

Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования «Уральский федеральный университет имени первого Президента России Б. Н. Ельцина» (УрФУ)

Институт радиоэлектроники и информационных технологий – РТФ

ОТЧЕТ

о проектной работе

по теме: Нормализация данных при помощи LLM

по дисциплине: Проектный практикум

Команда:

Королёв Семён РИ-231003

Верховых Савелий РИ-231003

Нажимов Игорь РИ-231003

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc165369900)

[1. Целевая аудитория 4](#_Toc165369901)

[2. Определение проблемы 6](#_Toc165369902)

[3. Подходы к решению проблемы 9](#_Toc165369903)

[4. Анализ аналогов 11](#_Toc165369904)

[5. Календарный план проекта 12](#_Toc165369905)

[6. Сценарии использования 14](#_Toc165369906)

[7. Требования к продукту и к MVP 15](#_Toc165369907)

[8. Стек для разработки 17](#_Toc165369908)

[9. Прототипирование 19](#_Toc165369909)

[10. Проектирование и разработка системы 21](#_Toc165369910)

[ЗАКЛЮЧЕНИЕ 22](#_Toc165369911)

[СПИСОК ЛИТЕРАТУРЫ 23](#_Toc165369912)

[ПРИЛОЖЕНИЕ 26](#_Toc165369913)

ВВЕДЕНИЕ

С развитием технологий появляется всё больше требований к качеству и скорости получаемой информации.

Ежегодно объёмы данных растут с немыслимой скоростью, а компании уделяют много внимания автоматизации процессов и повышению точности и эффективности обработки данных.

Чтобы частично решить данную проблему наша команда создала программу, которая занимается нормализацией данных при помощи LLM.

**Цель нашего проекта:** разработать программный продукт, занимающийся нормализацией необработанных текстовых данных, в виде товарных записей, при помощи большой языковой модели (LLM).

Под нормализацией понимается разбиение записей на компоненты, относящихся к определенному атрибуту (характерному свойству объекта).

**Задачи нашего проекта:**

* Определить способ взаимодействия с LLM.
* Определить модель LLM.
* Выбрать способ обработки нормализованных данных.
* Определить метрики.
* Выбрать способ вывода обработанных, нормализованных данных.

1. Целевая аудитория

Для определения целевой аудитории мы использовали методику 5W Марка Шеррингтона. Это наиболее распространенный способ определения целевой аудитории и психологических характеристик, которыми обладают потенциальные потребители.

Сегментация рынка проводилась по 5 вопросам:

1. Что? – мы предлагаем пользователям программу, которая поможет нормализировать необработанные текстовые данные, в виде товарных записей, при помощи LLM.
2. Кто? – нашей целевой аудиторией являются люди\компании, которым необходимо быстро и просто преобразовать входные данные в определённый формат.
3. Почему? – мотивация клиента на использование нашего продукта заключается в желании клиента получить результат, быстро, точно и бесплатно.
4. Когда? – поскольку это не коммерческий проект, то потребители могут получить данную программу бесплатно в любое удобное для них время.
5. Где? – наш продукт находится в открытом доступе на сайте GitHub.

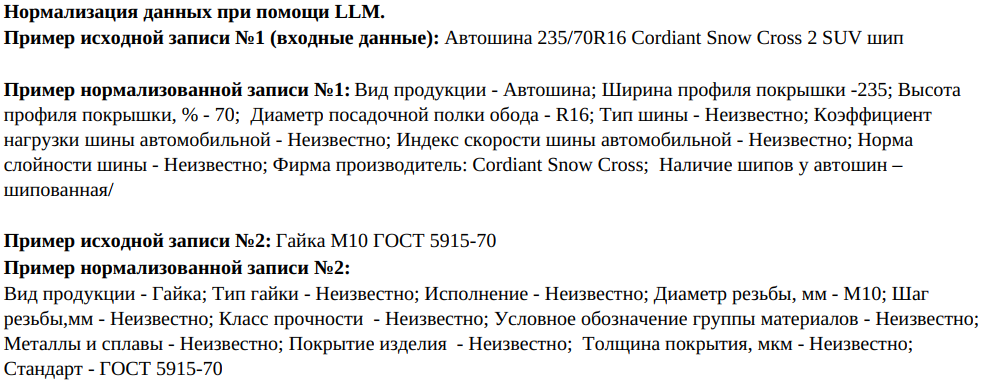
Поскольку наш продукт не нацелен на получение материальной выгоды, то основными вопросами, которыми мы пользовались для определения нашей целевой аудитории были «Что?», «Кто?», «Почему?».

