Санкт-Петербургский Национальный Исследовательский Университет

Информационных Технологий, Механики и Оптики

МФКТиУ, кафедра ПИиКТ

Лабораторная работа №3 по "Информатике"

Исследование языков разметки документов.

Выполнил: Романов Артём

Группа: Р3110

Вариант: 23

Санкт-Петербург

2020 г.

**Цель**

Получить знания о различных языках разметки и хранения данных. Научиться обработке данных с помощью языка Python 3.x

**Задание**

* Исходя из структуры расписания конкретного дня, сформировать файл с расписанием в формате, указанном в задании в качестве исходного.
* Написать программу на языке Python 3.x, которая бы осуществляла парсинг и конвертацию исходного файла в новый.
* Нельзя использовать готовые библиотеки, кроме re (регулярные выражения в Python) и библиотеки для загрузки XML-файлов.
* a) Найти готовые библиотеки, осуществляющие аналогичный парсинг и конвертацию файлов. b) Сравнить полученные результаты и объяснить их сходство/различие.
* a) Используя свою программу и найденные готовые библиотеки, сравнить десятикратное время выполнения парсинга + конвертации в цикле. b) Проанализировать полученные результаты и объяснить их сходство/различие.

**Выполнение**

**XML-файл**

1. **<?xml** version="1.0" encoding="UTF-8"**?>**
2. **<shedule>**
3. **<thursday>**
4. **<day**
5. day="Чт"**>**
6. **<subject1**
7. time="08:20-09:50"
8. room="466 ауд."
9. location="Кронверкский пр., д.49, лит.А"
10. subject="Математика (базовый уровень)(Лек)"
11. parity="True"
12. teacher="Возианова Анна"
13. format="Очно - дистанционный"**>**
14. **</subject1>**
15. **<subject2**
16. time="10:00-11:30"
17. room="Актовый зал"
18. location="Кронверкский пр., д.49, лит.А"
19. subject="Программирование(Лек)"
20. parity="True"
21. teacher="Письмак Алексей Евгеньевич"
22. format="Очно - дистанционный"**>**
23. **</subject2>**
24. **<subject3**
25. time="11:40-13:10"
26. room="Актовый зал"
27. location="Кронверкский пр., д.49, лит.А"
28. subject="Основы профессиональной деятельности(Лек)"
29. parity="True"
30. teacher="Клименков Сергей Викторович"
31. format="Очно - дистанционный"**>**
32. **</subject3>**
33. **</day>**
34. **</thursday>**
35. **</shedule>**

**Python-парсер**

1. **import** xml.dom.minidom
2. doc = xml.dom.minidom.parse("xml\_lab3.xml");
3. day = doc.getElementsByTagName("day")
4. **for** item **in** day:
5. json = "{\n\"shedule\" : {\n\"thursday\" : {\n\"day\" : {\n\"day\" : \""+item.getAttribute('day')+"\",\n"
6. **for** i **in** range(4):
7. **for** sub **in** item.getElementsByTagName('subject'+str(i)):
8. list = sub.attributes.items()
9. json += "\"" + sub.tagName + "\" : {\n"
10. **for** key **in** list:
11. json += "\""+key[0]+"\" : "+"\""+key[1]+"\", \n"
12. **if** i == 3:
13. json += "}"
14. **else**:
15. json += "},"
16. json += "\n}\n}\n}\n}"
17. **print**(json)

**Вывод парсера**

1. {
2. "shedule" : {
3. "thursday" : {
4. "day" : {
5. "day" : "Чт",
6. "subject1" : {
7. "time" : "08:20-09:50",
8. "room" : "466 ауд.",
9. "location" : "Кронверкский пр., д.49, лит.А",
10. "subject" : "Математика (базовый уровень)(Лек)",
11. "parity" : "True",
12. "teacher" : "Возианова Анна",
13. "format" : "Очно - дистанционный"
14. },"subject2" : {
15. "time" : "10:00-11:30",
16. "room" : "Актовый зал",
17. "location" : "Кронверкский пр., д.49, лит.А",
18. "subject" : "Программирование(Лек)",
19. "parity" : "True",
20. "teacher" : "Письмак Алексей Евгеньевич",
21. "format" : "Очно - дистанционный"
22. },"subject3" : {
23. "time" : "11:40-13:10",
24. "room" : "Актовый зал",
25. "location" : "Кронверкский пр., д.49, лит.А",
26. "subject" : "Основы профессиональной деятельности(Лек)",
27. "parity" : "True",
28. "teacher" : "Клименков Сергей Викторович",
29. "format" : "Очно - дистанционный"
30. }
31. }
32. }
33. }
34. }

**Метод парсинга и конвертации с помощью библиотек xmltodict и json**

1. **import** xmltodict
2. **import** json
4. with open('xml\_lab3.xml') as fd:
5. doc = xmltodict.parse(fd.read())
6. **print**(json.dumps(doc, indent=2))

**Вывод альтернативного метода**

1. {
2. "shedule": {
3. "thursday": {
4. "day": {
5. "@day": "Чт",
6. "subject1": {
7. "@time": "08:20-09:50",
8. "@room": " 466 ауд..",
9. "@location": "Кронверкский пр., д.49, лит.А",
10. "@subject": "Математика (базовый уровень)(Лек)",
11. "@parity": "True",
12. "@teacher": "Возианова Анна",
13. "@format": "Очно - дистанционный"
14. },
15. "subject2": {
16. "@time": "10:00-11:30",
17. "@room": "Актовый зал",
18. "@location": "Кронверкский пр., д.49, лит.А",
19. "@subject": "Программирование(Лек)",
20. "@parity": "True",
21. "@teacher": "Письмак Алексей Евгеньевич",
22. "@format": "Очно - дистанционный"
23. },
24. "subject3": {
25. "@time": "11:40-13:10",
26. "@room": "Актовый зал",
27. "@location": "Кронверкский пр., д.49, лит.А",
28. "@subject": "Основы профессиональной деятельности(Лек)",
29. "@parity": "True",
30. "@teacher": "Клименков Сергей Викторович",
31. "@format": "Очно - дистанционный"
32. }
33. }
34. }
35. }

**Сравнение**

Вариант с использованием готовых библиотек является существенно более универсальным, потому что работает по принципу перевода XML-файла в словарь и дальнейшей его конвертации в файл JSON. Написанный же мной метод работает с конкретной структурой файла, от которой не может отклониться.

**Сравнение времени выполнения**

|  |  |
| --- | --- |
| **Собственный метод** | **Библиотека** |
| 0.002753019332885742 | 0.0004131793975830078 |
| 0.0025436878204345703 | 0.00040602684020996094 |
| 0.002650022506713867 | 0.0005960464477539062 |
| 0.003119945526123047 | 0.0004050731658935547 |
| 0.003031015396118164 | 0.00043511390686035156 |
| 0.003718852996826172 | 0.0004153251647949219 |
| 0.0030341148376464844 | 0.0004038810729980469 |
| 0.0032308101654052734 | 0.00042891502380371094 |
| 0.0034978389739990234 | 0.0004088878631591797 |
| 0.0033888816833496094 | 0.00042891502380371094 |
| **≈0.03092** | **≈0.004337** |

Время десятикратного выполнения методов отличается примерно в 7 раз в пользу метода, где используются библиотеки xmltodict и json. Измерения проводились с помощью Python-библиотеки time. Различие

**Вывод**

Данная лабораторная научила меня обрабатывать данные с использованием языка Python, а также познакомила с форматами хранения данных XML и JSON.