



Информационные ресурсы в финансовом мониторинге

НИЯУ МИФИ, КАФЕДРА ФИНАНСОВОГО МОНИТОРИНГА

КУРС ЛЕКЦИЙ

В.Ю. РАДЫГИН. Д.Ю. КУПРИЯНОВ. ЛЕКЦИЯ 1. ЧАСТЬ 2.

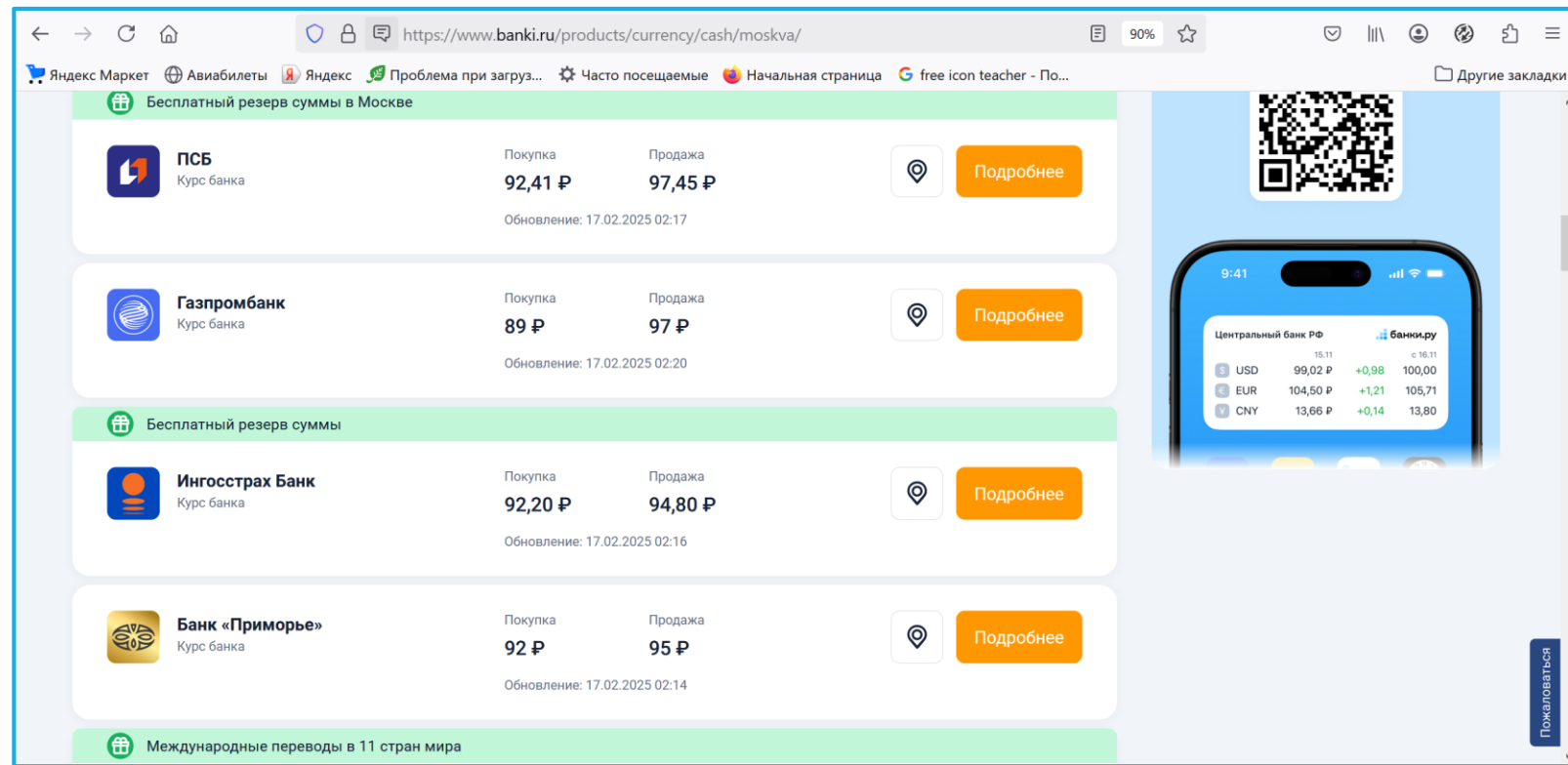
Часть 2

ВЫБОРКА ДАННЫХ ИЗ HTML-СТРАНИЦ

Не всё Excel под силу!

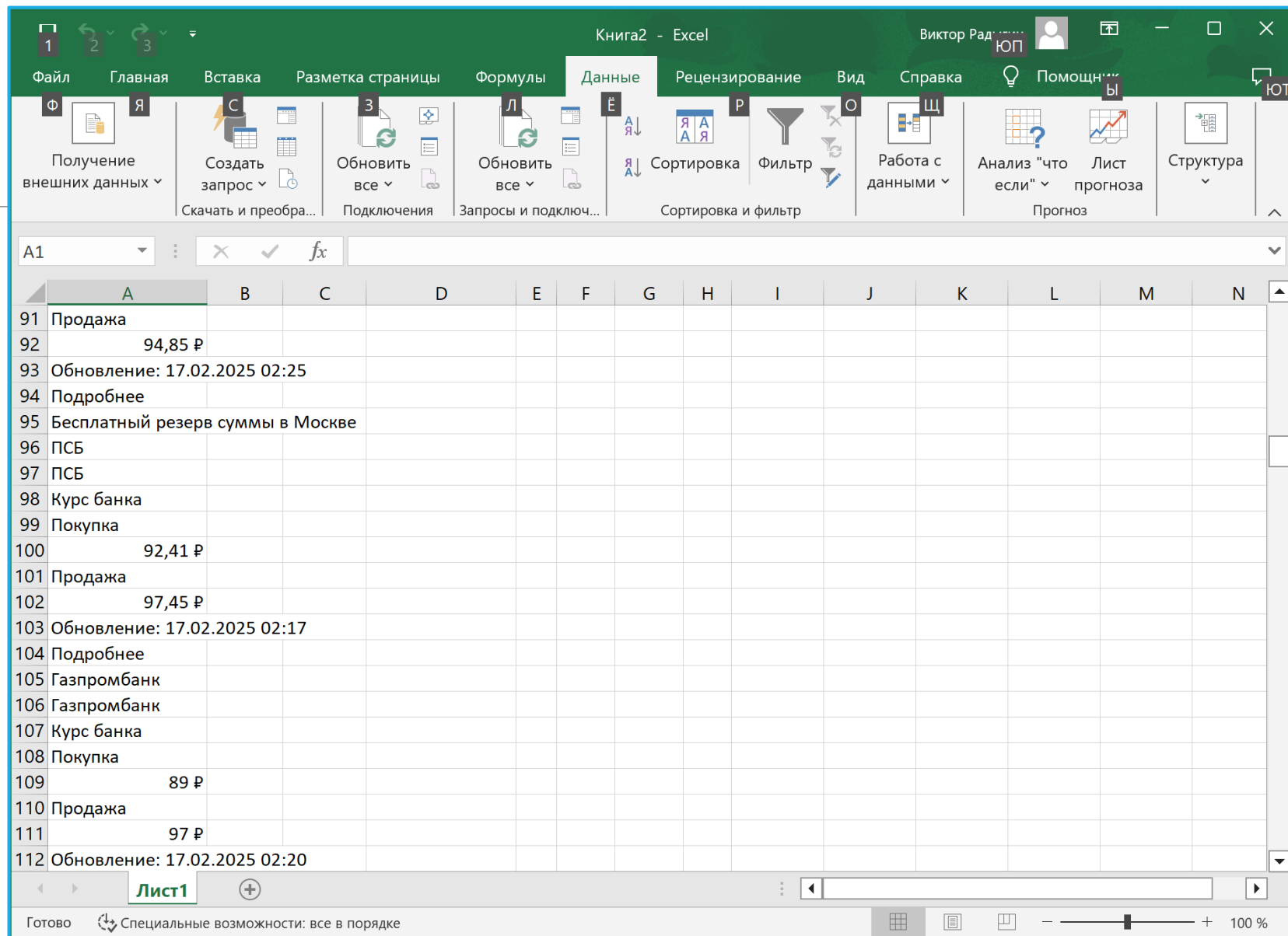
Существуют сайты, вёрстка которых устроена в современном адаптивном дизайне и не использует таблиц для представления информации.

В качестве примера можем рассмотреть страницу сайта <https://www.banki.ru/products/currency/cash/moskva/>, на которой показана информация о курсах покупки и продажи валюты в московских банках.



Не всё Excel под силу!

