C3 Page Event API token example 38jq 1.0

[default-match]
@m1=/auto_scrollnews/

[default-pattern]
@aa=/\>\<span\>(?P<title>.+)\<\/span\>\<\/h3\>/

[default-classify]
@c1=20160326-155548-fbd3

[event-match]
@m2=/C3EVENT/
[event-pattern]
@aa2=/\<li\>\<span\>\((?P<key>.+)\)\<\/span\>\<a
href=\"(?P<link>.+)\"\>(?P<linktitle>.+)\<\/a\>\<\/li\>/

[DATE-format]
@date=/(\d{4})-(\d{2})-(\d{2}) (\d{2}):(\d{2})/
@format=\1-\2-\3 \4:\5

[page]
@index=http://auto.163.com/special/00083TL0/auto_scrollnews.html

API 说明

API c3_http_form 模拟单个 HTTP FORM POST 操作
API c3_http_form_session 用于生成 login 后的 cookie 数据，用于后续 session 方式的访问
参数：在 BPA 界面的 payload 输入 form=j3k2， 其中 form 后面的字串为 token,可从
https://api.conby.com/ 生成

详细格式如下:

[action]
@url=http://www.abc.com/hello.html 表示启始的 url
@method=POST 表示 HTTP GET 或 POST
@encoding=y 表示将 data 中数据转为 utf-8
@json=y 表示返回的数据格式，存在则返回 json 格式，否则按行按本 INI 格式返回数据(无 section 部分)

[data] 表示 HTTP POST 的数据
@aa=message

@bb=message

[header] 表示 HTTP POST 的头信息
@aa=message
@bb=message

[cookie] 表示 HTTP POST 的 cookie 信息
@uid=message
@pwd=message

API c3_page 对一个或几个 URL 进行爬虫操作
参数：在 BPA 界面的 payload 输入 page=j3k2&w=200&d=5&t=2
page 后面的字串为 token,可从 https://api.conby.com/ 生成
w 为爬虫的宽度，默认 200
d 为爬虫的深度度，默认 5
t 为爬虫的线程并发数，默认无，即单线程

page token 格式如下:

[default-match] 默认的
@m1=/auto_scrollnews/ 正则表达式
@m2=/auto_scrollnews/ 正则表达式

[default-pattern] 默认的
@aa=/\>\<span\>(?P<title>.+)\<\/span\>\<\/h3\>/ 正则表达式，含别名变量定义是需要输出的数据
@aa2=/\>\<span\>(?P<title>.+)\<\/span\>\<\/h3\>/ 正则表达式，含别名变量定义是需要输出的数据

[default-classify] 默认的基于 GPU 深度学习人工智能识别标识
@c1=20160326-155548-fbd3

[event-match]
@m3=/C3EVENT/

[event-pattern] 默认的正则表达式提取规则
@aa2=/\<li\>\<span\>\((?P<key>.+)\)\<\/span\>\<a
href=\"(?P<link>.+)\"\>(?P<linktitle>.+)\<\/a\>\<\/li\>/
@aa3=/\<li\>\<span\>\((?P<key>.+)\)\<\/span\>\<a
href=\"(?P<link>.+)\"\>(?P<linktitle>.+)\<\/a\>\<\/li\>/

[DATE-format]
@date=/(\d{4})-(\d{2})-(\d{2}) (\d{2}):(\d{2})/
@format=\1-\2-\3 \4:\5

[page]
@index=http://auto.163.com/special/00083TL0/auto_scrollnews.html
@form=j3k2 (比上面的@index 优先，覆盖@index，格式按 c3_http_form 定义
@mypage1=http://www.abc.com/hello.html 自定义 URL
@mypage2=http://www.abc.com/hello.html 自定义 URL

说明：
1 输出的 json 格式可参考： https://api.conby.com/c3/c3_page_json.js
2 系统自动对每个 page 进行生成 C3HREF C3EVENT 事件
3 pattern 如果为 C3_PATTERN，则自动提交到 C3 平台的 pattern 库自动查询
4 -match 部分和-pattern 部分为成对的定义，缺一无效
5 正则表达式的定义需要尽量规范，这些表达式将在 Python,PHP 和 JAVA 开发的系统中使用，不规范定义可能导致问题
6 输出 JSON 数据在 HOOK 回调中，可以用 maillog 进行监控，或专门定义 HOOK TASK 实时监控