

Università degli studi di Catania

Corso di Laurea Magistrale in informatica Compilatori A.A 2014 - 2015

Antonio Fischetti W82-000021

Realizzazione di un web crawler in nodejs

Introduzione

Il progetto realizzato è un crawler, meglio conosciuto come web crawler, spider o robot. Si occupa di analizzare una lista di URL fornita in input, identificando tutti gli hyperlink(Url) contenuti nel documento ed salvando questi URL in un database, per poi successivamente visitarli in modo ricorsivo. Il web crawler che ho implementato è stato sviluppato utilizzando il linguaggio **nodeJS** ed un database non relazionale (**MongoDB**). Questo nasce da un progetto universitario dell'Università degli Studi di Catania. Lo scopo del progetto è quello di realizzare un crawler che analizzi gli URL a partire da seeds iniziali, memorizzando su file solo le pagine che contengono almeno una delle keyword date in input dall'utente.

Download ed Installazione (Mac Osx 10.10.3)

Prima di spiegare il funzionamento del progetto, mi soffermo sulla parte relativa all'installazione delle componenti utilizzate. E' stato necessario installare:

- Ambiente di sviluppo per **nodejs** disponibile su: <http://nodejs.org>
- Database non relazionale **mongodb** disponibile su: <http://mongodb.com>
- Libreria Mongoose di interfaccia tra db e nodejs disponibile su: <http://mongoosejs.com>
- Alcuni moduli npm di nodejs: "cheerio" - "express" - "async" - "request" - "mongoose" - "String" - "body parser" - "fs" - "events" - "mk dir" i cui download sono disponibili da terminale.
Es \$ npm install express.

Esecuzione del Database

Dopo aver correttamente installato **mongodb** nella nostra macchina, ed aver settato opportunamente tutti i file ed i path necessari, il database non relazionale deve essere mandato in esecuzione. Questo viene effettuato direttamente da terminale utilizzando il comando **mongod - - path** (il path di riferimento che abbiamo scelto durante la creazione del database).

Funzionamento del Crawler

Il software prende in input alcuni parametri :

- Uno o più seeds iniziali.
- Un tempo di esecuzione espresso in minuti.
- Una o più keywords.

```
curl --data "url1=http://it.wikipedia.org/wiki/Biologia_di_sintesi
&url2=http://en.wikipedia.org/wiki/Synthetic_biology
&url3=http://syntheticbiology.org/
&url4=http://www.systemsbiology.org/
&url5=http://www.synbioproject.org/
&url6=http://www.synberc.org/what-is-synbio
&url7=https://synbio.mit.edu/
&url8=http://www.igem.org/Main_Page
&url9=http://biobricks.org/
&url10=http://synbioconference.org/2014
&url11=http://synbio.berkeley.edu/
&url12=http://synbioconference.org/2015
&url13=http://www.globalengage.co.uk/synthetic-biology.html
&url14=http://www.embl.de/training/events/2015/SYN15-01/
&key1=PhD student&key2=synthetic biology&durata=60" http://localhost:8081/crawler
```

Questa richiesta al server (in questo caso in localhost) viene effettuata verso la porta 8081, che è in ascolto per la ricezione. Per avviare il servizio sul server basta aprire il terminale, posizionarsi all'interno della cartella che contiene il file **crawler.js** e avviarlo tramite il comando **node crawler.js**.

Nel nostro caso i seed iniziali sono quelli mostrati precedentemente, e le chiavi ricercate sono state ***"PhD student & synthetic biology"*** e ***"Post doc & synthetic biology"***.

Quando viene effettuata la richiesta al server con i seed iniziali e le parole chiave da trovare, inizialmente viene creata una cartella che ha come nome, la stringa della simulazione (ad esempio Simulation1433164201801 , che viene generata a partire dalla data in cui questa viene effettuata).

Quindi per ogni simulazione che andremo a fare, avremo diverse directory.

Successivamente vengono memorizzate all'interno del database, gli urls (seed) iniziali.