В результате анализа мы убедились, что наша целевая аудитория — это лица (как физические, так и юридические), которым требуется быстрая, точная и бесплатная нормализация данных.

1. Определение проблемы

Для определения проблемы нашей команде были предоставлены примеры исходных и нормализованных данных:

Рисунок 1 – Примеры исходных и нормализованных данных



Определив формат входных и выходных данных, мы столкнулись со следующими вопросами:

* Каким методом взаимодействия с LLM мы будем пользоваться?
* Как выбрать подходящую LLM модель?
* Как мы будем обрабатывать нормализованные данные?
* Как мы собираемся выводить обработанные, нормализованные данные?

Определив ключевые проблемы, мы можем переходить к формулированию подходов к их решению.

1. Подходы к решению проблемы

Чтобы добиться поставленной цели мы рассмотрели каждый вопрос по отдельности:

* **Каким методом взаимодействия с LLM мы будем пользоваться?**

Для решения данной проблемы мы изучили различные интернет-источники и перед нами встал выбор:

1. **Fine-Tuning** — это процесс дообучения заранее обученной модели на специализированных наборах данных для конкретных задач.
2. **Prompt Engineering** — это процесс создания и оптимизации запросов для языковых моделей.
3. **Prompt Tuning** — это метод, при котором к входу языковой модели добавляют небольшой набор обучаемых параметров (так называемых «мягких промптов»).

Рассмотрев каждый вариант, мы остановились на Prompt Engineering из-за простоты его реализации, т.к. для других вариантов требовалось больше времени для изучения и более мощное железо, что являлось нашими ограничениями при выполнении данного проекта.

Следующим шагом мы решали следующую проблему:

* **Как выбрать подходящую LLM модель?**

В качестве вариантов мы рассматривали такие модели как:

1. **Qwen** — китайская нейросеть с открытым исходным кодом от компании Alibaba.
2. **Deepseek** — языковая нейросеть в виде чат-бота. Работает на базе технологии LLM и понимает и генерирует естественный язык.
3. **LLaMA** — серия мультимодальных моделей искусственного интеллекта с открытым исходным кодом от компании Meta.

Протестировав каждую из них, мы решили остановиться на **Qwen** из учета скорости и точности.

Далее мы сформировали prompt и столкнулись со следующий проблемой:

* **Как мы будем обрабатывать нормализованные данные?**

В качестве решения мы сформулировали поэтапную инструкцию:

1. На вход подаются файлы csv/xlcx формата, которые содержат несколько запросов.
2. После получения ответа от LLM мы форматируем ответ, который получаем в виде строки, так чтобы на выходе можно было получить таблицу в формате csv/xlcx.

Получив обработанные данные нам оставалось решить последнюю проблему.

* **Как мы собираемся выводить обработанные, нормализованные данные?**

Изучив различные интернет-источники, мы остановились на 2-х вариантах:

1. **Gradio** —это открытая библиотека на языке Python, созданная для упрощения разработки веб-приложений.
2. **tkinter** — это кроссплатформенная событийно-ориентированная графическая библиотека на основе средств Tk для языка программирования Python.

Мы остановились на tkinter из-за его более простого изучения и наличия русской документации.

1. Анализ аналогов

Прямые конкуренты – это аналогичное программное обеспечение на аналогичном рынке и работающий с целевой аудиторией, что и разрабатываемый продукт.

Косвенные конкуренты – это похожее программное обеспечение с другими характеристиками или абсолютно другой продукт, но работающие с целевой аудиторией, что и разрабатываемый продукт.