Как видно из примера, надстройка Power Query Microsoft Excel не позволяет загрузить табличные данные со страницы, использующей блочную верстку с помощью элементов <div> (верстку без таблиц).



Используем Python

Современные языки программирования содержат специальные библиотеки для загрузки документов из сети Интернет и анализа их содержимого. При этом анализ выполняется по аналогии с обработкой XML-документов.

В языке Python для первой задачи – задачи загрузки документов – используется встроенная библиотека `urllib`.

Для работы с содержимым HTML-страницы необходимо использовать XML-библиотеку. Рассмотренная ранее библиотека `xml.dom.minidom` в данной ситуации недостаточно функциональна и удобна. Поэтому мы познакомимся с библиотекой `lxml`.

Библиотека `lxml` устанавливается командой `python.exe -m pip install lxml` [12, 13].

Загрузка страницы

Если адрес страницы известен и не требует пересылки формы по протоколу POST, то она может быть загружена при помощи метода `request.urlopen()`:

```
import urllib.request

url = "https://www.banki.ru/products/currency/cash/moskva/"
page = urllib.request.urlopen(url).read()

print(page)
```

Загрузка страницы

Загрузка страницы без её дальнейшей обработки является бессмысленным процессом.

Результат работы программы показан на скриншоте и не очень поможет анализу данных.

[illegible]

Работа с содержимым страницы

Работу с содержимым загруженной HTML-страницы можно выполнить при помощи библиотеки `lxml`. Причём, первым шагом будет преобразование кода страницы, загруженного в виде одной большой строки в иерархию HTML-тегов. Это делается при помощи метода `fromstring()`:

```
import urllib.request
from lxml import html

url = "https://www.banki.ru/products/currency/cash/moskva/"
page_string = urllib.request.urlopen(url).read()
page = html.fromstring(page_string)
```


Вспомним, что такое тег

У тега в HTML бывают как собственное название, например, `div`, `ul`, `li`, так и атрибуты, например, у тега `<div>` в данном примере кода два атрибута: `class` со значением `'test'` и `data-info` со значением `'something'`.

Кроме того, у тега есть его содержимое. Это может быть просто текст, как у тегов `` или набор из одного или нескольких вложенных тегов, как у `<div>` или ``.

С каждой из этих составляющих мы должны уметь работать в программе.

```
<div class = "test" data-info = 'something'>
  <ul>
    <li class = "list">
      1
    </li>
    <li class = "list">
      2
    </li>
  </ul>
</div>
```

Немного про Python

В python различают три понятия, связанных с работой с объектом.

Пусть у нас есть объект `example`. Тогда в общем случае:

1. Объект можно обрабатывать функциями – `test_function(example)`.
2. У объекта можно вызывать методы – `example.test_method()`.
3. Можно обращаться к атрибутам объекта: `example.test_attribute`.

Разница, прежде всего, в форме записи. Использовании функции и метода записывается со скобками. Метод размещается через точку после названия объекта. Функция размещается до объекта, а сам объект внутри в скобках. Атрибут записывается без скобок через точку после названия объекта.

Работа с содержимым страницы

Полученная в примере с загрузкой страницы конструкция `page` – это, грубо говоря, первый тег HTML-страницы. Для любой страницы это будет тег `<HTML>`.

С тегом в библиотеке `lxml` можно делать следующие основные действия:

- узнать его имя – атрибут `tag` (`div`, `ul`, `li` из напоминания о тегах);
- узнать список его атрибутов и их значений – атрибут `attrib` (`class` и `test`, `data-info` и `something`);
- узнать его текст – атрибут `text` (1, 2 из тегов `li`);
- узнать вложенные в него теги: метод `getchildren()` (`ul` в `div`, `li` в `ul`).

Посмотрим, что есть в нашей загруженной странице.

Пример 3

```
import urllib.request
from lxml import html

url = "https://www.banki.ru/products/currency/cash/moskva/"
page_string = urllib.request.urlopen(url).read()
page = html.fromstring(page_string)
print("Название тега: ", page.tag)
print("Текст тега:", page.text)
print("Атрибуты тега:", page.attrib)
print("Вложенные теги:", page.getchildren())
```

Пример 3

File Edit Format Run Options Window Help

```
import urllib.request
from lxml import html

url = "https://www.banki.ru/products/currency/cash/moskva/"
page_string = urllib.request.urlopen(url).read()
page = html.fromstring(page_string)
print("Название тега: ", page.tag)
print("Текст тега:", page.text)
print("Атрибуты тега:", page.attrib)
print("Вложенные теги:", page.getchildren())
```

Результат

```
IDLE Shell 3.13.2
File Edit Shell Debug Options Window Help
Python 3.13.2 (tags/v3.13.2:4f8bb39, Feb  4 2025, 15:23:48) [MSC v.1942 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: C:/Users/dukme/Desktop/IR 2025/ex_1_2_2.py =====
Название тега:  html
Текст тега:

Атрибуты тега: {'lang': 'ru', 'data-device-type': 'desktop', 'class': 'env-js-of
f'}
Вложенные теги: [<Element head at 0x27ed026a210>, <Element body at 0x27ed026a6c0
>]
>>>
```

Поиск нужного тега

Рассмотренные конструкции позволяют полностью обойти страницу и добраться до нужных тегов. Но если решать задачу загрузки информации о курсах валют, то такой подход будет очень неудобен, так как нам надо будет понять всю структуру тегов данной страницы.

Зачем это делать, если нужна только определённая часть страницы с нужными тэгами?

Для решения данной задачи можно применить метод поиска тегов на странице или внутри другого тега по технологии XPath.

XPath

XPath – это язык описания шаблонов для нахождения нужных тегов в XML-документе. Так как HTML-страницы в некотором смысле являются подмножеством множества XML-документов, то к ним тоже можно применять данный подход.

XPath мы уже обсуждали при изучении XSLT-преобразований. Тем не менее, кратко опишем его основные элементы.

Базовые шаблоны

nodename	Все теги типа nodename
/	Вершина иерархии тегов (сама по себе тегом не является)
/nodename	Выбор самых верхних в иерархии тегов, но только с именем nodename.
//	Поиск среди всех тегов документа (отдельно смысла не имеет)
//nodename	Поиск среди всех тегов документа тега типа nodename
nodename//nodename2	Поиск тегов типа nodename2 среди всех тегов, находящихся внутри тегов типа nodename.
.	Текущий тег
..	Тег, внутри которого находится текущий тег
@	Поиск среди атрибутов
@attrname	Атрибут с именем attrname

Пример

Если считать, что кроме данных тегов ничего больше нет (div – корневой тег), то:

`page.xpath("//li")` – все теги `` (два элемента);

`page.xpath("/div/div")` – тег `<div>` строго вложенный в корневой тег `<div>`;

`page.xpath("//*[@class]")` – атрибуты `class` всех тегов.

```
<div class = "test" data-info = 'something'>
  <ul>
    <li class = "list">
      1
    </li>
    <li class = "list">
      2
    </li>
  </ul>
  <div>3</div>
</div>
```

Выбор с условием

[number]	Выбор из найденных тегов только тега, номер которого по порядку равен number (нумерация с 1)
[last()]	Выбор из найденных тегов только тега, последнего по порядку.
[last() - number]	Выбор из найденных тегов только тега, номер которого по порядку с конца равен number + 1
[position() < number]	Поиск среди всех найденных тегов, у которых номер по счету меньше number
[position() > number]	Поиск среди всех найденных тегов, у которых номер по счету больше number
[position() = number]	Поиск среди всех найденных тегов, у которых номер по счету равен number
[@attname]	Поиск среди всех найденных тегов, у которых есть атрибут с именем attname
[@attname > value]	Поиск среди всех найденных тегов, у которых значение атрибута attname > value
[@attname < value]	Поиск среди всех найденных тегов, у которых значение атрибута attname < value
[@attname = value]	Поиск среди всех найденных тегов, у которых значение атрибута attname = value

Пример

Если считать, что кроме данных тегов ничего больше нет (div – корневой тег), то:

`page.xpath("//li[2]")` – второй тег ``;

`page.xpath("//*[@class = 'list']")` – все теги ``;

`page.xpath("//*[@class][1]")` – корневой тег `<div>`.

```
<div class = "test" data-info = 'something'>
  <ul>
    <li class = "list">
      1
    </li>
    <li class = "list">
      2
    </li>
  </ul>
  <div>3</div>
</div>
```

Выбор с указанием направления

ancestor::	Поиск среди всех тегов, стоящих в иерархии выше данного (родитель, «дед», «прадед» и т.д.)
ancestor-or-self::	Поиск среди всех тегов, стоящих в иерархии выше данного (родитель, «дед», «прадед» и т.д.) плюс сам текущий тег
child::	Поиск среди прямых потомков данного тега («детей»)
descendant::	Поиск среди всех потомков данного тега
descendant-or-self::	Поиск среди всех потомков данного тега плюс сам текущий тег
другие	Подробнее о других можно прочесть на сайте W3C[14].

