

Мегафакультет компьютерных управлений и технологий

Факультет ПИ и КТ

Лабораторная работа №4 по ИКТ

Вариант 19

**Выполнил:**

Болорболд Аригуун

группа P3111

**Преподаватель:**

Малышева Татьяна Алексеевна

г. Санкт-Петербург

2022 год

Оглавление

1. Задание…………………………………………………………………………….……..3
2. Решение к заданиям………………………………………………………………....4
3. Вывод……………………………………………………………………………….........9

Источники…………………………………………………………………………........9

**Задание:** 1. Определить номер варианта как остаток деления на 36 порядкового номера в списке группы в ISU. В случае, если в данный день недели нет занятий, то увеличить номер варианта на восемь.

2. Изучить форму Бэкуса-Наура.

3. Изучить особенности языков разметки/форматов JSON, YAML, XML.

4. Понять устройство страницы с расписанием для своей группы: <http://itmo.ru/ru/schedule/0/P3111/schedule.htm>

5. Исходя из структуры расписания конкретного дня, сформировать файл с расписанием в формате, указанном в задании в качестве исходного. При этом необходимо, чтобы в выбранном дне 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.

6. Обязательное задание (позволяет набрать до 65 процентов от максимального числа баллов БаРС за данную лабораторную): написать программу на языке Python 3.x, которая бы осуществляла парсинг и конвертацию исходного файла в новый.

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

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

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

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

c) Сравнить полученные результаты и объяснить их сходство/различие.

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

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

b) Сравнить полученные результаты и объяснить их сходство/различие.

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

a) Используя свою исходную программу из обязательного задания, программу из дополнительного задания №1 и программу из дополнительного задания №2, сравнить стократное время выполнения парсинга + конвертации в цикле.

b) Проанализировать полученные результаты и объяснить их сходство/различие. 11.Дополнительное задание №4 (позволяет набрать +5 процентов от максимального числа баллов БаРС за данную лабораторную).

c) 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.

d) Проанализировать полученные результаты, объяснить особенности использования формата. 12.Проверить, что все пункты задания выполнены и выполнены верно.

13.Написать отчёт о проделанной работе.

14.Подготовиться к устным вопросам на защите.

**Обязательное задание:**

def Parse\_default():  
 inputAddress = r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.xml"  
 outputAddress = r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.json"  
 inputFile = open(inputAddress, 'r', encoding='utf-8')  
 outputFile = open(outputAddress, 'w', encoding='utf-8')  
 outputFile.write('{\n')  
 inputFile.readline()  
 count = 0  
 lineDel = inputFile.read().split("\n")  
 for j in range(len(lineDel)):  
 lineSh = lineDel[j].strip().replace('<', '>').split('>')[1:-1]  
 if j == len(lineDel) - 1 or lineDel[j + 1].strip()[:2] == '</':  
 token = ''  
 else:  
 token = ','  
 if len(lineSh) > 2:  
 lineSh = lineSh[:-1]  
 if len(lineSh) == 1:  
 if lineSh[0][0] == '/':  
 outputFile.write(' ' \* count + '}' + token + '\n')  
 count -= 1  
 else:  
 outputFile.write(' ' \* count + '"' + lineSh[0] + '": {\n')  
 count += 1  
 else:  
 outputFile.write(' ' \* count + '"' + lineSh[0] + '" : "' + lineSh[1] + '" ' + token + '\n')  
 outputFile.write('}')  
 outputFile.close()  
 inputFile.close()

**Дополнительное задание №1:**

def Parse\_lib():  
 import xmltodict  
 import json  
 input\_address = r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.xml"  
 output\_address = r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.json"  
 input\_file = open(input\_address, 'r', encoding='utf-8')  
 output\_file = open(output\_address, 'w', encoding='utf-8')  
 inter\_input = input\_file.read()  
 inter\_output = xmltodict.parse(inter\_input)  
 preOutput = json.dumps(inter\_output, indent=4, ensure\_ascii=False)  
 output\_file.write(preOutput)  
 output\_file.close()  
 input\_file.close()

**Дополнительное задание №2:**

def Parse\_Regex():  
 import re  
 inputAddress = r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.xml"  
 outputAddress = r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.json"  
 inputFile = open(inputAddress, 'r', encoding='utf-8')  
 outputFile = open(outputAddress, 'w', encoding='utf-8')  
 inputFile.readline()  
 sumLines = 0  
 for s in inputFile:  
 countTab = 4  
 while s[0] == ' ':  
 s = s[1::]  
 countTab+=1  
 xmltag = str(\*(re.findall(r'^<\w\*[^\\]>', s)))[1:-1] + '\" : '  
 token = ""  
 if '<' in s and (len(re.findall(r">.\*<", s)) > 0):  
 token = "\"" + str(\*(re.findall(r'>.\*<', s)))[1:-1] + "\""  
 if countTab > sumLines:  
 outputFile.write('{' + '\n' + countTab \* " " + "\"" + xmltag + token)  
 elif countTab < sumLines:  
 outputFile.write('\n' + " " \* countTab + "}")  
 else:  
 outputFile.write(',' + '\n' + countTab \* ' ' + "\"" + xmltag + token)  
 sumLines = countTab  
 outputFile.write("\n}")  
 outputFile.close()  
 inputFile.close()

