THE DEVELOPER'S
CONFERENCE

Web scraping em GO!

Quais as vantagens em utilizar essa linguagem para coleta de dados na Web

Rafael Calixto

About me

THE DEVELOPER'S CONFERENCE

Licenciatura em Música

UFRJ

Análise e Desenvolvimento de Sistemas

UEZO

Ciência de Dados (Big Data)

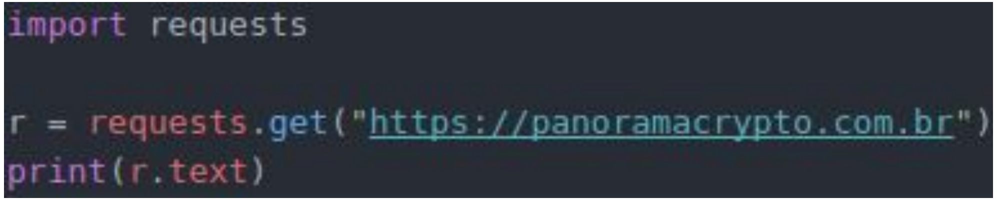
IGTI

Cientista de Dados

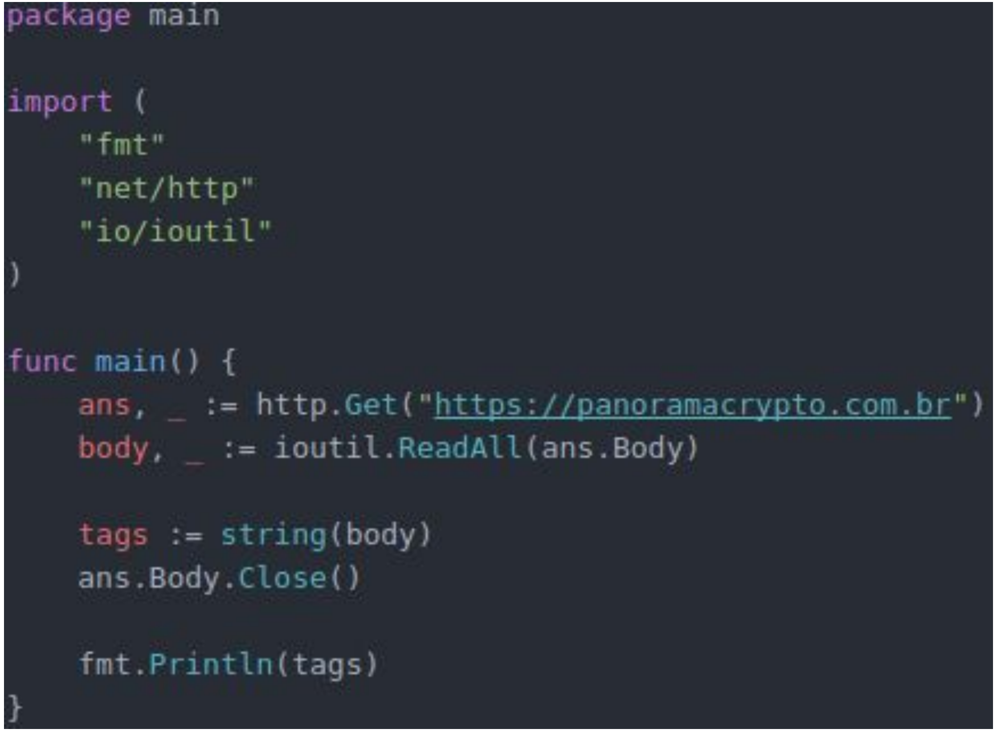
Wise&Trust

Redator

PanoramaCrypto



```python
import requests

r = requests.get("https://panoramacrypto.com.br")
print(r.text)
```

