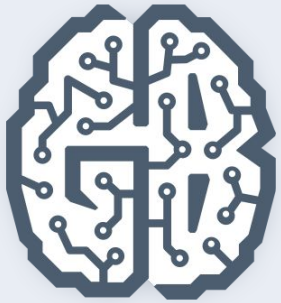
Методы сбора и обработки данных при помощи Python

Урок 6



Парсинг HTML

Краткий обзор технологий для понимания сбора и обработки данных

План урока

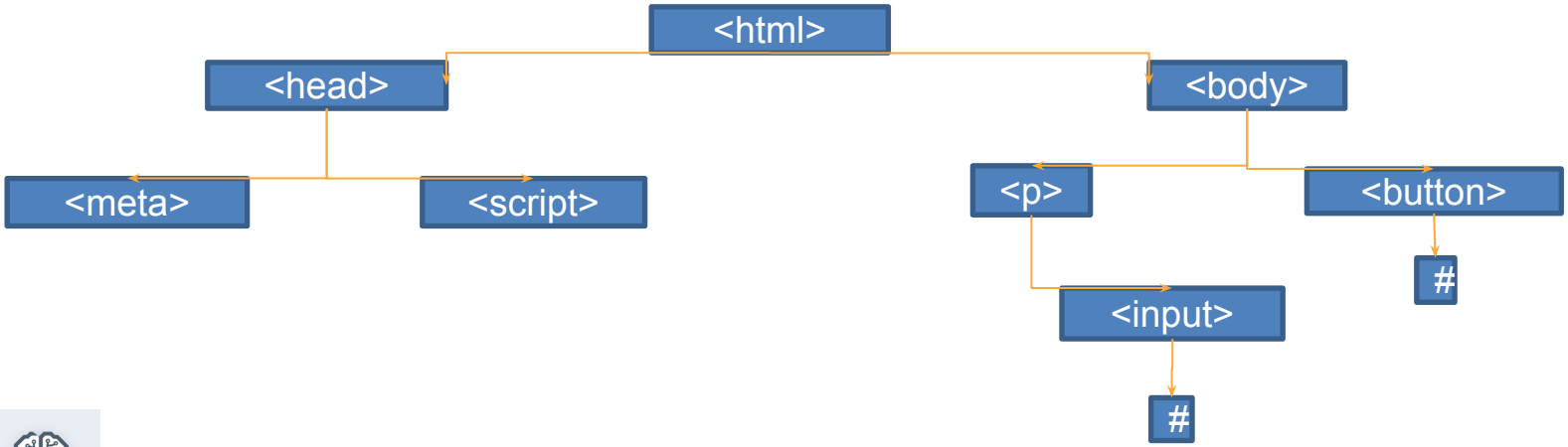
1) Структура страницы HTML - DOM
2) Язык запросов Xpath
3) Парсинг HTML в Python – библиотека lxml

Структура страницы HTML

Структура страницы HTML

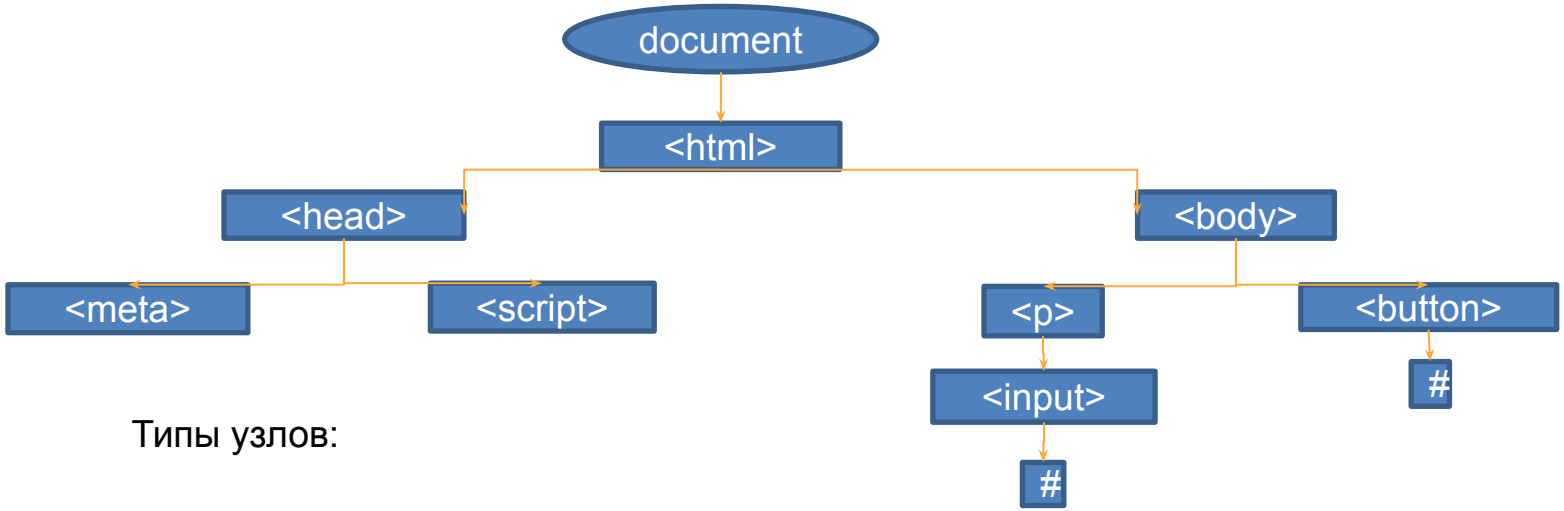
DOM

(от англ. *Document Object Model* — «объектная модель документа») —программный интерфейс, позволяющий программам и скриптам получить доступ к содержимому HTML-документов, а также изменять их содержимое, структуру и оформление таких документов.