I campi del database sono:

- **path** (che contiene il path del file se questo è stato memorizzato, altrimenti il suo valore sarà '-1').
- **urlParse** (che contiene l'url da memorizzare nel database).
- **father** (contiene l'url padre, nel caso iniziale conterrà "Initial Seed Url").
- **simulation** (che contiene l'id della simulazione, utile se più richieste contemporanee vengono effettuate sul server).
- **visited** (Un flag che viene settato a 'yes' se l'url memorizzata sul db è stata schedulata, altrimenti 'no').
- **depth** (contiene la profondità per ogni Url trovato).

Dopo di che viene impostato un timeout che servirà a far capire al software quando è il momento di terminare il processo. (come visto in precedenza viene impostato dall'utente il numero di minuti della simulazione).

Descrizione delle funzioni del Crawler

Per lo sviluppo del Crawler ho utilizzato una gestione dei processi basata su una coda, precisamente **Async.queue** , un oggetto che permette di costruire una coda, impostando un **worker** da eseguire ogni volta che viene effettuata una push. Questo mi ha permesso di evitare la ricorsione dovuta al lavoro del crawler. Il worker si occupa di chiamare in modo asincrono le funzioni di:

-**getUrl** (Restituisce una url ancora non visitata).

-**crawlingUrl** (Effettua una richiesta all'url restituita dalla funzione getUrl, effettuare un controllo delle parole chiave contenute nell' html, e nel caso positivo salvare la pagina su un file txt, ed effettuare lo scarping della pagina, trovando url valide da essere inserite nel db).

-**saveUrlFound** (Salvataggio su db di tutte le url valide trovate).

Effettuo lo scarping della pagina attraverso il metodo **cheerio.load(html)**, ottenendo tutti i tag **'a href'** , poichè nel linguaggio HTML questi contengono i collegamenti e quindi gli url ad altre pagine. Per ogni url trovata viene effettuato un controllo, utilizzando un regger (espressione regolare).

Simulazioni Effettuate

Le simulazioni sono state effettuate su una VPS con le seguenti caratteristiche:

- AMD Opteron(tm) Processor 4284 3GHZ
- 4 GB RAM

Seed Iniziali:

http://it.wikipedia.org/wiki/Biologia_di_sintesi
http://en.wikipedia.org/wiki/Synthetic_biology
<http://syntheticbiology.org/>
<http://www.systemsbiology.org/>
<http://www.synbioproject.org/>
<http://www.synberc.org/what-is-synbio>
<https://synbio.mit.edu/>
http://www.igem.org/Main_Page
<http://biobricks.org/>
<http://synbioconference.org/2014>
<http://synbio.berkeley.edu/>
<http://synbioconference.org/2015>
<http://www.globalengage.co.uk/synthetic-biology.html>
<http://www.embl.de/training/events/2015/SYN15-01/>

Keyword Iniziali:

- “PhD student” and “synthetic biology”
- “Post doc” and “synthetic biology”

Durata in ore della simulazione:

- 1 ora di CPU Time
- 24 ore di CPU Time
- 7*24 ore di CPU Time

```
AntonioFischetti - root@vps174580: /Crawler - ssh - 158x21
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~28-9-2014/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~29-9-2014/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~30-9-2014/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-month/exact_date~1438412400/request_format~html/cat_ids~19/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-month/exact_date~1438412400/request_format~html/tag_ids~51/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-agenda/exact_date~1438412400/request_format~html/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~1438412400/request_format~html/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-week/exact_date~1438412400/request_format~html/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-month/exact_date~1478034800/request_format~html/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~1-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~2-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~3-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~4-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~5-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~6-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~7-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~8-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~9-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~10-8-2015/ Deph:4
StartParsURL: http://pbd.lbl.gov/calendar/action-oneday/exact_date~11-8-2015/ Deph:4

1 [|||||] 18.6% Tasks: 21, 117 thr; 5 running
2 [|||||] 6.7% Load average: 0.94 0.56 0.50
3 [|||||] 3.8% Uptime: 22:17:18
Mem[|||||] 286/4096MB
Swp[|||||] 0/128MB

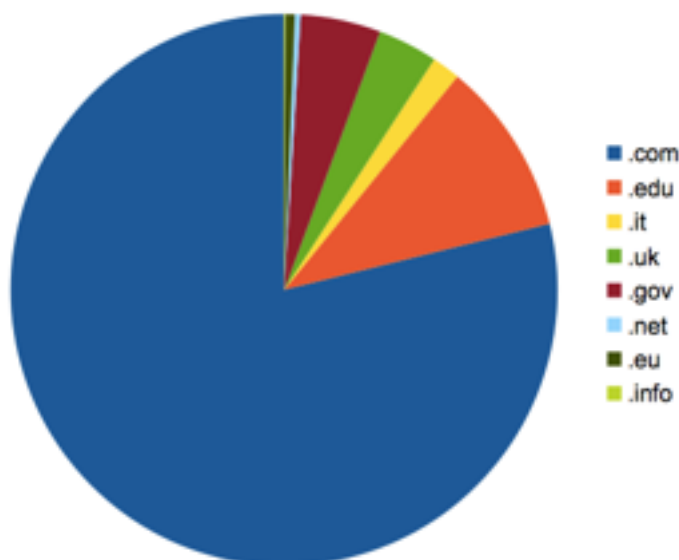
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
5590 root 20 0 656M 21428 1680 S 0.0 0.5 0:07.18 /usr/bin/nodejs /usr/lib/node_modules/forever/bin/monitor server.js
5592 root 20 0 1836M 121M 3920 S 12.4 3.0 21:52.36 /usr/bin/nodejs /Crawler/server.js
5597 root 20 0 1836M 121M 3920 S 0.0 3.0 0:03.39 /usr/bin/nodejs /Crawler/server.js
5596 root 20 0 1836M 121M 3920 S 0.7 3.0 0:03.67 /usr/bin/nodejs /Crawler/server.js
5595 root 20 0 1836M 121M 3920 S 0.0 3.0 0:03.49 /usr/bin/nodejs /Crawler/server.js
5594 root 20 0 1836M 121M 3920 S 0.0 3.0 0:03.89 /usr/bin/nodejs /Crawler/server.js
5593 root 20 0 1836M 121M 3920 S 0.7 3.0 1:20.06 /usr/bin/nodejs /Crawler/server.js
5591 root 20 0 656M 21428 1680 S 0.0 0.5 0:01.19 /usr/bin/nodejs /usr/lib/node_modules/forever/bin/monitor server.js

F1Help F2Setup F3Search F4Filter F5Sorted F6Colmap F7Nice F8Nice F9Kill F10Quit
```

Statistiche delle simulazioni

Sono state effettuate 3 simulazioni per ogni coppia di Keyword iniziali. L'output ottenuto dall'esecuzione del crawler è stato passato al MailParser che ha permesso di estrarre le mail valide presenti che corrispondono agli URL analizzati nella quale è presente almeno una delle due parole chiave.

Simulazione 1 ora - Keys "PhD student" - "synthetic biology"



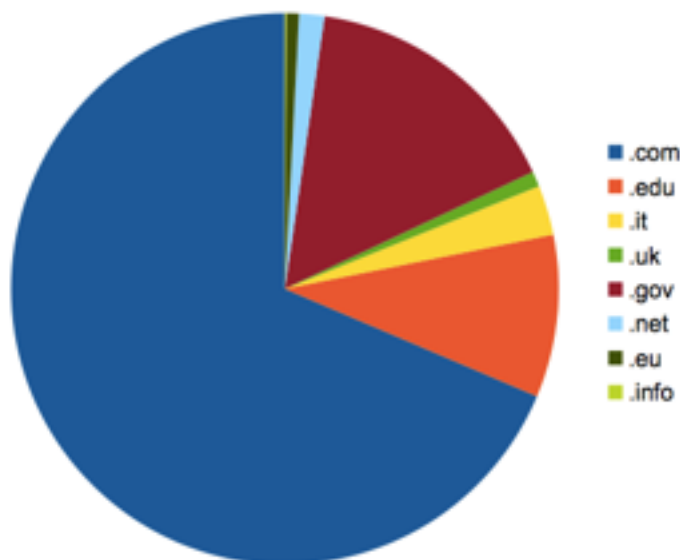
```
CrawlerNodeJS ANTONIO FISCHETTI
Test 24 hours - Keywords "PhD student" - "synthetic biology"
Number of HTML file saved is 343.
Number of Mail/Url founded is 50.

Start Simulation at:Fri Jun 12 2015 06:43:56 GMT-0400 (EDT)
Stop Simulation at:Fri Jun 12 2015 06:48:56 GMT-0400 (EDT)

Domain Found in : Simulation1434105716563

Number of Url Visited:1762
Domain (.org): 319
Domain (.com): 1134
Domain (.edu): 147
Domain (.it): 24
Domain (.uk): 51
Domain (.gov): 68
Domain (.net): 5
Domain (.eu): 8
Domain (.info): 1
```

Simulazione 1 ora - Keys “Post doc” - “synthetic biology”



CrawlerNodeJS ANTONIO FISCHETTI
 Test 24 hours - Keywords “Post doc” - “synthetic biology”
 Number of HTML file saved is 302.
 Number of Mail/Url founded is 40.

