Национальный исследовательский университет
Высшая школа экономики
Факультет компьютерных наук

Отчет по домашнему заданию №3

Синтаксический анализ, коллокации

студент 1 курса магистратуры
Зуев Кирилл Александрович

Москва, 2019

Содержание

1 Постановка задачи выбранного варианта (C)

Составить (с использованием любого модуля морфоанализа) программу, выполняющую синтаксическую сегментацию текста на русском языке на базе выбранных правил:

- сегментацию на простые предложения по знакам пунктуации **и/или**
- выделение неразрывных синтаксически связанных групп слов на основе локальных высоковероятных связей.

Протестировать программу на нескольких текстовых фрагментах (не менее 1-2 страниц).

2 Уточнение постановки задачи

В процессе выполнения задания я решил использовать два морфологических анализатора: **Mystem** (для токенизации текстов) и **PyMorphy** (для получения морфологических признаков каждого токена). **Mystem** тоже выдает морфологические признаки, но он снимает омонимию, оставляя только один вариант признаков, который в некоторых случаях оказывается неправильным и не позволяет корректно выполнить сегментацию. Поэтому я решил воспользоваться **PyMorphy**, так как он выдает все варианты разбора словоформы.

Мной выполнены оба пункта задания: и сегментация текста на простые предложения по знакам пунктуации, и выделение неразрывных синтаксически связанных групп слов. Для выделения групп использовались следующие высоковероятные связи (и их комбинация):

- прилагательное/причастие и существительное, если у них совпадает падеж;
- краткое прилагательное/причастие и существительное;
- прилагательное/причастие и существительное, если существительное является неизменяемым;

- предлог и существительное;
- частица не/ни и прилагательное/причастие/существительное.

Сегментация на простые предложения выполнена примитивно: если встречаем знак препинания, то отделяем эту часть. Это очень простой способ, который порождает много ошибок и неточностей. Например, разделение однородных членов предложения, вводные слова, причастные и деепричастные обороты и др. Можно ориентироваться по союзам или членам потенциального простого предложения. В этом плане есть еще огромный простор для улучшения работы программы.

Для тестирования и демонстрации работы программы я взял тексты Тотального диктанта за 2015 – 2018 гг.

3 Пример вывода

Ниже представлен пример обработанного абзаца текста:

```
[{Профессорская дача} {на берегу} {Финского залива}]. [{В отсутствие} хозяина],
[друга {моего отца}], [{нашей семье} позволялось там жить].
[Даже {спустя десятилетия} помню], [как {после утомительной дороги}
{из города} меня обволакивала прохлада {деревянного дома}],
[как собирала растрясшееся], [распавшееся {в экипаже} тело].
[{Эта прохлада} не была связана {со свежестью}], [скорее], [как ни странно],-
[{с упоительной затхлостью}], [в которой слились ароматы {старых книг} и
{многочисленных океанских трофеев}], [непонятно как
{доставшихся профессору-юристу}]. [Распространяя {солоноватый запах}],
[{на полках} лежали {засушенные морские звёзды}], [{перламутровые раковины}],
[{резные маски}], [{пробковый шлем} и даже игла рыбы-иглы].
```

Полный разбор текстов можно найти в конечном выводе работы программы.

4 Код программы

Реализация программы находится в приложенных файлах **HW3.pdf** и **HW3.ipynb**.

5 Выводы по исследованию

В результате работы программы довольно хорошо выделяются именные группы с одним или несколькими эпитетами, допускающие наличие предлога и/или частицы не/ни в начале группы. В программе не хватает обработки глагольных групп, наречий и др. Кроме того, стоит строже разделять простые предложения, добавив дополнительные условия, так как обработка по знакам препинания допускает много случаев, не соответствующих верному разделению. Но в целом в результате получаются вполне цельные части, которые могут упростить синтаксический анализ большого текста.