

Large Scale Crawling & Scraping Best Practices

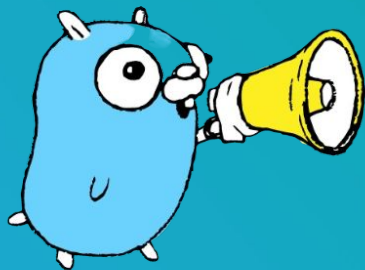
Kevser Sırça Alkış
Devops Engineer at seo.do



kevsersrca



kev_src



Large Scale Crawling **01**

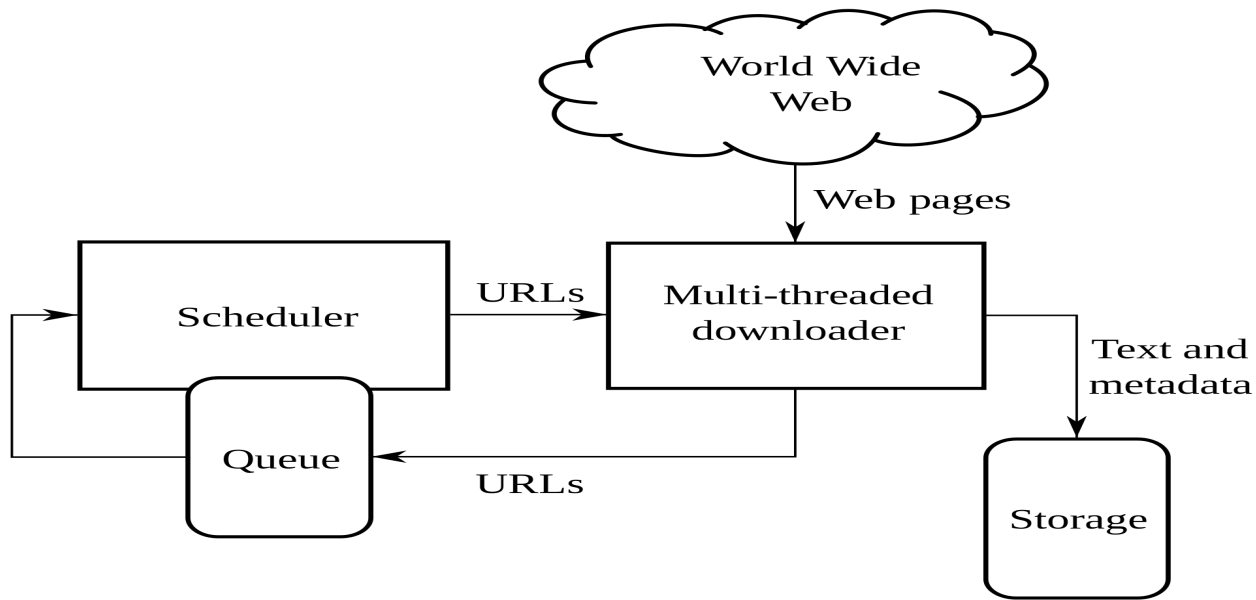
Scraping Best Practices **02**

Golang Scraping Practices **03**

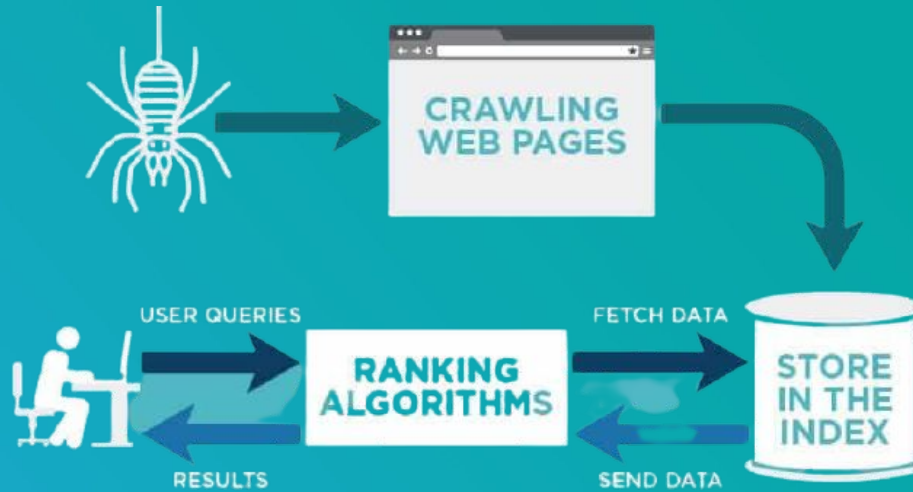
Conclusion **04**

Crawling

about discovering URLs or links



How search engines work?