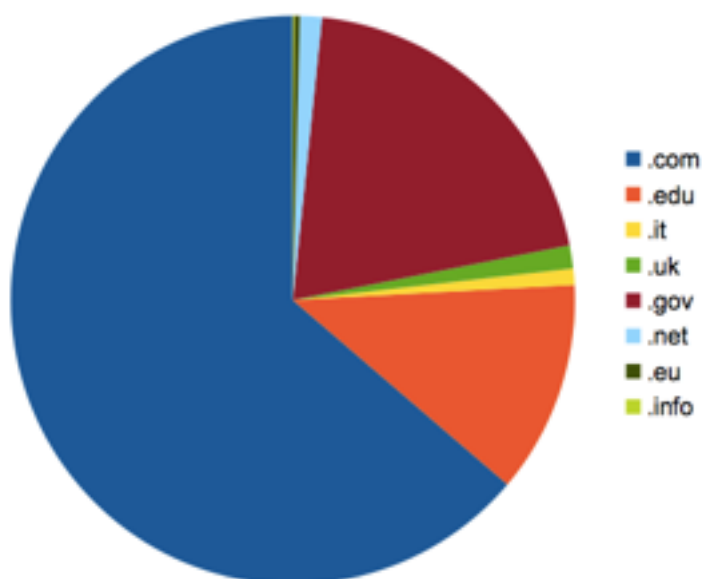
Start Simulation at: Fri Jun 12 2015 07:48:36 GMT-0400 (EDT)
 Stop Simulation at: Fri Jun 12 2015 08:48:36 GMT-0400 (EDT)

Domain Found in : Simulation1434109711544

Number of Url: 1672

Domain (.org): 313
 Domain (.com): 914
 Domain (.edu): 127
 Domain (.it): 39
 Domain (.uk): 12
 Domain (.gov): 209
 Domain (.net): 20
 Domain (.eu): 10
 Domain (.info): 1

Simulazione 24 ore - Keys “PhD student” - “synthetic biology”



CrawlerNodeJS ANTONIO FISCHETTI
 Test 24 hours - Keywords “PhD student” - “synthetic biology”
 Number of HTML file saved is 2153.
 Number of Mail/Url founded is 300.

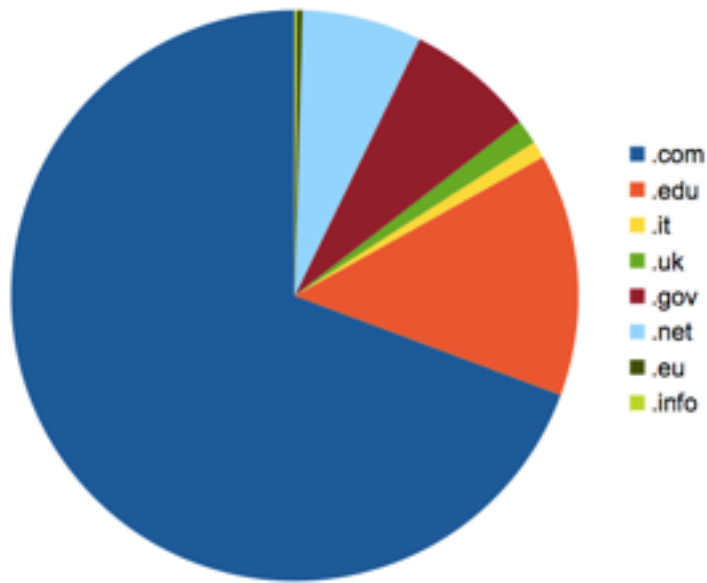
Start Simulation at: Thu Jun 11 2015 04:48:06 GMT-0400 (EDT)
 Stop Simulation at: Fri Jun 12 2015 04:48:06 GMT-0400 (EDT)