Структура страницы HTML

DOM, как дерево тэгов

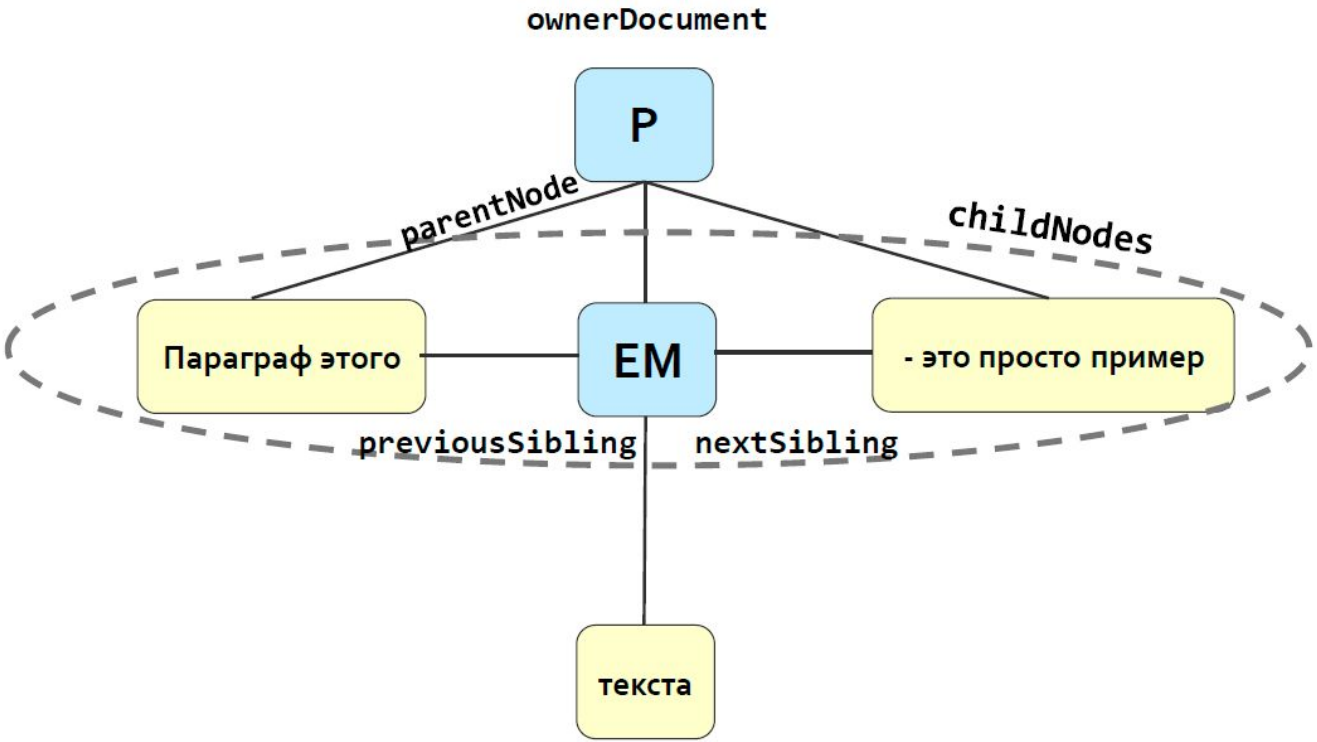


Типы узлов:

- документ
- элементы
- текстовые узлы
- комментарии

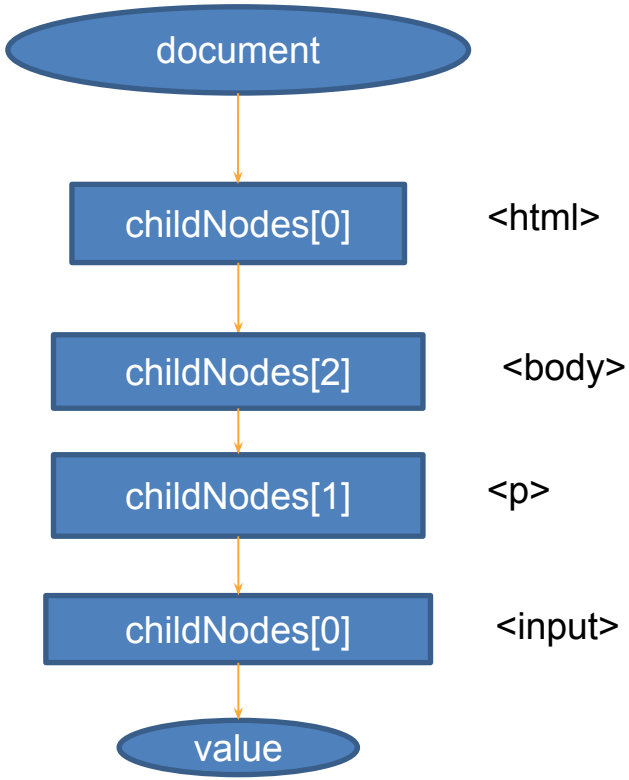
Связи между объектами

```
<p>Параграф этого <em>текста</em> - это просто пример</p>
```



