Министр науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

ИТМО

Факультет информационных технологий и программирования

Отчет по исследовательскому проекту

“Сравнительный анализ активности пользователей на Habr.com”

Выполнили студенты групп № M3106, № M3107, № M3108
Волков Никита Андреевич
Волков Фёдор Владимирович
Волков Глеб Романович
Бронштейн Михаил Александрович
Гайворон Алексей Александрович

Санкт-Петербург
2026

Содержание

---

1. Общее описание проекта

1.1. Инициатор, заказчик проекта

Инициатором и заказчиком проекта является образовательная программа НИУ ИТМО в рамках учебной дисциплины, ориентированной на проектную и исследовательскую деятельность студентов в области анализа данных.

1.2. Тип проекта

Проект является исследовательским, так как направлен на сбор реальных неструктурированных данных (Web Scraping), построение аналитической модели и получение выводов о поведенческих паттернах аудитории на основе интерпретации результатов.

---

2. Содержательная часть

2.1. Цели и задачи проекта

Цель проекта: Выявить ключевые факторы, влияющие на популярность контента, и определить поведенческие профили пользователей на профессиональном IT-ресурсе Habr.com.

Задачи проекта:
Сбор и предварительная обработка данных с веб-ресурса (Web Scraping).

1. Проектирование структур данных для хранения метрик статей и авторов.
2. Построение визуализаций распределения активности во времени и по пользователям.
3. Интерпретация аномалий и выявление зависимостей типа «формат-успех».

Источник данных: Открытые данные ресурса Habr.com (статьи, метрики, профили пользователей).

2.2. Описание данных и их обработка

Исходные данные представляют собой HTML-страницы хабов, содержащие неструктурированную информацию о публикациях. В ходе обработки данные преобразуются в структурированный табличный вид.

2.2.1. Производные файлы

В ходе работы скрипта формируется итоговый датасет:

Файл	Назначение
habr_big_data_20k.csv	Основной массив данных, содержащий 20 000+ записей с атрибутами: дата, автор, просмотры, время чтения, сложность, хабы.

Файл формируется автоматически с использованием Python-скрипта (parser.py).

2.3. Модель данных

Используется реляционная модель событий.

2.3.1. Объекты анализа

- **Статья (Article):** Основная сущность. Атрибуты: title, reading_time_min, date.
- **Автор (Author):** Сущность, генерирующая контент. Атрибуты: author.
- **Хаб (Hub):** Категория контента. Атрибуты: hub_name.

2.3.2. Метрики активности

- **Просмотры (Views):** Ключевая метрика охвата.

2.4. Используемые структуры данных и алгоритмы

2.4.1. Структуры данных (Python)

- list: Хранение списка хабов для обхода (HUBS) и накопления строк перед записью.
- set: Хранение уникальных ID статей (seen_ids) для исключения дубликатов при парсинге.
- dict: Использование словарей для передачи заголовков (HEADERS) в HTTP-запросах.

2.4.2. Алгоритмы

- Итеративный обход веб-страниц (циклы по хабам и страницам).
- Парсинг с помощью BeautifulSoup для извлечения текстовых метрик.
- Очистка и нормализация числовых данных (преобразование «1.5К» → 1500).
- Агрегация данных в BI-системе (расчет среднего AVG и суммы SUM).

2.5. Результаты анализа

2.5.1. Активность пользователей (Кто пишет?)

Анализ выполнен на основе матрицы эффективности авторов.

- **Доминирование «Среднего класса»:** Основной объем полезного контента генерируют не единичные контентмейкеры, а широкая прослойка авторов (10–20 статей) с суммарными просмотрами до 150 000. Это фундамент платформы.

- **Стратегии лидеров:** Топ-авторы делятся на «Top-Left» (мало статей, миллионные просмотры вирусных лонгридов) и «Top-Right» (высокая частота публикаций).

2.5.2. Временная динамика (Когда читают?)

Анализ тепловой карты (Heatmap) выявил аномалию, отличающую Хабр от деловых СМИ.

- **Феномен:** Пик активности приходится на **Субботу и Воскресенье**.
- **Вывод:** Аудитория использует ресурс для саморазвития, отдыха и хобби-проектов (Pet-projects) в личное время, а не для рабочих задач.

2.5.3. Тематические предпочтения (Что читают?)

Сравнительный анализ хабов показал приоритет практической пользы над “хайпом”.

- **Лидеры:** *network_technologies* и *arduino* (решение инженерных задач).
- **Аномалия:** Темы *AI/Machine Learning* проигрывают «прикладному железу» и сетям из-за перенасыщения рынка ИИ контентом.

2.6. Визуализация

Визуализация данных выполнена с использованием инструмента Yandex DataLens. Построены: Scatter Plot (матрица авторов), Heatmap (активность по дням), Bar Charts (рейтинги тем) и Line Charts (тренды).

---

3. Заключение

В результате выполнения проекта были достигнуты поставленные цели. В ходе работы были развиты навыки сбора неструктурированных данных, проектирования структур для их хранения и интерпретации поведенческих метрик.

Проект способствовал формированию компетенций в области бизнес-аналитики и работы с BI-инструментами.

---

4. Результат проекта

Результатом проекта являются:

- аналитический отчёт;
- набор данных `habr_big_data_20k.csv`;
- скрипт сбора данных `parser.py`;
- интерактивный дашборд с визуализациями в Yandex DataLens.

---

5. Приложения

В приложениях в репозитории на онлайн-ресурсе GitHub (https://github.com/TsNo0/dzen-analytics-project) представлены дополнительные материалы, подтверждающие получение результатов проекта:

- файл с исходным кодом парсера;
- скриншоты визуализаций из Yandex DataLens (Матрица авторов, Тепловая карта, Рейтинги).