1

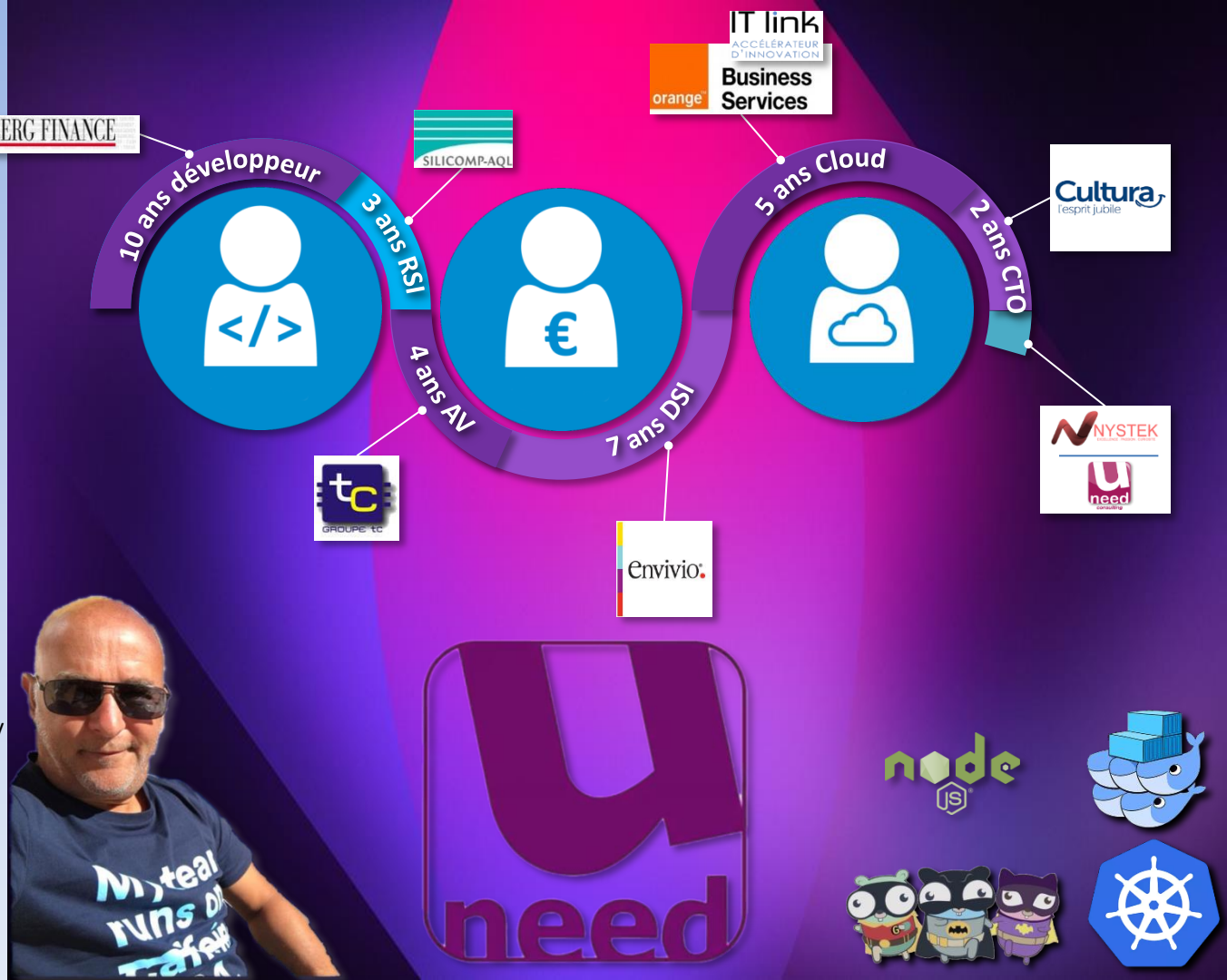
@itwars

http://it-wars.com

itwars

itwars

https://www.linkedin.com/in/vrabah/



Présentation

2

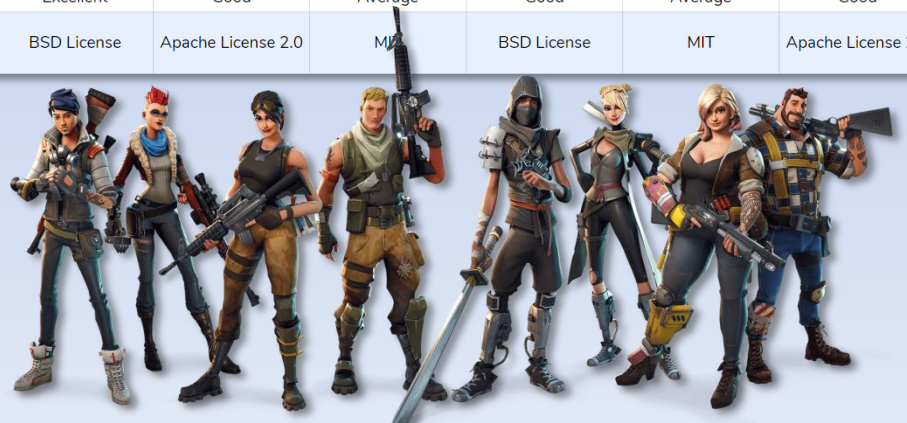
Pages HTML
de sites web

Framework
de scraping

Données
structurées

3

Feature / Framework	Scrapy	PySpider	MechanicalSoup	Portia	NodeCrawler	Selenium	Puppeteer	Webscraper.io
Built In Data Storage Suports	JSON, XML, CSV	CSV, JSON	CSV, JSON, XML	CSV, JSON, XML	CSV, JSON, XML	Customizable	JSON	CSV
Suitable for Broad Crawling	Yes	No	No	No	Yes	No	No	No
Built In Scaling	Yes	Yes	No	No	No	No	No	No
Direct Support to scrape AJAX Heavy Websites	No	Yes	No	Yes	No	Yes	Yes	Yes
Available Selectors	CSS,Xpath	CSS, Xpath	CSS, Xpath	CSS, Xpath	CSS	CSS, Xpath	CSS	CSS
Built in Interface for Periodic Jobs	No	Yes	No	No	No	No	No	No
Point-and-Click-Interface	No	Yes	No	Yes	No	No	No	Yes
Speed (Fast, Medium, Slow)	Fast	Medium	Medium	Slow	Medium	Slowest	Medium	Medium
CPU Usage (Fast, Medium, Slow)	Medium	Medium	Medium	Medium	Medium	High	High	Medium
Memory Usage (High, Medium, Low)	Medium	Medium	Medium	Medium	Medium	High	High	Medium
Github Forks	6,827	2,857	183	961	635	3,991	2,531	264
Github Open Issues	420	171	15	72	11	358	256	10
Github Stars	27,481	11,356	2,694	6,071	3,770	10,643	32,755	757
Last Updated (from published date of blog)	May 23, 2018	May 12, 2018	Feb 14, 2018	April 24, 2018	May 29, 2018	May 31, 2018	May 31, 2018	May 29, 2017
Documentation	Excellent	Good	Average	Good	Average	Good	Good	Excellent
License	BSD License	Apache License 2.0	MI	BSD License	MIT	Apache License 2.0	Apache License 2.0	GNU Lesser General Public License v3.0