Навигация по DOM

```
document.childNodes[0].childNodes[2].childNodes[1].childNodes[0].value
```



Язык запросов XPath

XPath (XML Path Language) — язык запросов к элементам XML-документа. Разработан для организации доступа к частям документа XML в файлах трансформации XSLT и является стандартом консорциума W3C. XPath призван реализовать навигацию по DOM в XML. В XPath используется компактный синтаксис, отличный от принятого в XML.

```
<html>
 <body>
    <div>Первый слой
      <span>блок текста в первом слое</span>
    </div>
    <div>Второй слой</div>
    <div>Третий слой
      <span class="text">первый блок в третьем слое</span>
      <span class="text">второй блок в третьем слое</span>
      <span>третий блок в третьем слое</span>
    </div>
    <span>четвёртый слой</span>
    <img />
 </body>
</html>
```

XPath-путь **/html/body/*/span[@class]** будет соответствовать в нём двум элементам исходного документа — `<span class="text">первый блок в третьем слое</span>` и `<span class="text">второй блок в третьем слое</span>`

Язык запросов XPath

Выражение XPath	Результат
messages	Выбирает все узлы с именем "messages"
/messages	Выбирает корневой элемент сообщений **Примечание**: Если путь начинается с косой черты (/), то он всегда представляет абсолютный путь к элементу!
messages/note	Выбирает все элементы note, являющиеся потомками элемента messages
//note	Выбирает все элементы note независимо от того, где в документе они находятся
messages//note	Выбирает все элементы note, являющиеся потомками элемента messages независимо от того, где они находятся от элемента messages
//@date	Выбирает все атрибуты с именем date

Язык запросов XPath

Выражение XPath	Результат
/messages/note[1]	Выбирает первый элемент note, который является прямым потомком элемента messages. **Примечание**: В IE 5,6,7,8,9 первым узлом будет [0], однако согласно W3C это должен быть [1]. Чтобы решить эту проблему в IE, нужно установить опцию SelectionLanguage в значение XPath. В JavaScript: xml.setProperty("SelectionLanguage","XPath");
/messages/note[last()]	Выбирает последний элемент note, который является прямым потомком элемента messages.
/messages/note[last()-1]	Выбирает предпоследний элемент note, который является прямым потомком элемента messages.
/messages/note[position()	Выбирает первые два элемента note, которые являются прямыми потомками элемента messages.
//heading[@date]	Выбирает все элементы heading, у которых есть атрибут date
//heading[@date="10/01/2008"]	Выбирает все элементы heading, у которых есть атрибут date со значением "10/01/2008"

Язык запросов XPath

Выбор неизвестных заранее узлов

Чтобы найти неизвестные заранее узлы XML документа, XPath позволяет использовать специальные символы.

Спецсимвол	Описание
*	Соответствует любому узлу элемента
@*	Соответствует любому узлу атрибута
node()	Соответствует любому узлу любого типа

В следующей таблице приводятся некоторые выражения XPath со спецсимволами, позволяющие сделать выборки по демонстрационному XML документу:

Выражение XPath	Результат
/messages/*	Выбирает все элементы, которые являются прямыми потомками элемента messages
//*	Выбирает все элементы в документе
//heading[@*]	Выбирает все элементы heading, у которых есть по крайней мере один атрибут любого типа

Язык запросов XPath

Выбор нескольких путей

Использование оператора | в выражении XPath позволяет делать выбор по нескольким путям.

В следующей таблице приводятся некоторые выражения XPath, позволяющие сделать выборки по демонстрационному XML документу:

Выражение XPath	Результат
//note/heading | //note/body	Выбирает все элементы heading И body из всех элементов note
//heading | //body	Выбирает все элементы heading И body во всем документе

Язык запросов Xpath и python

```
from lxml import html


links = html.fromstring(html_page).xpath('//ul[28]//li/a/@href')
```

Домашнее задание

1) Необходимо собрать информацию о вакансиях на должность программиста или разработчика с сайта job.ru или hh.ru. (Можно с обоих сразу) Приложение должно анализировать несколько страниц сайта. Получившийся список должен содержать в себе:

*Наименование вакансии,
*Предлагаемую зарплату
*Ссылку на саму вакансию

2) Доработать приложение таким образом, чтобы можно было искать разработчиков на разные языки программирования (Например Python, Java, C++)

Ваши вопросы?