Пример

Если считать, что кроме данных тегов ничего больше нет (div – корневой тег), то:

`page.xpath("//ul/child::*")` – все теги ``;

`page.xpath("//li[1]/parent::*")` – тег ``;









`page.xpath("/div/descendant::div")` –
вложенный тег `<div>`.

```
<div class = "test" data-info = 'something'>
  <ul>
    <li class = "list">
      1
    </li>
    <li class = "list">
      2
    </li>
  </ul>
  <div>3</div>
</div>
```

Вернёмся к нашему примеру

Как получить нужные нам данные. Для однозначности, пусть нам нужны название банка, цена продажи доллара и цена покупки доллара.

Замечание. Курс валюты на скриншотах и при обработке с помощью программы может отличаться в связи с тем, что меняется со временем.

Бесплатный резерв суммы в Москве				
	ПСБ Курс банка	Покупка 92,13 ₽	Продажа 96,36 ₽	 Подробнее
Обновление: 17.02.2025 20:25				
	Газпромбанк Курс банка	Покупка 89,90 ₽	Продажа 96,90 ₽	 Подробнее
Обновление: 17.02.2025 20:25				
Бесплатный резерв суммы				
	Ингосстрах Банк Курс банка	Покупка 92,20 ₽	Продажа 94,80 ₽	 Подробнее
Обновление: 17.02.2025 20:22				
	Банк «Приморье» Курс банка	Покупка 93 ₽	Продажа 94 ₽	 Подробнее
Обновление: 17.02.2025 20:21				


«Инструменты разработчика» в браузере

Используем встроенные в браузер «Инструменты разработчика», чтобы понять в каком теге, какая информация лежит.

Для этого на строке с описанием информации банка нажмем правую кнопку мыши и в контекстном меню браузера выберем пункт «Просмотреть код».


Примечание. В разных браузерах пункт меню вызывающий просмотр выбранного элемента в виде html-кода может называться по-разному, например «Просмотреть код», «Исследовать», «Исследовать элемент».

Тэг с описанием информации по банку

**ПСБ**
Курс банка


Покупка
92,13 ₽

Продажа
96,36 ₽




Подробнее

Обновление: 17.02.2025 20:25

**Газпромбанк**
Курс банка

Покупка
89,90 ₽

Продажа
96,90 ₽



Подробнее

Обновление: 17.02.2025 20:25

Элементы

Консоль

Источники

Сеть

Производительность

Память

Приложение

Безопасность

Lighthouse

Регистратор

<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 jkNAek">

<div data-test="currency_rates-form_result-item">

<div data-test="currency_rates-form_result-item">

<div>

<div data-test="currency_rates-form_result-item">

<div data-test="currency_rates-form_result-item"> == \$0

<section class="Panel__sc-1g68tnu-1 hRExxV resultItemstyled__StyledPanel-sc-qb3d7j-0 fK1kyc">

<div class="Panel__sc-1g68tnu-0 cieTjP">

<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 jeAtcS">

Тэг с названием банка

Газпромбанк

Курс банка

Покупка

89,90 ₽

Продажа

96,90 ₽

Подробнее

Обновление: 17.02.2025 20:25

Элементы

Консоль

Источники

Сеть

Производительность

Память

Приложение

Безопасность

Lighthouse

Регистратор

<div data-test= flexbox-grid direction= row class= FlexboxGrid__sc-akw000-0 b1GLZK resultItemStyled__StyledWrapper
1 gcnFhb"> flex

<div data-test="flexbox-grid" width="40" class="FlexboxGrid__sc-akw86o-0 eNzgjv resultItemStyled__StyledWidth-sc-q
direction="row"> flex

- <a data-test="currency--result-item--logo" href="https://www.banki.ru/products/currency/bank/gazprombank/USD/mosk
nk" rel="nofollow noopener"> ...
- <a href="https://www.banki.ru/products/currency/bank/gazprombank/USD/moskva/" target="_blank" rel="nofollow noop

- <div data-test="currentc--result-item--name" class="Text__sc-vycpdy-0 OiTuY">Газпромбанк</div> == \$0
- <div data-test="currency--result-item--address" class="Text__sc-vycpdy-0 dQFICT">Курс банка</div>

Тэг с информацией о покупке валюты

The screenshot displays a web interface for Gazprombank. At the top left is the bank's logo and name. To the right, the exchange rate for the Russian Ruble is shown as 116.8 x 48. Below this, two buttons indicate the purchase and sale rates: 89.90 RUB for purchase and 96.90 RUB for sale. A location pin icon and a 'Подробнее' (More details) button are also present. The update time is noted as 17.02.2025 20:25. The bottom portion of the image shows the Chrome DevTools console with the 'Elements' tab selected. The DOM tree is expanded to show a flexbox grid structure, with the element containing the purchase rate highlighted.

Гazпромбанк
Курс банка

116.8 × 48

Покупка
89,90 ₺


Продажа
96,90 ₺

Обновление: 17.02.2025 20:25

Элементы Консоль Источники Сеть Производительность Память Приложение Безопасность Lighthouse Регистратор

```
</div>  
▼ <div data-test="flexbox-grid" direction="vert" width="30" class="FlexboxGrid__sc-akw86o-0 ipZNf resultItemstyled__3d7j-3 kJVkgx"> flex  
  ▼ <div data-test="flexbox-grid" direction="row" class="FlexboxGrid__sc-akw86o-0 eNzgjv"> flex  
    ▼ <div class="FlexboxGridItem__sc-1crr98y-0 fcKIDw">  
      ▼ <div data-test="currency--result-item---rate-buy"> == $0  
        ▼ <div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 dELgzc"> flex  
          <div data-test="text" class="Text__sc-vycpdy-0 iLRvuw">Покупка</div>  
          <div data-test="text" class="Text__sc-vycpdy-0 cQqMIr">89,90 ₺</div>  
        </div>
```

Тэг со значением курса покупки валюты



Газпромбанк
Курс банка

div.Text_sc-vycpdy-0.cQqMlr 116.8 × 28


Покупка

89,90 ₽

Продажа

96,90 ₽

Обновление: 17.02.2025 20:25



Подробнее

Элементы

Консоль

Источники

Сеть

Производительность

Память

Приложение

Безопасность

Lighthouse

Регистратор

</div>

<div data-test="flexbox-grid" direction="vert" width="30" class="FlexboxGrid__sc-akw86o-0 ipZNf resultItemstyled__93d7j-3 kJVkgx"> flex

<div data-test="flexbox-grid" direction="row" class="FlexboxGrid__sc-akw86o-0 eNzgjv"> flex

<div class="FlexboxGridItem__sc-1crr98y-0 fcKIDw">

<div data-test="currency--result-item---rate-buy">

<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 dElgzc"> flex

<div data-test="text" class="Text__sc-vycpdy-0 iLRvuw">Покупка</div>

<div data-test="text" class="Text__sc-vycpdy-0 cQqMlr">89,90 ₽</div> == \$0

</div>

28

Аналогично получаем информацию о продаже валюты

Газпромбанк

Курс банка

Покупка

89,90 ₽

Продажа

96,90 ₽

Подробнее

Обновление: 17.02.2025 20:25

Элементы

Консоль

Источники

Сеть

Производительность

Память

Приложение

Безопасность

Lighthouse

Регистратор

<div data-test="currency--result-item---rate-buy">

<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 dELgzc"> flex

<div data-test="text" class="Text__sc-vycpdy-0 iLRvuW">Покупка</div>

<div data-test="text" class="Text__sc-vycpdy-0 cQqMlR">89,90 ₽</div>

</div>

</div>

<div class="FlexboxGridItem__sc-1crr98y-0 fcKIDw">

<div data-test="currency--result-item---rate-sell"> == \$0


<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 dELgzc"> flex

<div data-test="text" class="Text__sc-vycpdy-0 iLRvuW">Продажа</div>

<div data-test="text" class="Text__sc-vycpdy-0 cQqMlR">96,90 ₽</div>

</div>

Аналогично получаем информацию о продаже валюты



Газпромбанк

Курс банка

Покупка

89,90 ₹

Обновление: 17.02.2025 20:25

div.Text__sc-vycpdy-0.cQqMlr

116.8 × 28

96,90 ₹

Покупка

📍

Подробнее

Элементы

Консоль

Источники

Сеть

Производительность

Память

Приложение

Безопасность

Lighthouse

Регистратор

<div data-test="currency--result-item---rate-buy">

<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 dElgzc"> flex

<div data-test="text" class="Text__sc-vycpdy-0 iLRvuw">Покупка</div>

<div data-test="text" class="Text__sc-vycpdy-0 cQqMlr">89,90 ₹</div>

</div>

</div>

<div class="FlexboxGridItem__sc-1crr98y-0 fcKIDw">

<div data-test="currency--result-item---rate-sell">

<div data-test="flexbox-grid" direction="vert" class="FlexboxGrid__sc-akw86o-0 dElgzc"> flex

<div data-test="text" class="Text__sc-vycpdy-0 iLRvuw">Продажа</div>

<div data-test="text" class="Text__sc-vycpdy-0 cQqMlr">96,90 ₹</div> == \$0

</div>

Что выяснили?

