

ЛАБОРАТОРНАЯ РАБОТА №5

Дата сдачи - 7 декабря 2022

Общая постановка

Разработать/выбрать классификатор лексико-грамматических свойств для английского языка.

Разработать инструментарий с GUI-интерфейсом для работы с ЛБЗ, обладающий следующей функциональностью:

1) По классификатору должна вызываться справка с пояснением значений кодов.

2) Аннотирование корпуса текстов с разработкой соответствующего инструментария для выполнения аннотирования и просмотра текстов:

<https://www.nltk.org/book/ch05.html>

https://www.tutorialspoint.com/python_text_processing/python_tagging_words.htm#:~:text=Tagging%20is%20an%20essential%20feature,the%20tags%20for%20each%20word.

Промежуточная сдача по 2-м пунктам - 16 ноября 2022

3) Получение и отображение аннотированного словаря (на основе полученного аннотированного корпуса), где каждому слову приписаны все возможные вне контекста лексико-грамматические коды (ЛГК). Каждому словоупотреблению сопоставить его каноническую форму (лемму) с соответствующим ЛГК.

Реализовать весь функционал ЛБ №1:

– сортировку словаря по алфавиту как по возрастанию, так и убыванию;

– поиск и отображение слов по введенному шаблону;

– исправление слова с возможностью корректировки его в текстах, где оно встречается, и с суммированием частот, если есть правильное написание;

– удаление слова из словаря (с уточняющим вопросом) – небезопасная операция, чтобы не удалить случайно нужное;

– добавление нового слова с возможностью приписать ручную ЛГК и каноническую форму.

4) Снять следующую статистику с корпуса текстов:

- абсолютная частота встречаемости каждого кода;
- абсолютная частота встречаемости пары *слово_код*;
- абсолютная частота встречаемости пар кодов (рядом стоящих).

Если слова из словаря нет в тексте, но значения остаются нулевыми.

5) Пополнение корпуса и словаря новыми текстами с пересчетом статистики.