

Методы и инструменты анализа текста для метрик сходства в больших национальных текстовых корпусах: на примере Латвийской национальной электронной библиотеки и Национального репозитория академических текстов Украины



Латвийская часть проекта финансируется Латвийским государственным агентством развития образования.

Проектная группа на латвийской стороне

Валтерс Щербинскис	исследования
Катрина Тейване-Корпа	исследование, тестирование
Анда Баклане	исследование методов, координация разработок, тестирование
Валдис Саулеспуренс	исследование и разработка среды и инструментов, исследование методов
Артурс Жогла	извлечение и создание текстовых корпусов из электронной библиотеки, поддержка в разработке
Карина Бандере	управление проекта
Янис Крейцбергс	координация сотрудничества, организация мероприятий, коммуникация результатов

Цели проекта

1. Найти и выбрать наиболее подходящий метод анализа текста для выявления сходства в больших текстовых корпусах
 - а. Для украинской стороны существует текстовый корпус научных текстов относительно последних лет.
 - б. Для латвийской стороны у нас есть корпус исторических газет.
2. Разработать среду тестирования микросервиса для анализа текста в выбранных корпусах.
3. Провести пару кейсовых исследований для тестирования разработанного инструмента.

Цели проекта

4. Привлечь местное сообщество исследователей и энтузиастов гуманитарных и социальных наук к тестированию и улучшению микросервиса, а также к распространению информации о нашей деятельности и намерениях.

5 Распространение опыта и результатов через научное сообщество - научные публикации, семинары, доклады на конференциях.

6. Опыт сотрудничества и обмена знаниями, а также укрепления связей между Украиной и Латвией в области цифровых исследований, анализа текста, открытой науки, репозиторий и академической честности.

Возможности применения

Мы ищем не решение для обнаружения плагиата, а инструмент исследования и расследования распространения и развития повествований, идей, историй, интерпретаций через различные источники и периоды времени. Например:

- Изменения в языке для лингвистики
- Контент-анализ, распространение мнений, интерпретаций, клише
- для журналистских исследований - качество журналистики, распространение новостей и повествований
- Последствия цензуры
- может быть еще больше способов использования, которые мы еще не можем предсказать