Robots.txt

```
User-agent: Googlebot
Disallow: /print/
Disallow: /listing/*/print
Disallow: /ilan/*/yazdir
Disallow: /api/
Disallow: /req/
Disallow: /m/req/
Disallow: /de/
Disallow: /ru/
```

```
User-agent: Yandex
Disallow: /
```

User Agents

Mozilla/5.0 (iPad; U; CPU OS 3_2_1 like Mac OS X; en-us) AppleWebKit/531.21.10 (KHTML, like Gecko) Mobile/7B405

Indicates compatibility
with the Mozilla
rendering engine

Details of the
system in which the
browser is running

The platform the
browser uses

Browser platform
details

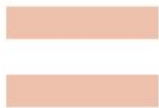
Additional details
specific to the browser

Scraping



Scraping

Art of collecting data
from the web



Parsing

Page analysis and
data extraction



Crawling

Navigating from pages
to pages on the web.

Large Scale Crawling - Proxy

Shared Proxies



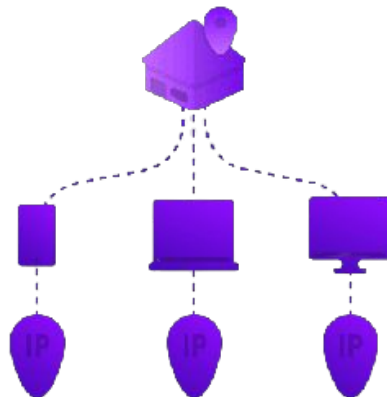
Private Proxies

- Dedicated access
- Unshared bandwidth
- Fast Speed
- Security

Large Scale Crawling - Proxy

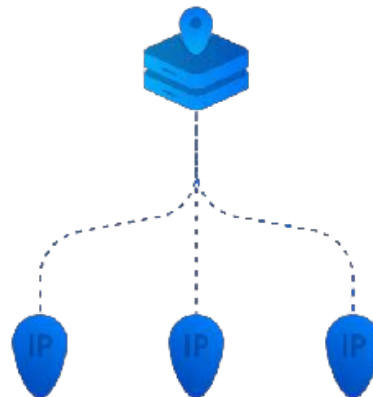


Residential Proxies



Residential Proxies

Datacenter Proxies



Datacenter Proxies

Large Scale Crawling - Proxy



The World's #1 Web Data Platform

From data collection infrastructure to ready-made datasets, Bright Data allows you to retrieve the web data you care about.

[Start now >](#)

[Request a demo >](#)

Data collection

Proxy

Resources

Pricing

[Sign in](#)

[Sign up](#)

Proxy Infrastructure

Utilize the biggest IP Network in the world



Proxy Manager

Manage all proxies using one open-source interface

[Learn more >](#)



Data Center Proxies

700,000+ shared data-center IPs from any geolocation

[Learn more >](#)



Residential Proxies

72 million+ IPs rotated from real-peer devices in 195 countries

[Learn more >](#)



ISP Proxies

160,000+ real home IPs across the globe, for long-term use

[Learn more >](#)



Mobile Proxies

7 million+ IPs forming the largest real-peer 3G/4G mobile network

[Learn more >](#)



Proxy
Infrastructure

Large Scale Crawling - Proxy

Mobile Proxies - litport.net

[Pricing ▾](#)[Locations ▾](#)[Resources ▾](#)[Use Cases ▾](#)[🔥 News](#)[Log in](#)

Buy residential mobile proxy

Currently we offer mobile proxies from 15 countries. The cheapest plan is \$2.99 per day.

Our huge pool of fresh mobile devices is more than 55+ million unique IPs. Rotation time varies from 2 minutes to 12 hours or manually by API.

Click on the country to see full list of available packages and pricing.

