

Лабораторная работа 4. Исправление ошибок

В вашу тему на Piazza загружен текст.

На странице Resources загружен словарь (dict1.txt)

Ход работы

1. Предобработка текста

1.1 Текст нужно разделить на слова.

1.2 Удалить следующие знаки препинания: ! ? , ; . : « () »

Не удаляйте из слов дефисы.

1.3 Перевести все буквы в строчные (маленькие). Например, "Средний" - заменить на "средний".

Приводить слова к нормальной форме не нужно, так как в словаре присутствуют различные словоформы.

Например, которая которого которое которой которым которому которую которые который которым которыми которых. Это всё разные словоформы. Всего в словаре 4773 разных словоформ, отсортированных по алфавиту.

2. Первичные расчёты

2.1 Посчитайте словоформы в своём тексте

2.2 Посчитайте разные словоформы

2.3 Посчитайте сколько разных словоформ из вашего текста присутствуют в словаре

Обратите внимание, что в словаре после слова через пробел написано число – это частота встречаемости во всём тексте.

3. Поиск и исправление ошибок

3.1 Посчитайте, сколько словоформ не присутствует в словаре ("потенциальные ошибки")

3.2 Найдите для каждого из них редакторское расстояние до ближайшего слова.

Редакторское расстояние – это минимальное количество разрешённых операций, необходимых для превращения одной строки в другую. В настоящем задании разрешены следующие операции: вставка одного символа, удаление одного символа и замена одного символа на другой. Допустимо в строку вставить символ «пробел», превратив строку в две.

3.3 В настоящем задании, если редакторское расстояние равняется 1 или 2, то словоформа в вашем тексте признаётся ошибочной и её нужно заменить словоформу на соответствующую словоформу из словаря.

Если в словаре оказалось несколько словоформ с одинаковым редакторским расстоянием до ошибочной словоформы из текста, то нужно заменить на то, у которого частота выше.

4. После поиска и исправления ошибок повторите расчёты:

4.1 Посчитайте словоформы в своём тексте

4.2 Посчитайте разные словоформы

4.3 Посчитайте сколько разных словоформ из вашего текста присутствуют в словаре

5. Выведите все "потенциальные ошибки" в порядке встречаемости в тексте в следующем виде: словоформа из текста - словоформа из словаря - редакторское расстояние.

Если удалось исправить не все "потенциальные ошибки", то нужно вывести только неисправленное слово из текста с пометкой "не найдено".

Например,

1) Сридний - средний - 1

2) гепаталамус - гипоталамус - 2

3) гепотоламмуc - не найдено - >2

Ответы на вопросы нужно опубликовать в ответном сообщении на Piazza.

Также к ответам нужно прикрепить отчёт в формате pdf. Имя файла должно включать номер группы, ФИО и номер работы.

В отчёт включить.

1. Задание на лабораторную работу

2. Выданный вам текст

3. Код функции (метода) для подсчёта редакторского расстояния

4. Ответы на вопросы задания 2-5 и программный код, с помощью которого эти ответы были получены.

5. Текст с исправленными ошибками