XML ( extensibal markup language ) là ngôn ngữ đánh dấu siêu văn bản mở rộng .

Với khả năng mô tả nhiều loại dữ liệu khác nhau , XML được ứng dụng nhiều để đơn giản hóa việc chia sẻ dữ liệu giữa các hệ thống có kết nối Internet .

XML được sử dụng chủ yếu trong việc định dạng tài liệu nên 1 số công cụ văn phòng đều sử dụng ngôn ngữ xml thành định dạng mặc định.

Xuất bản web và tác vụ web

Cấu trúc XML :

1. được dựa trên các NODE lồng vào nhau . Mỗi NODE có đủ 1 thẻ đóng và mở . Ex : <NodeName> Content </NodeName>
2. file XML gồm 4 thành phần bao gồm :

* header file ( thẻ đầu tiên trong file XML nhằm khai báo version và chứa các thông tin về mã hóa . Ex : <?xml version =”1.0” encoding =”utf-8”?> . Ở đây , giá trị encoding ( kiểu mã hóa ký tự ) thuộc 1 trong những dạng : : UTF-8, UTF-16, ISO-10646-UCS-2, ISO-10646-UCS-4, ISO-2022-JP . Với utf-8 ( 8-bit Unicode Tranformation Format ) định dạng chuyển đổi Unicode 8-bit
* root node : là phần tử gốc , là phần tử ở bên ngoài cùng . Phần tử này thường là tên tài liệu hoặc đại diện cho thông tin được lưu trữ trong XML .
* element : là 1 dạng tên gọi nhưng sẽ dùng để chỉ các tệp con nằm trong Root node
* Attribute : là thuộc tính của thẻ , được dùng để phân biệt đặc tính giữa 2 element trùng tên

Ex : <?xml version =’1.0’ encoding =’utf-8’?> //đây là header file

<employees> //root node nè

<employee gender =”male”> // employee là element và gender là attribute

<Id> ..... </Id> // đây là child element

</employee>

</employees>

Thư viện hỗ trợ XML trong python

1. Thư viện BeautifulSoup hỗ trợ lxml giúp đọc file xml
2. Module ElementTree cung cấp hàm ElementTree.parse() để bắt đầu phân tích cú pháp của file XML
3. Import xml.dom.minidom để sử dụng module minidom hỗ trợ hàm parse() để đọc file XML

HTML ( Hyper Text Markup language) ngôn ngữ đánh dấu siêu văn bản với những ưu điểm như :

1. Là ngông ngữ được sử dụng rộng lớn có rất nhiều tài nguyên hỗ trợ và cộng đồng sử dụng lớn
2. Markup gọn gàng và đồng nhất
3. Chuẩn chính của web được vận hành bởi W3C
4. Dễ dàng tích hợp với các ngôn ngữ back-end như PHP và Node.js

Từ những ưu điểm trên , ứng dụng và vai trò của HTML rất quan trọng . HTML được sử dụng để thiết kế Web vì HTML là ngôn ngữ duy nhất mà trình duyệt có thể hiểu và biên dịch được

Cấu trúc cơ bản của HTML gồm các phần sau :

1. <!Doctype> : là phần khai báo chuẩn của HTML hay XHTML
2. <html> ... </html> : Nội dung của 1 tập tin HTML phải được bắt đầu bằng thẻ <html> và kết thúc bằng thẻ </html> . Bên trong cặp thẻ <html> </html> được chia làm 2 thẻ chính là :

* <head> </head> : khai báo về meta , title , css , js , ...
* <body> </body> : nơi hiện thị nội dung .

Ex :

<!DOCTYPE html>   
<html>  
 <head>  
 <title>Tiêu đề trang web</title>  
 </head>  
 <body>  
 ...Phần thân viết ở đây...  
 </body>  
</html>

Thư viện hỗ trợ HTML trong python :

1 ) BeautifulSoup và requests