Поскольку мы не стремимся занять определённую нишу в данной сфере, а наш продукт является абсолютно бесплатным и общедоступным, мы решили отказаться от анализа аналогов.

1. Календарный план проекта

Название проекта:

Руководитель проекта:

Таблица 1 – Календарный план

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **№** | **Название** | **Ответственный** | **Длительность** | **Дата начала** | **Временные рамки проекта** | | | |
| 1 нед | 2 нед | 3 нед | 4 нед |
| ***Анализ возможных алгоритмов, поиск решений для прототипа. Анализ и подготовка данных.*** | | | | | | | | |
| *1.1* | *Выбор решений для прототипа* | Верховых Савелий | Неделя | 21.03.2025 | + |  |  |  |
| *1.2* | *Подготовка данных* | Королёв Семён | Неделя | 21.03.2025 | + |  |  |  |
| *1.3* | *Выбор между самописной нейронной сетью и готовыми LLM* | Верховых Савелий | Неделя | 21.03.2025 | + |  |  |  |
| *1.4* | *Изучение доступных LLM (выбор подходящей)* | Нажимов Игорь | Неделя | 21.03.2025 | + |  |  |  |
| ***Разработка кодовой базы. Адаптация алгоритмов. Поиск метрик качества работы выбранных решений. Оценка качества работы.*** | | | | | | | | |
| *2.1* | *Подготовить скрипты для запросов, отправляемых в AI* | Верховых Савелий | 3 недели | 03.04.2025 |  | + |  |  |
| *2.2* | *Провести ретроспективу по пройденной итерации* | Королёв Семён | 3 недели | 03.04.2025 |  | + |  |  |
| *2.3* | *Изучить способы подключения данных скриптов для взаимодействия с AI* | Верховых Савелий | 3 недели | 03.04.2025 |  | + |  |  |
| ***Оптимизация алгоритмов. Улучшение работы и повышение уровня метрик. Доработка и оптимизация сборки.*** | | | | | | | | |
| *3.1* | *Разобраться с Gradio (или иными средствами\библиотеками создания GUI)* | Королёв Семён | 3 недели | 25.04.2025 |  |  | + |  |
| *3.2* | *Разобраться с работой в git* | Нажимов Игорь | 3 недели | 25.04.2025 |  |  | + |  |
| *3.3* | *Подготовить код для работы с файлом* | Нажимов Игорь | 3 недели | 25.04.2025 |  |  | + |  |
| *3.4* | *Доработать LLM запросы (скрипт) и взаимодействие с AI* | Верховых Савелий | 3 недели | 25.04.2025 |  |  | + |  |
| ***Представление разработанного решения. Презентация и показ заказчику.*** | | | | | | | | |
| *4.1* | *Оптимизировать работу нейронки* | Верховых Савелий | 2 недели | 15.05.2025 |  |  |  | + |
| *4.2* | *Подготовить отчёт* | Королёв Семён | 2 недели | 15.05.2025 |  |  |  | + |
| *4.3* | *Подготовить презентацию* | Королёв Семён | 2 недели | 15.05.2025 |  |  |  | + |
| *4.4* | *Завершить разработку приложения (подготовить метрики)* | Нажимов Игорь | 2 недели | 15.05.2025 |  |  |  | + |
|  | *Защита проекта* |  | Месяц | 30.05 - 26.06 |  |  |  | + |

1. Сценарии использования

Сценарий использования (use case) — описание поведения системы, когда она взаимодействует с кем-то (или чем-то) из внешней среды.

1. **Запускается исполняемый exe файл (перед вами откроется GUI)**
2. **Добавляете необходимые для обработки файлы в соответствующее окно (поддерживаются только форматы csv, xlcx)**

**Важно:** если был подан файл с иным расширением программа не перейдёт к 3 пункту, а выдаст ошибку и попросит загрузить файл с верным расширением.

