Лабораторная работа №2

Реализация алгоритмов Стемминга

Цель работы: научиться реализовывать на выбранном языке программирования алгоритмы поиска по тексту однокоренных слов на основе поиска Стеммера Портера.

Краткие теоретические сведения

Стеммер Портера – алгоритм стемминга (нахождения основы слова для заданного исходного слова), опубликованный Мартином Портером. Алгоритм не использует баз основ слов, а работает, последовательно применяя ряд правил отсечения окончаний и суффиксов.

Вводим ряд определений:

Гласные буквы – а, е, и, о, у, ы, э, ю, я (буква ё считается равнозначной букве е).

RV – область слова после первой гласной. Она может быть пустой, если гласные в слове отсутствуют.

R1 – область слова после первого сочетания "гласная-согласная".

R2 – область R1 после первого сочетания "гласная-согласная".

Пример

В слове противоестественном:

RV = тивоестественном.

R1 = ивоестественном.

R2 = остественном.

Теперь определим несколько классов окончаний слова (табл. 1).

Таблица 1. Классы окончаний слова

Название класса	Окончания слов

PERFECTIVE	Группа 1*: в, вши, вшись.

GERUND	Группа 2: ив, ивши, ившись, ыв, ывши, ывшись.
ADJECTIVE	ее, ие, ые, ое, ими, ыми, ей, ий, ый, ой, ем, им, ым, ом, его, ого, ему, ому, их, ых, ую, юю, ая, яя, ою, ею.
PARTICIPLE	Группа 1*: ем, нн, вш, ющ, щ.
	Группа 2: ивш, ывш, ующ.
REFLEXIVE	ся, сь.
VERB	Группа 1*: ла, на, ете, йте, ли, й, л, ем, н, ло, но, ет, ют, ны, ть, ешь, нно.
	Группа 2: ила, ыла, ена, ейте, уйте, ите, или, ыли, ей, уй, ил, ыл, им, ым, ен, ило, ыло, ено, ят, ует, уют, ит, ыт, ены, ить, ыть, ишь, ую, ю.
NOUN	а, ев, ов, ие, ье, е, иями, ями, ами, еи, ии, и,ией, ей, ой, ий, й, иям, ям, ием, ем, ам, ом, о, у, ах, иях, ях, ы, ь, ию, ью, ю, ия, ья, я.
SUPERLATIVE	ейш, ейше.
DERIVATIONAL	ост, ость.
ADJECTIVAL	ADJECTIVAL определяется как ADJECTIVE или PARTICIPLE + ADJECTIVE. *Например: бегавшая = бега + вш + ая.*

* Окончаниям из группы 1 должна предшествовать буква а или я.

Правила:

1. При поиске окончания из всех возможных выбирается наиболее длинное. Например, в слове величие выбираем окончание ие, а не е.

2. Все проверки производятся над областью RV. Так, при проверке на PERFECTIVE GERUND предшествующие буквы а и я также должны быть внутри RV. Буквы перед RV не участвуют в проверках вообще.

Шаг 1

Найти окончание PERFECTIVE GERUND. Если оно существует – удалить его и завершить этот шаг.

Иначе, удаляем окончание REFLEXIVE (если оно существует). Затем в следующем порядке пробуем удалить окончания: ADJECTIVAL, VERB, NOUN. Как только одно из них найдено – шаг завершается.

Шаг 2

Если слово оканчивается на и – удаляем и.

Шаг 3

Если в R2 найдется окончание DERIVATIONAL – удаляем его.

Шаг 4

Возможен один из трех вариантов:

1. Если слово оканчивается на нн – удаляем последнюю букву.
2. Если слово оканчивается на SUPERLATIVE – удаляем его и снова удаляем последнюю букву, если слово оканчивается на нн.
3. Если слово оканчивается на ь – удаляем его.

Задание к лабораторной работе

Написать программу на выбранном языке программирования, реализующую описанный выше алгоритм для поиска по тексту однокоренных слов. Программа должна запрашивать имя входного файла и слово для поиска. Результатом работы программы должен быть файл, содержащий список однокоренных слов.

Контрольные вопросы

1. Что такое Стемминг?
2. Что такое Стеммер Портера?
3. Принципы работы алгоритма стемминга.

4. Роль морфологического анализа слов в системах поиска.
5. Аналоги Стеммер Портера.
6. Преимущества алгоритма Стемминга перед алгоритмами со словарем.
7. Недостатки алгоритма Стемминга.