

Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»

Факультет программной инженерии и компьютерной техники

Лабораторная работа по информатике №4

Исследование протоколов, форматов обмена информацией и языков разметки
документов

Вариант №18

Выполнил:

Студент группы Р3106

Мельник Фёдор Александрович

Проверил:

Балакшин П.В.,

Кандидат технических наук, доцент ФПиКТ

Санкт-Петербург, 2024

Оглавление

| | |
|---|----|
| Оглавление | 2 |
| Задание..... | 3 |
| Решение | 5 |
| Задание 1 | 5 |
| Исходник JSON..... | 5 |
| Задание 2..... | 6 |
| Результат..... | 6 |
| Задание 3..... | 7 |
| Результат..... | 7 |
| Сравнение результатов задания 2 и задания 3 | 7 |
| Задание 4..... | 8 |
| Результат..... | 8 |
| Сравнение результатов задания 2 и задания 4 | 8 |
| Задание 5..... | 9 |
| Исходник JSON..... | 9 |
| Код | 9 |
| Результат..... | 9 |
| Сравнение результатов задания 2 и задания 5 | 9 |
| Задание 6..... | 10 |
| Результат..... | 10 |
| Объяснение результата..... | 10 |
| Задание 7..... | 11 |
| Код | 11 |
| Результат..... | 11 |
| Описание формата..... | 11 |
| Заключение..... | 12 |
| Список источников | 13 |

Задание

1. Исходя из структуры расписания конкретного дня, сформировать файл с расписанием в формате, указанном в задании в качестве исходного. При этом необходимо, чтобы хотя бы в одной из выбранных дней было не менее двух занятий (можно использовать своё персональное). В случае, если в данный день недели нет таких занятий, то увеличить номер варианта ещё на восемь.

2. Обязательное задание (позволяет набрать до 45 процентов от максимального числа баллов БаРС за данную лабораторную): написать программу на языке Python 3.x или любом другом, которая бы осуществляла парсинг и конвертацию исходного файла в новый путём простой замены метасимволов исходного формата на метасимволы результирующего формата.

Нельзя использовать готовые библиотеки, в том числе регулярные выражения в Python и библиотеки для загрузки XML-файлов.

3. Дополнительное задание №1 (позволяет набрать +10 процентов от максимального числа баллов БаРС за данную лабораторную).

а) Найти готовые библиотеки, осуществляющие аналогичный парсинг и конвертацию файлов.

б) Переписать исходный код, применив найденные библиотеки. Регулярные выражения также нельзя использовать.

с) Сравнить полученные результаты и объяснить их сходство/различие. Объяснение должно быть отражено в отчёте.

4. Дополнительное задание №2 (позволяет набрать +10 процентов от максимального числа баллов БаРС за данную лабораторную).

а) Переписать исходный код, добавив в него использование регулярных выражений.

б) Сравнить полученные результаты и объяснить их сходство/различие. Объяснение должно быть отражено в отчёте.

5. Дополнительное задание №3 (позволяет набрать +25 процентов от максимального числа баллов БаРС за данную лабораторную).

а) Переписать исходный код таким образом, чтобы для решения задачи использовались формальные грамматики.

То есть ваш код должен уметь осуществлять парсинг и конвертацию любых данных, представленных в исходном формате, в данные, представленные в результирующем формате: как с готовыми библиотеками из дополнительного задания №1.

- b) Проверку осуществить как минимум для расписания с двумя учебными днями по два занятия в каждом.
- с) Сравнить полученные результаты и объяснить их сходство/различие. Объяснение должно быть отражено в отчёте.

6. Дополнительное задание №4 (позволяет набрать +5 процентов от максимального числа баллов БаРС за данную лабораторную).

- a) Используя свою исходную программу из обязательного задания и программы из дополнительных заданий, сравнить стократное время выполнения парсинга + конвертации в цикле.
- b) Проанализировать полученные результаты и объяснить их сходство/различие. Объяснение должно быть отражено в отчёте.

7. Дополнительное задание №5 (позволяет набрать +5 процентов от максимального числа баллов БаРС за данную лабораторную).

- a) Переписать исходную программу, чтобы она осуществляла парсинг и конвертацию исходного файла в любой другой формат (кроме JSON, YAML, XML, HTML): PROTOBUF, TSV, CSV, WML и т.п.
- b) Проанализировать полученные результаты, объяснить особенности использования формата. Объяснение должно быть отражено в отчёте.

Решение

Задание 1

Исходник JSON

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/in.json

Задание 2

Репозиторий с кодом: https://github.com/ldpst/itmo/tree/main/sem-1_inf/labs/lab4/tasks/main_task

Результат

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/out.xml

Задание 3

Репозиторий с кодом: https://github.com/ldpst/itmo/tree/main/sem-1_inf/labs/lab4/tasks/additional1

Результат

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/out-add1.xml

Сравнение результатов задания 2 и задания 3

Результаты отличаются форматом вывода элементов массива. Если в случае задания 2 тэги элемента массива назывались `<(list_name)_element>`, то в задании 3 теги элемента массива называются `<item>`. Это происходит из-за того, что во втором задании мы сами выбрали название тэгов, а в третьем библиотека сама поставила тегу название по умолчанию.

