
بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه مهندسی نرم‌افزار

سند توضیحات تاکتیک های اعمال شده در پروژه

بخش اول

اعضا گروه:

مهرداد قصابی

نوید شاقوزائی

محمد قربانپور

فاطمه ابراهیم زاده

فاطمه کریمی

نرگس غریبی

استاد راهنما:

دکتر رضا رمضانی

بهار ۱۴۰۱

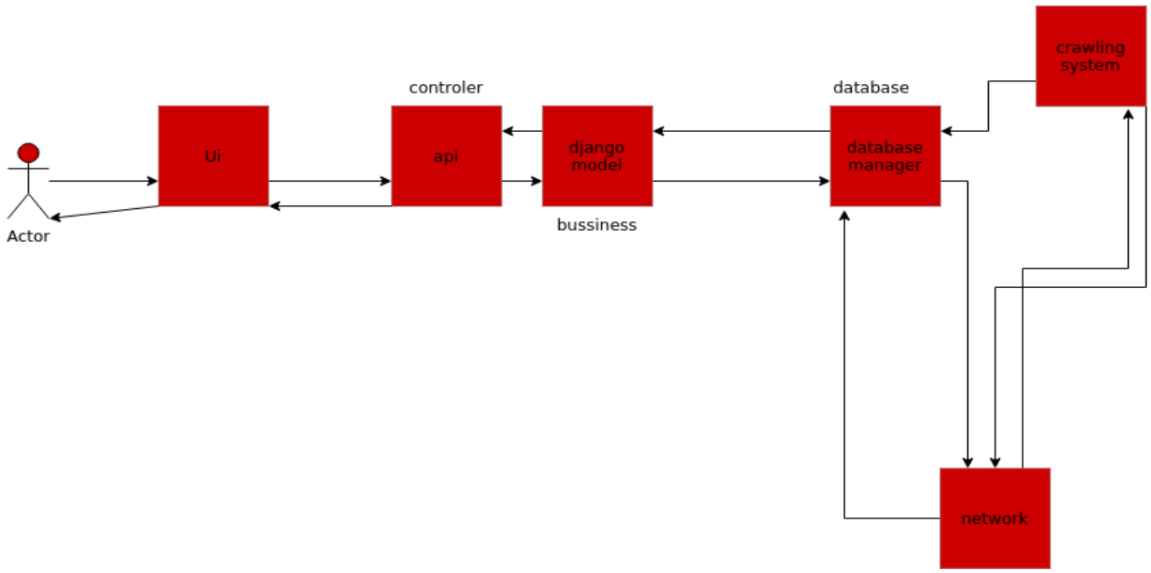
فهرست مطالب

مقدمه

ما در این پروژه یک سایت جمع آوری آگهی را به وجود آوردیم که از سایت های مختلف آگهی های استخدام را جمع آوری می کند و در یک سایت تجمیع می کند. هدف از به وجود آوردن این سایت این است که کاربرها دسترسی راحتتری به آگهی های استخدام داشته باشند و همه ی آگهی های استخدام را به صورت تجمیع شده مشاهده کنند.

معماری سیستم

در این سیستم از یک معماری n tier استفاده شده است.



تاتیک های به کارگرفته شده در معماری سیستم

۱- دسترسی پذیری(availability)

برای این صفت کیفی از تاکتیک های مونیتورینگ و retry استفاده کردیم.(تاکتیک های دیگر مانند exeption handling و rollback از قبل در سیستم وجود داشتند).

در تاکتیک مونیتورینگ یک وب سرویس یه نام scrapydweb را اجرا کردیم که در این سرویس می توانیم کدهای هر کرولر را بدون مراجعه به سرور آپلود و دیپلوی کنیم. در این وب سرویس لاگ عملکرد هر کدام از کرولرها از سرور کرولینگ اسکرپی (scrapy) به صورت لحظه ای و آماری نشان داده می شود و برای خطاهای critical تنظیم شده تا پس از رخداد به تلگرام یکی از اعضای گروه ارسال شود تا برای برطرف کردن آن اقدام شود.

