Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет Программной инженерии и компьютерной техники

**Лабораторная работа №4.**

**«Исследование протоколов, форматов обмена информацией и языков разметки документов»**

Вариант №5

Группа: P3131

Выполнил: Воронин И.А.

Проверил:

к.т.н. преподаватель Белозубов А.В.

Санкт-Петербург

2021г

Оглавление

[**Обязательное задание** 3](#_Toc88493528)

[**Дополнительное Задание №1** 5](#_Toc88493529)

[**Дополнительное Задание №2** 6](#_Toc88493530)

[**Дополнительное Задание №3** 7](#_Toc88493531)

[**Дополнительное Задание №4** 7](#_Toc88493532)

[**Заключение** 9](#_Toc88493533)

**Обязательное задание**

Написать программу на языке Python 3.x, которая бы осуществляла парсинг и конвертацию исходного файла в новый.

Вариант №5 : Конвертация из XML в YAML.

**Исходный XML файл:**

**<schedule>  
 <day>**Пн**</day>  
 <classes>  
 <class\_1>  
 <time>**11:40-13:10**</time>  
 <week>**четная неделя**</week>  
 <auditory>**303\2 (УСЛ) АУД.**</auditory>  
 <place>**Кронверкский пр., д.49, лит.А**</place>  
 <subject>**ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ)**</subject>  
 <teacher>**Ларочкин Глеб Игоревич**</teacher>  
 <format>**Очно - дистанционный**</format>  
 </class\_1>  
 <class\_2>  
 <time>**13:30-15:00**</time>  
 <week>**четная неделя**</week>  
 <auditory>**303\2 (УСЛ) АУД.**</auditory>  
 <place>**Кронверкский пр., д.49, лит.А**</place>  
 <subject>**ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ)**</subject>  
 <teacher>**Ларочкин Глеб Игоревич**</teacher>  
 <format>**Очно - дистанционный**</format>  
 </class\_2>  
 </classes>  
</schedule>**

**Код программы:**

xml = open("schedule.xml", encoding="utf-8")

text = xml.read().split('**\n**')

yaml = open("schedule.yaml", "w", encoding="utf-8")

**print**(text)

**for** i **in** range(1,len(text)):

line = text[i]

**if** line.count('</') != 0 **and** line.count('<') > 1:

yaml.write(' ' \* line.count('**\t**') +line[line.find('<')+1:line.find('>')] + ': '+ line[line.find('>')+1:line.rfind('<')] +'**\n**')

**elif** line.count('</') == 0:

yaml.write(' ' \* line.count('**\t**') + line[line.find('<')+1:line.find('>')] + ':' + '**\n**')

yaml.close()

xml.close()

**Конечный YAML файл:**

**schedule:  
 day:** Пн **classes:  
 class\_1:  
 time:** 11:40-13:10  
 **week:** четная неделя  
 **auditory:** 303\2 (УСЛ) АУД.  
 **place:** Кронверкский пр., д.49, лит.А  
 **subject:** ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ) **teacher:** Ларочкин Глеб Игоревич **format:** Очно - дистанционный **class\_2:  
 time:** 13:30-15:00 **week:** четная неделя **auditory:** 303\2 (УСЛ) АУД. **place:** Кронверкский пр., д.49, лит.А **subject:** ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ)  
 **teacher:** Ларочкин Глеб Игоревич **format:** Очно - дистанционный

**Дополнительное Задание №1**

Найти готовые библиотеки, осуществляющие аналогичный парсинг и конвертацию файлов. Переписать исходный код, применив найденные библиотеки. Регулярные выражения также нельзя использовать. Сравнить полученные результаты и объяснить их сходство/различие.

**Код программы с использованием библиотек xmlplain:**

**import** xmlplain

root = xmlplain.xml\_to\_obj(open("schedule.xml", encoding="utf-8"), strip\_space=True, fold\_dict=True)

**with** open("result.yml", "w") **as** out:

xmlplain.obj\_to\_yaml(root, out)

**Конечный YAML файл:**

**schedule:  
 day:** Пн **classes:  
 class\_1:  
 time:** 11:40-13:10  
 **week:** четная неделя  
 **auditory:** 303\2 (УСЛ) АУД.  
 **place:** Кронверкский пр., д.49, лит.А  
 **subject:** ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ) **teacher:** Ларочкин Глеб Игоревич **format:** Очно - дистанционный **class\_2:  
 time:** 13:30-15:00 **week:** четная неделя **auditory:** 303\2 (УСЛ) АУД. **place:** Кронверкский пр., д.49, лит.А **subject:** ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ)  
 **teacher:** Ларочкин Глеб Игоревич **format:** Очно – дистанционный

Разницы между результатами работы с библиотекой и без нет.

**Дополнительное Задание №2**

Переписать исходный код, добавив в него использование регулярных выражений. Сравнить полученные результаты и объяснить их сходство/различие.

