Команда "Капибары"

СОЗДАНИЕ КОРПУСА ОСЕТИНСКОГО ЯЗЫКА В МАШИНОЧИТАЕМОМ ВИДЕ

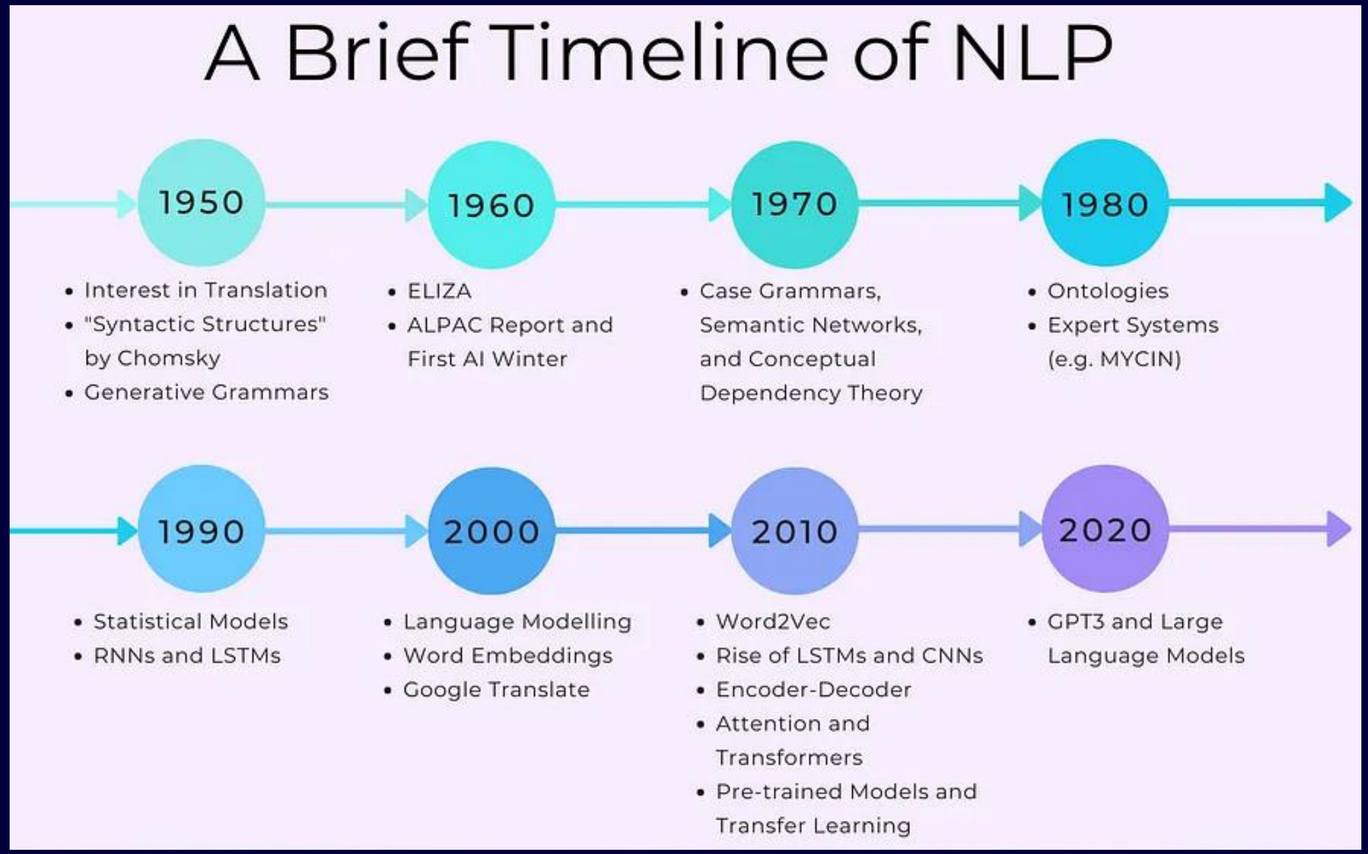
Игитян А. Г.

Роговецкий Р. С.

Лалиев Д. А.

Северо-Осетинский государственный университет
имени К. Л. Хетагурова

- NLP - область искусственного интеллекта, связанная с обработкой естественного языка
- NLP технологии стремительно развиваются
- NLP технологии находят применение в анализе малоресурсных языков, что помогает сохранять культурное наследие



Эволюция NLP технологий

Постановка задачи

Необходимы данные на языке, который будет анализироваться.
Сбор достаточного количества данных является важной задачей.

Сбор данных для создания корпуса осетинского языка.