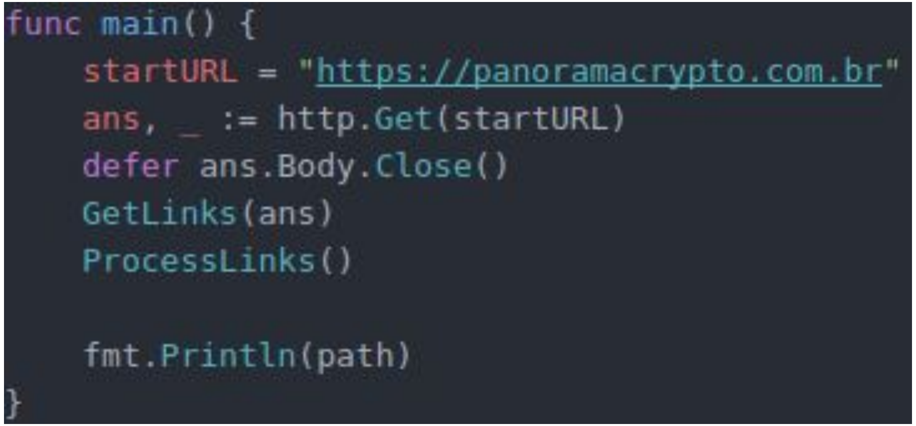


```go
package main

import (
    "fmt"
    "net/http"
    "io/ioutil"
)

func main() {
    ans, _ := http.Get("https://panoramacrypto.com.br")
    body, _ := ioutil.ReadAll(ans.Body)

    tags := string(body)
    ans.Body.Close()

    fmt.Println(tags)
}
```

Construindo um Spider

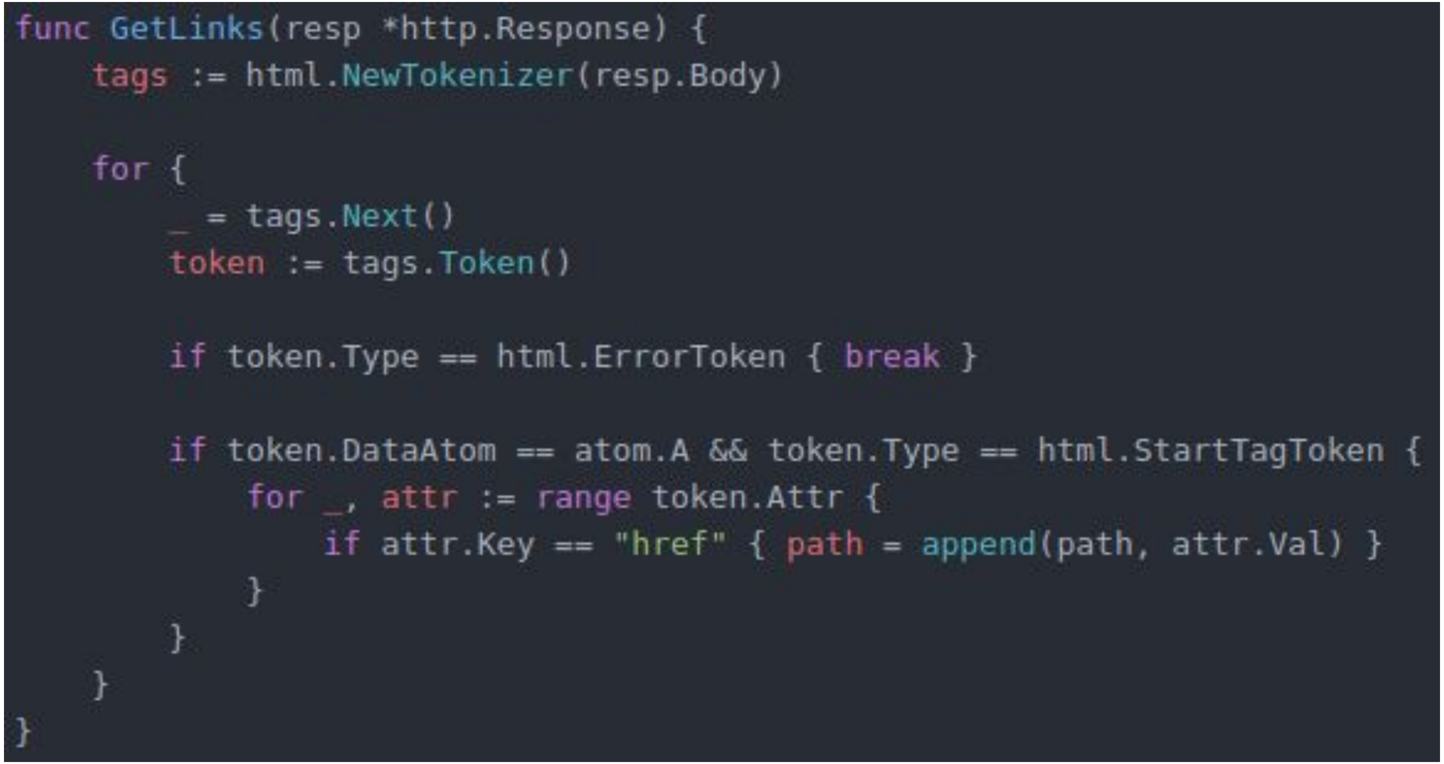
“net/http”