Panorama

4

Les outils

regular expressions 101

@regex101 $ donate contact bug reports & feedback wiki

SAVE & SHARE

save regex ctrl+s

FLAVOR

pcre (php)

javascript

python

golang

TOOLS

code generator

regex debugger

SPONSOR

DreamHost

Your website is more than code. Power your purpose with DreamHost.

REGULAR EXPRESSION

no match

/ insert your regular expression here / gm

TEST STRING

SWITCH TO UNIT TESTS

insert your test string here

SUBSTITUTION

https://regex101.com/

EXPLANATION

An explanation of your regex will be automatically generated as you type.

MATCH INFORMATION

Detailed match information will be displayed here automatically.

QUICK REFERENCE

Search reference

all tokens

common tokens

general tokens

anchors

meta sequences

quantifiers

group constructs

A single character of: a, b or c [abc]

A character except: a, b or c [^abc]

A character in the range: a-z [a-z]

A character not in the range: a-z [^a-z]

A character in the range: a-z or A-Z [a-zA-Z]

Any single character .

Any whitespace character \s

Any non-whitespace character \S

Colly

4

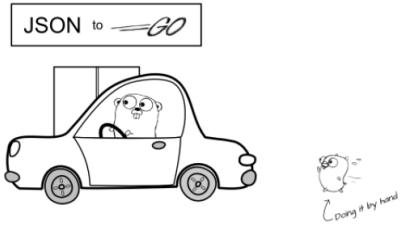
Les outils

JSON-to-Go

Convert JSON to Go struct

This tool instantly converts JSON into a Go type definition. Paste a JSON structure on the left and the equivalent Go type will be generated to the right, which you can paste into your program. The script has to make some assumptions, so double-check the output!

For an example, try converting JSON from the SmartyStreets API or the GitHub API.