- Определить источники текстов на осетинском языке
- Собрать данные с выбранных источников
- Структурировать данные
- Создать корпус для дальнейшего использования в NLP задачах

Методы решения

Ручной парсинг

Процесс сбора информации путем ручного копирования нужной информации.

- Времязатратность
- Сложность работы с большими объемами данных
- Высокая вероятность ошибок при копировании информации
- Невозможность автоматизировать процесс сбора данных

API

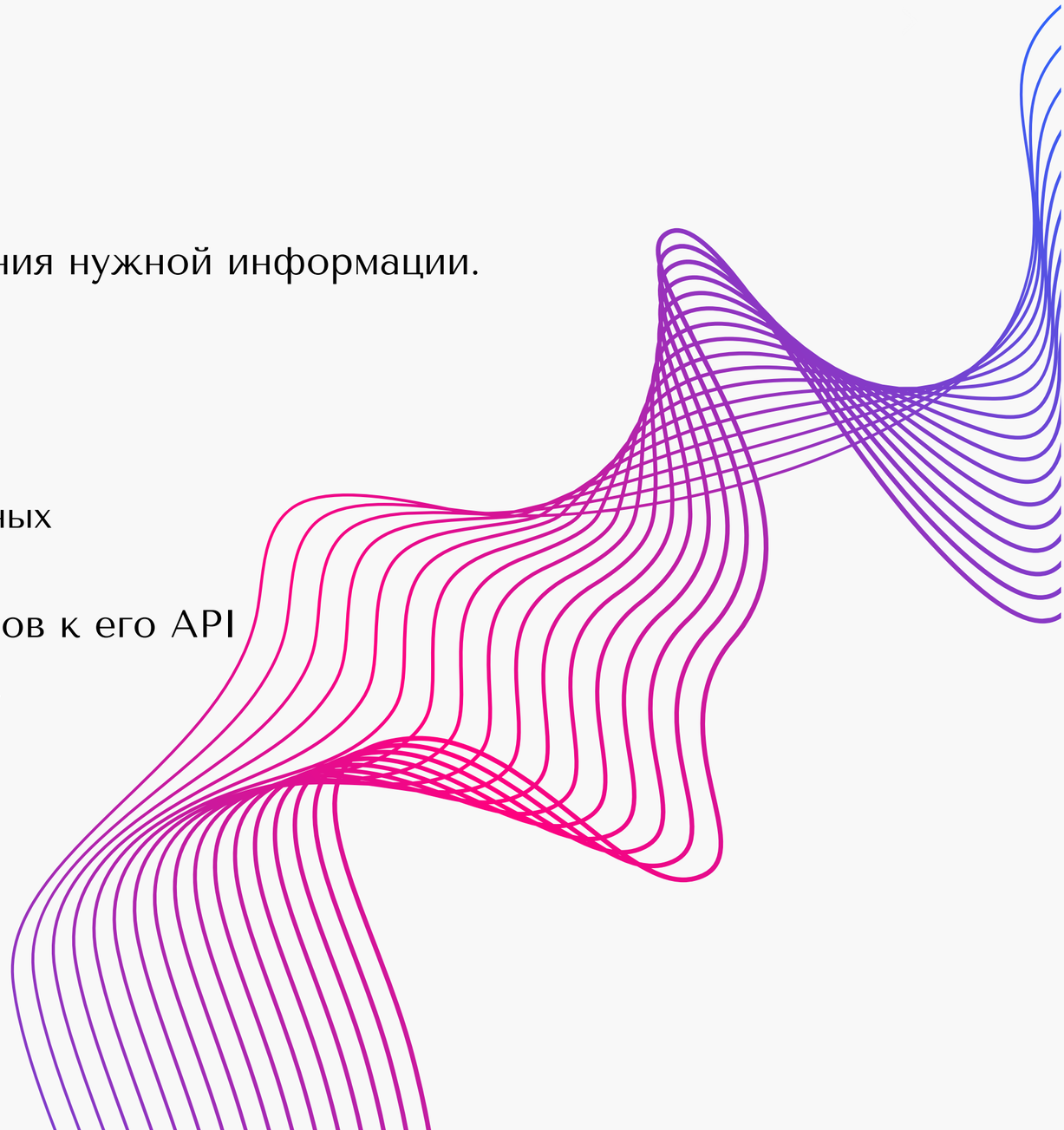
Способ получения данных из источника путем запросов к его API