1. **Ожидайте окончания обработки данных**

* В этот момент программа отправляет prompt в Qwen
* Получив ответ, форматирует данный ответ в csv или xlcx (в зависимости от расширения исходного файла)
* Загружает файл в GUI

1. **Скачивается готовый файл**
2. **Закрывается GUI или нажимается кнопку «Попробовать снова»**

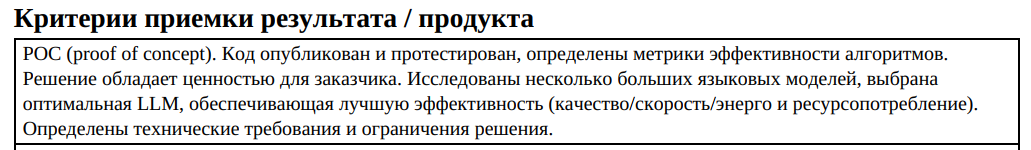
**Важно:** если была нажата кнопка **«Попробовать снова»**, программа возвращается к исходному состоянию, как при запуске.

1. Требования к продукту и к MVP

Функциональные требования

В качестве функциональных требований используется сценарий использования, если один из пунктов сценария использования не выполняется, то MVP не достигнуто.

Требования заказчика



1. Стек для разработки

Стек для разработки — это набор инструментов и технологий, которые используются для создания продуктов и приложений.

В качестве стека разработки нами были взяты:

* Python
* Библиотеки Python: transformers, tkinter, os, shutil, subprocess, threading, uuid, pathlib, sys
* Платформа GitHub
* Qwen

1. Прототипирование

Для дальнейшей разработки и теста выбранного стека разработки нами создавались следующие прототипы:

* Прототип скрипта (prompt)
* Прототип GUI
* Прототип

Все прототипы можно посмотреть на нашем GitHub, ссылка на него находится во вкладке «**ПРИЛОЖЕНИЕ**».

1. Проектирование и разработка системы

На этом этапе формируется общая структура программного комплекса. В соответствии с технологией нисходящего структурного программирования программный комплекс разбивается на небольшие части – программные модули (блоки). Для каждого программного модуля формулируются требования по реализуемым функциям и разрабатывается алгоритм, реализующий эти функции. Необходимо привести описание этих модулей, а также задач, которые они решают.

Необходимо привести алгоритм работы приложения, т.е. точного предписания и последовательности действий, заложенных в логике продукта.

Также на этом этапе можно привести схему взаимодействия программных модулей, т.е. схема потоков данных программы.

В этой части можно привести описание самого процесса разработки разнотипных сложных блоков программного обеспечения.

ЗАКЛЮЧЕНИЕ

Заключение должно содержать общие выводы, обобщенное изложение основных проблем. Могут быть указаны положительные и отрицательные стороны разработанного решения. Должны быть указаны перспективы дальнейшей разработки темы. Примерный объем заключения 2-3 страницы.

В заключительной части работы должно быть дано последовательное, логическое изложение полученных результатов в соответствие поставленными изначально задачами.

На данный момент сервис готовится к запуску и тестированию на реальных пользователях.

