Оценка письменной части ЕГЭ по английскому языку

Годовой проект

Application Programming Interface

Telegram bot

Streamlit Application

КОМАНДА

Вольф Елена

Куратор

Андрейко Максим

Data Parsing
Machine Learning
Data Processing
Deployment

API Development

Василенко Павел

Data Parsing
Machine Learning
Data Processing
Deployment

Streamlit Development

Бурлова Альбина

Data Parsing
Machine Learning
Data Processing
Deployment

Telegram Bot Development

Письменная часть ЕГЭ по английскому языку

1. Ежегодно экзамен сдает большое количество выпускников:

Количество выпускников, проходящих экзамен, в зависимости от года

Год	Количество сдающих
2023	85 796
2022	92 804
2021	89 770

2. Для организации работы экспертной комиссии требуется:

- Организация регулярного обучения проверяющих
- Предоставление размещения и командировки

3. Проверка экзаменационных работ должна осуществляться минимум двумя экспертами

Для проверки работ требуется квалифицированные специалисты. Однако, этот процесс можно автоматизировать с помощью искусственного интеллекта.

Письменное задание - Email

Письмо ЕГЭ по английскому языку — одно из заданий с развернутым ответом письменной части экзамена.
Данное задание содержит отрывок письма от друга по переписке. Предлагается написать ответ с соблюдением определенных критериев.

Письмо следует структурировать в соответствии **с определенными критериями (К1, К2, К3)**, по которым оно оценивается (от 0 до 2 баллов по каждому критерию - в сумме 6 баллов максимум).

К1 - Решение коммуникативной задачи:

- Умение дать четкие и полные ответы на заданные вопросы в задании;
- Умение задать встречные вопросы по предложенной теме;
- Поддержание вежливого тона общения и выбор стиля, соответствующего контексту общения.

К2 - Организация текста:

- Правильное разделение на абзацы;
- Логичность изложения.

К3 - Языковое оформление текста:

- Правильность использования грамматических конструкций;
- Адекватность выбора лексики;
- Соблюдение правил пунктуации.

Письменное задание - Email

37 You have received an e-mail message from your English-speaking pen friend, Ronny.

From: Harry@mail.uk

To: Russian_friend@ege.ru

Subject: Animal Protection

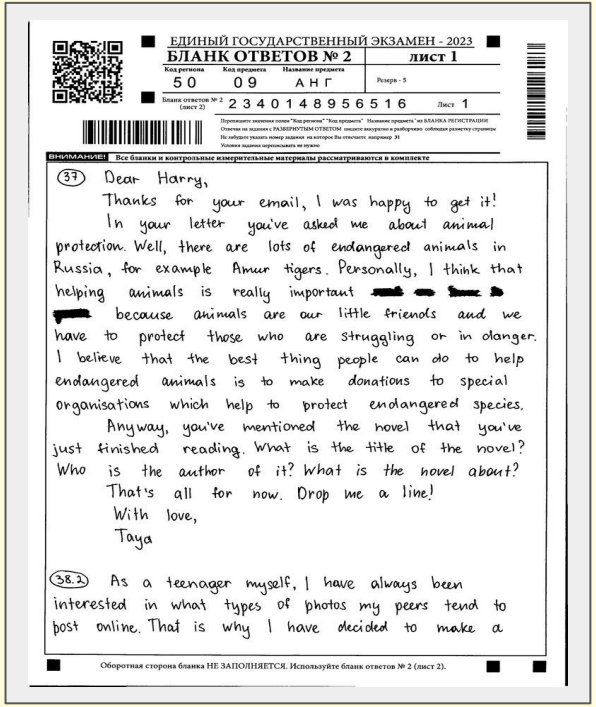
Last month our class went to the zoo to find out about the animal conservation programme. Are there any endangered animals in Russia and what are these? Do you consider helping animals important, why or why not? What can people do to help endangered species? I've just finished reading an interesting novel ...

Write a message and answer the 3 questions.

Write **100-120** words. Remember the rules of letter writing.

Задание письменной части ЕГЭ

[Более подробно про критерии оценивания письма (37 задание)]



ЕДИНЫЙ ГОСУДАРСТВЕННЫЙ ЭКЗАМЕН - 2023

БЛАНК ОТВЕТОВ № 2 лист 1

Код региона 50 Код предмета 09 Название предмета АНГ

Резерв - 5

Бланк ответов № 2 (лист 2) 2340148956516 Лист 1

ВНИМАНИЕ! Все бланки и контрольные измерительные материалы рассматриваются в комплекте

37 Dear Harry,

Thanks for your email, I was happy to get it! In your letter you've asked me about animal protection. Well, there are lots of endangered animals in Russia, for example Amur tigers. Personally, I think that helping animals is really important [redacted] because animals are our little friends and we have to protect those who are struggling or in danger. I believe that the best thing people can do to help endangered animals is to make donations to special organisations which help to protect endangered species.

Anyway, you've mentioned the novel that you've just finished reading. What is the title of the novel? Who is the author of it? What is the novel about?

That's all for now. Drop me a line!

With love,

Taya

38.2 As a teenager myself, I have always been interested in what types of photos my peers tend to post online. That is why I have decided to make a

Оборотная сторона бланка НЕ ЗАПОЛНЯЕТСЯ. Используйте бланк ответов № 2 (лист 2).

Пример ответа на задание

Парсинг данных

* Процент относительно количества собранных работ

Реализация парсеров

Парсер РешуЕГЭ

Итерация парсера по ссылкам для извлечения текста задания и ответов. Удаление служебных символов из текста и присвоение максимального балла полученным ответам по всем критериям.

