ОПИСАНИЕ

Кросс-лингвистический морфологический анализатор на основе использования аннотированных данных языков N для анализа другого родственного языка X. При этом не используются аннотированные данные для целевого языка X; все языки N + X являются родственными друг другу славянскими языками.

Языки N:

1. Russian
2. Czech
3. Polish
4. Ukrainian
5. Belarusian
6. Bulgarian
7. Macedonian
8. Slovenian
9. Serbian
10. Croatian
11. Silesian

Язык X: не определен

ЦЕЛЬ И ЗАДАЧИ

- **Исследовательские задачи:**

❏ роль когнатов в проекте;
❏ поиск инструментов для обнаружения когнатов;
❏ поиск методов и инструментов для эффективного использования пословных и побуквенных эмбеддингов в проекте;
❏ какие методы и алгоритмы можно использовать для обеспечения лучшего результата морфологического анализатора;

- **Инженерные задачи:**

❏ сбор и обработка данных;
❏ обучение модели;
❏ конечный продукт: кросс-лингвистический морфологический анализатор для славянских языков.

ДАННЫЕ И МЕТОДЫ

Данные: Wiki Dumps для языков N (https://dumps.wikimedia.org/backup-index.html)

Обработка данных, использованные методы и инструменты, результаты:

- ❏ Извлечение текстовых данных из дампов википедии с помощью WikiExtractor;
- ❏ Написание уникальных регулярных выражений и создание списков частотностей для каждого языка: https://github.com/NIS-2018-CROSS-M/cross-lingual-morph-analysis/blob/master/DATA/ANALYSIS.md
- ❏ Транслитератор для сербского языка из кириллицы в латиницу – разметка списка частотности с помощью модуля apertium-hbs: https://github.com/NIS-2018-CROSS-M/cross-lingual-morph-analysis/tree/master/DATA/sr_transliteraton
- ❏ Разметка списков частотностей с помощью модулей Apertium для каждого языка: https://github.com/NIS-2018-CROSS-M/cross-lingual-morph-analysis/blob/master/DATA/ANALYSIS.md
- ❏ Извлечение из размеченных списков частотностей первых 10 000 словоформ, относящихся к открытому лексическому классу <n>, <vblex>, <adj>: https://github.com/NIS-2018-CROSS-M/cross-lingual-morph-analysis/blob/master/DATA/selector.py
- ❏ Конвертирование размеченных данных для каждого языка в файлы формата .conllu

ОБЯЗАННОСТИ

- Oleg:
  - calma code reimplementation
  - multilingual embeddings approaches exploring
- Vlad:
  - frequency lists
  - morphological analysis
- Nastya:
  - cognate research
  - Serbian transliterator
- Lorenzo:
  - cognate research
  - open-category forms extraction

ЭТАПЫ

	Этап	Влад	Лоренцо	Олег	Настя
Nov Dec	Подготовка данных (1)	1) Извлечение текстовых данных из дампов википедии 2) Написание уникальных регулярных выражений и создание списков частотностей для каждого языка 3) Разметка списков частотностей для каждого языка	1) Извлечение из размеченных списков частотностей первых 10 000 словоформ, относящихся к открытому лексическому классу <n>, <vblex>, <adj> 2) Поиск инструментов для обнаружения когнатов	1) Воспроизведение кода, описанного в опорной статье 2) Исследование методов использования эмбеддингов в кроссязыковых задачах	1) Транслитератор для сербского языка 2) Поиск инструментов для обнаружения когнатов
Jan	Создание начальной модели (2)	To be filled	To be filled	To be filled	To be filled
Feb-April	Циклическая доработка модели (3)	To be filled	To be filled	To be filled	To be filled
April-May	Написание статьи (4)				