СПИСОК ЛИТЕРАТУРЫ

1. Веб-технологии для разработчиков Обзор протокола HTTP / Веб-технологии для разработчиков. – 2022. – URL: https://developer.mozilla.org/ru/docs/Web/HTTP/Overview (дата обращения: 30.04.2022).
2. Вигерс К. Разработка требований к программному обеспечению / К. Вигерс, Дж. Битти. – 3-е изд. – Москва: Русская редакция, 2014. – 736 с. – ISBN 978-5-9909805-3-2.
3. Евлантьев С.Карта эмпатии клиента / Станислав Евлантьев. – 2022. – URL: https://in-scale.ru/blog/karta-empatii/ (дата обращения: 01.05.2022).
4. Ильяхов М. Пиши, сокращай: как создать сильный текст / М. Ильяхов, Л. Сарычева. – 3-е изд. – Москва: Альпина Паблишер, 2022. – 440 с. – ISBN 978-5-9614-6526-6.
5. Кинзябулатов Р. IDEF0. Знакомство с нотацией и пример использования / Рамиль Кинзябулатов. – 2022. – URL: https://trinion.org/blog/idef0-znakomstvo-s-notaciey-i-primer-ispolzovaniya (дата обращения: 30.04.2022).
6. Коберн А. Современные методы описания функциональных требований к системам / А. Коберн. – Москва: Издательство «Лори», 2012. – 264 с. – ISBN 978-5-85582-326-4.
7. Мидоус Ф. Введение в протоколы HTTP и HTTPS / Филлип Мидоус. – 2022. – URL: https://docs.microsoft.com/ru-ru/azure/rtos/netx-duo/netx-duo-web-http/chapter1 (дата обращения: 05.05.2022).
8. Сырцев В. Использования диаграммы вариантов использования UML при проектировании программного обеспечения / Вадим Сырцев. – 2021. – URL: https://habr.com/ru/post/566218/ (дата обращения:30.04.2022).
9. Фирстов Л. О развитии навыка говорения / Леонид Фирстов. – 2020. – URL: https://habr.com/ru/post/494424/ (дата обращения: 30.04.2022).
10. Фитцпатрик Р. Спроси маму: Как общаться с клиентами и подтвердить и подтвердить правоту своей бизнес-идеи, если все кругом врут? / Р. Фитцпатрик. – Москва : Альпина Паблишер, 2021. – 160 с. – ISBN 978-5-9614-3045-5.
11. Шейнблат В. Естественный подход к освоению языка / Владимир Шейнблат. – 2022. – URL: https://englishsimple.ru/articles/natural-approach-to-language-learning/ (дата обращения: 25.04.2022).
12. Шторкин С. Средневзвешенная система голосования / Семен Шторкин – 2022. – URL: https://habr.com/ru/post/63664/ (дата обращения: 30.04.2022).
13. Apple MacOS / Apple. – 2022. – URL: https://www.apple.com/za/macos/what-is/ (дата обращения: 30.04.2022).
14. Best Programmer Функциональные и нефункциональные требования: полное руководство / Best Programmer. – 2021. – URL: https://bestprogrammer.ru/izuchenie/funktsionalnye-i-nefunktsionalnye-trebovaniya-polnoe-rukovodstvo (дата обращения: 15.05.2022).
15. Google Chrome Браузер от Google / Google Chrome. – 2022. – URL: https://www.google.ru/chrome/ (дата обращения: 30.04.2022).
16. HeyLady Женское сообщество для создания конференций и практики английского языка / HeyLady. – 2022. – URL: https://www.heylady.io/ (дата обращения: 30.04.2022).
17. Lingbe Бесплатные звонки с носителями в любое время / Lingbe. – 2022. – URL: https://lingbe.com/ (дата обращения: 30.04.2022).
18. PostgrePro Что такое PostgreSQL? / PostgrePro. – 2020. – URL: https://postgrespro.ru/docs/postgresql/12/intro-whatis (дата обращения: 30.04.2022).
19. Proglib Об Agile / Denver. – 2022. – URL: https://proglib.io/p/klyuchevye-razlichiya-mezhdu-agile-scrum-i-kanban-2022-02-24 (дата обращения 05.04.2022).
20. SimilarWeb Аналитика трафика Tandem / SimilarWeb. – 2022. – URL: https://www.similarweb.com/ru/website/tandem.net/#traffic (дата обращения: 30.04.2022).
21. Tandem / Tandem.net. – 2022. – URL: https://www.tandem.net/ru (дата обращения: 30.04.2022).
22. JavaScript WebSocket / JavaScript. – 2022. – URL: https://learn.javascript.ru/websocket (дата обращения: 02.05.2022).
23. VC.RU Анализ конкурентов на практике / VC.RU. – 2020. – URL: https://vc.ru/marketing/181297-analiz-konkurentov-na-praktike-10-shagov-shablon (дата обращения: 30.04.2022).
24. VC.RU Как составить портрет клиента (целевой аудитории) / VC.RU. – 2020. – URL: https://vc.ru/marketing/156147-kak-sostavit-portret-klienta-celevoy-auditorii-instrukciya-s-primerami (дата обращения: 30.04.2022).

ПРИЛОЖЕНИЕ

GitHub – ­ URL: https://github.com/NEONWolfYT/LLM-Project.git