Ссылка на код в репозитории

Парсер Cloudtext

Использование Selenium для парсинга из-за динамической подгрузки страниц с помощью JavaScript и требования авторизации.
Итерация по ссылкам работ с сохранением текста работы и выставленных баллов.

Ссылка на код в репозитории

Парсер сообщества VK

Использование API VK для извлечения с помощью requests и BeautifulSoup комментариев с фотографиями работ в тематических группах.
Распознавание рукописного текста на фотографиях посредством pytesseract. Последующая "ручная" обработка.

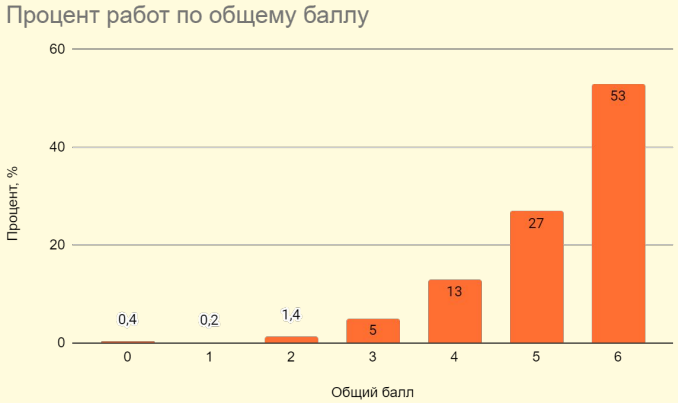
Ссылка на код в репозитории

EDA

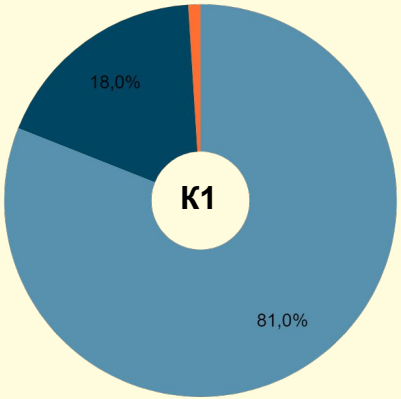
Количество реальных писем:

453

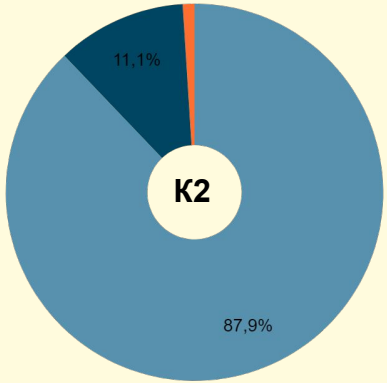
Дисбаланс распределения по баллам по всем критериям



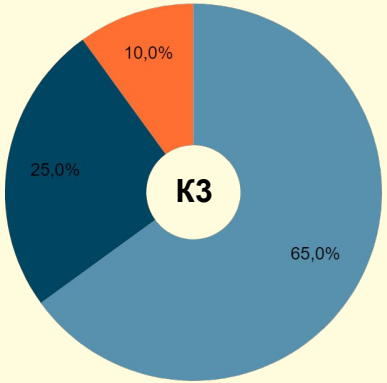
Решение коммуникативной задачи



Организация текста



Языковое оформление текста



EDA

Вывод

> **Без сгенерированных данных** преобладают работы с наивысшим баллом (6 баллов) с большим отрывом. Малое количество работ с баллами меньше 4;
> Заметно неравномерное распределение по баллам для всех критериев К1, К2, К3.
> Наибольшая часть работ оценена в 2 балла.

Генерация данных

OpenaAI API

1. Chat API

- **Запросы являются независимыми;**
- Низкая стоимость за токен;
- Малое время ожидания ответа.

2. Assistant API

- **Сохранение контекста**;
- Расширенная настройка;
- Бóльшая стоимость за токен по сравнению с Chat API;
- Дольше обрабатывается запрос.

Цель генерации - увеличение количества данных с целью улучшения распределения баллов по критерию К1 для дальнейшего обучения моделей.

Генерация данных

Благодаря сохранению контекстной целостности новые данные на основе старых генерировались через Assistant API. Генерировались как сами ответы, так и задания, на основе заданий генерировались ответы.

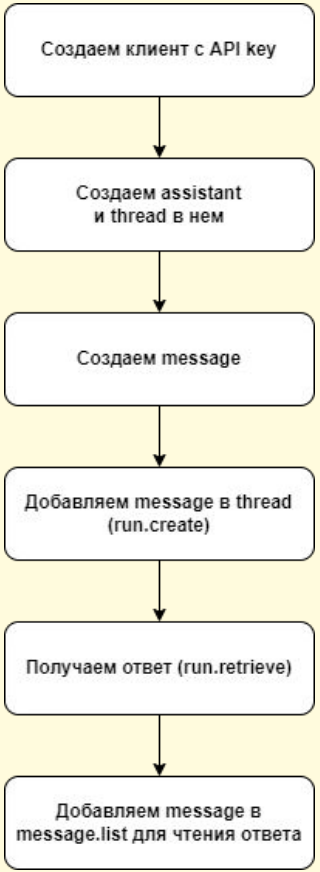
До

…What is the last book you've read? What books do you have to read for your literature classes? How often do you spend time reading a book for pleasure?