Domain Found in : Simulation1434012486102

Number of Url Visited: 78183

Domain (.org): 19512
 Domain (.com): 36679
 Domain (.edu): 6952
 Domain (.it): 533
 Domain (.uk): 760
 Domain (.gov): 11644
 Domain (.net): 723
 Domain (.eu): 192
 Domain (.info): 40

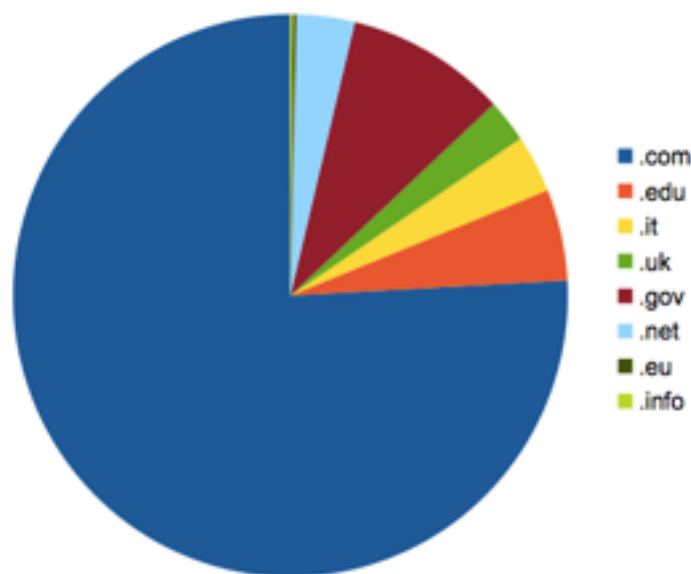
Simulazione 24 ore - Keys “Post doc” - “synthetic biology”



CrawlerNodeJS ANTONIO FISCHETTI
 Test 24 hours - Keywords "Post doc" - "synthetic biology"
 Number of HTML file saved is 1.384.
 Number of Mail/Url founded is 207.
 Start Simulation at: Fri Jun 12 2015 10:28:16 GMT-0400 (EDT)
 Stop Simulation at: Sat Jun 13 2015 10:28:16 GMT-0400 (EDT)

Domain Found in : Simulation1434119294855
 Number of Url Visited: 67594
 Domain (.org): 9848
 Domain (.com): 48042
 Domain (.edu): 7927
 Domain (.it): 555
 Domain (.uk): 815
 Domain (.gov): 4217
 Domain (.net): 3936
 Domain (.eu): 180
 Domain (.info): 74

Simulazione 7*24 ore - Keys “PhD student” - “synthetic biology”



CrawlerNodeJS ANTONIO FISCHETTI
 Test 24 hours - Keywords "PhD student" - "synthetic biology"
 Number of HTML file saved is 3047.
 Number of Mail/Url founded is 570.

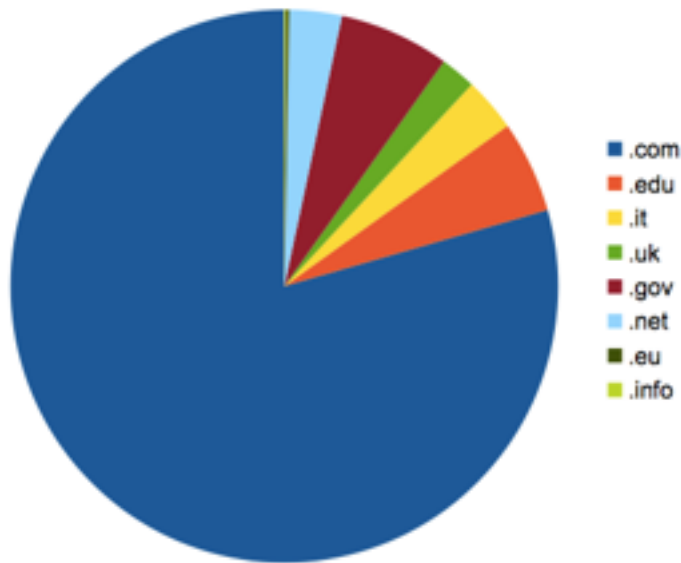
Start Simulation at: Mon Jun 15 2015 14:44:02 GMT-0400 (EDT)
 Stop Simulation at: Mon Jun 22 2015 14:44:02 GMT-0400 (EDT)