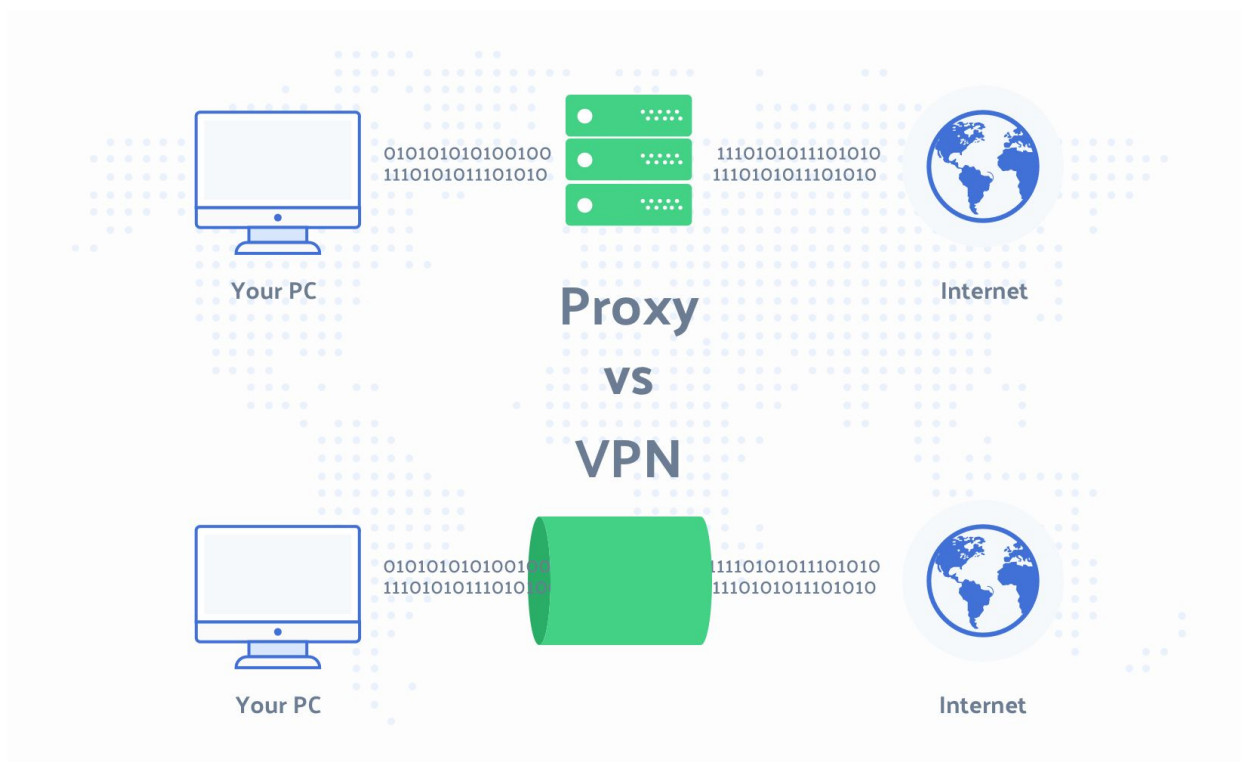


Our process is simple and easy. Get your proxy in 60 seconds.

For most of our proxy packages we have automated system that will create an account and issue a new proxy for you. Typically we process your order under 60 seconds. Yes, this is that easy and fast.



Large Scale Crawling - VPN



Large Scale Crawling - Data warehouse



If you are parsing small volumes, a simple spreadsheet or text file may be enough

If you are parsing large-scale volumes, use databases

- Oracle, MySQL, MongoDB

If you want more effective and fast,

- In-memory database (Redis, Memcached)

Large Scale Crawling - Extra Tips



- Save the URLs to queue
- Don't use panic!
- Cache
- Logs
- Keep websites from overloading

Is Web Scraping legal?

Web scraping and crawling aren't illegal.

- Don't send too many requests at the same time.
- Look at crawl-delay.
- Pretend to be a real user

LinkedIn sues anonymous data scrapers (2015)

Source: <https://techcrunch.com/2016/08/15/linkedin-sues-scrapers/>

Preventing Web Scraping



- Rate Limit Individual IP Addresses
- Web Application Firewall (WAF)
- Require a Login for Access
- Change Your Website's HTML Regularly
- Use CAPTCHAs When Necessary
- Create “Honeypot” Pages