**XML–файл(зад. 1-3):**

<?xml version="1.0" encoding="UTF-8"?>  
<schedule>  
 <day>  
 <dayShorthand>Чт</dayShorthand>  
 <subjects>  
 <subject1>  
 <time>11:40-13:10</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2(усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Программирование (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Харитонова Анастасия Евгеньевна</teacher>  
 <format>Очно-дистанционный</format>  
 </subject1>  
 <subject2>  
 <time>13:30-15:00</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2 (усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Программирование (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Харитонова Анастасия Евгеньевна</teacher>  
 <format>Очно-дистанционный</format>  
 </subject2>  
 <subject3>  
 <time>15:20-16:50</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2 (усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Основы профессиональной деятельности (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Сорокин Роман Борисович</teacher>  
 <format>Очно-дистанционный</format>  
 </subject3>  
 <subject4>  
 <time>17:00-18:30</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2 (усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Основы профессиональной деятельности (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Сорокин Роман Борисович</teacher>  
 <format>Очно-дистанционный</format>  
 </subject4>  
 </subjects>  
 </day>  
</schedule>

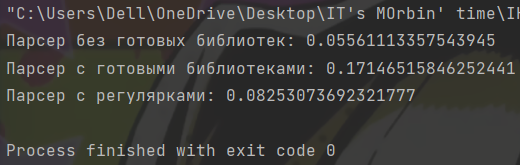
**Вывод (JSON-файл, зад. 1-3):**

{  
 "schedule" : {  
 "day" : {  
 "dayShorthand" : "Чт",  
 "subjects" : {  
 "subject1" : {  
 "time" : "11:40-13:10",  
 "weeks" : "2, 4, 6, 8, 10, 12, 14, 16",  
 "room" : "2305/1 (бывш. 303/2(усл.)) ауд.",  
 "location" : "Кронверкский пр., д.49, лит.А",  
 "name" : "Программирование (лаб)",  
 "week" : "Чётная неделя",  
 "teacher" : "Харитонова Анастасия Евгеньевна",  
 "format" : "Очно-дистанционный"  
 },  
 "subject2" : {  
 "time" : "13:30-15:00",  
 "weeks" : "2, 4, 6, 8, 10, 12, 14, 16",  
 "room" : "2305/1 (бывш. 303/2 (усл.)) ауд.",  
 "location" : "Кронверкский пр., д.49, лит.А",  
 "name" : "Программирование (лаб)",  
 "week" : "Чётная неделя",  
 "teacher" : "Харитонова Анастасия Евгеньевна",  
 "format" : "Очно-дистанционный"  
 },  
 "subject3" : {  
 "time" : "15:20-16:50",  
 "weeks" : "2, 4, 6, 8, 10, 12, 14, 16",  
 "room" : "2305/1 (бывш. 303/2 (усл.)) ауд.",  
 "location" : "Кронверкский пр., д.49, лит.А",  
 "name" : "Основы профессиональной деятельности (лаб)",  
 "week" : "Чётная неделя",  
 "teacher" : "Сорокин Роман Борисович",  
 "format" : "Очно-дистанционный"  
 },  
 "subject4" : {  
 "time" : "17:00-18:30",  
 "weeks" : "2, 4, 6, 8, 10, 12, 14, 16",  
 "room" : "2305/1 (бывш. 303/2 (усл.)) ауд.",  
 "location" : "Кронверкский пр., д.49, лит.А",  
 "name" : "Основы профессиональной деятельности (лаб)",  
 "week" : "Чётная неделя",  
 "teacher" : "Сорокин Роман Борисович",  
 "format" : "Очно-дистанционный"  
 }  
 }  
 }  
 }  
}

**Дополнительное задание №3:**

timeDef = time.time()  
for u in range(100):  
 Parse\_default()  
timeDef = time.time() - timeDef  
  
timeLib = time.time()  
for u in range(100):  
 Parse\_lib()  
timeLib = time.time() - timeLib  
  
timeRe = time.time()  
for u in range(100):  
 Parse\_Regex()  
timeRe = time.time() - timeRe  
print("Парсер без готовых библиотек: " + str(timeDef) + '\n' + "Парсер с готовыми библиотеками: " + str(timeLib) + '\n' + "Парсер с регулярками: " + str(timeRe))

**Результат:**

****

**Дополнительное задание №4: конвертация в CSV–файл:**

def xml\_to\_csv():  
 import xmltodict, csv  
 xmlFile = open(r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.xml", 'r', encoding='utf-8')  
 xml = xmltodict.parse(xmlFile.read())  
 csvFile = open(r"C:\Users\Dell\OneDrive\Desktop\IT's MOrbin' time\IKT\_17starplatinum\I\I semester\lab4\schedule.csv", 'w', encoding='utf-8', newline='')  
 csvFile\_writer = csv.writer(csvFile)  
 csvFile\_writer.writerow(["time", "weeks", "room", "location", "name", "week", "teacher", "format"])  
 for subject in xml["schedule"]["day"]["subjects"]["subject"]:  
 csv\_line = [subject["time"], subject["weeks"], subject["room"], subject["location"], subject["week"], subject["name"], subject["teacher"], subject["format"]]  
 csvFile\_writer.writerow(csv\_line)