Строки с информацией о покупке и продаже валюты каждым банком описываются тэгом `<div>` с атрибутом `data-test`, у которого выставлено значение `'currency__rates-form__result-item'`.

Название банка лежит во вложенном к строке банка тэге `<div>` с атрибутом `data-test`, у которого выставлено значение `'current--result-item--name'`.

Курс покупки валюты лежит во вложенном к строке банка тэге `<div>` с атрибутом `data-test`, у которого выставлено значение `'currency--result-item---rate-buy'`, а далее во втором по счету вложенном к строке банка тэге `<div>` с атрибутом `data-test`, у которого выставлено значение `'text'`.

Что выяснили?

Аналогично, курс продажи валюты лежит во вложенном к строке банка тэге `<div>` с атрибутом `data-test`, у которого выставлено значение `'currency--result-item---rate-sell'`, а далее во втором по счету вложенном к строке банка тэге `<div>` с атрибутом `data-test`, у которого выставлено значение `'text'`.

Таким образом получаем следующую структуру необходимых к обработке тегов.

Программа подготовки и записи данных

Зная структуру, включая вложенность тэгов с искомыми значениями, можем написать программу, считывающую искомые данные.

Программа будет включать в себя три блока:

- загрузку страницы;
- анализ html-кода с помощью языка запросов XPath;
- запись результатов в csv-файл.

Пример 4. Загрузка страницы

```
import urllib.request
from lxml import html

url = "https://www.banki.ru/products/currency/cash/moskva/"
page_string = urllib.request.urlopen(url).read()
page = html.fromstring(page_string)

result = []
```

Пример 4. Анализ HTML

```
for item in page.xpath("//div[@data-test='currency__rates-form__result-item']"):
    bank_name = item.xpath("descendant::div[@data-test='current--result-item--name']")[0]
    buy_tag = item.xpath("descendant::div[@data-test='currency--result-item---rate-buy']")[0]
    buy_val = buy_tag.xpath("descendant::div[@data-test='text']")[1]
    sell_tag = item.xpath("descendant::div[@data-test='currency--result-item---rate-sell']")[0]
    sell_val = sell_tag.xpath("descendant::div[@data-test='text']")[1]
    result.append([bank_name.text,
                   buy_val.text.split()[0].replace(",", "", "."),
                   sell_val.text.split()[0].replace(",", "", ".")])
for row in result: print(result)
```

Пример 4. Результаты анализа HTML

Распечатаем загруженную информацию в терминал.

```
===== RESTART: C:/Users/dukme/Desktop/IR 2025/ex_1_2_3.py =====  
['КАМКОМБАНК', '94.13', '93.99']  
['Банк Казани', '92.75', '94.75']  
['ПСБ', '92.13', '96.36']  
['Газпромбанк', '90.20', '98']  
['Ингосстрах Банк', '92.20', '94.80']  
['Банк «Приморье»', '93', '94']  
['Т-Банк', '89.65', '100.65']  
['Агророс Банк', '93.21', '95.04']  
['Реалист Банк', '93', '97.50']  
['Внешфинбанк', '93.50', '96.80']
```

>>>

Пример 4. Запись результатов в csv-файл

```
import csv
```

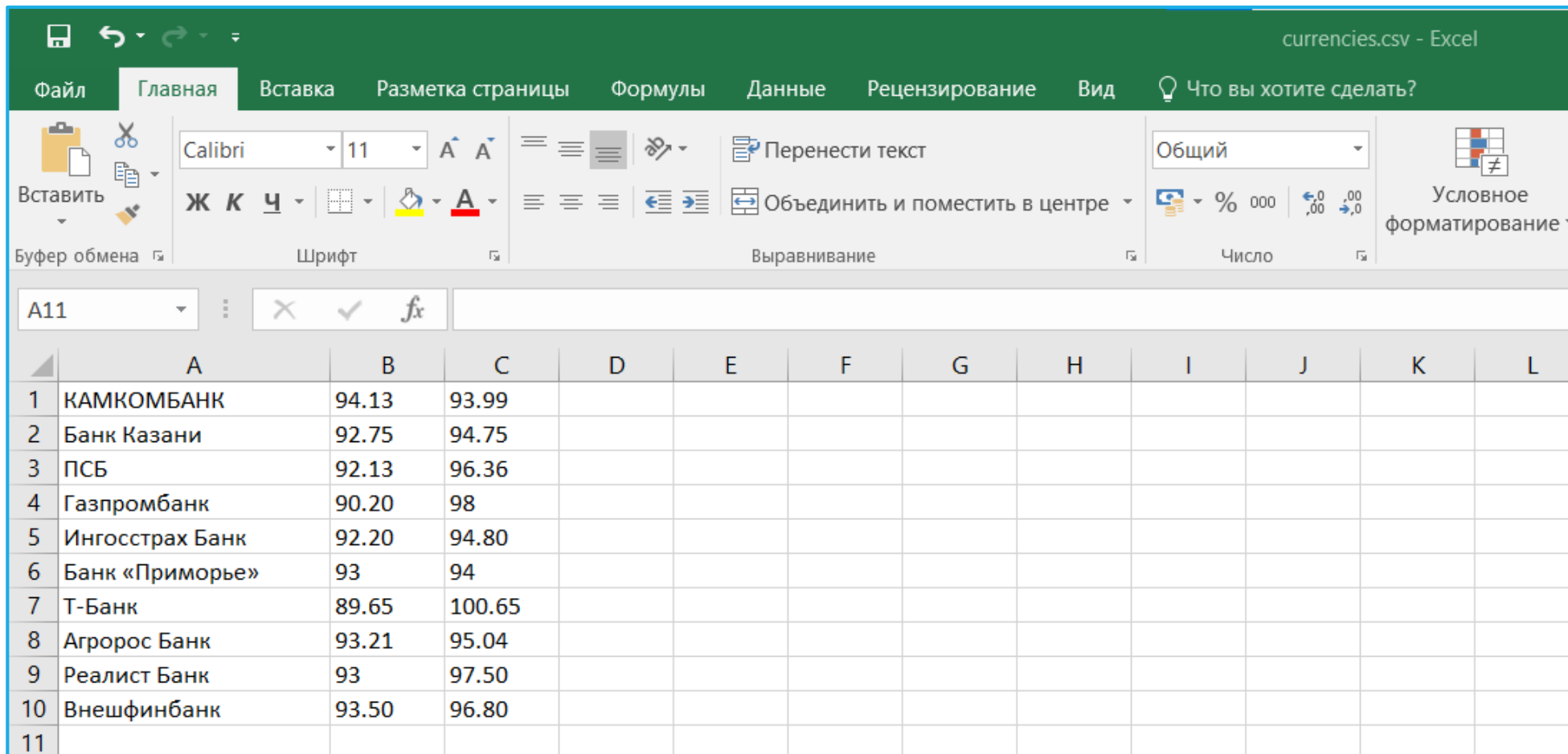
```
with open('currencies.csv', mode = 'w', newline='') as file:
```

```
    writer = csv.writer(file, delimiter=';', quotechar='"', quoting = csv.QUOTE_MINIMAL)
```

```
    for row in result:
```

```
        writer.writerow(row)
```

Полученный csv-файл с курсом валют по банкам



The screenshot shows the Microsoft Excel interface with the 'Главная' (Home) tab selected. The ribbon includes options for Font (Шрифт), Paragraph (Выравнивание), Numbers (Число), and Conditional Formatting (Условное форматирование). The active cell is A11. The data is organized in columns A through L, with rows 1 through 11. The data represents exchange rates for various banks.

