Телеграм-бот для удобной работы со словарями малоресурсных языков

Авторы:
Ознобихин Арсений Романович, БПМИ213
Зайцев Федор Васильевич, БПМИ213

Higher School of Economics
Faculty of Computer science

Словари малоресурсных языков

Проблема:

- Есть собранные данные (машиночитаемый словарь)
- Нет возможности удобной работы с ними

Решение:

- Фреймворк для обработки набора данных и работы с ними в формате телеграм-бота

Цели и задачи

Цель – создание удобного фреймворка для работы со словарями малоресурсных языков

Задачи:

- Предобработка:
  Парсинг словаря
- Бэкенд:
  Реализация поиска по словарю
- Фронтенд:
  Реализация телеграм-бота

Предобработка

- Мы попытались обработать Чамалинско-русский словарь[1] – у нас ничего не получилось((
- По итогу для работы были взяты словари из проекта ***lang-tasks***[2] в формате .xml

[1] Магомедова П. Т. Чамалинско-русский словарь / П. Т. Магомедова. – Махачкала : Дагестанский научный центр РАН, 1999.
[2] https://github.com/kod-odin/lang-tasks

Бэкенд

- Ищем нужный результат итеративно, используя встроенные методы библиотеки python3 для работы с .xml файлами
- При этом обрабатываем различные варианты структуры .xml файла

Проблемы:

- Отсутствие единообразия разметки

Фронтенд: Telegram Bot API

Плюсы:

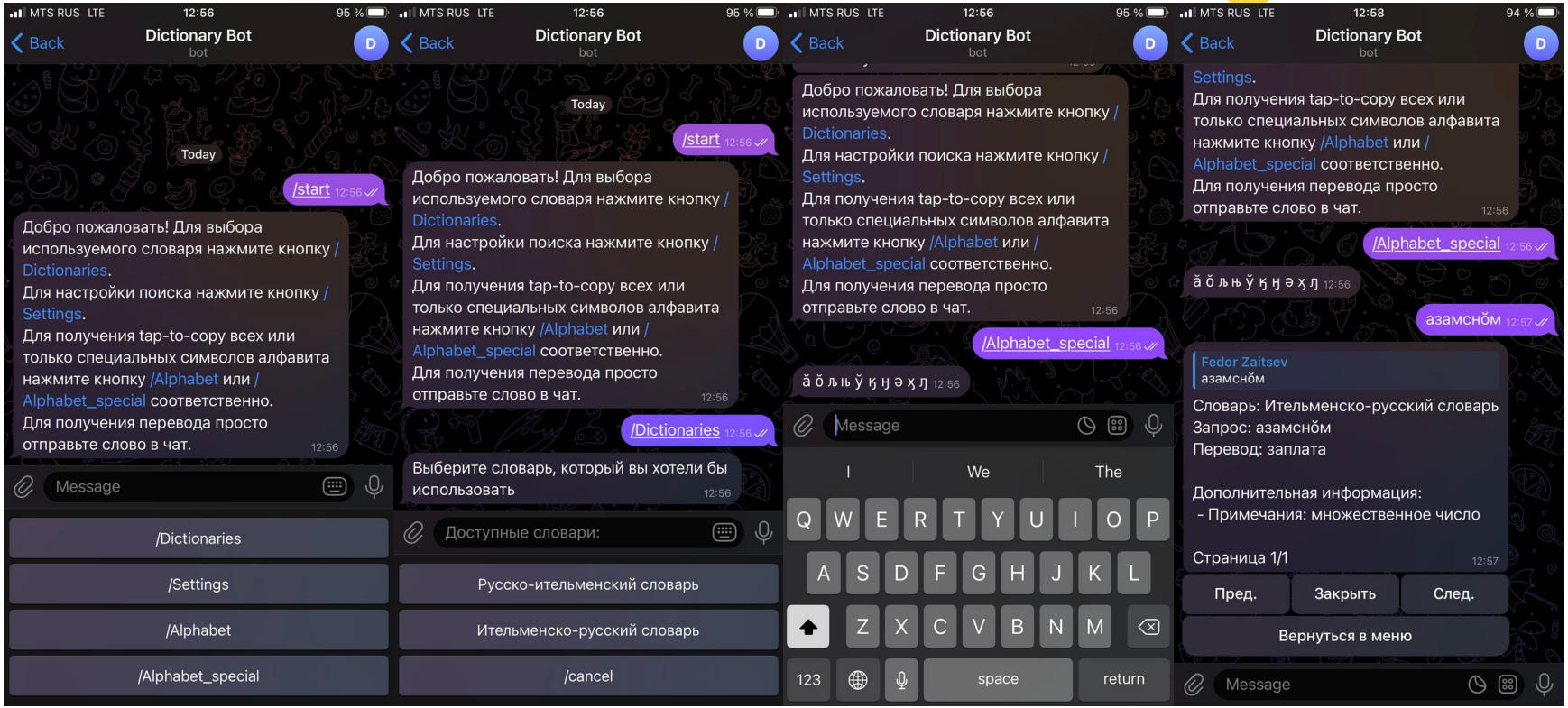
- Универсальность и единообразие пользовательского опыта на любом девайсе
- Удобство и доступность для пользователей вследствие популярности мессенджера Telegram
- Удобство разработки и запуска готового бота.

Фронтенд: Telegram Bot API

Использованные концепты:

- Асинхронность (***aiogram***)
- Система диалогов для выбора пользовательских конфигураций (***Finite State Machine***)
- Удобные клавиатурные и инлайн кнопки
- Частичный или полный вывод алфавита в *tap-to-copy* формате для упрощенного ввода специальных символов

Фронтенд: Telegram Bot API



Результаты

Реализована система для работы со словарями в формате телеграм-бота.

Преимущества фреймворка:

- Легкая расширяемость на любые словари нужного формата
- Простой запуск системы в работу
- Юзер-френдли интерфейс приложения

Перспективы развития проекта:

- Оптимизация поисковых алгоритмов и методов хранения данных
- Добавление возможности работы с новыми форматами словарей
- Улучшение пользовательского опыта
- Добавление проверки на правописание
- Добавление новых фичей (TTS, STT, поиск по корпусу, etc)

Спасибо за внимание!

- https://github.com/ARS404/dictionary_project