“golang.com/x/net/html”



```
func main() {
    startURL = "https://panoramacrypto.com.br"
    ans, _ := http.Get(startURL)
    defer ans.Body.Close()
    GetLinks(ans)
    ProcessLinks()

    fmt.Println(path)
}
```

Encontrando links

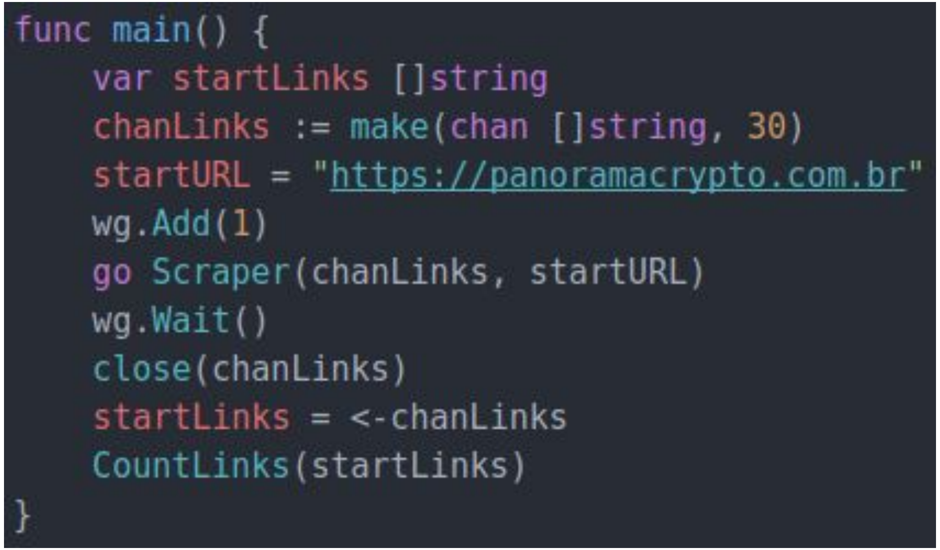


```go
func GetLinks(resp *http.Response) {
    tags := html.NewTokenizer(resp.Body)

    for {
        _ = tags.Next()
        token := tags.Token()

        if token.Type == html.ErrorToken { break }

        if token.DataAtom == atom.A && token.Type == html.StartTagToken {
            for _, attr := range token.Attr {
                if attr.Key == "href" { path = append(path, attr.Val) }
            }
        }
    }
}
```

Podemos ir além!!!