	A	B	C	D	E	F	G	H	I	J	K	L
1	КАМКОМБАНК	94.13	93.99									
2	Банк Казани	92.75	94.75									
3	ПСБ	92.13	96.36									
4	Газпромбанк	90.20	98									
5	Ингосстрах Банк	92.20	94.80									
6	Банк «Приморье»	93	94									
7	Т-Банк	89.65	100.65									
8	Агророс Банк	93.21	95.04									
9	Реалист Банк	93	97.50									
10	Внешфинбанк	93.50	96.80									
11												

Генерируемые страницы

Не все данные доступны со статических HTML-страниц. Некоторые массивы данных получаются после указания ряда параметров. Фактически, такие страницы генерируются «на лету» в соответствии с выбранными значениями HTML-формы.

Например, поисковая система yandex передаёт строку поиска как параметр. В результате отображается список найденных страниц.

Другой пример – это сайт <http://pogoda-service.ru>. Он позволяет получить архивные данные о погоде в мире. Но для генерации нужно заполнить форму.

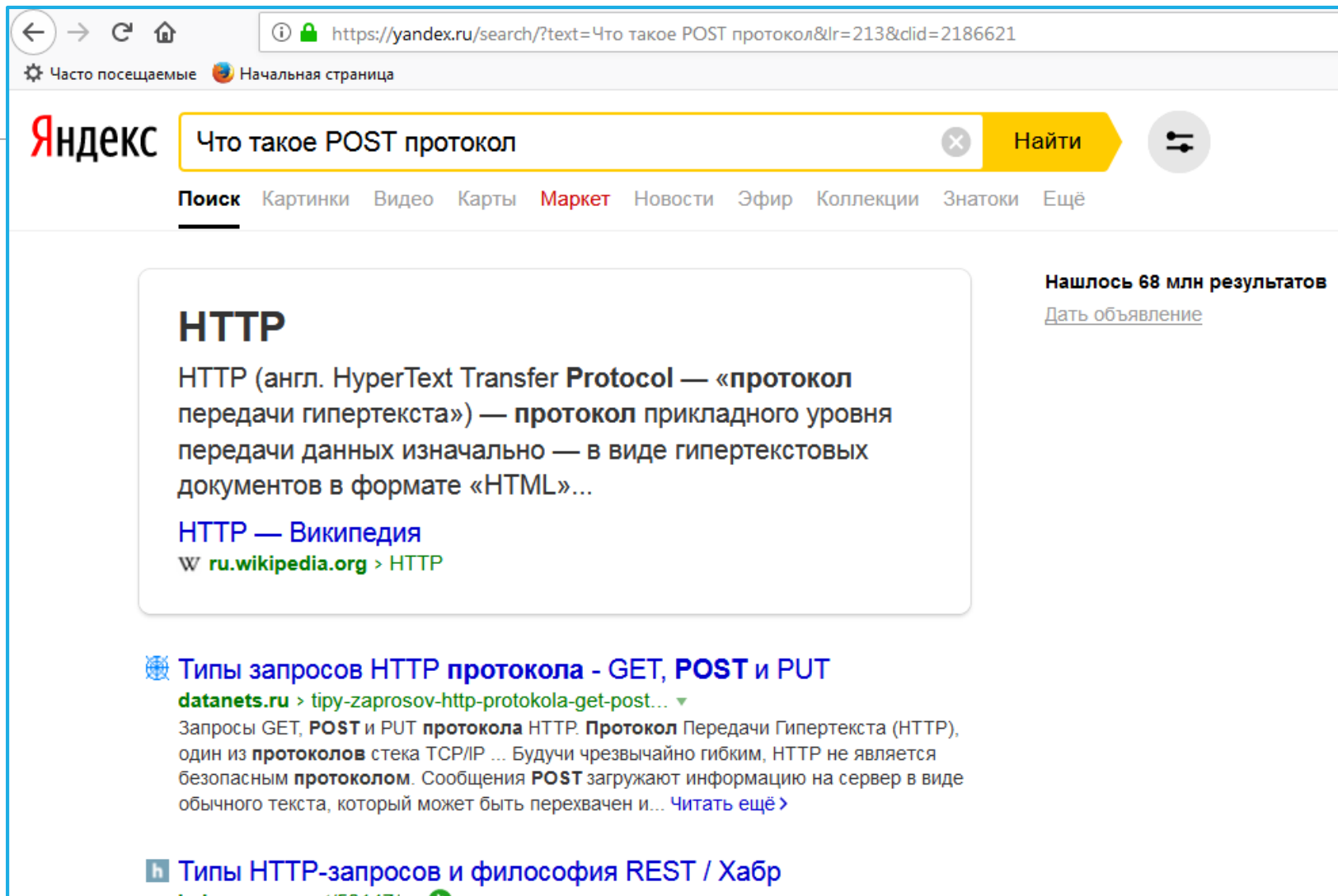
POST и GET протоколы

Для передачи параметров от одной страницы к другой в Интернет применяют два основных протокола: POST и GET.

Протокол GET подразумевает, что параметры передаются посредством адресной строки. После адреса сайта ставится символ '?', а затем записываются пары «ключ=значение», разделённые символом '&'.

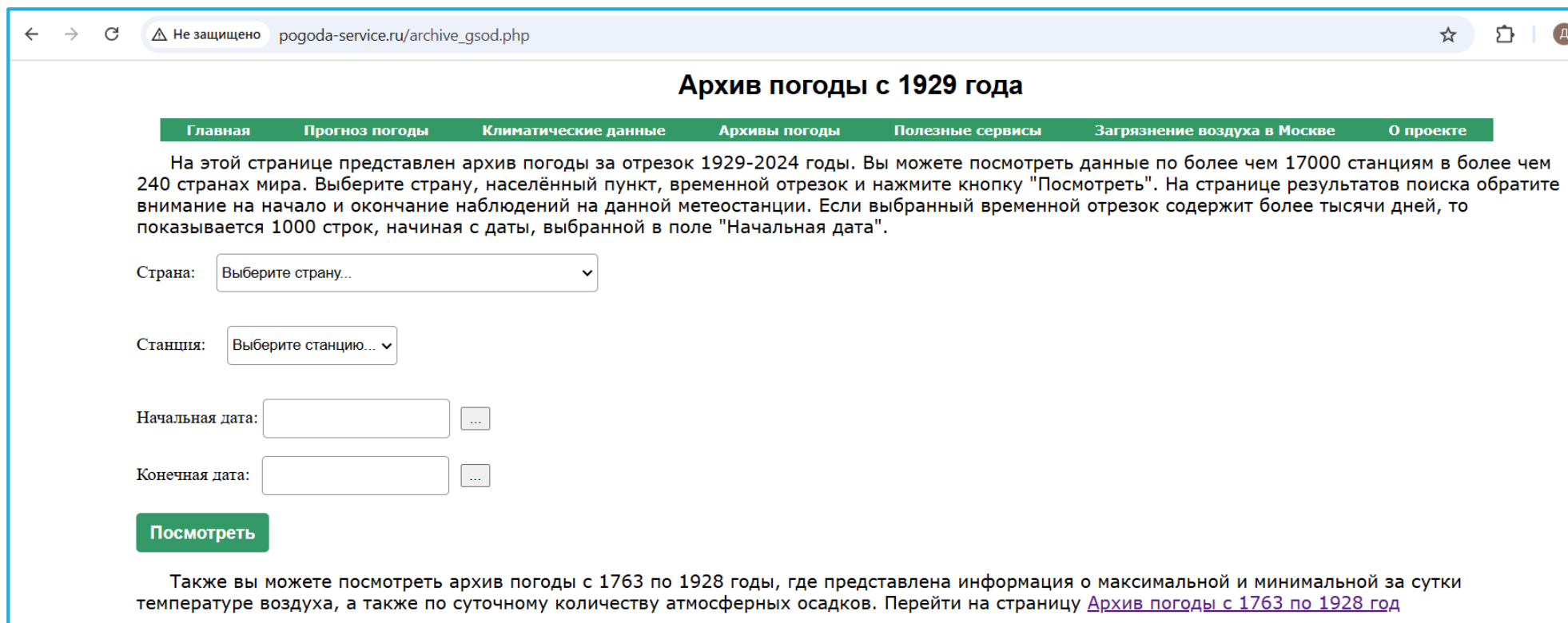
Поэтому, если странице нужно передавать параметры и используется протокол передачи – GET, то мы можем использовать уже известную нам схему. Для этого необходимо, просто, усложнить адрес.

Пример GET-запроса



Передача параметров GET-запроса из формы на сайте

На сайте http://pogoda-service.ru/archive_gsod.php с помощью специальной формы можно получить данные архива погоды.