Assistant API request for reformulation

После

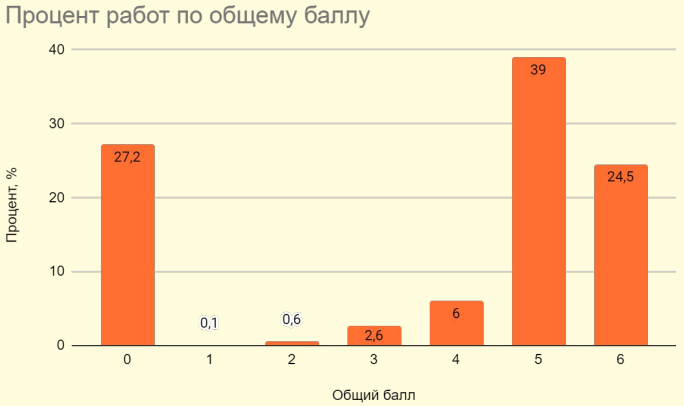
…Do you prefer attending book festivals or book signings? Which books are included in your required reading list? How do you balance your time between reading for pleasure and other activities?



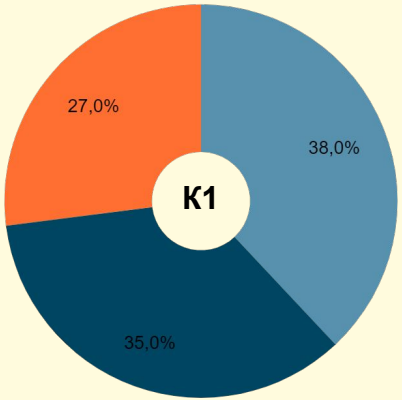
EDA

Количество реальных писем: **453** + **Количество синтетических писем:** **514** = **Общее количество:** **967**

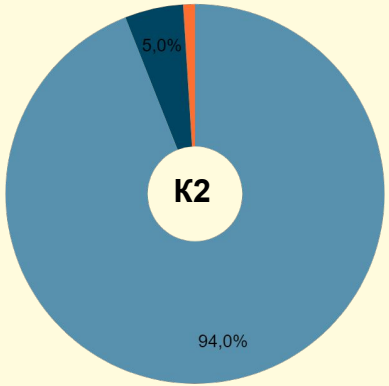
С помощью добавления синтетических данных удалось достичь равномерного распределения по К1



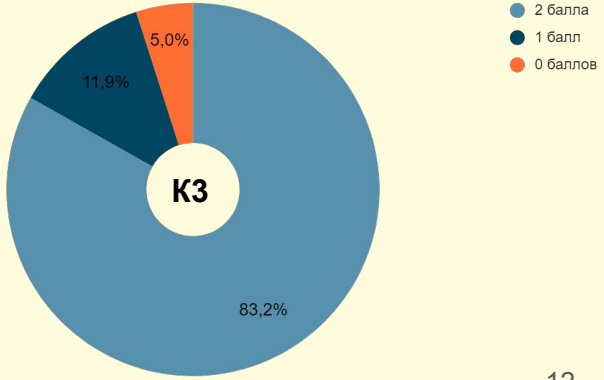
Решение коммуникативной задачи



Организация текста



Языковое оформление текста



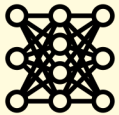
EDA

Вывод

> **Добавление сгенерированных данных** способствовало более равномерному распределению между оценками в 4, 5 и 6 баллов. Распределение по работам от 1 до 4 баллов не изменилось.
> Достигнуто равномерное распределение по баллам для критерия К1. Критерии К2 и К3 остались не тронуты

3 критерия

К1

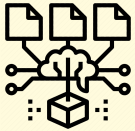


Для оценки текста по критерию К1 была дообучена модель BERT.

К2

Написан ру-скрипт, который подсчитывает ключевые слова и структурные элементы, такие как абзацы, для выставления оценки по К2.

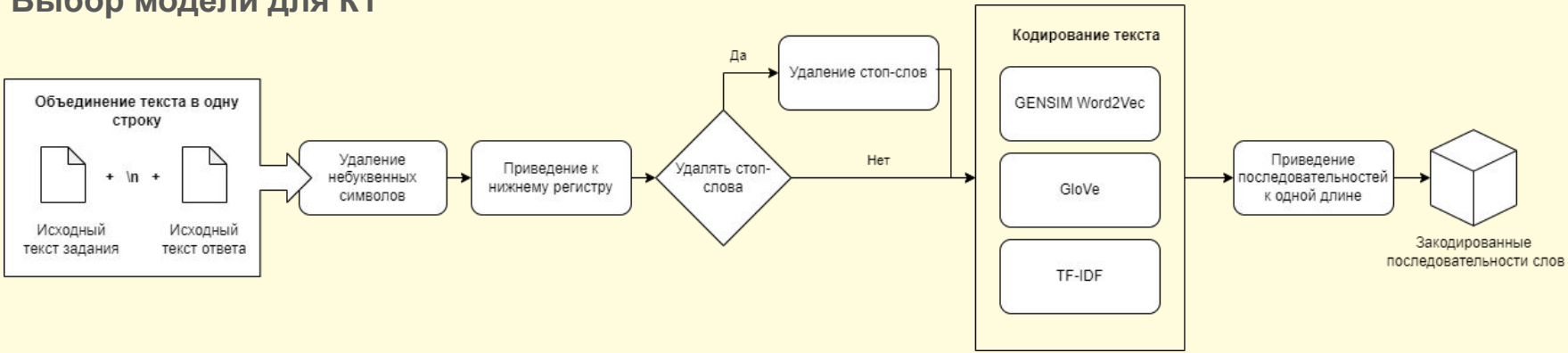
К3



Оценивание по К3 осуществляется с помощью предобученной модели FLAN-T5, включая генерацию комментариев по возможным исправлениям.