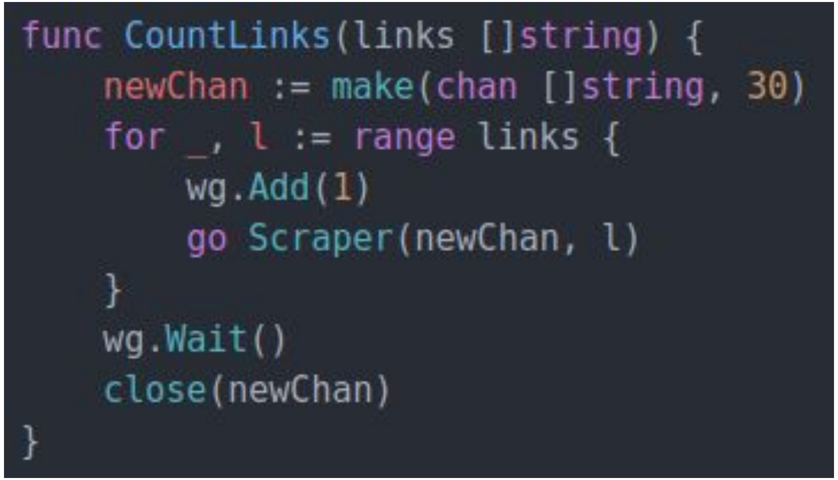
Channel

Chamadas assíncronas

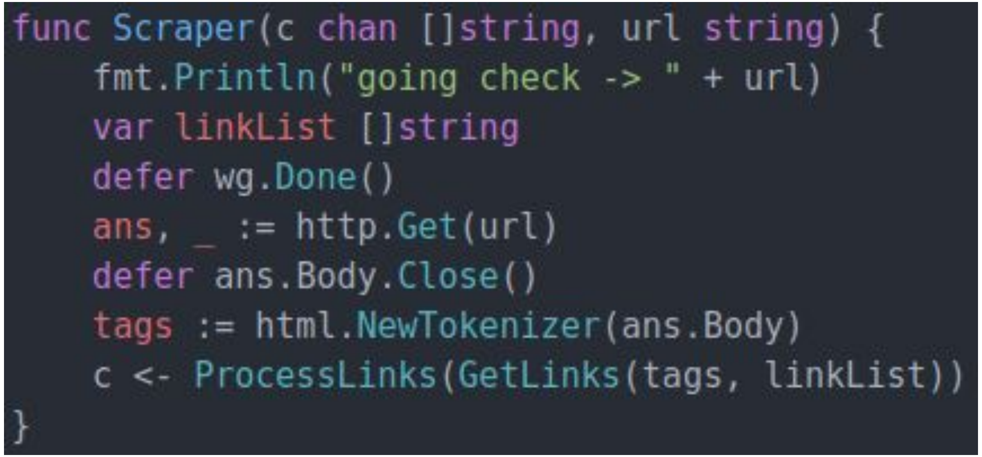


```
func main() {
    var startLinks []string
    chanLinks := make(chan []string, 30)
    startURL = "https://panoramacrypto.com.br"
    wg.Add(1)
    go Scraper(chanLinks, startURL)
    wg.Wait()
    close(chanLinks)
    startLinks = <-chanLinks
    CountLinks(startLinks)
}
```

Chamada Assíncrona



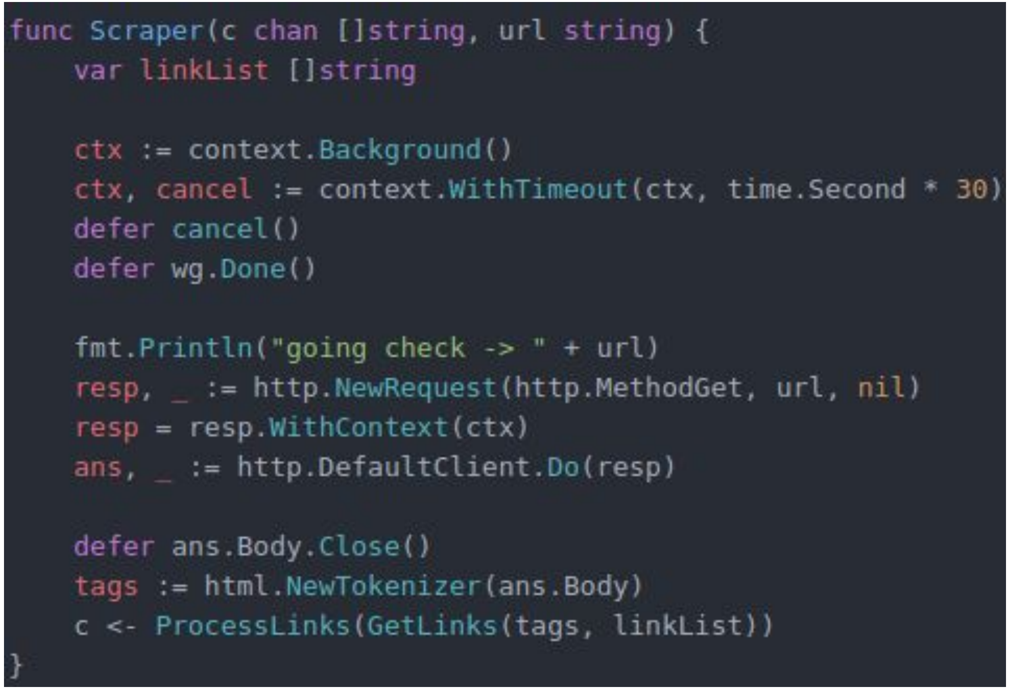
```go
func CountLinks(links []string) {
	newChan := make(chan []string, 30)
	for _, l := range links {
		wg.Add(1)
		go Scraper(newChan, l)
	}
	wg.Wait()
	close(newChan)
}
```



```go
func Scraper(c chan []string, url string) {
	fmt.Println("going check -> " + url)
	var linkList []string
	defer wg.Done()
	ans, _ := http.Get(url)
	defer ans.Body.Close()
	tags := html.NewTokenizer(ans.Body)
	c <- ProcessLinks(GetLinks(tags, linkList))
}
```