**Код программы с использованием регулярных выражений:**

**import** re

xml = open("schedule.xml", encoding="utf-8")

text = xml.read().split('**\n**')

yaml = open("schedule.yaml", "w")

**for** i **in** range(1,len(text)):

line = text[i]

line = re.sub('**\t**', ' ', line)

line = re.sub('/**\S**\*', '', line)

line = re.sub('<', '', line)

line = re.sub('>', ': ', line)

yaml.write(line + '**\n**')

yaml.close()

xml.close()

**Конечный YAML файл:**

**schedule:  
 day:** Пн **classes:  
 class\_1:  
 time:** 11:40-13:10  
 **week:** четная неделя  
 **auditory:** 303\2 (УСЛ) АУД.  
 **place:** Кронверкский пр., д.49, лит.А  
 **subject:** ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ) **teacher:** Ларочкин Глеб Игоревич **format:** Очно - дистанционный **class\_2:  
 time:** 13:30-15:00 **week:** четная неделя **auditory:** 303\2 (УСЛ) АУД. **place:** Кронверкский пр., д.49, лит.А **subject:** ОСНОВЫ ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ(ЛАБ)  
 **teacher:** Ларочкин Глеб Игоревич **format:** Очно – дистанционный

Разницы между результатами работы с регулярными выражениями и без нет.

**Дополнительное Задание №3**

Используя свою исходную программу из обязательного задания, программу из дополнительного задания №1 и программу из дополнительного задания №2, сравнить десятикратное время выполнения парсинга + конвертации в цикле. Проанализировать полученные результаты и объяснить их сходство/различие.

Десятикратное время выполнения кода из заданий:

Основное задание: ≈0.03 секунд

Доп. Задание №1: ≈0.12 секунд

Доп. Задание №2: ≈0.05 секунд

Программа из основного задания имеет наивысшую скорость из-за отсутствия библиотек, программа из доп. задания №2 использует встроенную библиотеку и оптимизированную библиотеку, программа из доп. задания №1 самая медленная из-за использования сторонней библиотеки, алгоритм работы которой является не самым оптимальным.

**Дополнительное Задание №4**

Переписать исходную, чтобы она осуществляла парсинг и конвертацию исходного файла в любой другой формат. Проанализировать полученные результаты, объяснить особенности использованного формата.

**Программа, осуществляющая парсинг из XML в CSV:**

**import** re

xml = open("schedule.xml", encoding="utf-8")

text = xml.read().split('**\n**')

csv = open("schedule.csv", "w")

**print**(text)

line1 = ''

line2 = ''

path = ''

urok = ''

k = 1

**for** i **in** range(1,len(text)):

line = text[i]

**if** line.count('</') != 0 **and** line.count('<') > 1:

line1 += path + line[line.find('<') + 1:line.find('>')] + ';'

line2 += (line[line.find('>')+1:line.rfind('<')] +';')

**elif** line.count('</') == 0 **and** k == 1:

path += line[line.find('<')+1:-1] + '/'

**elif** line.count('</') == 1:

k += 1

path = re.sub('\_**\d**','\_'+str(k),path)

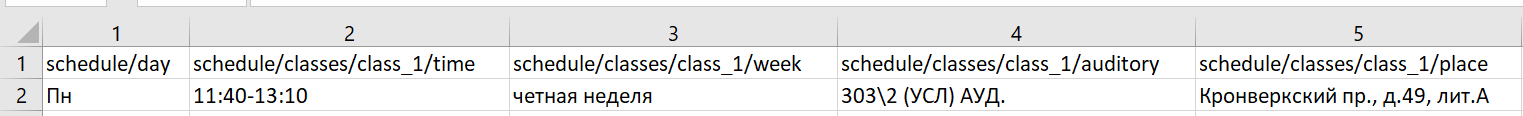
csv.write(line1 + '**\n**')

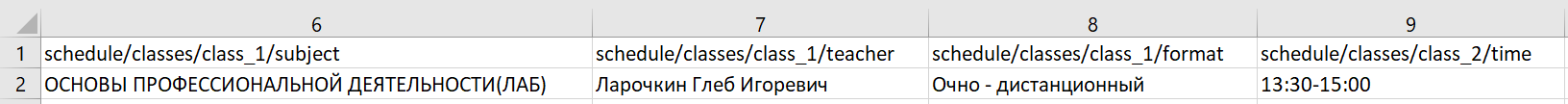
csv.write(line2)

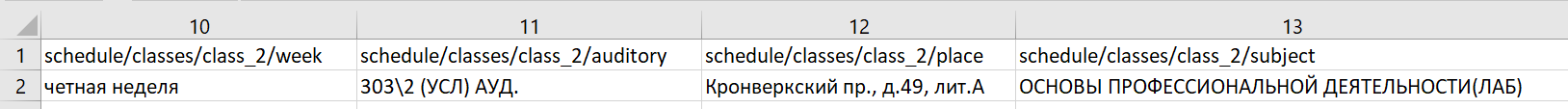
csv.close()

xml.close()

**Конечный файл CSV:**







Изображение выглядит как стол

Автоматически созданное описание

**Особенности формата CSV:**

Предназначен для представления табличных данных, строка текста соответствует строке таблицы, а столбцы разделяются с помощью « ; » и « , ».

**Заключение**

В ходе выполнения данной лабораторной работы я познакомился с различными языками разметки и научился конвертировать их между собой с помощью Python.