Проект “ANTIDICT”

База морфологических структур, допустимых в современном русском языке

Студенты:
Алексей Доркин
Владислава Смирнова
Антон Вахранев

Куратор: Варвара Магомедова

Цели проекта и задачи на 2019-2020 год

Цель 1 (академическая)
Изучить:
– заимствования
– экспрессивные формы
– ошибки

Цель 2 (образовательная)
Изучить разные библиотеки для обработки текста (mystem, fasttext, pymorphy, polyglot)

Предмет изучения
Продуктивная морфология

Цели проекта и задачи на 2019-2020 год

Задача 0 (глобальная)
Определить формальные критерии, по которым мы будем разграничивать слова, относящиеся к разным категориям

Задача 1 (попроще)
Выяснить, что характерно для заимствований (корни/морфемы)

Задача 2 (посложнее)
Попробовать разграничить слова с ошибками, слова с опечатками и слова с намеренными искажениями

Задача 3
Выявить недокументированные корни и морфемы, характерные для экспрессивных форм

Бэкграунд

Подобные проекты

Пока что не обнаружено аналогичных проектов.
Есть корпуса ошибок, но они основаны на текстах.
Мы изучаем слова вне контекста и их морфологическую структуру.

Статьи

Статья, написанная предыдущей командой

Актуальность, новизна и практическая значимость

Для науки

Исследование молодой области знаний с новой стороны:
– Как классифицировать необычные объекты?
– Как устроены современные заимствования, экспрессивная лексика?
– Как отличить ошибку от опечатки и намеренного искажения?

Заинтересованность в проекте со стороны иностранных коллег

Практическое применение

Создание сайта с возможностью поиска по словам и морфемам

Данные

Было на момент начала работы:
Массив из 8 млн слов, полученных из ГИКРЯ после удаления дубликатов.
Предыдущая команда разметила слова (но им это не помогло).

Есть в данный момент:
Отсортированный массив слов

Что мы успели сделать

Для академической цели

1. Отсортировали слова (убрали дубликаты, но не убрали одинаковые слова)
2. Выкачали словарь заимствований Дьякова

Для образовательной цели

1. Начали разбираться с pandas

Что нам предстоит сделать в ближайшее время

1. Добыть доступ к суперкомпьютеру для получения больших вычислительных мощностей

2. Использовать возможности pandas и mystem на наших данных

Репозиторий

https://github.com/wksmirnowa/antidict