

Hyper-Text Query Language Python Interface

2011/03/24

1. About

Hyper-Text Query Language (HTQL) is a language for the query and extraction HTML data. This guide explains the use of HTQL COM interface for use in Python applications. HTQL can be used to:

- 1) Extract content from HTML pages
- 2) Retrieve HTML page through HTTP protocol
- 3) Modify HTML pages

HTQL syntax can be found at <http://htql.net/htql-manual.pdf>.

2. Installation

Download the [htql.zip \(v2.7\)](#) or [htql.zip \(v3.2\)](#) and extract the htql.pyd to Python's DLLs directory, such as in 'C:\Python27\DLLs\' or 'C:\Python32\DLLs\'.

3. Simple Examples

A simple example to extract url and text from links.

```
import htql;
page="<a href=a.html>1</a><a href=b.html>2</a><a href=c.html>3</a>";
query="<a>:href,tx";

for url, text in htql.HTQL(page, query):
    print(url, text);
```

An example using htql.Browser to bing search and get the first link page:

```
import htql;
a=htql.Browser();
b=a.goUrl("http://www.bing.com/");
c=a.goForm("<form>1", {"q":"test"});
for d in htql.HTQL(c[0], "<a (tx like '%test%')>"):
    print(d);
e=a.click("<a (tx like '%test%' and not (href like '/search%'))>1");
```

4. HTQL Python Interface

htql.HTQL(page, query) returns tuple iterator

Query the page with HTQL

htql.Browser(type) returns *htql.Browser*

Create an HTQL browser. When type==1, returns a socket browser; when type==2, returns an IRobot browser

htql.Browser.goUrl(url, {name:value}, {cookie:value}, {http_header:value}) returns (page, url)

Go to URL page.

htql.Browser.goForm(form_htql, {name:value}, {cookie:value}, {http_header:value}) returns (page, url)

Submit a form.

htql.Browser.click(item_htql, wait) returns (page, url)

Click an item.

htql.Browser.getPage() returns (page, url)

Get the current page in browser.

htql.Browser.getUpdatedPage() returns (page, url)

Get the current page in browser after javascript has executed.

htql.Browser.runCommand(command) returns (page, url)

Run a function command. Refer IRobot manual for available functions.

htql.Browser.setTimeout(connect, transfer)

Set browser timeout for connect and transfer, in seconds.

5. Reference

HTQL was designed and created by Dr. Liangyou Chen as part of his Ph.D. dissertation work:

Ad Hoc Integration and Querying of Heterogeneous Online Distributed Databases. 2004. Mississippi State University.

Please cite the dissertation for references.