Domain Found in : Simulation1434551604740

Number of Url Visited: 209.051

Domain (.org): 25.998
 Domain (.com): 136.606
 Domain (.edu): 9.458
 Domain (.it): 5.974
 Domain (.uk): 4.486
 Domain (.gov): 16.822
 Domain (.net): 6.107
 Domain (.eu): 360
 Domain (.info): 267

Simulazione 7*24 ore - Keys “Post doc” - “synthetic biology”



CrawlerNodeJS ANTONIO FISCHETTI
 Test 7*24 hours - Keywords “Post doc” - “synthetic biology”
 Number of HTML file saved is 2.306.
 Number of Mail/Url founded is 376.

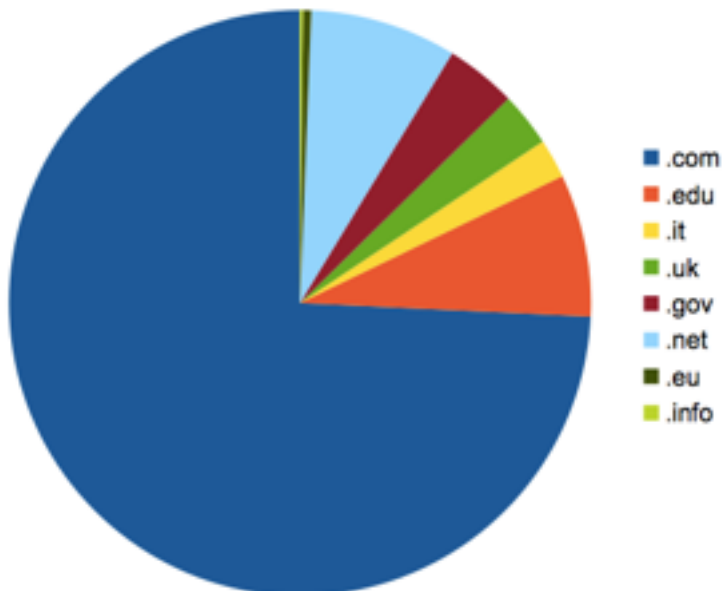
Start Simulation at: Mon Jun 15 2015 18:28:16 GMT-0400 (EDT)
 Stop Simulation at: Mon Jun 22 2015 18:28:16 GMT-0400 (EDT)

Domain Found in : Simulation1434551722493

Number of Url: 21.0793

Domain (.org): 19.234
 Domain (.com): 15.0337
 Domain (.edu): 10.199
 Domain (.it): 6.007
 Domain (.uk): 3.953
 Domain (.gov): 12.330
 Domain (.net): 5.864
 Domain (.eu): 352
 Domain (.info): 189

Simulazione 1 Mese - Keys “PhD student” - “synthetic biology”



CrawlerNodeJS ANTONIO FISCHETTI
 Test 1 month - Keywords “PhD student” - “synthetic biology”
 Number of HTML file saved is 3782.
 Number of Mail/Url founded is 812.

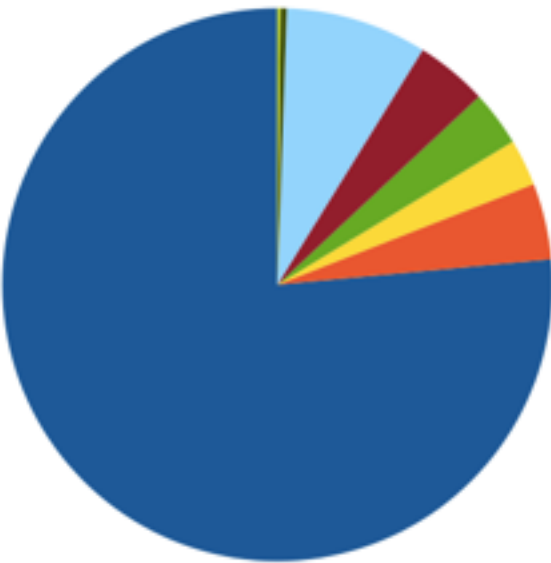
Start Simulation at: Mon Jun 15 2015 10:02:22 GMT-0400 (EDT)
 Stop Simulation at: Wen Jul 15 2015 10:02:22 GMT-0400 (EDT)