Задание 4

Репозиторий с кодом: https://github.com/ldpst/itmo/tree/main/sem-1_inf/labs/lab4/tasks/additional2

Результат

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/out-add2.xml

Сравнение результатов задания 2 и задания 4

Результаты полностью совпадают, так как изменен лишь принцип парсинга JSON файла, а конвертация не изменялась.

Задание 5

Исходник JSON

Json файл, имеющий большой объем данных

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/in-add3.json

Код

Репозиторий с кодом: https://github.com/ldpst/itmo/tree/main/sem-1_inf/labs/lab4/tasks/additional3

Результат

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/out-add3.xml

Сравнение результатов задания 2 и задания 5

Входной и выходной файл для задания 5 содержат в себе больше данных, чем файлы задания 2. Парсер задания 5 в отличие от парсера задания 2 может обработать пары ключ-значение, значение которых равно null или имеет тип bool

Задание 6

Код: https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/additional4/main.py

Результат

```
Собственный: 0.12338447570800781
Собственный с регулярными: 1.3031673431396484
С библиотеками: 0.22119426727294922
Собственный полный: 0.22249460220336914
```

Объяснение результата

Первый код выполняется быстрее всего, так как он обладает меньшими возможностями, чем коды 3 и 4 и запущен на тесте меньшего размера. Второй код обрабатывается дольше чем первый (хоть у них одинаковые возможности), так как каждое регулярное выражение медленно обрабатывает всю строку на поиск всех “мэтчей”. Третий код и запущен на меньшем объеме данных, поэтому он выполняется быстрее кода 4. Четвертый код выполняется дольше всех, так как обладает самыми большими возможностями (=3) и запущен на самом большом наборе данных.

Задание 7

Выбранный формат: TSV

Код

Репозиторий с кодом: https://github.com/ldpst/itmo/tree/main/sem-1_inf/labs/lab4/tasks/additional5

Результат

https://github.com/ldpst/itmo/blob/main/sem-1_inf/labs/lab4/tasks/data/out-add5.tsv

В виде таблицы:

Таблица разделена на несколько изображений в связи с размером: рис. 1, рис. 2 и рис. 3

| A | B | C | D | E | F | G | H | I | J |
|------------------|-------------|---------------------------|------------|----------|--------------|--------------|--------------|--------------|--------------|
| title | group | teacher/name | teacher/id | date/day | date/weeks/0 | date/weeks/1 | date/weeks/2 | date/weeks/3 | date/weeks/4 |
| Линейная алгебра | Лин Алг 7.1 | Тушавин Глеб Владимирович | 207357 | 2 | 2 | 3 | 4 | 5 | 6 |
| Линейная алгебра | Лин Алг 7 | Тушавин Глеб Владимирович | 207357 | 2 | 2 | 3 | 4 | 5 | 6 |
| Волейбол | Волейбол | Северюгин Владимир | 269329 | 2 | 2 | 3 | 4 | 5 | 6 |

Рисунок 1 - Таблица. Часть 1

| K | L | M | N | O | P | Q | R | S | T | U |
|--------------|--------------|--------------|--------------|--------------|---------------|---------------|---------------|---------------|---------------|---------------|
| date/weeks/5 | date/weeks/6 | date/weeks/7 | date/weeks/8 | date/weeks/9 | date/weeks/10 | date/weeks/11 | date/weeks/12 | date/weeks/13 | date/weeks/14 | date/weeks/15 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |

Рисунок 2 - Таблица. Часть 2

| V | W | X | Y | Z | AA | AB | AC | AD | AE | AF |
|--------------------|------------|----------|------------------|---------|--------------------------------|-----------|-----------------------|---------------------|---------------------------|-------------------------|
| time/lesson_number | time/start | time/end | room/name | room/id | campus/name | campus/id | lesson_type/type_name | lesson_type/type_id | lesson_format/format_name | lesson_format/format_id |
| 1 | 8:20 | 9:50 | ауд. 4210 | 4210 | ул. Ломоносова, д.9, лит. Б | 0 | Практика | 1 | Очно - дистанционный | 1 |
| 2 | 10:00 | 11:30 | ауд. 2202 | 2202 | ул. Ломоносова, д.9, лит. А | 1 | Лекция | 0 | Очно - дистанционный | 1 |
| 3 | 11:40 | 13:10 | ауд. Большой зал | 4405 | ул. Ломоносова, д.9, лит. Б | 0 | Спорт | 3 | Очно | |

Рисунок 3 - Таблица. Часть 3

Описание формата

TSV - табличный формат, разделителем которого служит символ табуляция

Заключение

В процессе выполнения лабораторной работы я познакомился с такими форматами, как JSON, XML, YAML, CSV, TSV. Я написал собственный парсер из JSON файла и конвертер в XML и CSV. Я изучил библиотеки, с помощью которых можно с легкостью обрабатывать данные файлы и углубился в познании регулярных выражений, используя их на практике. Также я получил опыт работы с хронометражом и анализа полученных данных.

СПИСОК ИСТОЧНИКОВ

1. Лямин А.В., Череповская Е.Н. Объектно-ориентированное программирование. Компьютерный практикум. – СПб: Университет ИТМО, 2017. – 143 с. – Режим доступа: <https://books.ifmo.ru/file/pdf/2256.pdf>.
2. Пишем изящный парсер на Питоне [Электронный ресурс] Хабр : [сайт]. — URL: <https://habr.com/ru/articles/309242> (дата обращения: 28.10.2024).