The screenshot shows a web browser window with the address bar displaying "Не защищено" and "pogoda-service.ru/archive_gsod.php". The page title is "Архив погоды с 1929 года". Below the title is a navigation bar with links: Главная, Прогноз погоды, Климатические данные, Архивы погоды, Полезные сервисы, Загрязнение воздуха в Москве, and О проекте. The main content area contains a paragraph explaining the archive's scope (1929-2024) and the number of stations (over 17,000). Below this is a form with the following fields: "Страна:" with a dropdown menu showing "Выберите страну...", "Станция:" with a dropdown menu showing "Выберите станцию...", "Начальная дата:" with a text input and a calendar icon, and "Конечная дата:" with a text input and a calendar icon. A green button labeled "Посмотреть" is positioned below the date fields. At the bottom of the page, there is additional text about the archive from 1763 to 1928, including a link to "Архив погоды с 1763 по 1928 год".

Архив погоды с 1929 года

Главная Прогноз погоды Климатические данные Архивы погоды Полезные сервисы Загрязнение воздуха в Москве О проекте

На этой странице представлен архив погоды за отрезок 1929-2024 годы. Вы можете посмотреть данные по более чем 17000 станциям в более чем 240 странах мира. Выберите страну, населённый пункт, временной отрезок и нажмите кнопку "Посмотреть". На странице результатов поиска обратите внимание на начало и окончание наблюдений на данной метеостанции. Если выбранный временной отрезок содержит более тысячи дней, то показывается 1000 строк, начиная с даты, выбранной в поле "Начальная дата".

Страна:

Станция:

Начальная дата: ...

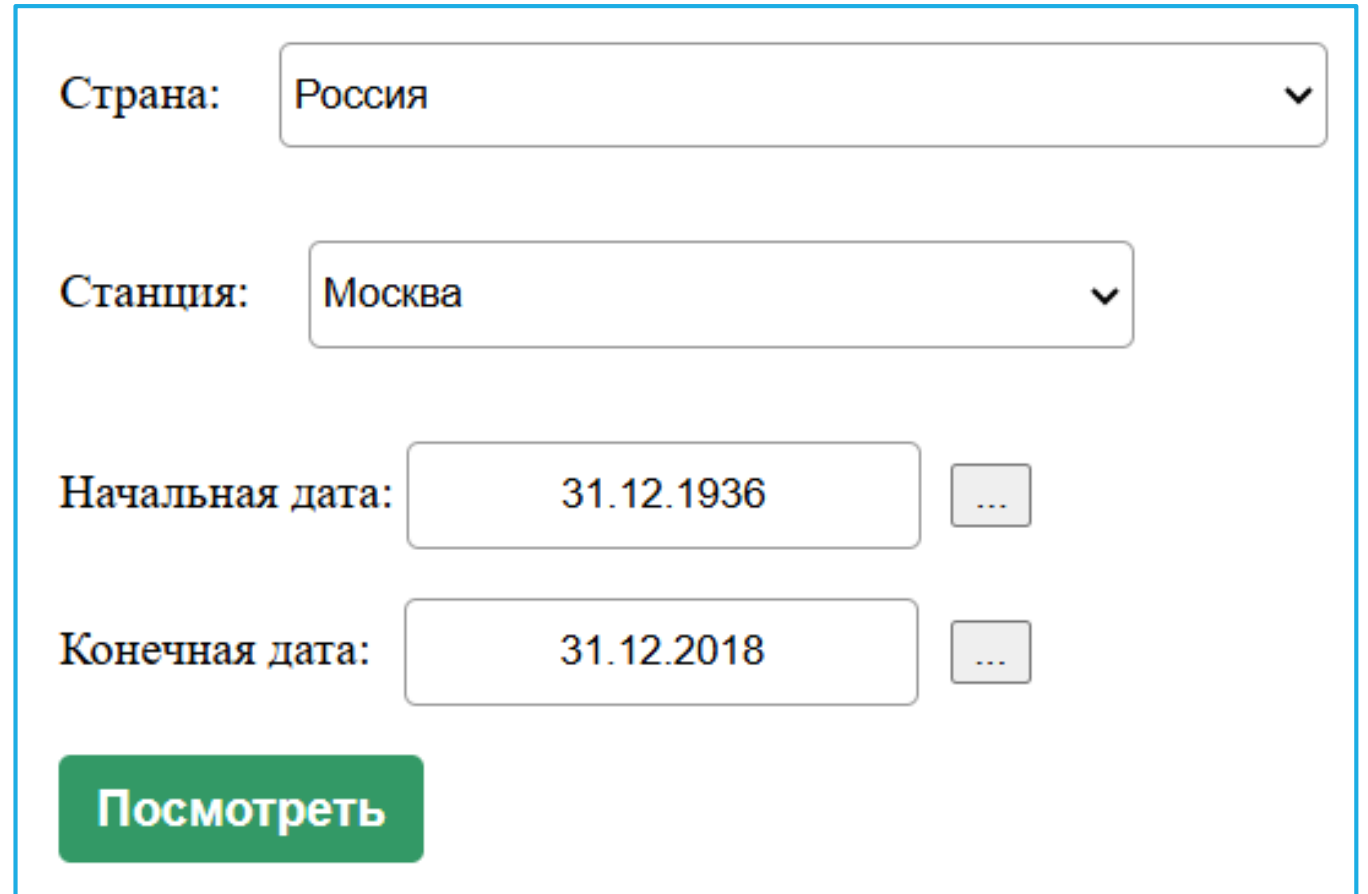
Конечная дата: ...

Также вы можете посмотреть архив погоды с 1763 по 1928 годы, где представлена информация о максимальной и минимальной за сутки температуре воздуха, а также по суточному количеству атмосферных осадков. Перейти на страницу [Архив погоды с 1763 по 1928 год](#)

Заполнение формы для передачи параметров

Разберём, как автоматически, минуя данную форму, добраться до данных.

Для этого сначала заполним форму и посмотрим какой GET-запрос формирует страница.



The image shows a web form with the following fields:

- Страна:** A dropdown menu with "Россия" selected and a downward arrow icon.
- Станция:** A dropdown menu with "Москва" selected and a downward arrow icon.
- Начальная дата:** A text input field containing "31.12.1936" and a small grey button with three dots "..." to its right.
- Конечная дата:** A text input field containing "31.12.2018" and a small grey button with three dots "..." to its right.

At the bottom of the form is a green button with the text "Посмотреть" in white.

Результат передачи запроса

<div>← → ↻ ⚠ Не защищено pogoda-service.ru/archive_gsod_res.php?country=RS&station=276120&datepicker_beg=31.12.1936&datepicker_end=31.12.2018 ☆</div>							
Данные из архива погоды. Москва							
<div>ГлавнаяПрогноз погодыКлиматические данныеАрхивы погодыПолезные сервисыЗагрязнение воздуха в МосквеО проекте</div>							
<div>Географические координаты: 55.833,37.617 Первое наблюдение: 31.12.1936 Последнее наблюдение: 15.01.2022</div>							
Дата	Максимальная температура	Минимальная температура	Средняя температура	Атмосферное давление	Скорость ветра	Осадки	Эффективная температура
31.12.1936	-2.8	-15.0	-8.9		5	0	
01.01.1937	1.1	-7.8	0.8		7		
02.01.1937	0.0	-2.8	-1.8			0	
03.01.1937	-1.1	-2.8	-1.7		9		
04.01.1937	0.0	-2.2	-1.1		7		
05.01.1937	0.0	-5.0	-2.2		6	0	
06.01.1937	-2.2	-6.1	-3.1		7		
07.01.1937	-1.1	-2.8	-1.4		6		
08.01.1937	1.1	-1.1	0.3		7		
09.01.1937	1.1	-2.8	-0.2		2		
10.01.1937	-5.0	-11.1	-7.2		5		
11.01.1937	-10.0	-16.1	-12.9		3	0	
12.01.1937	-15.0	-18.9	-16.5		3	0	
13.01.1937	-11.1	-17.2	-14.0		2		
14.01.1937	-8.9	-17.2	-12.5		3		
15.01.1937	-6.1	-10.0	-7.5		3		

Параметры GET-запроса в «Инструментах разработчика»

