Морфологический анализатор и корпус для азербайджанского языка

Этот проект представляет собой приложение на Python с графическим интерфейсом для морфологического анализа и аннотирования предложений. Он позволяет:

- Выполнять автоматический морфоанализ слов;

- Редактировать теги вручную;

- Добавлять аннотированные предложения в корпус;

- Экспортировать корпус в форматы CoNLL-U, CSV, JSONL и Excel.

📦 Структура проекта

morpho\_corpus/

├── annotator\_ui.py # UI-приложение (Gradio)

├── text\_input.py # Токенизация и подготовка текста

├── ml\_models.py # Предсказание морфологических тегов

├── corpus\_db.py # Работа с файлом корпуса

├── export\_corpus.py # Экспорт в CoNLL-U

├── export\_csv.py # Экспорт в CSV

├── export\_jsonl.py # Экспорт в JSONL

├── export\_excel.py # Экспорт в Excel (.xlsx)

└── corpus/

└── corpus.json # Основной файл с корпусом

🚀 Установка

pip install gradio pandas openpyxl

🖥 Запуск приложения

python annotator\_ui.py

🧠 Обучение и настройка модели

Файл ml\_models.py содержит базовый шаблон для морфологического предсказания.

1. Соберите размеченный корпус;

2. Экспортируйте в corpus/corpus.conllu;

3. Обучите свою модель;

4. Подключите её в ml\_models.py.

🧩 Расширение графа (NeoFUGI или Neo4j)

Используйте графовую модель с узлами (корни, аффиксы) и рёбрами (морфологические переходы).

Поддерживаются: Neo4j, NetworkX. Граф можно использовать для предсказания форм или генерации слов.

📤 Экспорт корпуса

python export\_corpus.py # CoNLL-U

python export\_csv.py # CSV

python export\_jsonl.py # JSONL

python export\_excel.py # Excel (.xlsx)

🌍 Публикация корпуса

1. Проверьте CoNLL-U формат;

2. Добавьте метаданные;

3. Опубликуйте на GitHub, Zenodo или UD;

4. Лицензия: CC BY-SA 4.0 или MIT.

🤝 Сотрудничество

Проект открыт для лингвистов и NLP-исследователей.

Контакт: [вставьте email/Telegram]

📜 Лицензия

MIT License — свободное использование для научных и образовательных целей.