**XML–файл:**

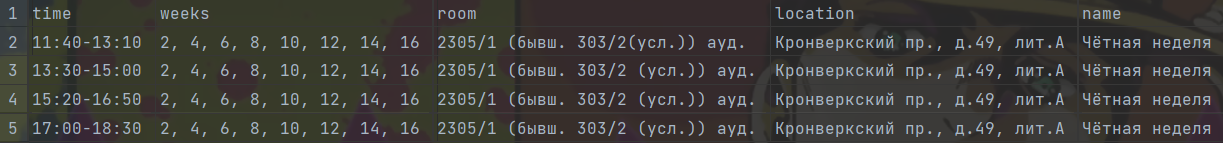
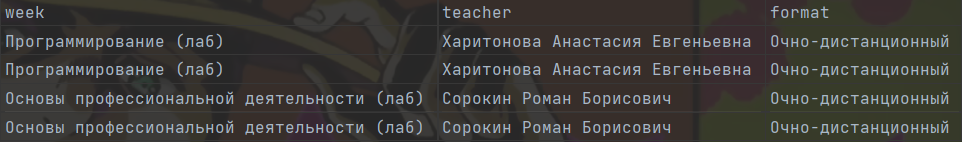
<?xml version="1.0" encoding="UTF-8"?>  
<schedule>  
 <day>  
 <subjects>  
 <subject>  
 <time>11:40-13:10</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2(усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Программирование (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Харитонова Анастасия Евгеньевна</teacher>  
 <format>Очно-дистанционный</format>  
 </subject>  
 <subject>  
 <time>13:30-15:00</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2 (усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Программирование (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Харитонова Анастасия Евгеньевна</teacher>  
 <format>Очно-дистанционный</format>  
 </subject>  
 <subject>  
 <time>15:20-16:50</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2 (усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Основы профессиональной деятельности (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Сорокин Роман Борисович</teacher>  
 <format>Очно-дистанционный</format>  
 </subject>  
 <subject>  
 <time>17:00-18:30</time>  
 <weeks>2, 4, 6, 8, 10, 12, 14, 16</weeks>  
 <room>2305/1 (бывш. 303/2 (усл.)) ауд.</room>  
 <location>Кронверкский пр., д.49, лит.А</location>  
 <name>Основы профессиональной деятельности (лаб)</name>  
 <week>Чётная неделя</week>  
 <teacher>Сорокин Роман Борисович</teacher>  
 <format>Очно-дистанционный</format>  
 </subject>  
 </subjects>  
 </day>  
</schedule>

**CSV­­–файл:**

1. **Текст:**

time,weeks,room,location,name,week,teacher,format  
11:40-13:10,"2, 4, 6, 8, 10, 12, 14, 16",2305/1 (бывш. 303/2(усл.)) ауд.,"Кронверкский пр., д.49, лит.А",Чётная неделя,Программирование (лаб),Харитонова Анастасия Евгеньевна,Очно-дистанционный  
13:30-15:00,"2, 4, 6, 8, 10, 12, 14, 16",2305/1 (бывш. 303/2 (усл.)) ауд.,"Кронверкский пр., д.49, лит.А",Чётная неделя,Программирование (лаб),Харитонова Анастасия Евгеньевна,Очно-дистанционный  
15:20-16:50,"2, 4, 6, 8, 10, 12, 14, 16",2305/1 (бывш. 303/2 (усл.)) ауд.,"Кронверкский пр., д.49, лит.А",Чётная неделя,Основы профессиональной деятельности (лаб),Сорокин Роман Борисович,Очно-дистанционный  
17:00-18:30,"2, 4, 6, 8, 10, 12, 14, 16",2305/1 (бывш. 303/2 (усл.)) ауд.,"Кронверкский пр., д.49, лит.А",Чётная неделя,Основы профессиональной деятельности (лаб),Сорокин Роман Борисович,Очно-дистанционный

1. **Данные:**

**** ****

**Вывод: это было однозначно самой трудной работой. И это не мудрено, потому что почти вся ИТ-сфера опирается на обработку данных. Парсинг встречается везде, где причастны электронные устройства. Поэтому я лично считаю, что время тратил в сторону пользы, а не зря.**

**Источники:**

# [**https://habr.com/ru/post/309242/**](https://habr.com/ru/post/309242/) **– Пишем изящный парсер на Питоне ­**

1. <https://ru.wikipedia.org/wiki/Форма_Бэкуса_—_Наура> **­­­­–** страница Википедии про БНФ
2. <https://pypi.org/project/xmltodict/> – страница с описанием про библиотеки xmltodict
3. <https://docs.python.org/3/library/csv.html> – страница с описанием про библиотеки csv
4. <https://appmaster.io/ru/blog/json-protiv-xml> – XML против JSON