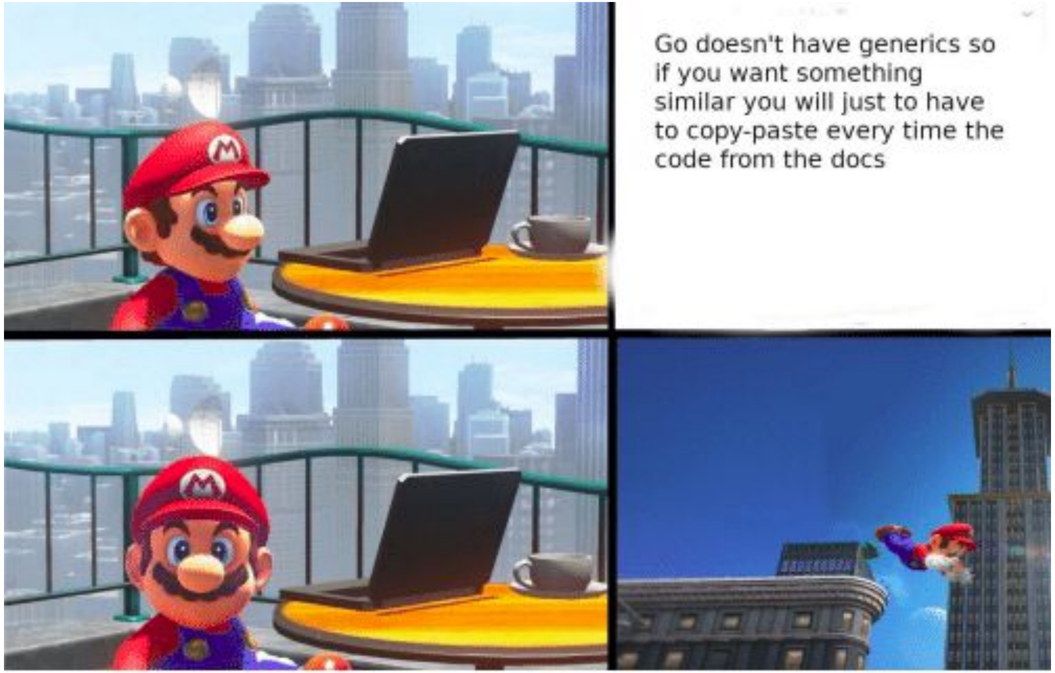
Usando Context

O Context permite cancelar uma operação desfazendo os passos anteriores



```
func Scraper(c chan []string, url string) {
    var linkList []string

    ctx := context.Background()
    ctx, cancel := context.WithTimeout(ctx, time.Second * 30)
    defer cancel()
    defer wg.Done()

    fmt.Println("going check -> " + url)
    resp, _ := http.NewRequest(http.MethodGet, url, nil)
    resp = resp.WithContext(ctx)
    ans, _ := http.DefaultClient.Do(resp)

    defer ans.Body.Close()
    tags := html.NewTokenizer(ans.Body)
    c <- ProcessLinks(GetLinks(tags, linkList))
}
```

Nem tudo são flores...



Go doesn't have generics so if you want something similar you will just to have to copy-paste every time the code from the docs

All good nothing wrong

THE DEVELOPER'S CONFERENCE

Perguntas?

Contatos

Twitter
@rafaelcalixtopy

LinkedIn
https://www.linkedin.com/in/rafael-calixto-9a11936b/

Github
https://github.com/rafaelcalixto

Telegram
@**rafaelcalixto**

THE DEVELOPER’S
CONFERENCE