JSON	→	Go
		Go will appear here

https://mholt.github.io/json-to-go/

Colly

5

Colly

Fast and Elegant Scraping Framework for Gophers

Features

- Clean API
- Fast (>1k request/sec on a single core)
- Manages request delays and maximum concurrency per domain
- Automatic cookie and session handling
- Sync/async/parallel scraping
- Distributed scraping
- Caching
- Automatic encoding of non-unicode responses
- Robots.txt support
- Google App Engine support

6

```go
func main() {
    c := colly.NewCollector(
        colly.AllowedDomains("www.monsiteweb.com", "monsiteweb.com"),
        colly.DisallowedDomains("media.monsiteweb.com"),
        colly.UserAgent(""),
        colly.MaxDepth(2),
        colly.DisallowedURLFilters("\.*\/admin"),
        colly.URLFilters(""),
        colly.AllowURLRevisit(),
        colly.MaxBodySize(128),
        colly.CacheDir("~/.colly"),
        colly.IgnoreRobotsTxt(),
        colly.Async(true),
    )

    c.OnError(func(_ *colly.Response, err error) {
    })

    c.OnRequest(func(r *colly.Request) {
    })

    c.OnHTML("xxx", func(e *colly.HTMLElement) {
    })

    c.OnScraped(func(s *colly.Response) {
    })

    c.Visit("https://www.monsiteweb.com")
}
```

Colly

7

Les exemples

TJM json
+
Web App

Sport CLI

Meteo Web

https://github.com/itwars/golang-scraping-colly

TJM json

8

freelance


```
{"Administrateur BD":540,"Administrateur ERP":510,"Administrateur produits":470,"Administrateur réseaux":390,"Administrateur système":430,"Analyste":450,"Analyste d'exploitation":340,"Analyste programmeur":390,"Analyste réalisateur":410,"Architecte":620,"Architecte réseaux":590,"Assistant à maîtrise d'ouvrage":570,"Auditeur":660,"Chef de projet":570,"Concepteur BD":430,"Concepteur multimédia":380,"Concepteur télématique":450,"Consultant":600,"Consultant fonctionnel":650,"Consultant technique":540,"Consultant technique et formateur":640,"Directeur de projet":760,"Directeur informatique":790,"Développeur":410,"Expert":650,"Formateur":510,"Infographiste":330,"Ingénieur d'exploitation":430,"Ingénieur d'études":440,"Ingénieur de production":470,"Ingénieur réseaux":480,"Ingénieur système":500,"Maquettiste PAO":320,"Pupitreur/Pilote":300,"Responsable d'exploitation":550,"Responsable maintenance":390,"Rédacteur technique":340,"Support utilisateurs":310,"Technicien d'exploitation":280,"Technicien micro / réseaux":270,"Webmaster":320}
```


Colly

Freelance-info.fr Mon CV Missions

Base sur les tarifs des Freelances en informatique

Ce sondage permanent permet aux freelances en informatique et aux recruteurs de confronter leur tarifs de vente par rapport au marché.

▶ D'ou viennent ces chiffres ?

- Les tarifs de facturation jour présentés sont calculés à partir des contributions effectuées anonymement par les informaticiens indépendants membres de la communauté Freelance-info.
- Ces résultats vous sont communiqués à titre informatif, rien ne garanti qu'ils soient représentatifs du marché. Vous est seul responsable des conclusions que vous pourrez tirer de ces chiffres.

Directeur informatique	790 €/j	Concepteur télématique	450 €/j
Directeur de projet	760 €/j	Ingénieur d'études	440 €/j
Auditeur	660 €/j	Administrateur système	430 €/j
Expert	650 €/j	Concepteur BD	430 €/j
Consultant technique et formateur	650 €/j	Ingénieur d'exploitation	430 €/j
Consultant fonctionnel	650 €/j	Développeur	410 €/j
Architecte	620 €/j	Analyste réalisateur	410 €/j
Consultant	600 €/j	Responsable maintenance	390 €/j
Architecte réseaux	590 €/j	Analyste programmeur	390 €/j
Chef de projet	570 €/j	Administrateur réseaux	390 €/j
Assistant à maîtrise d'ouvrage	570 €/j	Concepteur multimédia	380 €/j
Responsable d'exploitation	550 €/j	Pupitreur/Pilote	370 €/j
Consultant technique	540 €/j	Analyste d'exploitation	340 €/j
Administrateur BD	540 €/j	Rédacteur technique	340 €/j
Formateur	510 €/j	Infographiste	320 €/j
Administrateur ERP	510 €/j	Webmaster	320 €/j
Ingénieur système	500 €/j	Maquettiste PAO	320 €/j
Ingénieur réseaux	480 €/j	Support utilisateurs	310 €/j
Administrateur produits	470 €/j	Technicien d'exploitation	280 €/j
Ingénieur de production	470 €/j	Technicien micro / réseaux	270 €/j
Analyste	460 €/j		

8

Sports CLI

france football

RUGBYRAMA

LE FIGARO.fr

Colly

8

Météo Web

METEO FRANCE

SURF-forecast.com

ALLOSURF

marée.info
horaires - hauteurs - coefficients

Colly

Web App

8

freelance

Colly

Freelance-info.fr

Mon CV

Missions

Base sur les tarifs des Freelances en informatique

Ce sondage permanent permet aux freelances en informatique et aux recruteurs de confronter leur tarifs de vente par rapport au marché.