Scraping Best Practices



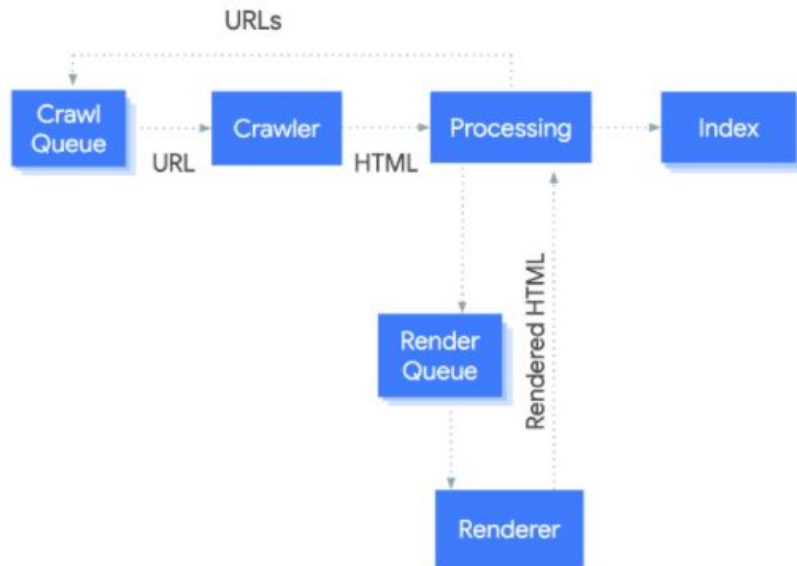
Data management

Data accuracy is the number one challenge when dealing with parsing a thousand pages per day

- Set Requirements
- Define the testing criteria
- Start testing

Scraping Best Practices

Dynamic Rendering Queue



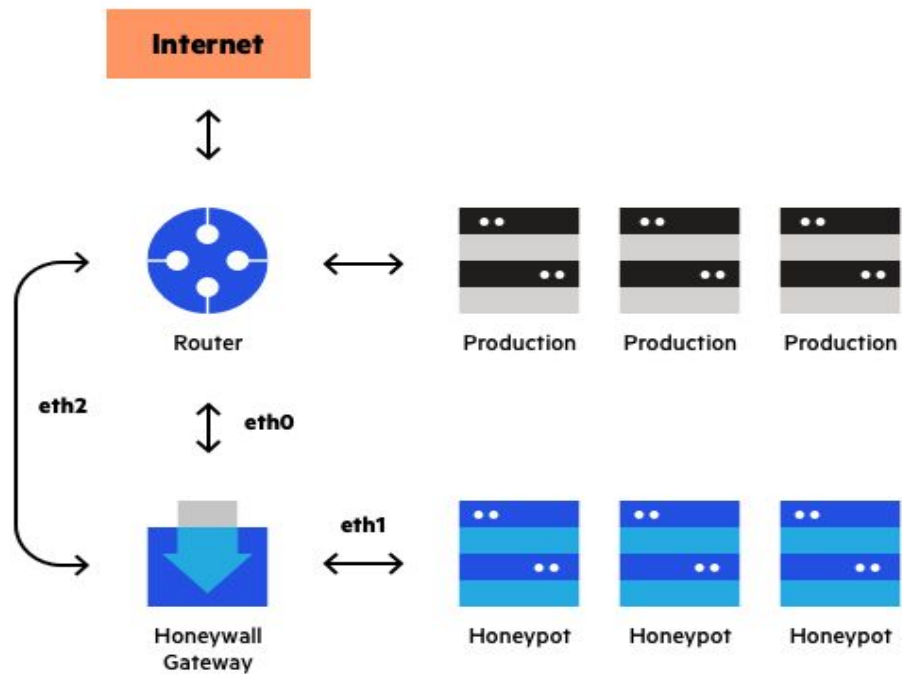
Scraping Best Practices

Captchas



Scraping Best Practices

Avoid Honeypot Traps



Scraping Best Practices

IP blocking

```
var proxies []string = []string{
    "http://207.154.231.208:8080",
    "http://138.68.230.88:8080",
}

func GetProxy() (*url.URL, error) {
    randomIndex := rand.Intn(len(proxies))
    randomProxy := proxies[randomIndex]
    return url.Parse(randomProxy)
}

func main() {
    proxyURL, err := GetProxy()
    transport := &http.Transport{Proxy: http.ProxyURL(proxyURL)}
    // Continue with your HTTP requests ...
}
```

Scraping Best Practices



Set Additional Request Headers

```
req, err := http.NewRequest("GET", "http://example.com", nil)

req.Header.Add("Accept",
`text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,
*/*;q=0.8`)

req.Header.Add("Accept-Encoding", `gzip, deflate`)

req.Header.Add("Accept-Language", `en-US,en;q=0.9`)

resp, err := client.Do(req)
```

Web Scraping Etiquettes

How to throttle your scraper?

As for scraping multiple pages from the same website, you should first follow the Crawl-Delay in a robots.txt file.