Domain Found in : Simulation1435046993906

Number of Url Visited: 355.208

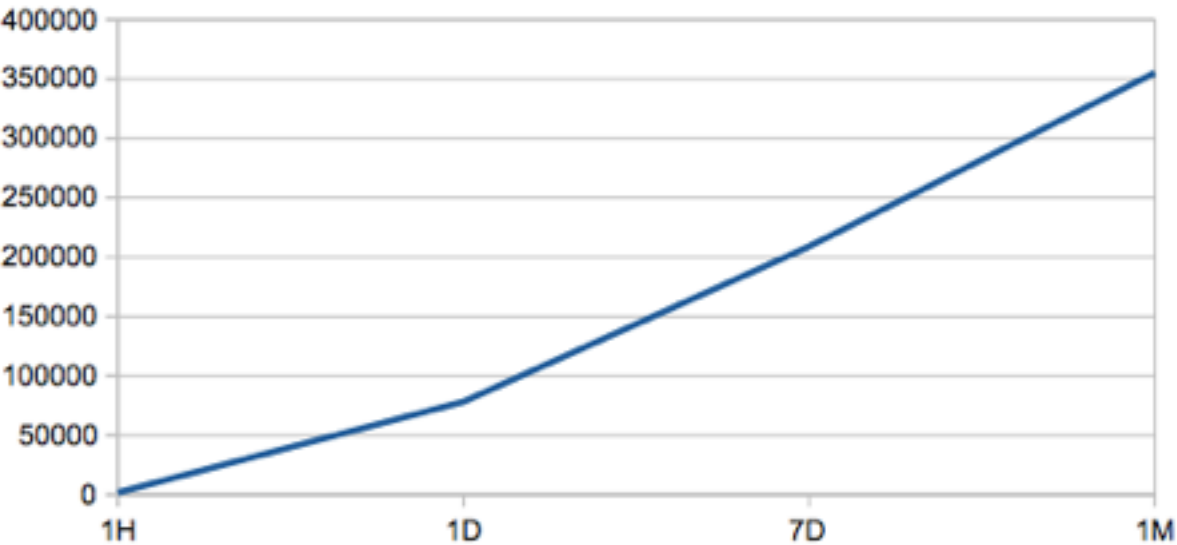
Domain (.org): 61.515
 Domain (.com): 210.087
 Domain (.edu): 23.121
 Domain (.it): 6.365
 Domain (.uk): 8.794
 Domain (.gov): 11.592
 Domain (.net): 23.918
 Domain (.eu): 1.327
 Domain (.info): 489

Simulazione 1 Mese - Keys “Post doc” - “synthetic biology”