- Сайт ironau.ru не предоставляет API для доступа к своим данным

Веб-парсинг

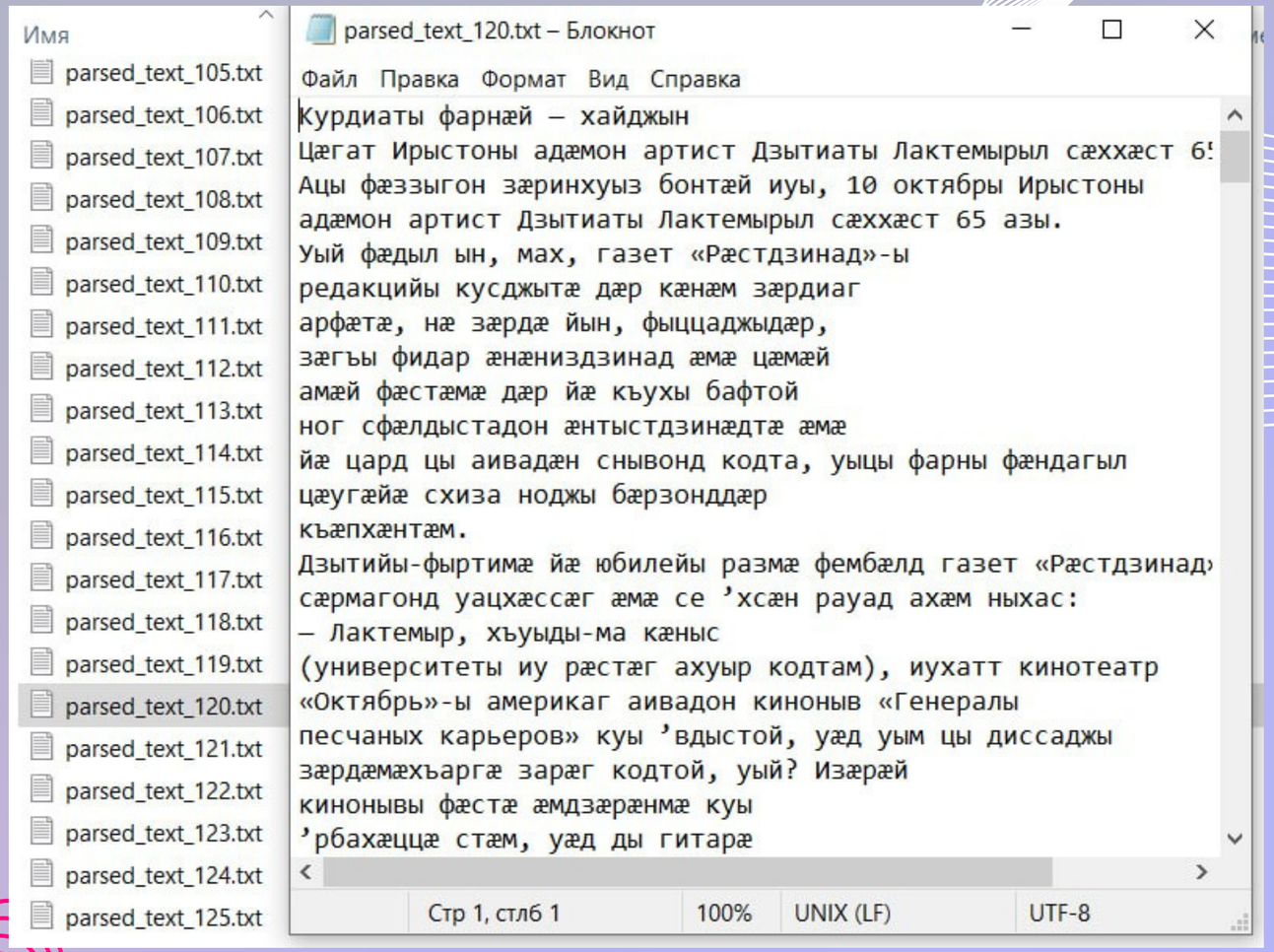
Процесс автоматического извлечения информации с веб-сайтов

- Быстрый и удобный доступ к данным



Результат работы

- Собран корпус текстов на осетинском языке
- 203 txt файла с текстами
- Разработан парсер для автоматического сбора данных
- Предоставлена возможность для дальнейших исследований в области NLP для осетинского языка



Дальнейшие возможности работы

- Расширение корпуса с помощью дополнительных источников данных
- Создание системы автоматического анализа осетинского языка
- Интеграция корпуса в различные NLP-приложения для осетинского языка

Спасибо за внимание!

Игитян А. Г.

https://github.com/MeikoFudo

Роговецкий Р. С.

https://github.com/Rogowiecki

Лалиев Д. А.

https://github.com/iddave

Ссылка на GitHub проекта

https://github.com/MeikoFudo/PROJECT_IT_SCHOOL

Команда "Капибары"