If there is no Crawl-Delay specified, then you should manually delay your requests by one second after every page.

```
import (  
    "fmt"  
    "time"  
  
    "go.uber.org/ratelimit"  
)  
  
func main() {  
    rl := ratelimit.New(100)  
  
    prev := time.Now()  
    for i := 0; i < 5; i++ {  
        now := rl.Take()  
        fmt.Println(i, now.Sub(prev))  
        prev = now  
    }  
  
    // Output:  
    // 0 0  
    // 1 10ms  
    // 2 10ms  
    // 3 10ms  
    // 4 10ms  
}
```

Web Scraping Etiquettes

How to use caching?

Cache Control

- private
- public
- no-store
- no-cache
- max-age

Last-Modified

- If-Modified-Since

Etag

- If-None-Match

HTTP/1.1 200 OK

Age: 409842

Cache-Control: max-age=604800

Content-Type: text/html;
charset=UTF-8

Date: Tue, 14 Dec 2021 16:26:49 GMT

Etag: "3147526947+gzip"

Expires: Tue, 21 Dec 2021 16:26:49
GMT

Last-Modified: Thu, 17 Oct 2019
07:18:26 GMT


Server: ECS (dcb/7F83)

Vary: Accept-Encoding

Web Scraping Etiquettes

How to use caching?

github.com/gregjones/httpcache













gregjones Update README adding project status (#100)

✓

901d907 on Jun 11, 2019

🕒

108 commits

 diskcache	test: Add a helper for testing cache implementations (#91)	3 years ago
 leveledbcache	test: Add a helper for testing cache implementations (#91)	3 years ago
 memcache	test: Add a helper for testing cache implementations (#91)	3 years ago
 redis	test: Add a helper for testing cache implementations (#91)	3 years ago
 test	test: Add a helper for testing cache implementations (#91)	3 years ago
 .travis.yml	.travis.yml: bump to Go 1.11, use 'gofmt -s' (#92)	3 years ago
 LICENSE.txt	oh yeah, license.	9 years ago
 <div>LICENSE.txt</div> README	Update README adding project status (#100)	3 years ago
 httpcache.go	.travis.yml: bump to Go 1.11, use 'gofmt -s' (#92)	3 years ago
 httpcache_test.go	Refrain from setting 200 OK on cached responses (#77)	4 years ago

Web Scraping Etiquettes



How to use caching?

```
import (  
    "github.com/gregjones/httpcache"  
    "github.com/gregjones/httpcache/diskcache"  
)  
  
// Set up the local disk cache  
storage := diskcache.New("./cache")  
cache := httpcache.NewTransport(storage)  
  
// Set this to true to inform us if the responses are being read from  
a cache  
cache.MarkCachedResponses = true  
cachedClient := cache.Client()  
  
// Make the initial request  
fmt.Println("Caching: http://www.example.com/index.html")  
resp, err := cachedClient.Get("http://www.example.com/index.html")
```

Web Scraping Etiquettes

How to use caching?

```
// httpcache requires you to read the body in order to cache the response
```

```
ioutil.ReadAll(resp.Body)
```

```
resp.Body.Close()
```

```
// Request index.html again
```

```
fmt.Println("Requesting: http://www.example.com/index.html")
```

```
resp, err = cachedClient.Get("http://www.example.com/index.html")
```

```
if err != nil {
```

```
    panic(err)
```

```
}
```

```
// Look for the flag added by httpcache to show the result is read from the
```

```
_, ok := resp.Header["X-From-Cache"]
```

```
if ok {
```

```
    fmt.Println("Result was pulled from the cache!")
```

```
}
```

Web Scraping Etiquettes



How to use caching?

