Commercio Elettronico

Web Crawling

Claudio Silvestri

HTML: un rapido riassunto

Markup language

<html>

<head> ….</head>

<body> HELLO WORLD</body>

</html>

HTML: alcuni tag

	paragrafo
	intestazione di livello 1 (2,3,4,...)
	contenitore ‘a blocco’
	contenitore ‘inline’
	link

Ogni tag può essere associato a degli attributi e ad uno stile.

Estrazione dati

Alcune possibilità:

- Script
- Espressioni regolari
- XPath

Espressioni regolari (1)

Es: codice=(\d+);

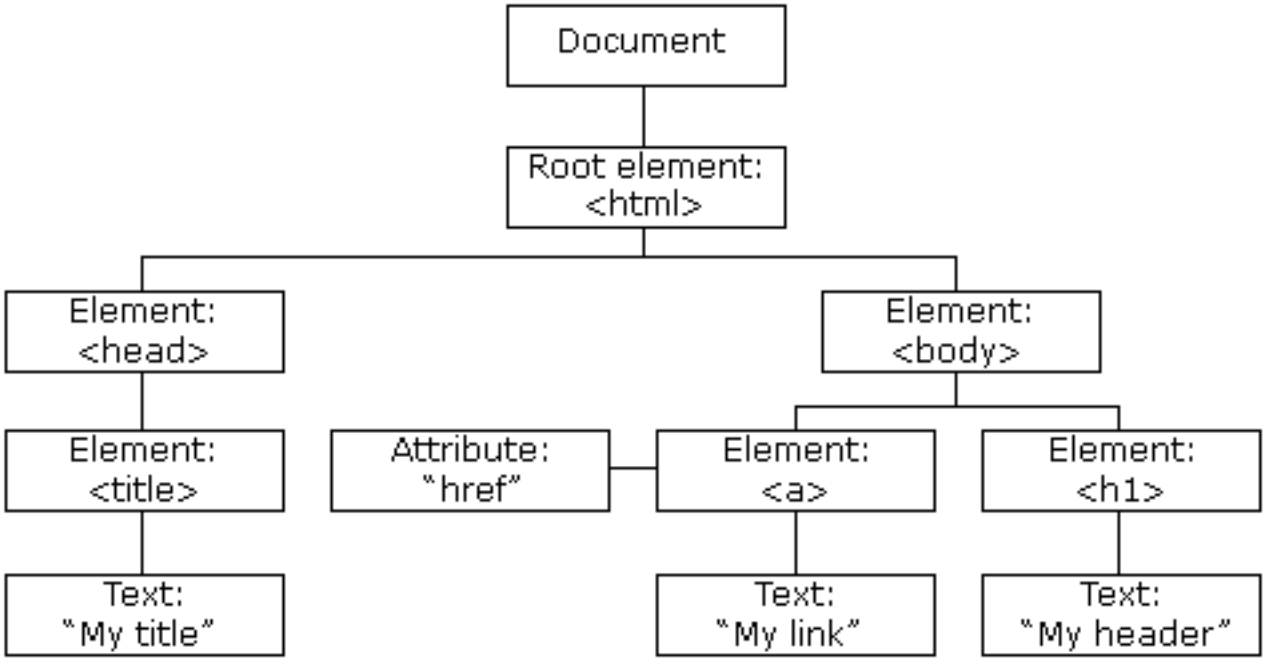
- Identifica un codice composto da sole cifre (0-9) e di lunghezza maggiore o uguale a 1 da ‘codice=‘ e seguito da ‘;’
- Possono essere ‘ancorate’ alla fine ($) o inizio del testo (^)
- Gli spazi sono caratteri come tutti gli altri (ad esempio ‘codice = 9;’ non è riconosciuto)

Espressioni regolari (2)

Poco adatte a identificare dati basandosi sulla struttura della pagina

PERCHE'?

Document Object Model - DOM



XPath

Esempi:

/bookstore/book[1]/title

/bookstore/book/price/text()

/bookstore/book[price>35]/title/text()

Sintassi:

http://www.w3schools.com/xpath/xpath_syntax.asp