Предобработка данных

Выбор модели для К1



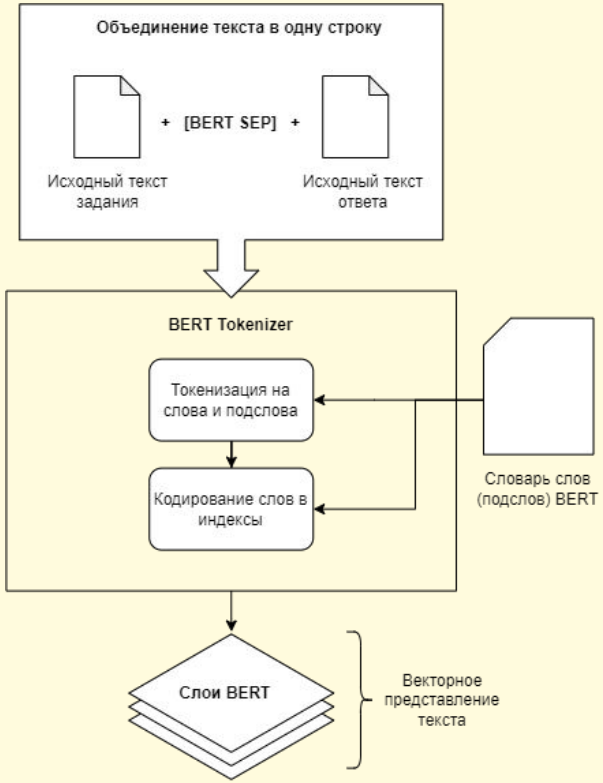
Текст задания и ответа объединялись в одну строку. Во время подбора классической модели для кодирования текста использовалось 3 метода векторного представления: GENSIM (Word2Vec), GloVe, TF-IDF. Для этих методов текст был минимально обработан:

1. Удалены все небуквенные символы
2. Символы приведены к нижнему регистру
3. (Опционально) удалены стоп-слова (“I”, “me”, “myself”, и т.д.)
4. Токенизация по словам (разделитель - пробелы).

Предобработка данных

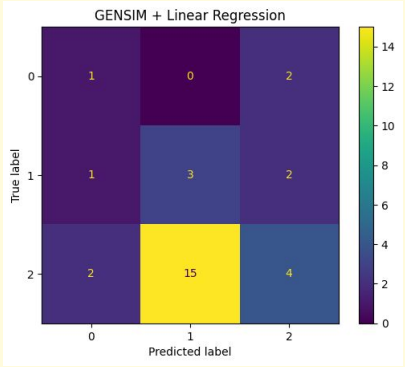
Выбор модели для К1

Для использования модели BERT текст задания и текст ответа был объединен в одну строку через специальный разделитель [BERT SEP]. Далее текст подавался на BertTokenizer - особый токенайзер для разбиения текста на токены согласно словарю BERT. Каждому слову (подслову) присваивается индекс, которые затем представляются в виде векторов в скрытом пространстве слоев BERT с размерностью 768.



Базовые модели

Качество на отложенной выборке (лучшая из Word2Vec/TF-IDF/GloVe)

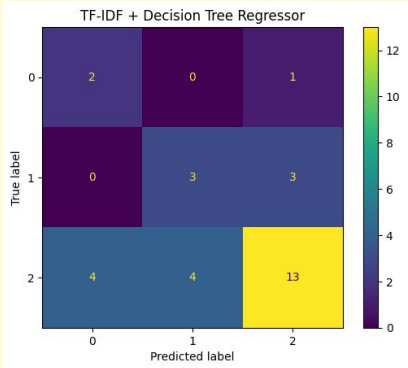


Linear Regression

Kappa=-0.09

RMSE=1.065

F1:
0: 0.29
1: 0.25
2: 0.28

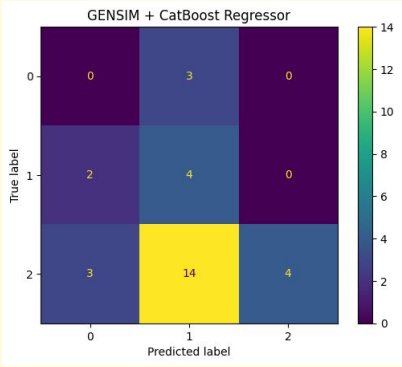


Decision Tree

Kappa=0.201

RMSE=0.949

F1:
0: 0.44
1: 0.46
2: 0.68

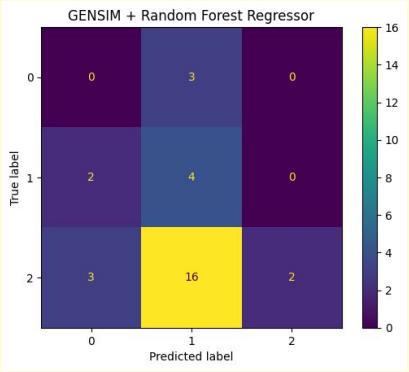


CatBoost

Kappa=0.094

RMSE=1.017

F1:
0: 0
1: 0.30
2: 0.32

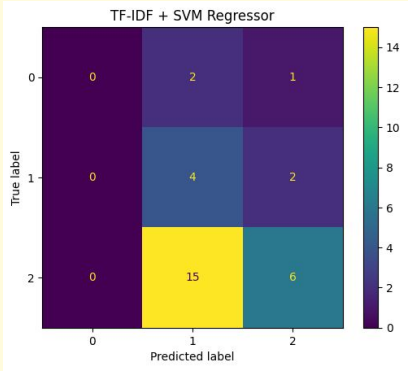


Random Forest

Kappa=0.046

RMSE=1.049

F1:
0: 0
1: 0.28
2: 0.17



SVM

Kappa=-0.036

RMSE=0.876

F1:
0: 0
1: 0.30
2: 0.40

Базовые модели

Для сравнения моделей между собой была использована метрика Каппа Коэна (Cohen's Kappa), позволяющая измерять меру согласия между двумя оценщиками.