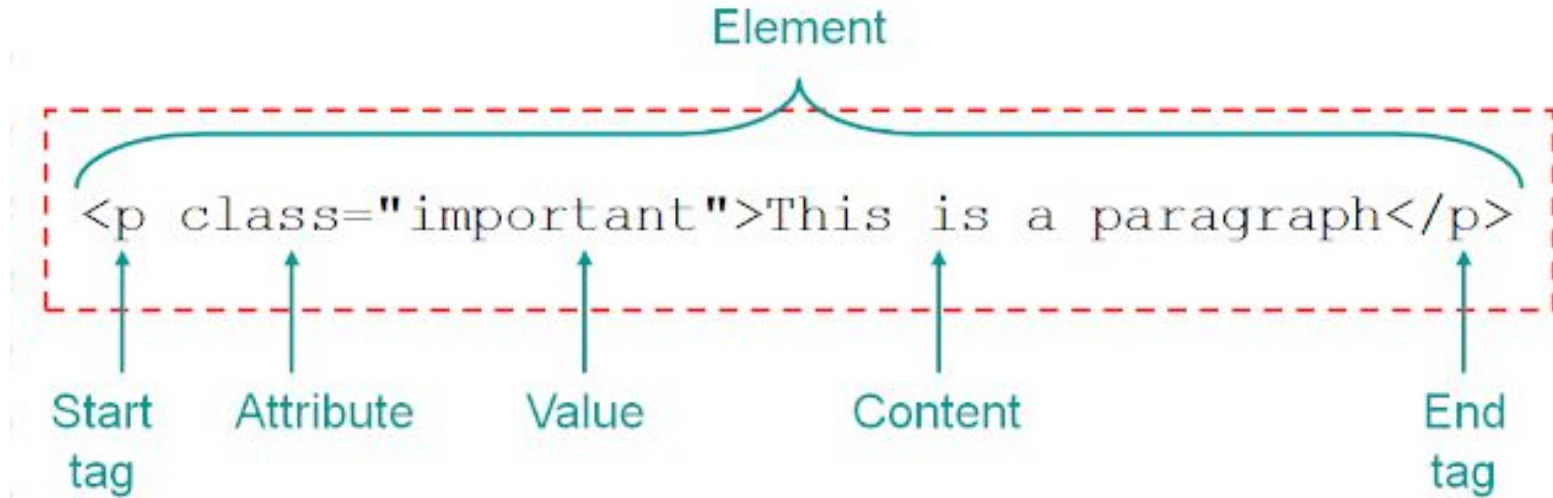
```
X go run main.go
```

```
Caching: http://www.example.com/index.html
```

```
Requesting: http://www.example.com/index.html
```

```
Result was pulled from the cache!
```

Parsing HTML



Class ".important"
ID "#important"

Parsing HTML (strings)



```
func main() {  
    resp, err := http.Get("https://www.amazon.com")  
    if err != nil {  
        panic(err)  
    }  
    data, err := ioutil.ReadAll(resp.Body)  
    if err != nil {  
        panic(err)  
    }  
    stringBody := string(data)  
    numLinks := strings.Count(stringBody, "<a")  
    fmt.Printf("Amazon homepage has %d links!\n", numLinks)  
  
    isTitles := strings.Contains(stringBody, "<title")  
    if isTitles {  
        fmt.Printf("Amazon homepage has title")  
    }  
}
```

Parsing HTML (golang.org/x/net/html)

```
z := html.NewTokenizer(response.Body)

for {
    tt := z.Next()
    switch {
    case tt == html.ErrorToken:
        return
    case tt == html.StartTagToken:
        t := z.Token()

        isAnchor := t.Data == "a"
        if isAnchor {
            for _, a := range t.Attr {
                if a.Key == "href" {
                    fmt.Println("Found href:", a.Val)
                    break
                }
            }
        }
    }
}
```

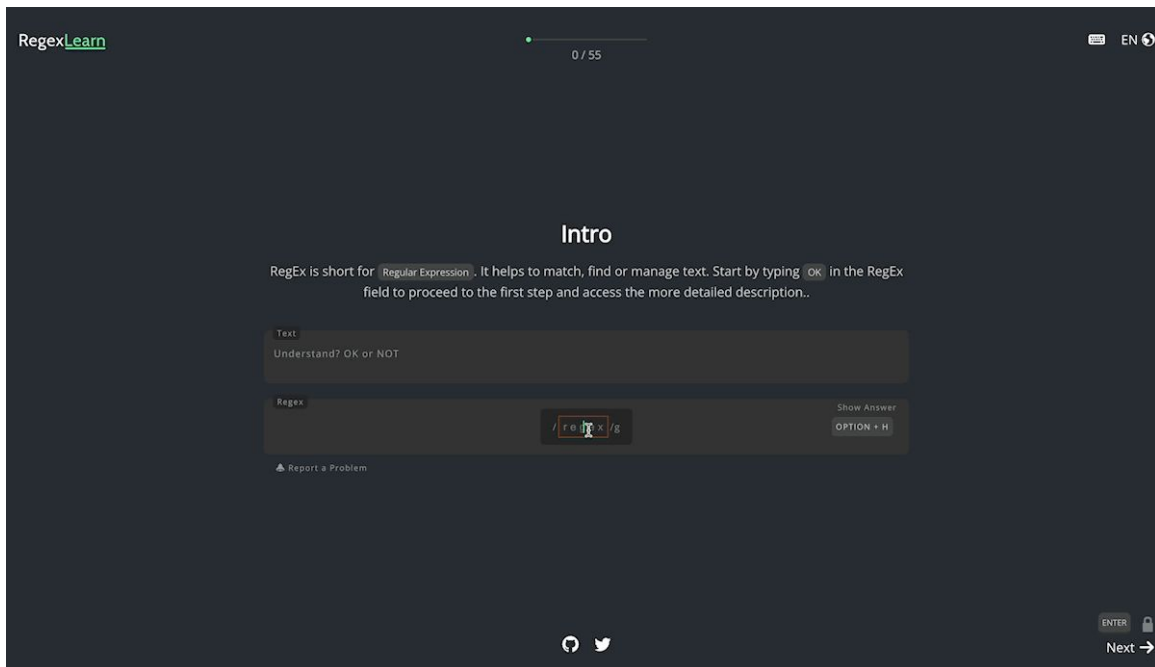
Parsing HTML (regexp)



```
re := regexp.MustCompile(`  
linkMatches := re.FindAllStringSubmatch(stringBody, -1)  
  
fmt.Printf("Found %d links:\n", len(linkMatches))  
  
for _, linkGroup := range(linkMatches){  
    fmt.Println(linkGroup)  
}
```