▶ D'ou viennent ces chiffres ?

- Les tarifs de facturation jour présentés sont calculés à partir des contributions effectuées anonymement par les informaticiens indépendants membres de la communauté Freelance-info.
- Ces résultats vous sont communiqués à titre informatif, rien ne garanti qu'ils soient représentatifs du marché. Vous est seul responsable des conclusions que vous pourrez tirer de ces chiffres.

Directeur informatique	790 €/j	Concepteur télématique	450 €/j
Directeur de projet	760 €/j	Ingénieur d'études	440 €/j
Auditeur	660 €/j	Administrateur système	430 €/j
Expert	650 €/j	Concepteur BD	430 €/j
Consultant technique et formateur	650 €/j	Ingénieur d'exploitation	430 €/j
Consultant fonctionnel	650 €/j	Développeur	410 €/j
Architecte	620 €/j	Analyste réalisateur	410 €/j
Consultant	600 €/j	Responsable maintenance	390 €/j
Architecte réseaux	590 €/j	Analyste programmeur	390 €/j
Chef de projet	570 €/j	Administrateur réseaux	390 €/j
Assistant à maîtrise d'ouvrage	570 €/j	Concepteur multimédia	380 €/j
Responsable d'exploitation	550 €/j	Pupitreur/Pilote	370 €/j
Consultant technique	540 €/j	Analyste d'exploitation	340 €/j
Administrateur BD	540 €/j	Rédacteur technique	340 €/j
Formateur	510 €/j	Infographiste	320 €/j
Administrateur ERP	510 €/j	Webmaster	320 €/j
Ingénieur système	500 €/j	Maquettiste PAO	320 €/j
Ingénieur réseaux	480 €/j	Support utilisateurs	310 €/j
Administrateur produits	470 €/j	Technicien d'exploitation	280 €/j
Ingénieur de production	470 €/j	Technicien micro / réseaux	270 €/j
Analyste	460 €/j		

Conclusions

9

- Structuration de données non-structurées
- Agrégations de données de plusieurs sites
- Mise en base de données (statistique, ML, enrichissement de données existantes, …)
- Réalisation de maquette à partir d'un site existant

Colly

Regardez !

```go
package main

import (
    "fmt"
    "math"
)

var π = math.Pi
var φ = .0084
var z = 3.0

func main() {
    π′ := π * (math.Sin(z) + φ*math.Sin(2*z))
    Δz := π′ * (math.Sin(φ) + φ*math.Sin(2*z))
    fmt.Println("π  =", π)
    fmt.Println("φ  =", φ)
    fmt.Println("π′ =", π′)
    fmt.Println("Δz =", Δz)
}
```

https://github.com/soniakeys/meeus

```go
package main

import (
    "encoding/json"
    "fmt"

    "github.com/pulumi/pulumi-aws/sdk/go/aws/s3"
    "github.com/pulumi/pulumi/sdk/go/pulumi"
    "github.com/pulumi/pulumi/sdk/go/pulumi/asset"
)

func main() {
    pulumi.Run(func(ctx *pulumi.Context) error {
        // Create a bucket and expose a website index document
        siteBucket, err := s3.NewBucket(ctx, "s3-website-bucket", &s3.BucketArgs{
            Website: map[string]interface{}{
                "indexDocument": "index.html",
            },
        })
        if err != nil {
            return err
        }

        siteDir := "www" // directory for content files

        // For each file in the directory, create an S3 object stored in `siteBucket`
        files, err := ioutil.ReadDir(siteDir)
        if err != nil {
            return err
        }
        for _, item := range files {
            name := item.Name()
            filePath := filepath.Join(siteDir, name)
            if _, err := s3.NewBucketObject(ctx, name, &s3.BucketObjectArgs{
                Bucket:      siteBucket.ID(),                       // reference to the s3.Bucket object
                Source:      asset.NewFileAsset(filePath),          // use FileAsset to point to a file
                ContentType: mime.TypeByExtension(path.Ext(name)), // set the MIME type of the file
            }); err != nil {
                return err
            }
        }
    return nil
    })
}
```

A surveiller

GO CLOUD pulumi

Questions ?

This is the end