Вывод

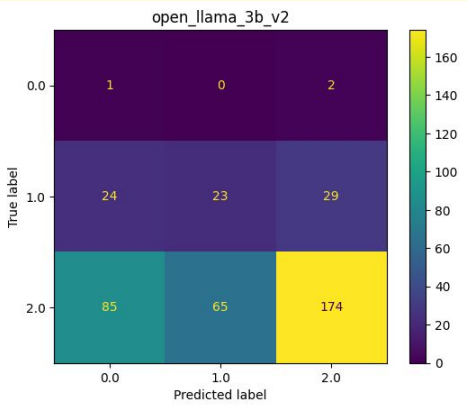
> **Наилучшее качество показала модель решающего дерева, обученная на выборке с кодированием текста методом TF-IDF (kappa=0.201).** Однако видно, что даже сравнив самые лучшие результаты среди разных семейств базовых моделей, целевая метрика (kappa) получается низкой, а зачастую и отрицательной, что говорит о нецелесообразном использовании базовых моделей для оценки К1.

Попробуем более сложные модели …

Языковые модели

Модель: llama2 **Проблема:** низкие метрики даже на обучающей выборке

3 млрд. параметров

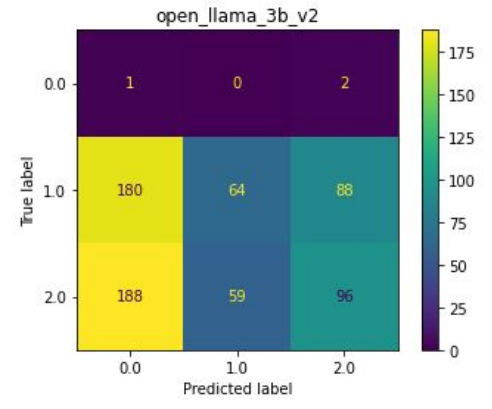


Качество на исходных данных (CV)

RMSE=1.1

F1:
0: 0.01
1: 0.28
2: 0.66

3 млрд. параметров

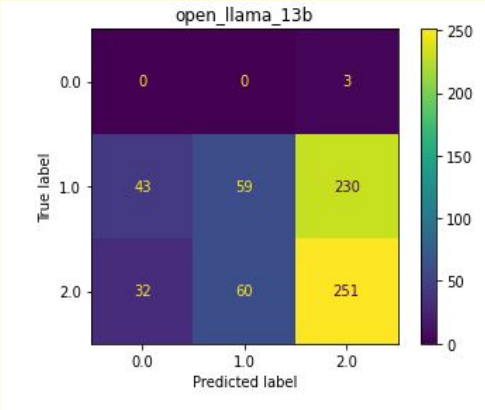


Качество на сгенерированных данных (CV)

RMSE=1.4

F1:
0: 0.01
1: 0.28
2: 0.36

13 млрд. параметров



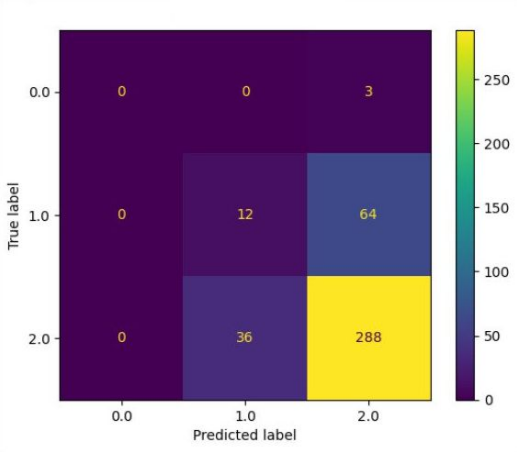
Качество на сгенерированных данных (CV)

RMSE=1.16

F1:
0: 0
1: 0.26
2: 0.61

Языковые модели

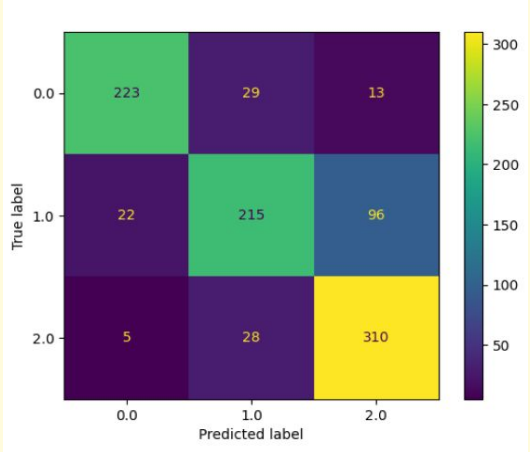
Модель: bert-base-uncased **Проблема**: склонность к предсказанию 2 баллов



Качество на исходных данных (CV)

RMSE=0.45

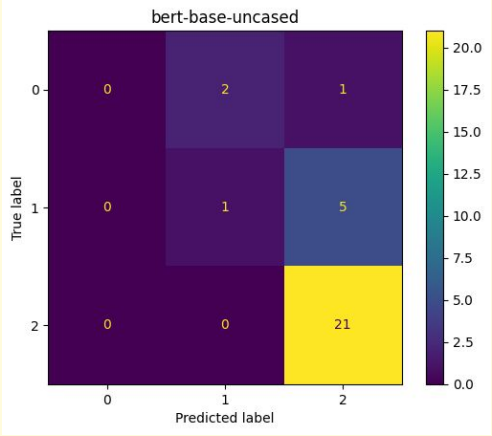
F1:
0: 0
1: 0.19
2: 0.85



Качество на сгенерированных данных (CV)