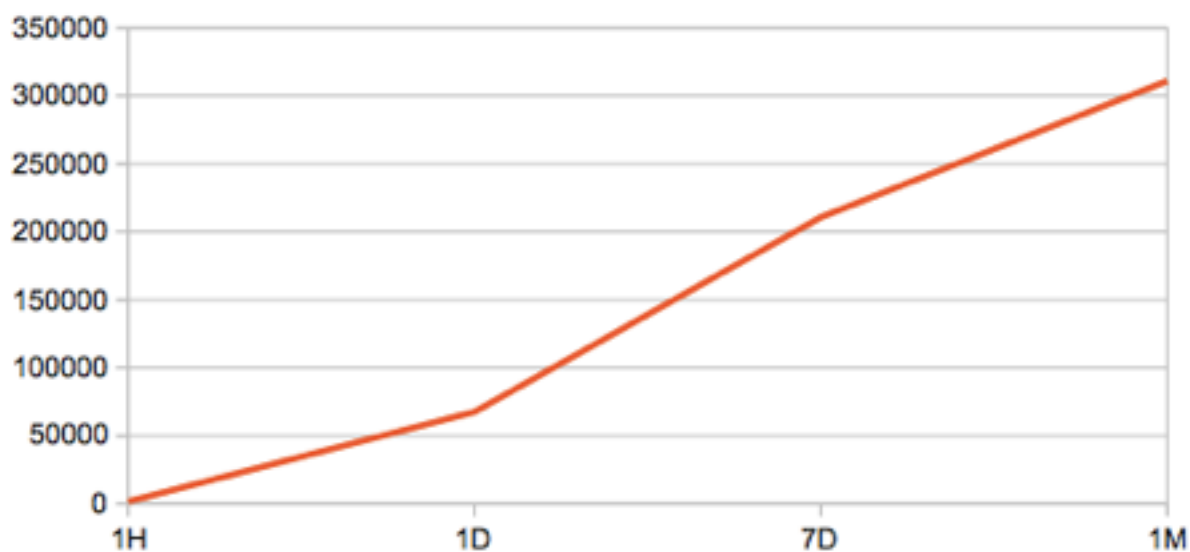


CrawlerNodeJS ANTONIO FISCHETTI	
Test 1 month - Keywords "Post doc" - "synthetic biology"	
Number of HTML file saved is 2800.	
Number of Mail/Url founded is 500.	
.com	Start Simulation at:Mon Jun 15 2015 10:12:41 GMT-0400 (EDT)
.edu	Stop Simulation at:Wen Jul 15 2015 10:12:41 GMT-0400 (EDT)
.it	Domain Found in : Simulation1435047262201
.uk	Number of Url Visited: 311.008
.gov	Domain (.org): 41.515
.net	Domain (.com): 286.087
.eu	Domain (.edu): 12.021
.info	Domain (.it): 7.355
	Domain (.uk): 8.768
	Domain (.gov): 11.392
	Domain (.net): 22.418
	Domain (.eu): 987
	Domain (.info): 465

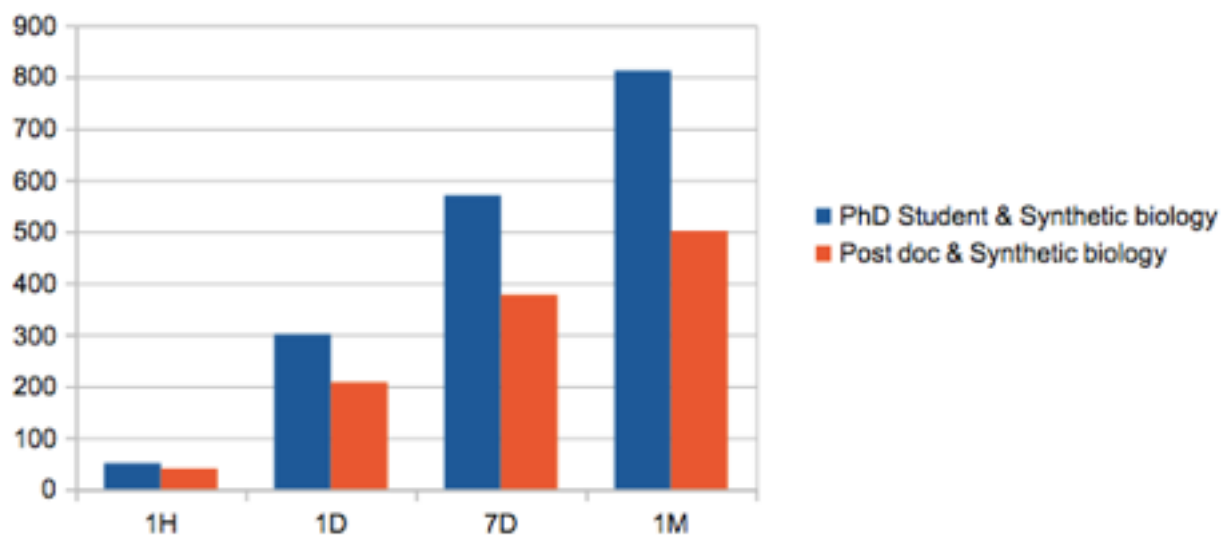
Url Visitate - Keywords "Phd Student" & "Synthetic biology"



Url Visitate - Keywords "Post doc" & "Synthetic biology"



Mail Valide Trovate



Credits

Il Crawler è disponibile su Github alla pagina:

<http://github.com/Jacitano87/Crawler-Web-Nodejs>

I Risultati ottenuti dalle varie simulazioni sono disponibili all'interno della directory "Risultati Crawler" - all'interno del progetto su Github.