


Lab8

NYCU Go Programming 2024

2024/12/03

Colly(doc.)

 README.md

Colly

Lightning Fast and Elegant Scraping Framework for Gophers

Colly provides a clean interface to write any kind of crawler/scrapper/spider.

With Colly you can easily extract structured data from websites, which can be used for a wide range of applications, like data mining, data processing or archiving.

[GO](#) [reference](#) [backers](#) 7 [sponsors](#) 11 [CI](#) [passing](#) [report card](#) [a+](#) [learn by](#) [examples](#) [coverage](#) 54% [license scan](#) [passing](#)

[twitter](#) [follow](#)

推	LineageM	: 讚 禮拜一漲就是強 跌就是套	01/08 18:35
推	b23058179	: 39被甩下去QQ	01/08 18:35
噓	iDoDo	: 丸子	01/08 18:35
推	ntupean	: 樓下支援v7馬頭丁人	01/08 18:35
噓	ping860622	: 新聞可以不要那麼快出嗎	01/08 18:35
推	Coffeewater	: 38.3砍掉的舉個手	01/08 18:36

```
<span class="f2">...</span>
<div class="push">
  <span class="h1 push-tag">推 </span>
  <span class="f3 h1 push-userid">LineageM </span>
  <span class="f3 push-content">: 讚 禮拜一漲就是強 跌就是套</span>
  <span class="push-ipdatetime"> 01/08 18:35 </span>
</div>
<div class="push">...</div>
<div class="push">...</div>
<div class="push">...</div>
<div class="push">
  <span class="f1 h1 push-tag">噓 </span>
  <span class="f3 h1 push-userid">ping860622 </span>
  <span class="f3 push-content">: 新聞可以不要那麼快出嗎</span>
  <span class="push-ipdatetime"> 01/08 18:35 </span>
</div>
```

goquery (doc.)

CSS Selector

⋮ README.md

goquery - a little like that j-thing, only in Go

test **passing** reference * used by **1.7k projects**

goquery brings a syntax and a set of features similar to [jQuery](#) to the [Go language](#). It is based on Go's [net/html package](#) and the CSS Selector library [cascadia](#). Since the net/html parser returns nodes, and not a full-featured DOM tree, jQuery's stateful manipulation functions (like `height()`, `css()`, `detach()`) have been left off.

Also, because the net/html parser requires UTF-8 encoding, so does goquery: it is the caller's responsibility to ensure that the source document provides UTF-8 encoded HTML. See the [wiki](#) for various options to do this.

Syntax-wise, it is as close as possible to jQuery, with the same function names when possible, and that warm and fuzzy chainable interface. jQuery being the ultra-popular library that it is, I felt that writing a similar HTML-manipulating library was better to follow its API than to start anew (in the same spirit as Go's `fmt` package), even though some of its methods are less than intuitive (looking at you, [index\(\)](#)...).

select an element in the page to inspect it

F12

The image shows a web browser window displaying a forum post on a stock market discussion board. The forum post is in Chinese and discusses stock market movements. A specific line of text in the forum post is highlighted with a green box: "讚 禮拜一漲就是強 跌就是套".

Overlaid on the right side of the browser window is the Chrome DevTools interface. The "Elements" panel is active, showing the DOM tree. The selected element is a `span` with the class `f3 push-content`. The text content of this element is "讚 禮拜一漲就是強 跌就是套". The DevTools interface also shows the "Styles" panel at the bottom.

flag

- Variables

- `var CommandLine = NewFlagSet(os.Args[0], ExitOnError)`
 - `CommandLine` is the default set of command-line flags, parsed from `os.Args`. The top-level functions such as `BoolVar`, `Arg`, and so on are wrappers for the methods of `CommandLine`.

- `func Int(name string, value int, usage string) *int`

- `Int` defines an `int` flag with specified name, default value, and usage string. The return value is the address of an `int` variable that stores the value of the flag.

- `func Parse()`

- `Parse` parses the command-line flags from `os.Args[1:]`. Must be called after all flags are defined and before flags are accessed by the program.

Lab8: Web Crawler

- 爬取這篇 PTT (<https://www.ptt.cc/bbs/joke/M.1481217639.A.4DF.html>) 的留言
- 只抓文字, 忽略圖片或影片
- 用 Flag 控制印出留言的數量
 - [-max]: 限制印出資料的數量, 預設為10。 E.g. -max 10

Lab8: Web Crawler

```
axelhowe@DESKTOP-85LD9SI:/mnt/c/Users/USER/Desktop/312552019-Go-2024/lab8$ go run lab8.go
```

1. 名字 : Ommmmmm5566, 留言: 哈哈哈哈哈厂厂厂厂哈哈哈哈哈, 時間: 12/09 03:18
2. 名字 : Ommmmmm5566, 留言: 笑死, 時間: 12/09 03:18
3. 名字 : husky01, 留言: 樓上..., 時間: 12/09 03:27
4. 名字 : bakapika, 留言: 厂厂, 時間: 12/09 06:25
5. 名字 : scmdwyam, 留言: XDD, 時間: 12/09 08:48
6. 名字 : L2e2o4, 留言: 厂厂, 時間: 12/09 09:21
7. 名字 : NobleDino, 留言: 厂厂, 時間: 12/09 09:34
8. 名字 : tottoko0908, 留言: 厂厂, 時間: 12/09 09:45
9. 名字 : lmh911152, 留言: 厂厂, 時間: 12/09 09:56
10. 名字 : monicamomo, 留言: 厂厂, 時間: 12/09 10:25

```
axelhowe@DESKTOP-85LD9SI:/mnt/c/Users/USER/Desktop/312552019-Go-2024/lab8$ go run lab8.go -max 4
```

1. 名字 : Ommmmmm5566, 留言: 哈哈哈哈哈厂厂厂厂哈哈哈哈哈, 時間: 12/09 03:18
2. 名字 : Ommmmmm5566, 留言: 笑死, 時間: 12/09 03:18
3. 名字 : husky01, 留言: 樓上..., 時間: 12/09 03:27
4. 名字 : bakapika, 留言: 厂厂, 時間: 12/09 06:25

Lab8: Web Crawler

- 如果程式輸入沒有定義的 flag, 或是錯誤的 flag, 要印出 Usage
- “Usage of” 後面的 path 不用特別處理(testing 不檢查)

```
axelhowe@DESKTOP-85LD9SI:/mnt/c/Users/USER/Desktop/312552019-Go-2024/Lab8$ go run . -hello
flag provided but not defined: -hello
Usage of /tmp/go-build2921271239/b001/exe/lab8:
  -max int
        Max number of comments to show (default 10)
exit status 2
axelhowe@DESKTOP-85LD9SI:/mnt/c/Users/USER/Desktop/312552019-Go-2024/Lab8$ go run . -max
flag needs an argument: -max
Usage of /tmp/go-build2698444471/b001/exe/lab8:
  -max int
        Max number of comments to show (default 10)
exit status 2
axelhowe@DESKTOP-85LD9SI:/mnt/c/Users/USER/Desktop/312552019-Go-2024/Lab8$ go run . -max abcd
invalid value "abcd" for flag -max: parse error
Usage of /tmp/go-build1402000172/b001/exe/lab8:
  -max int
        Max number of comments to show (default 10)
exit status 2
axelhowe@DESKTOP-85LD9SI:/mnt/c/Users/USER/Desktop/312552019-Go-2024/Lab8$
```

Lab8: Web Crawler

- .github
 - workflows
 - lab8.yml
- Lab8
 - go.mod
 - go.sum
 - lab8.go
 - validate.py

```
$ python validate.py
```