RMSE=0.46

F1:
0: 0.86
1: 0.67
2: 0.79



Качество на сгенерированных данных (отложенная выборка)

RMSE=0.606

Kappa=0.409

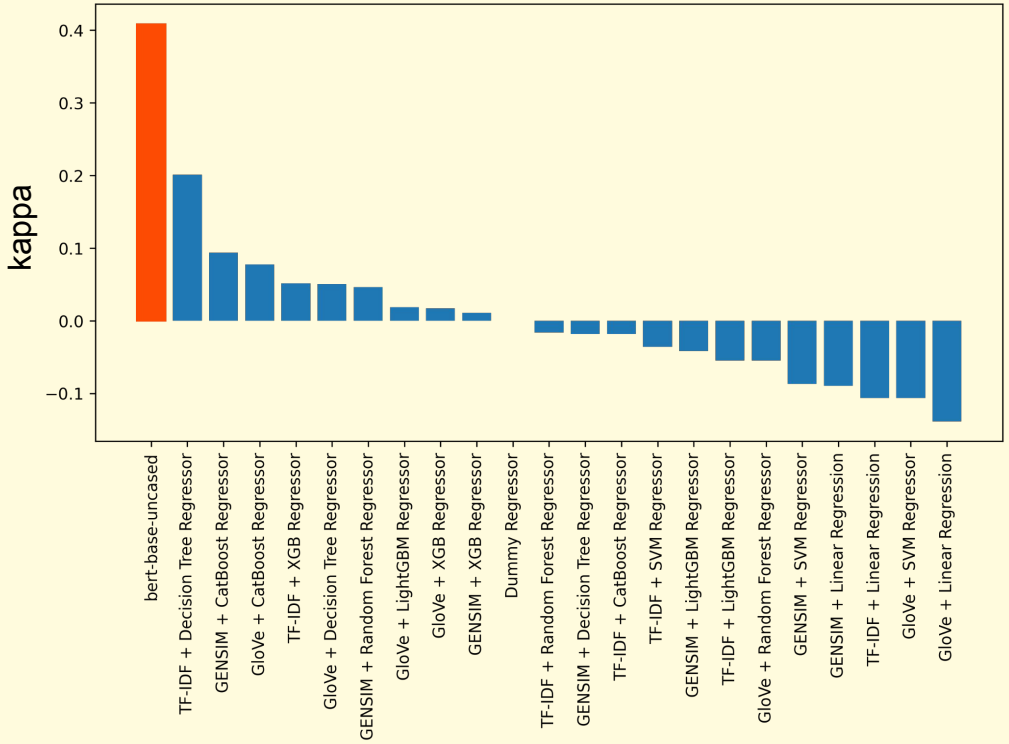
F1:
0: 0
1: 0.22
2: 0.88

Языковые модели

Вывод

> **Наилучшее качество из LLM показала модель bert-base-uncased (kappa=0.401).** Но видна склонность к предсказанию 2 баллов, что еще раз указывает на дисбаланс обучающей выборки. Для llama-2 эта метрика не измерялась, но по матрице ошибок видно, что на обучающей выборке качество оценки критерия очень низкое и близко к случайному угадыванию.

Сравнение моделей

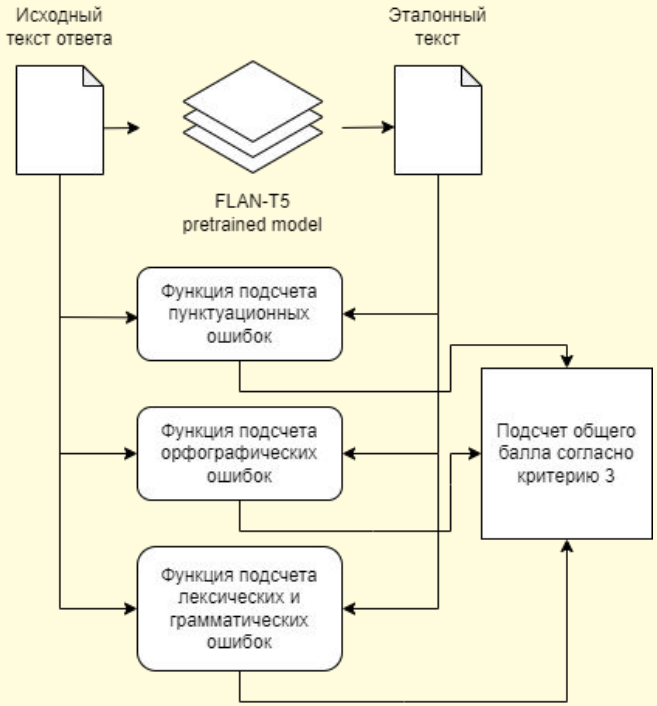


Вывод

Среди проведенных экспериментов с моделями наилучшее качество показала LLM модель BERT (kappa=0.4), однако при этом видна склонность к предсказанию максимального балла (2) по критерию 1. Стоит отметить, что решающее дерево, с применением TF-IDF лучше всего определяла работы для баллов 0 и 1, что наталкивает на мысль о возможном применении комбинации этих двух моделей.