How to learn regex?

regexlearn.com



Parsing HTML (goquery)

"github.com/PuerkitoBio/goquery"

```
// Load the HTML document
doc, err := goquery.NewDocumentFromReader(res.Body)
if err != nil {
    log.Fatal(err)
}

// Find the review items
doc.Find("#search").Each(func(i int, s *goquery.Selection) {
    // For each item found, get the title
    title := s.Find("a").Text()
    fmt.Printf("Review %d: %s\n", i, title)
})
```

Parsing HTML (CSS Selector)

```
<table class="a-lineitem">
  <tbody>
    <tr>
      <td class="a-span9 a-text-left">
        <span class="a-size-base a-color-secondary"> Price </span>
      </td>
      <td class="a-span1 a-text-right"> </td>
      <td class="a-span2 a-text-right">
        <span class="a-size-base a-color-base" id="price"> $34.99 </span>
      </td>
    </tr>
  </tbody>
</table>
...

```

```
doc.find("table.a-lineitem>tbody>tr>td>span.a-color-base")
```

Parsing HTML (CSS Selector)

```
<table class="a-lineitem">
  <tbody>
    <tr>
      <td class="a-span9 a-text-left">
        <span class="a-size-base a-color-secondary"> Price </span>
      </td>
      <td class="a-span1 a-text-right"> </td>
      <td class="a-span2 a-text-right">
        <span class="a-size-base a-color-base" id="price"> $34.99 </span>
      </td>
    </tr>
  </tbody>
</table>
...

```

```
doc.find("table.a-lineitem>tbody>tr>td>span.a-color-base")
```

Parsing HTML (CSS Selector)

```
<table class="a-lineitem">
  <tbody>
    <tr>
      <td class="a-span9 a-text-left">
        <span class="a-size-base a-color-secondary"> Price </span>
      </td>
      <td class="a-span1 a-text-right"> </td>
      <td class="a-span2 a-text-right">
        <span class="a-size-base a-color-base" id="price"> $34.99 </span>
      </td>
    </tr>
  </tbody>
</table>

...
```

```
doc.find("table span.a-color-base")
```

Parsing HTML (CSS Selector)

```
<table class="a-lineitem">
  <tbody>
    <tr>
      <td class="a-span9 a-text-left">
        <span class="a-size-base a-color-secondary"> Price </span>
      </td>
      <td class="a-span1 a-text-right"> </td>
      <td class="a-span2 a-text-right">
        <span class="a-size-base a-color-base" id="price"> $34.99 </span>
      </td>
    </tr>
  </tbody>
</table>

...
```

```
doc.find("span#price")
```

Parsing HTML Extension



AMD Ryzen 7 3700X 8-Core, 16-Thread Unlocked Desktop Processor with Wraith Prism LED Cooler

[Visit the AMD Store](#)

★★★★★ 21,166 ratings | 466 answered questions

Amazon's Choice in Computer CPU Processors by AMD

List Price: ~~\$329.00~~ [Details](#)

Price: **\$299.99** + \$85.33 Shipping & Import Fees Deposit to Turkey [Details](#)

You Save: **\$29.01 (9%)**

Available at a lower price from [other sellers](#) that may not offer free Prime shipping.

