D:/Творчество/git-analysis/MariUt005/magma
9	MersOff	D:/Творчество/git-analysis/MariUt005/MersOff
10	mqwerty	D:/Творчество/git-analysis/MariUt005/mqwerty
11	mqwerty_bot	D:/Творчество/git-analysis/MariUt005/mqwerty_bot
12	qq	D:/Творчество/git-analysis/MariUt005/qq
13	regex	D:/Творчество/git-analysis/MariUt005/regex
14	reports_i2z1	D:/Творчество/git-analysis/MariUt005/reports_i2z1
15	Students_db	D:/Творчество/git-analysis/MariUt005/Students_db
16	tetris	D:/Творчество/git-analysis/MariUt005/tetris
17	tic-tac-toe	D:/Творчество/git-analysis/MariUt005/tic-tac-toe
18	TMP	D:/Творчество/git-analysis/MariUt005/TMP

Результат

1. Датасет с краткой историей коммитов

	commit	parent_commit	author	date	message	repo
1	75f5da303f0baaf391961148af1690c5ac776ec2	adff96e6b01eef3e3d4dcc3b62457a6b51365522	MariUt005	2021-07-24 09:22:06 +0300	Some latest changes	book_reader
2	adff96e6b01eef3e3d4dcc3b62457a6b51365522	100b53956053f4d30bb9bae201c249a49c6688ce	MariUt005	2021-07-09 07:07:52 +0300	Update README.md	book_reader
3	100b53956053f4d30bb9bae201c249a49c6688ce	c7c02c6db07f809edbefac7b141ff458fbb867f3	MariUt005	2021-07-09 07:05:33 +0300	Created main menu page	book_reader
4	c7c02c6db07f809edbefac7b141ff458fbb867f3		MariUt005	2021-07-09 07:00:39 +0300	Initial commit	book_reader
5	d6b76d02fb5262c264000e3931e2e1989aaa5460	5a0119593db5ec60dff88c3f4cd719fda3d324f6	MariUt005	2021-06-27 07:09:22 +0300	Added rm_mysql.sh in ubuntu/	dotfiles
6	5a0119593db5ec60dff88c3f4cd719fda3d324f6	de87cef63c1d9ea6854efa7dc97b061d9c91441a	MariUt005	2021-04-14 23:41:27 +0300	Added LAMP and reorganized some files	dotfiles
7	de87cef63c1d9ea6854efa7dc97b061d9c91441a	a68e9787ff8f67cb6c53f70eda3e2adfe6460fb1	MariUt005	2021-03-14 10:03:29 +0300	Add files via upload	dotfiles
8	a68e9787ff8f67cb6c53f70eda3e2adfe6460fb1	b12375c0489e41017d5eb86eed046d1ef89fcec7	MariUt005	2021-03-14 10:03:17 +0300	Delete install_gnome.sh	dotfiles
9	b12375c0489e41017d5eb86eed046d1ef89fcec7	878b140ba0163722ffef8b6a0030c58954b8f05a	MariUt005	2021-03-14 10:02:27 +0300	Add files via upload	dotfiles
10	878b140ba0163722ffef8b6a0030c58954b8f05a	c8e479642315a7353998c706c838debd1dd8e222	MariUt005	2020-09-11 18:58:21 +0300	Update install.sh	dotfiles
11	c8e479642315a7353998c706c838debd1dd8e222	b326babcbf2514393a7bb08feb3410608b52ef68	MariUt005	2020-09-11 18:49:22 +0300	Create readme	dotfiles
12	b326babcbf2514393a7bb08feb3410608b52ef68	d4ca1080fe0e14bf51d5412b10c159bcc2d2ebbd	MariUt005	2020-05-08 17:52:39 +0300	Change the branch of the repository "lamp"	dotfiles
13	d4ca1080fe0e14bf51d5412b10c159bcc2d2ebbd	fc28c3f200f4ef85f43f5bc937a88c6f1d02b9f2	MariUt005	2020-05-08 17:18:01 +0300	upd2	dotfiles

2. Датасет с изменениями в файлах

	commit	src_file	dst_file	start_del	count_del	start_add	count_add	src	is_add	repo
31	75f5da303f0baaf391961148af1690c5ac776ec2	a/web/settings.html	b/web/settings.html	1	0	1	22	!DOCTYPEhtmlhtmlheadmetacharset"utf-8"title/title...	TRUE	book_reader
32	adff96e6b01eef3e3d4dcc3b62457a6b51365522	a/README.md	b/README.md	1	0	2	4	##Neededpipinstalleel	TRUE	book_reader
33	100b53956053f4d30bb9bae201c249a49c6688ce	/dev/null	b/main.py	0	0	1	5	Binaryfiles/dev/nullandb/web/favicon.pngdiffer	FALSE	book_reader
34	100b53956053f4d30bb9bae201c249a49c6688ce	/dev/null	b/main.py	0	0	1	5	importeeleel.init("web")eel.start("main.html")\None...	TRUE	book_reader
35	100b53956053f4d30bb9bae201c249a49c6688ce	/dev/null	b/web/main.css	0	0	1	38	*{font-family:'Roboto',sans-serif;}body{background...	TRUE	book_reader
36	100b53956053f4d30bb9bae201c249a49c6688ce	/dev/null	b/web/main.html	0	0	1	21	!DOCTYPEhtmlhtmlheadmetacharset"utf-8"title/title...	TRUE	book_reader
37	c7c02c6db07f809edbefac7b141ff458fbb867f3	/dev/null	b/LICENSE	0	0	1	21	MITLicenseCopyright(c)2021MariUt005Permissionis...	TRUE	book_reader
38	c7c02c6db07f809edbefac7b141ff458fbb867f3	/dev/null	b/README.md	0	0	1	1	#book_reader\Nonewlineatendoffile	TRUE	book_reader
39	d6b76d02fb5262c264000e3931e2e1989aaa5460	/dev/null	b/ubuntu/rm_mysql.sh	0	0	1	7	systemctlstopmysqldaptpurgemysql-servermysql-co...	TRUE	dotfiles
40	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/install.sh	0	0	1	20	#!/bin/bash#sudopacman-Syu#installsudopacman-...	TRUE	dotfiles
41	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/install_gnome.sh	0	0	1	17	#!/bin/bash#Source:https://wiki.manjaro.org/index....	TRUE	dotfiles
42	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/lamp/README.md	0	0	1	1	#lamp\Nonewlineatendoffile	TRUE	dotfiles
43	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/lamp/httpd.conf	0	0	1	544	##ThisisthemainApacheHTTPserverconfigurationfile...	TRUE	dotfiles
44	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/lamp/init.vim	0	0	1	169	"Pluginswillbedownloadedunderthespecifieddirector...	TRUE	dotfiles
45	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/lamp/install.sh	0	0	1	29	#!/bin/bash#installsudopacman-Sapachemysqlphp...	TRUE	dotfiles
46	5a0119593db5ec60dff88c3f4cd719fda3d324f6	/dev/null	b/manjaro/lamp/php.ini	0	0	1	1943	[PHP];;;;;;;;;;;;;;;;;;;;;;Aboutphp.ini;;;;;;;;;;;;;;;;;;;;;PHP'si...	TRUE	dotfiles

Авторы

Мария
@ MariUt005

Кристина
@ gigwrld

Александра
@ gusarovaal