Ссылка на ноутбук

Оценка языкового оформления текста (К3)

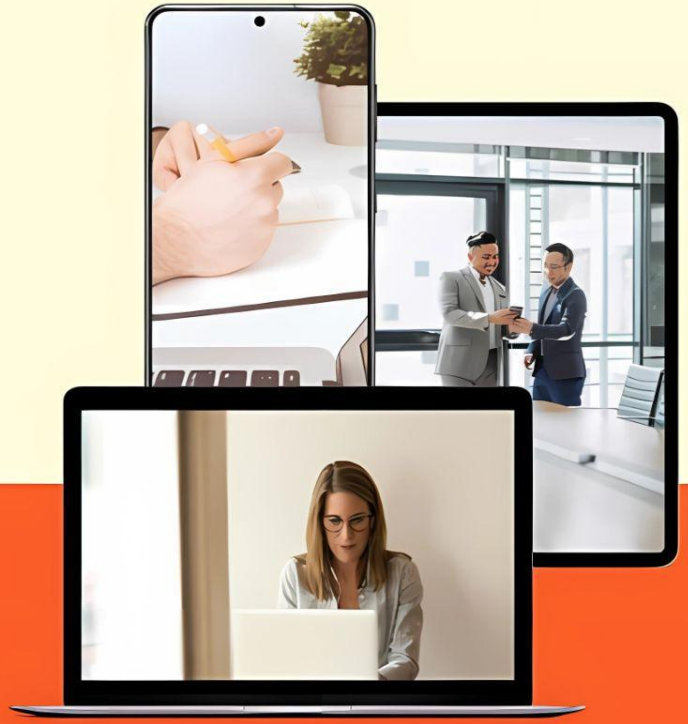


Для оценки критерия 3 был реализован скрипт, который использовал исправленный текст ответа, сгенерированный с помощью предобученной языковой модели FLAN-T5, и исходный текст письма для поиска различий. Основываясь на разнице между текстами, подсчитываются ошибки в соответствии с критерием, а именно:

1. Пунктуационные ошибки
2. Орфографические ошибки
3. Лексические и грамматические ошибки

Благодаря сгенерированному моделью “эталонному” тексту, пользователю в качестве обратной связи доступен просмотр неверно написанных предложений.

Ссылка на код в репозитории



Streamlit application

- Общая информация о задании
- Разведочный анализ данных
- Обращение к модели

Telegram bot

- Обращение к модели
- Обратная связь
- Статистика

API

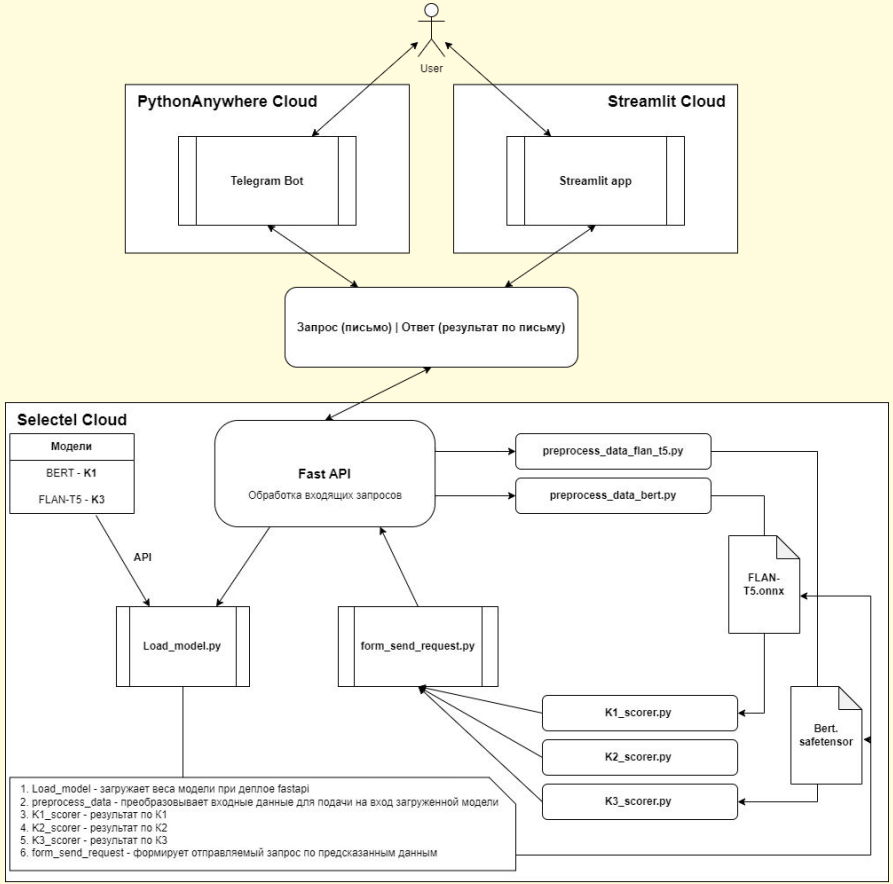
- Вся логика сервиса
- Предсказательная модель

Архитектура сервиса

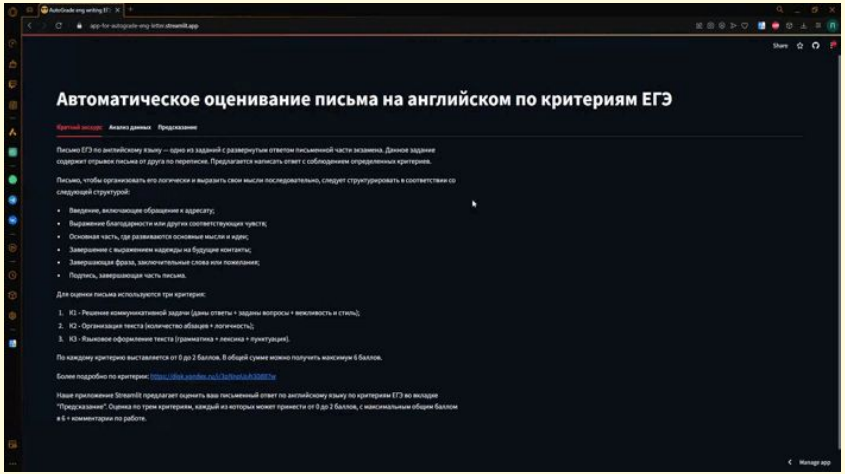
К1_scorer.py
1. На вход принимает обработанный текст;
2. По загруженной модели BERT делает оценку по К1
3. Возвращает цифру (0, 1, 2).