Brand	AMD
CPU	AMD
Manufacturer	
CPU Model	AMD Ryzen 7
CPU Speed	4.4 GHz
Platform	Windows

About this item

- The world's most advanced processor in the desktop PC gaming segment

<https://github.com/teamseodo/muninn-extension>

Best Room Price Scraper from Booking.com



Search

Destination/property name:

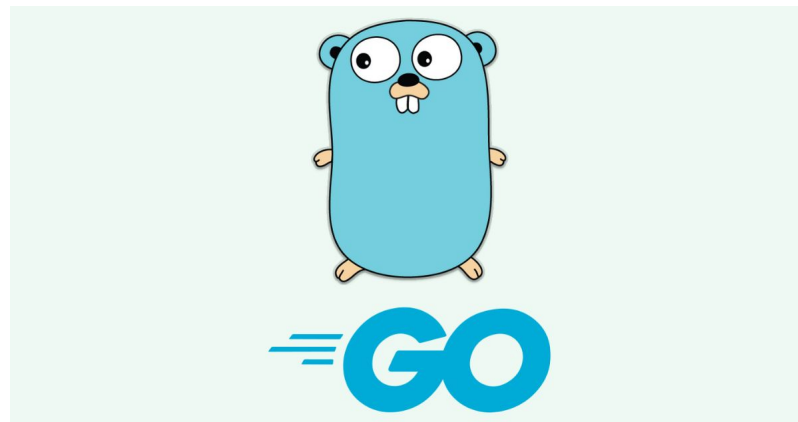
Check-in date

Check-out date

2-night stay

☐ Entire homes & apartments ?

☐ I'm travelling for work ?



—

9



 Travel Sustainable property

Breakfast inc

Breakfast included

FREE cancellation • No prepayment needed

You can cancel later, so lock in this great price today.

8.4

US\$276

Includes taxes and charges

[See availability >](#)

Only 1 room left at this price on our site

80

Location 9.3

US\$222

Includes taxes and charges

[See availability >](#)

Breakfast included

FREE cancellation • No prepayment needed

You can cancel later, so lock in this great price today.

Only 3 rooms left at this price on our site

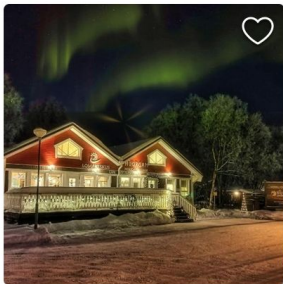
01

US\$249

Includes taxes and charges

[See availability >](#)

Best Room Price Scraper from Booking.com



Holiday Village Nuorgamin Lomakeskus 🏡

[Nuorgam](#) · [Show on map](#)

Double Room with Bathroom
2 single beds

Only 1 room left at this price on our site

Fabulous
326 reviews **8.9**

Location 9.3

2 nights, 2 adults

US\$222

Includes taxes and charges

[See availability](#) >

Set Requirements

- Hotel Name
- Room Type
- Review Score
- Price

```
type bookingRoom struct {  
    HotelName string  
    RoomType  string  
    Review    string  
    Price     string  
}
```

Testing Criteria

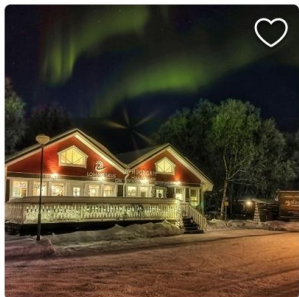
- Hotel Name
- Price

```
if room.HotelName != "" || room.Price != "" {  
    records = append(records, room)  
}
```

Best Room Price Scraper from Booking.com



div.fde444d7ef._c445487e2



Holiday Village Nuorgamin Lomakeskus 🏡

[Nuorgam](#) · [Show on map](#)

Double Room with Bathroom

2 single beds

Only 1 room left at this price on our site

Fabulous

326 reviews

8.9

Location 9.3

2 nights, 2 adults

US\$222

Includes taxes and charges

[See availability](#) >

div._9c5f726ff.bd528f9ea6

span._c5d12bf22

span.fde444d7ef._e885fdc12

Best Room Price Scraper from Booking.com



→ bookingroomscraper git:(main) X go run main.go

Scraping...

2021/12/17 23:54:59 Cheapest price is 136.00, Nikkilän Elämyskylä and Twin
Room. scraped from booking

github.com/kevsersrca/bookingroomscraper



THANK YOU!