МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ им. М.В. Ломоносова
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ
КАФЕДРА АЛГОРИТМИЧЕСКИХ ЯЗЫКОВ

Краткий отчёт студента о выполнении задания преддипломной практики

Ушивец Никита Алексеевич
Группа 424

Научный руководитель
Большакова Е. И.

2020

Задание преддипломной практики:

Создать программное средство обработки словосочетаний с существующими в русском языке именами числительных. На основе терминологии имён числительных, представленной в виде совокупности правил склонения с различного рода существительными русского языка, найти алгоритм обработки вводимых словосочетаний. Используя имеющиеся данные реализовать алгоритм, позволяющий автоматизировано склонять получаемые словосочетания. Так же с помощью обработки различного рода текстов исправлять неправильные словосочетания в них.

Путь решения

Выполнение задания преддипломной практики было разделено на три основные взаимосвязанные части: поиск и изучения необходимого теоретического материала, построение алгоритма относительно созданных правил, доработка и тестирование алгоритмов.

Теоретический материал из различных источников позволив выявить и разделить характерные черты склонения существительных с разными группами имен числительных русского языка, что и легло в основу разработки базы правил. Основной задачей на этапе программной реализации алгоритмов было правильное перенесения словесного описания базы правил в структурированный вид, который можно использовать в коде и с легкостью обновлять или модифицировать.

Вводимые данные

Для решения поставленной задачи необходимо было реализовать программу, которая принимала бы численное значение (ограниченное квинтиллионом) и словосочетание с обязательными параметром в виде существительного. Для работы я решил использовать язык программирование Python, с открытой библиотекой pymorphy2.
Если коротко описывать интерфейс - он содержит простой и понятный дизайн, необходимы были поля ввода текста и поле вывода массива с полученными после выполнения программы предложениями.

Этап реализации

Данная программа представляет с собой поэтапный разбор вводимых данных, а именно:

1. **Перевод численного значения в строку**

На вход программа получает любое число, ограниченное каким-то максимальным значением. Для начала определятся группа данного числительного (порядковый, количественный, собирательный), после чего вызывается отвечающая данной группе функция. При вызове первым делом идёт разбор числа на единицы, десятки, тысячи с помощью

написанных разделительных функций. Каждый разбор записывается в массив с котором и будет идти дальнейшая работа.

2. **Удаление лишних символов**

При работе с массивом разделенных значений остаются как ненужные символы, так и нужными по типу знака перед числом или запятой, ограничивающей целую часть числа. На данном этапе проходит глобальная отчистка, после чего стилистически корректный массив передается в работу со словарями.

3. **Использование словарей**

При посимвольном проходе массива ищется соответствующие значения в заранее определенных словарях. Существует несколько типов словарей, отвечающих группе числительного. Это может быть как окончание для порядковых чисел, так и список склоняемых групп. При нахождении значения в словаре, происходит замена изначального массива на строчные величины.

4. **Обработка числительного**

Однако массив состоит из взаимно несвязанных элементов. При очередном проходе устанавливается связь групп единиц, десятков, тысяч, которая изменяет строчный значения на правильные. Так учитывается род числительного, его месторасположение в массиве, изначально заданную группу (порядковое, количественное).

5. **Соответствие с существительным**

На данном этапе по средствам подключенной библиотеки pymorphy2, в частности используемого размеченного корпуса OpenCorpora, находится нужная информация о роде и числе существительного. Дополнительно проходя по массиву чисел, учитываются полученные данные и выдается конечная строка.

6. **Склонение по падежам (есть трудности с порядковыми числительными)**

Дополнительная функция работы с выданной строки состоит в том, что после сегментации - разделение на слова, происходит склонение каждого элемента по падежам. Полученные данные записываются в массив при нужном формате, который и выводится на экран.

7. **Поиск слов в заданном тексте (есть трудности в работе)**

С помощью библиотеки pymorphy2, по средству используемого tag'а и проверки на существование в тексте необычных нормальных форм слов, происходит выделение из всего текста слов представоющих собой численные значения. Этот массив строк приводится к нормальной форме для дальнейшего упрощения работы с ним. Далее, применяется

словарь наиболее распространённых имён числительных, по которому мы формируем массив чисел, закрепленных относительного своего места в тексте.

8. **Проверка на правильность использования словосочетаний (в разработке)**

Сложные случаи

При тестирование программной реализации алгоритмов были замечены сложные случаи словосочетаний, которые уступают ручному проведению операции:

- Существительное, не имеющее единственного числа (до сих пор есть трудности)

Когда существительное не имеет формы единственного числа (*ножницы, сутки и др.*), часто сложно образовать форму числительного, более 20, которое бы правильно согласовалось с таким числительным: (*сорок трое суток, сорок три суток*), правильный вариант: *сорок три дня*. С такими существительными сочетаются только числительные, оканчивающиеся на единицу или пятерку: *двадцать одни сутки*. Для того, чтобы обозначить количество других собирательных существительных (*например, ножниц, трусов, глаз*), употребляется слово «*штука*» или «*пара*»: *сорок восемь пар глаз, двадцать две пары трусов, семь штук ножниц.*

- Случаи, в которых употребляются собирательные существительные (до сих пор есть трудности)
    - С существительными люди, дети и названиями детёнышей животных: *семеро козлят, двое котят, пятеро детей*.
    - С существительными, означающими название лиц мужского пола: *четверо братьев, трое друзей*.

Вывод

В ходе работы были реализованы алгоритмы с помощью, которых решается часть поставленной задачи. Несмотря на то, что данная работа не может обработать идеально все случаи, она обрабатывает самые популярные из них, давая представлении о структуре подобных программ и тех сложностей, с которыми можно столкнуться. Итоговые результаты и оценки позволяют наметить дальнейшее усовершенствование и уточнение алгоритмов, а также закончить полноценный, заранее предусмотренный функционал.