К2_scorer.py
1. Подсчитывает количество ключевых слов и абзацев;
2. Возвращает цифру (0, 1, 2).

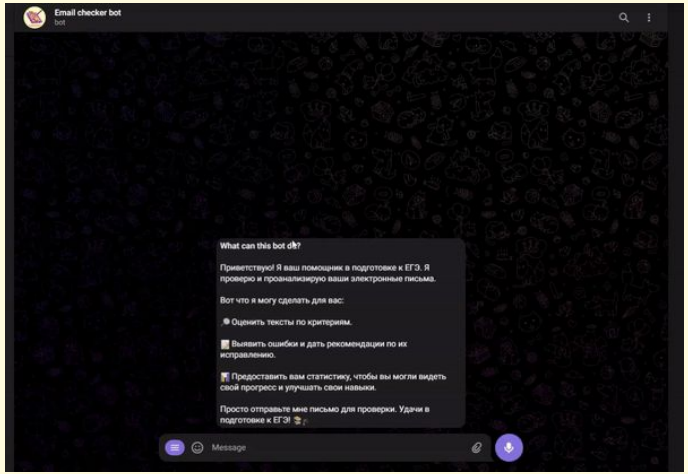
К3_scorer.py
1. На вход принимает обработанный текст;
2. По загруженной модели FLAN-T5 делает оценку по К3 и генерирует комментарии по исправлению;
3. Возвращает цифру (0, 1, 2) + комментарии.



Демонстрация работы сервиса



Streamlit-приложение



Telegram-бот

Ссылка на приложение

Ссылка на бота

Проверка работоспособности сервиса

Проверочные тексты для заданий

Для 1 задания

Dear Mark,
Thanks a lot for your recent email. I'm always happy to get messages from you!
In your email you asked me about housework. Well I always help my mother around the house. Personaly I am responsible for making the bed, doing the grocery shopping, and vacuuming the floor. To be honest, I hate washing the dishes and moping the floor. As for me, it takes too much time. However, I try to enjoy doing my household chones and I always listen to music at this time to keep posittive.
By the way, tell me more about your father's lawn mower. How does it function? What colour is it? What is the function warranty period of the lawn mower?
That's all for now. Drop me a line.
Best wishes,
Polina

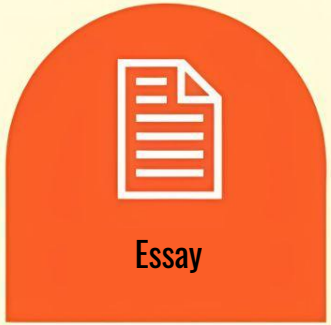
Для 3 задания

Dear Andy,
Thanks a lot for your recent email. I'm always happy to get messages from you.
In your email you asked me about my friends and my friend-making skills, well, i don't make new friends on purpose because i like my close friends. also i have a lot of acquaintances. i have a large number of cases then i got acquainted in social media. this happens most often on instagram. well, i made a good friend on the internet. therefore i consider that online friends are as good as real friends. just you can't see him often and there is a great distance between you.
by the way, tell me more about sister's wedding. where will the wedding take place? how many guests will be on the wedding? do your sister have the big wedding or small?
sorry, i have to go now. drop me a line.
best wishes,
Kate

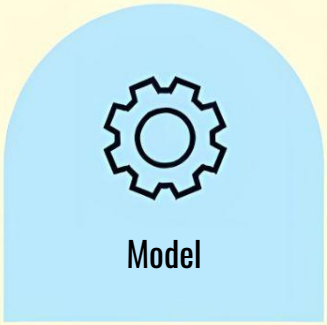
Для 2 задания

Dear Nora,
Thank you for your recent e-mail. I'm always glad to get messages from you.
In your e-mail you asked me about dreams. I'm eager to be successful and I'm crazy about my goals. Well, I want to live in Moscow and have my own flat.Besides, I'm want to graduate the best university in my country. In my opinion, I should be ambitious, patient, intelligent for my dreams. Moreover, I have to think outside the box. As for telling about dreams, I don't tell anyone. In my mind, you should achive your dreams and show results.
By the way, tell me more about your elder brother. How old is your sibling? What's he look like? Does he study at school or at university?
That's all for now. Looking forward to your answer!
Best wishes,
Nastya

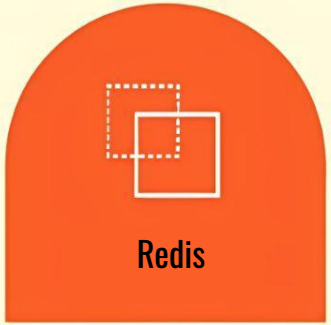
Future plans



Реализация полного функционала для второй письменной части ЕГЭ



Улучшение предсказательной модели



Обновление телеграмм-бота: реализация системы webhooks и интеграция с redis

Разработка приложения на android