The screenshot shows the Chrome DevTools Network tab. The top toolbar includes icons for Elements, Console, Sources, Network (active), Performance, Memory, Application, Security, Lighthouse, and Page. Below the toolbar, there are checkboxes for 'Сохранять журнал' (Save logs) and 'Отключить кеш' (Disable cache), and a dropdown menu set to 'Без ограничения' (No limit). A filter bar shows 'Фильтр' (Filter) and 'Инвертировать' (Invert) options, with a list of filter categories: 'Все' (All), 'Fetch/XHR', 'Документ' (Document), 'CSS', 'JS', 'Шрифт' (Font), 'Изображение' (Image), and 'Носитель' (Media). A timeline at the top shows a request taking approximately 250 ms. The main list of requests includes:

- archive_gsod_res.php?country=RS&station=276120&datepicker_beg=31.12.1936&datepicker_end=31.12.2018
- style.css
- watch.js
- watch.js
- injectExtensionId.js
- 27283160?wmode=7&page-url=http%3A%2F%2Fpogoda-serv...(0-0-0)rqn(1)aw(1)rcm(0)cdl(na)eco(3178752)ti(1)
- favicon.ico

The right-hand pane shows the details for the selected request, with tabs for 'Заголовки' (Headers) and 'Полезная нагрузка' (Payload). The 'Полезная нагрузка' tab is active, displaying the 'Параметры строки запроса' (Request parameters) section:

- country: RS
- station: 276120
- datepicker_beg: 31.12.1936
- datepicker_end: 31.12.2018

Передача GET-параметров в программе с помощью библиотеки urllib

```
import urllib.request
import urllib.parse
from lxml import html

url = 'http://pogoda-service.ru/archive_gsod_res.php'
values = {'country': 'RS',
          'station': '276120',
          'datepicker_beg': '31.12.1936',
          'datepicker_end': '31.12.2018',
          'bsubmit': 'Посмотреть'}
```

```
data = urllib.parse.urlencode(values)
page_string = urllib.request.urlopen(url +
                                     '?' + data).read()
page = html.fromstring(page_string)
```

Анализ структуры таблицы с погодой в «Инструментах разработчика»

tr 868.76 × 21.2	Максимальная температура	Минимальная температура	Средняя температура	Атмосферное давление	Скорость ветра	Осадки	Эффективная температура
31.12.1936	-2.8	-15.0	-8.9		5	0	
01.01.1937	1.1	-7.8	0.8		7		
02.01.1937	0.0	-2.8	-1.8			0	

Элементы

Консоль

Источники

Сеть

Производительность

Память

Приложение

Безопасность

Lighthouse

Регистратор

69

68

⚙

⋮

✕

▼ <table class="table_res" align="center" style="margin-top:20px; border-color: rgb(102,102,255)">

▶ <thead> ⋮ </thead>

<tbody align="center"></tbody>

▼ <tbody align="center">

▼ <tr> == \$0

<td class="td_res" style="border-color: rgb(102,102,255)">31.12.1936</td>

<td class="td_res" style="border-color: rgb(102,102,255)">-2.8</td>

<td class="td_res" style="border-color: rgb(102,102,255)">-15.0</td>

<td class="td_res" style="border-color: rgb(102,102,255)">-8.9</td>

<td class="td_res" style="border-color: rgb(102,102,255)"></td>

<td class="td_res" style="border-color: rgb(102,102,255)">5</td>

<td class="td_res" style="border-color: rgb(102,102,255)">0</td>

<td class="td_res" style="border-color: rgb(102,102,255)"></td>

</tr>

▶ <tr> ⋮ </tr>

Стили

Вычисленные

Макет >>

🔍 Фильтр

:hov .cls + 🗨 📄

element.style {

}

tr {

таблица стилей агента пользователя

display: table-row;

vertical-align: inherit;

unicode-bidi: isolate;

border-color: ▶ ■ inherit;

}

Унаследовано от table.table_res

.table_res {

border: ▶ 5px solid;

border-collapse: collapse;

style.css:75

ЧТО ВЫЯСНИЛИ?

Все необходимые для обработки данные лежат внутри таблицы (тег `<table>`), у которой значение атрибута `class` выставлено в значение `'table_res'`.

Полезные данные (не заголовки столбцов) находятся внутри тега `<tbody>`.

Каждой строке с наблюдением по конкретной дате соответствует один тег `<tr>`.

Каждой ячейке таблицы соответствует один тег `<td>`, данные в колонках не размечены (имеют одинаковое значение атрибута `class`, выставленное в `'td_res'`), и могут быть обработаны только по порядковому номеру внутри массива.

Найдём нужные данные и сохраним в CSV

```
result = []  
  
for row in page.xpath("//table[@class='table_res']/tbody//tr"):   
    result_row = [cell.text for cell in row]  
    result.append(result_row)  
  
import csv  
  
with open('weather.csv', mode = 'w', newline='') as file:  
    writer = csv.writer(file, delimiter=';', quotechar='"', quoting = csv.QUOTE_MINIMAL)  
    for row in result:  
        writer.writerow(row)
```

POST-протокол

GET-запрос имеет ряд ограничений. Во-первых его длина не бесконечна и, поэтому, много параметров с его помощью передать невозможно.

Во-вторых с помощью GET-запроса нельзя передавать сложные данные, например, файлы.

По этой причине разработан POST-протокол, позволяющий передавать параметры другим способом, не связанным с адресной строкой.

Как узнать, что было передано странице с помощью POST-запроса?

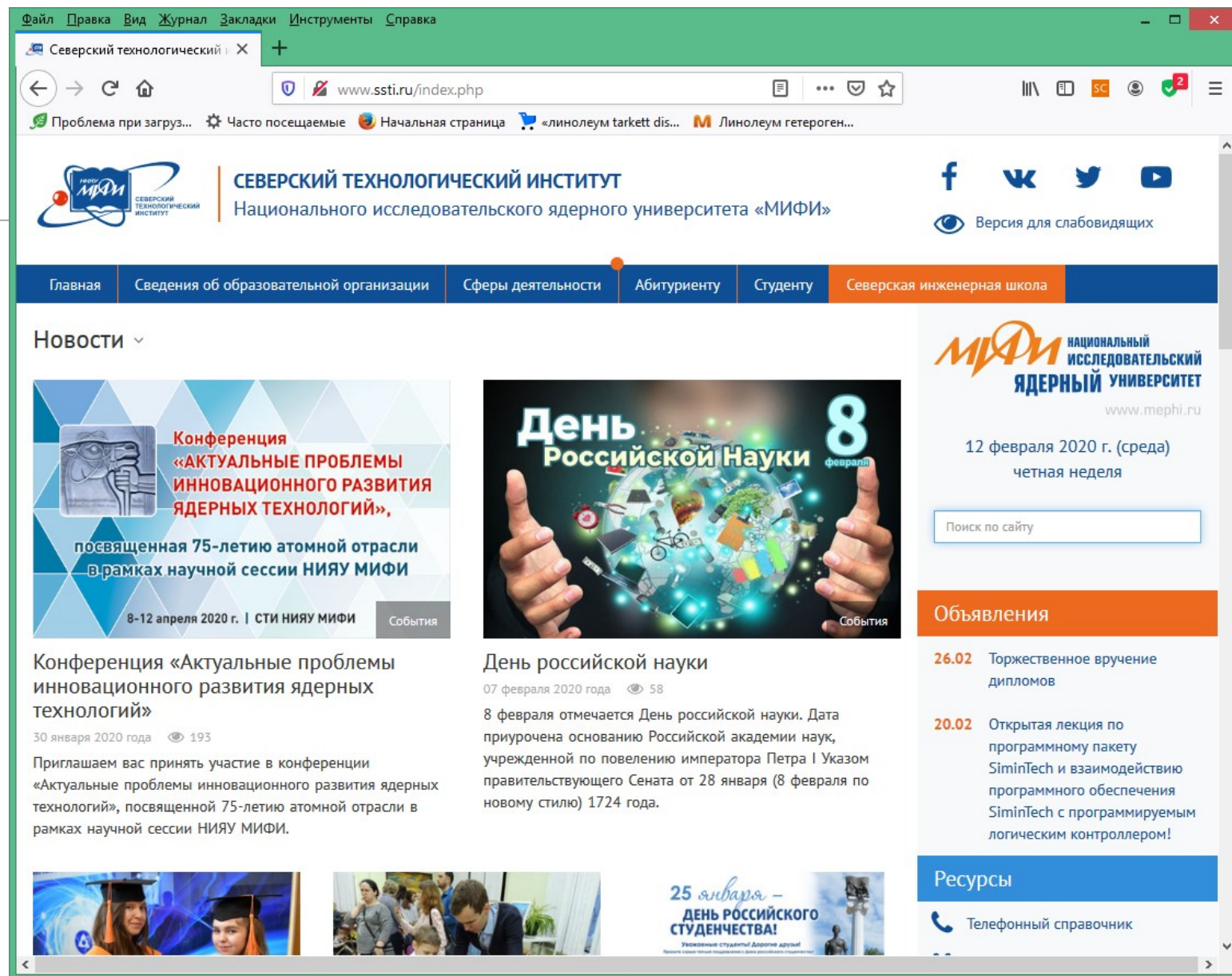
Воспользуемся «Инструментами разработчика», встроенными в браузер.

