МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА

Факультет информатики и систем управления

Кафедра теоретической информатики и компьютерных технологий

**Лабораторная работа № 2.**

**по курсу «Теория формальных языков»**

Выполнил: Иванов Г.М.

ИУ9-52

Проверил: Магазов С.C.

«26» ноября 2016 г.

Москва, 2016

**Задание:**

Подготовить текст (более чем на 10 страниц), в котором должен быть фрагмент шаблона A (см. таблицу). Написать программу, которая:

1. определяла бы существует ли А в тексте

2. выдавала XML файл, в котором фрагменты типа A выделены в виде стеков

***Вариант №6:***

Шаблон А:

По тексту выделить адрес сайта подготовить ХML

<txt></txt>

<a></a>

Для выполнения задания был использован язык Python 3.5 с помощью модуля re (регулярных выражений) и lxml (для формирования xml файла) и среда разработки Pycharm by JetBrains.

* Регулярное выражение имеет вид:

*Листинг 1. Задание регулярного выражения (паттерна).*

#!usr/bin/python3

# -\*- coding: utf-8 -\*-

**import** re

pattern = re.compile(r"(https?:\/\/)?([\w\.]+)\.([a-z]{2,6}\.?)(\/[-a-zA-Z0-9@:%\_\+.~#?&//=\*)\*\/?")

*Пояснение (правил регулярного выражения):*

? - используется один или ноль раз  
s? - присутствует буква s или нет

\/ - соответствует символу /

(https?:\/\/)? – присутствует протокол или нет (http:// или https://)

\w - соответствует выражению [A-Za-z0-9], то есть любой символ из этой группы (буквы английские и цифры)

\. - соответствует символу .

+ - одно или бесконечно много повторений

([\w\.]+)\. – некоторая неограниченная последовательность букв, цифр с точками (для доменов уровня > 1)

{2,6} – повторений минимум 2, максимум 6

[a-z]{2,6} – некоторая последовательность букв (a – z) длиной от 2 до 6 символов

([a-z]{2,6}\.?) – указание домена нулевого уровня

\* - ноль или бесконечно много повторений

(\/[[-a-zA-Z0-9@:%\_\+.~#?&//=]\*)\* - некоторая неограниченная последовательность букв, цифр с ?, #, &, . (для GET и POST-запросов на url)

* **Описание контрольного примера:**

Для тестирования программы использовались исходные коды страниц google поиска с абсолютными путями.

* **Контрольный пример (Часть):**

<!DOCTYPE html><html lang="en" class="chekov"><head><script>var a=window.devsite||{};window.devsite=a;a.readyCallbacks=[];window.devsite.readyCallbacks=a.readyCallbacks;a.ready=function(b){a.readyCallbacks.push(b)};window.devsite.ready=a.ready;

</script><meta charset="utf-8"><meta name="xsrf\_token" content="oUlLsgFj9bGYQ5pHLI2OXW1Vu4-rq8IVgAIJ6ppH0zU6MTQ4MTk1NTI1Njc1MDMzMA" /><link rel="canonical" href="<https://developers.google.com/search/docs/guides/create-URLs>"><link rel="alternate" href="<https://developers.google.com/search/docs/guides/create-URLs>" hreflang="en"><link rel="alternate" href="<https://developers.google.cn/search/docs/guides/create-URLs>" hreflang="en-cn"><link rel="alternate" href="<https://developers.google.com/search/docs/guides/create-URLs>" hreflang="x-default"><link rel="shortcut icon" href="<https://developers.google.com/_static/254aea64a1/images/favicon.png>"><link rel="apple-touch-icon" href="<https://developers.google.com/_static/254aea64a1/images/touch-icon.png>">

*Листинг 2. Разработка работоспособной программы.*

# !usr/bin/python3

# -\*- coding: utf-8 -\*-

**import** re

from lxml **import** etree

pattern = re.compile(r"(https?:\/\/)?([\w\.]+)\.([a-z]{2,6}\.?)([-a-zA-Z0-9@:%\_\+.~#?&//=]\*)\*\/?")

with open("test1.html") as file\_input:

test = pattern.finditer(file\_input.read())

root = etree.Element("txt")

**for** i in test:

element = etree.Element("a")

element.text = i.group(0)

root.append(element)

print(etree.tostring(root))

* **Результат выполнения программы:**

b'<txt>

<a>[https://developers.google.com/search/docs/guides/create-URLs </a](https://developers.google.com/search/docs/guides/create-URLs%20%3c/a)>

<a>[https://developers.google.com/search/docs/guides/create-URLs </a](https://developers.google.com/search/docs/guides/create-URLs%20%3c/a)>

<a>[https://developers.google.cn/search/docs/guides/create-URLs </a](https://developers.google.cn/search/docs/guides/create-URLs%20%20%3c/a)>

<a><https://developers.google.com/search/docs/guides/create-URLs </a>>

<a>[https://developers.google.com/\_static/254aea64a1/images/favicon.png </a](https://developers.google.com/_static/254aea64a1/images/favicon.png%20%3c/a)>

<a>[https://developers.google.com/\_static/254aea64a1/images/touch-icon.png </a](https://developers.google.com/_static/254aea64a1/images/touch-icon.png%20%3c/a)>

</txt>'