Сайт СТИ НИЯУ МИФИ

На сайте СТИ НИЯУ МИФИ есть поиск.

Он работает путем передачи параметров методом POST.

Как автоматически выполнить поиск и собрать его результаты?



Параметры POST-запроса

Воспользуемся средством «Инструменты веб-разработчика» в браузере Firefox для просмотра параметров, пересылаемых по протоколу POST из формы поиска сайта СТИ НИЯУ МИФИ.

Файл Правка Вид Журнал Закладки Инструменты Справка

Поиск по сайту » Северский т X

www.ssti.ru/index.php?do=search

Проблема при загруз... Часто посещаемые Начальная страница «линолеум tarkett dis... М Линолеум гетероген...

СЕВЕРСКИЙ ТЕХНОЛОГИЧЕСКИЙ ИНСТИТУТ
Национального исследовательского ядерного университета «МИФИ»

f vk tw yt

Версия для слабовидящих

Главная Сведения об образовательной организации Сферы деятельности Абитуриенту Студенту Северская инженерная школа

Поиск по сайту

MEPhi

Начать поиск Расширенный поиск

По Вашему запросу найдено 50 ответов (Результаты запроса 1 - 10) :

Конференция «Актуальные проблемы инновационного развития ядерных технологий»

Приглашаем вас принять участие в конференции «Актуальные проблемы инновационного развития ядерных технологий», посвященной 75-летию атомной отрасли в рамках научной сессии НИЯУ МИФИ.

Внимание студентов 2, 3, 4, 5 курсов, аспирантов и магистров 1 курса очной формы обучения!

Инспектор Консоль Отладчик Стили Профайлер Память Сеть Хранилище Поддержка доступности

Поиск URL Все HTML CSS JS XHR Шрифты Изображения Медиа WS Прочее [] Непрерывные логи [] Отключить кэш Без ограни... Н...

Статус	Метод	Домен	Файл	Причина	Тип	Передано	Размер
200	POST	www.ssti.ru	index.php?do=search	document	html	12,47 КБ	40,5...
200	GET	www.ssti.ru	jquery.js	script	js	кэшировано	90,6...
200	GET	www.ssti.ru	jqueryui.js	script	js	кэшировано	63,3...
200	GET	www.ssti.ru	dle_js.js	script	js	кэшировано	21,8...
200	GET	www.ssti.ru	theme.min.css	stylesheet	css	кэшировано	114,...
200	GET	www.ssti.ru	ssti.css	stylesheet	css	кэшировано	85,2...

27 запросов 822,66 КБ / 18,27 КБ передано Передано за: 15,89 с DOMContentLoaded: 411

Заголовки Куки Параметры Ответ Тайминги стек вызовов

Данные форм

do: search
subaction: search
search_start: 1
full_search: 0
result_from: 1
story: MEFhi

Полезная нагрузка запроса

Передача POST-параметров в программе с помощью библиотеки urllib

```
import urllib.request
import urllib.parse
from lxml import html
url = 'http://www.ssti.ru/index.php'
values = {'do': 'search',
          'subaction': 'search',
          'search_start': '1',
          'full_search': '0',
          'result_from': '1',
          'story': 'MEPhI'}
```

```
data = urllib.parse.urlencode(values).encode()
req = urllib.request.Request(url, data)
page_string = urllib.request.urlopen(req).read().decode(
    'windows-1251')
page = html.fromstring(page_string)
result = []
for row in page.xpath("//div[@class='left-news']//h2/a"):
    result_row = [row.text, row.attrib['href']]
    result.append(result_row)
print(result)
```


Результат

```
Python 3.7.2 Shell
File Edit Shell Debug Options Window Help
Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: E:\Works\Victor\Students\infres\2020\Lecton1\sti-post.py =====
[['Конференция «Актуальные проблемы инновационного развития ядерных технологий»',
  'http://www.ssti.ru/main/1527-konferenciya-aktualnye-problemy-innovacionnogo-r
azvitiya-yadernyh-tehnologiy.html'], ['Внимание студентов 2, 3, 4, 5 курсов, асп
ирантов и магистров 1 курса очной формы обучения!', 'http://www.ssti.ru/vote/15
21-vnimanie-studentov-2-3-4-5-kursov-aspirantov-i-magistrov-1-kursa-ochnoy-formy
-obucheniya.html'], ['Приглашаем принять участие в очно-заочном отборочном туре
олимпиады школьников «Росатом»!', 'http://www.ssti.ru/main/1480-priglashaem-pri
nyat-uchastie-v-ochno-zaochnom-otborochnom-ture-olimpiady-shkolnikov-rosatom.htm
l'], ['Объявлен конкурс на именные стипендии АО «СХК»!', 'http://www.ssti.ru/vot
e/1469-obyavlen-konkurs-na-imennye-stipendii-ao-shk.html'], ['Внимание студентов
очной формы обучения! Объявлен конкурс «Студент года»!', 'http://www.ssti.ru/vo
te/1461-vnimanie-studentov-ochnoy-formy-obucheniya-obyavlen-konkurs-student-goda
.html'], ['Внимание студентов очной формы обучения! Объявлен конкурс на стипенди
и имени В.Н.Меренкова!', 'http://www.ssti.ru/vote/1447-vnimanie-studentov-ochnoy
-formy-obucheniya.html'], ['Опубликованы приказы о зачислении 1 волны.', 'http:/
/www.ssti.ru/priemnaya-komissiya/1415-opublikovany-prikazy-o-zachislenii-1-volny
.html'], ['Информация для иногородних абитуриентов граждан РФ, желающих учиться
в НИЯУ МИФИ!', 'http://www.ssti.ru/priemnaya-komissiya/1408-informaciya-dlya-ino
gorodnih-abiturientov-grazhdan-rf-zhelayuschih-uchitsya-v-niyau-mifi.html'], ['В
нимание студентов 2, 3, 4, 5 курсов, аспирантов и магистров 1 курса очной формы
обучения!', 'http://www.ssti.ru/vote/1401-vnimanie-studentov-2-3-4-5-kursov-asp
irantov-i-magistrov-1-kursa-ochnoy-formy-obucheniya.html'], ['Объявлен конкурс н
а именные стипендии городского округа ЗАТО Северск!', 'http://www.ssti.ru/vote/1
368-obyavlen-konkurs-na-imennye-stipendii-gorodskogo-okruga-zato-seversk.html']]
>>> |
```

Полезные ссылки

1. <https://pandas.pydata.org/> – библиотека Pandas для Python.
2. <https://www.w3.org/TR/xml/> – рекомендации W3C по стандарту XML.
3. <https://www.w3.org/TR/1999/REC-xslt-19991116> – рекомендации W3C по стандарту XSLT 1.0 (поддерживаются большинством браузеров).
4. <https://www.w3.org/TR/2017/REC-xslt-30-20170608/> – рекомендации W3C по стандарту XSLT 3.0 (новейший стандарт).
5. <https://www.w3.org/TR/1999/REC-xpath-19991116/> – рекомендации W3C по стандарту XPATH 1.0 (поддерживаются большинством браузеров).
6. <https://gitlab.gnome.org/GNOME/libxslt/-/wikis/home> – страница libxslt.
7. <https://www.zlatkovic.com/pub/libxml/> – скачать libxslt для windows.
8. <https://www.w3.org/TR/1999/REC-xslt-19991116#format-number> – описание работы с функцией format-number.

Полезные ссылки

7. <https://docs.python.org/2/library/xml.etree.elementtree.html> – библиотека контейнер для представления иерархических структур в Python
8. <https://docs.python.org/2/library/xml.dom.html> – библиотека для работы с DOM-объектами в Python
9. <https://docs.python.org/2/library/xml.dom.minidom.html> – библиотека для минимальных действия с DOM-объектами в Python.
10. <https://pypi.org/project/lxml/> – библиотека lxml.
11. <https://python-scripts.com/parsing-lxml> – о библиотеке lxml.
12. https://www.w3schools.com/xml/xpath_intro.asp – основы XPath на сайте консорциума W3C.
13. <https://www.banki.ru/products/currency/cash/moskva/>
14. http://pogoda-service.ru/archive_gsod_res.php