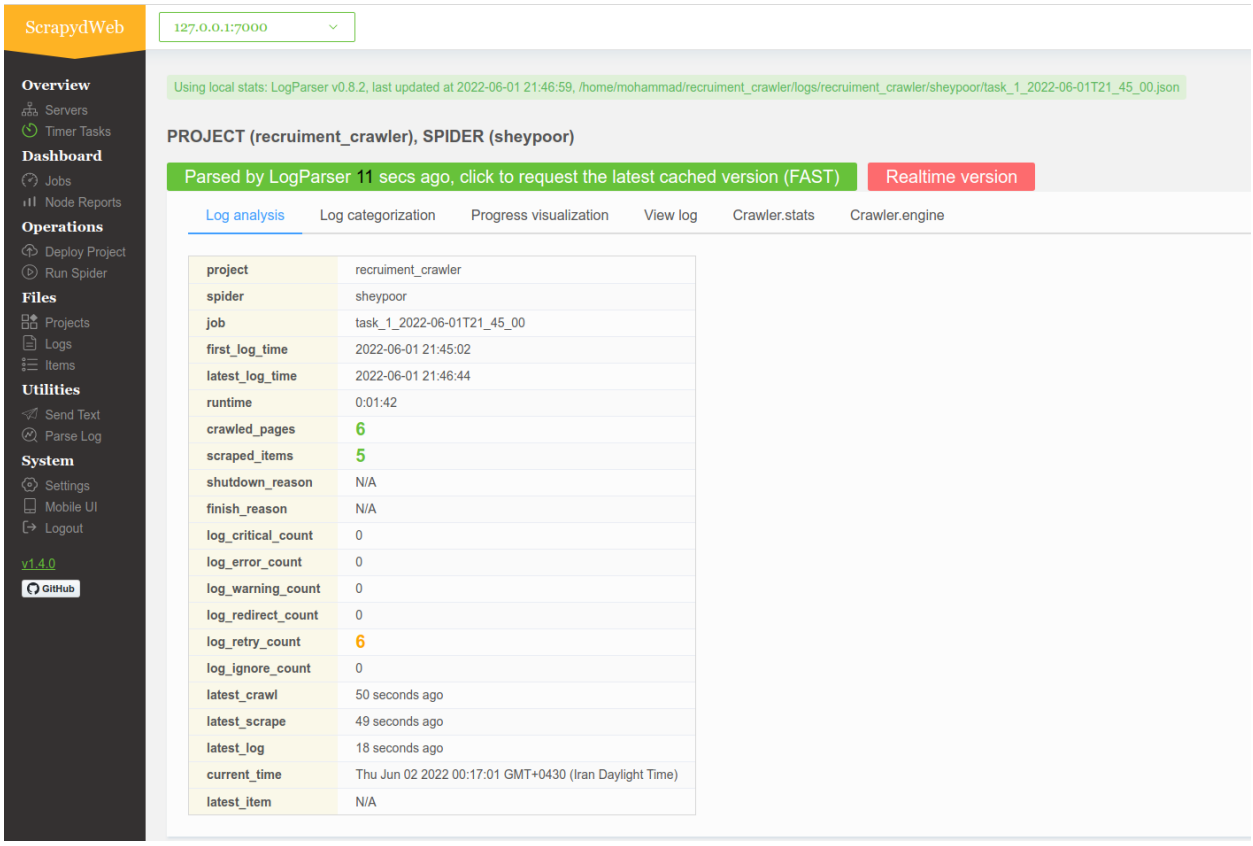
هنگام گرفتن یک صفحه ی وب ممکن است به خطاهای رایج سطح وب (مانند خطاهای دسته ی ۴۰۰ و ۵۰۰) برخورد کنیم. برای رفع حدودی این خطاها از تاکتیک retry استفاده می کنیم که پس برخورد به این خطا دوباره یک درخواست برای صفحه ی مورد نظر فرستاده شود. اگر در درخواست آخر مجدد صفحه مورد نظر (برای تعداد درخواست های مجدد یک عدد ثابت در نظر گرفته شده است) خطا رخ داد، از کرول کردن صفحه مورد نظر صرف نظر (ignore) می کنیم و به کرول بقیه صفحات ادامه می دهیم.

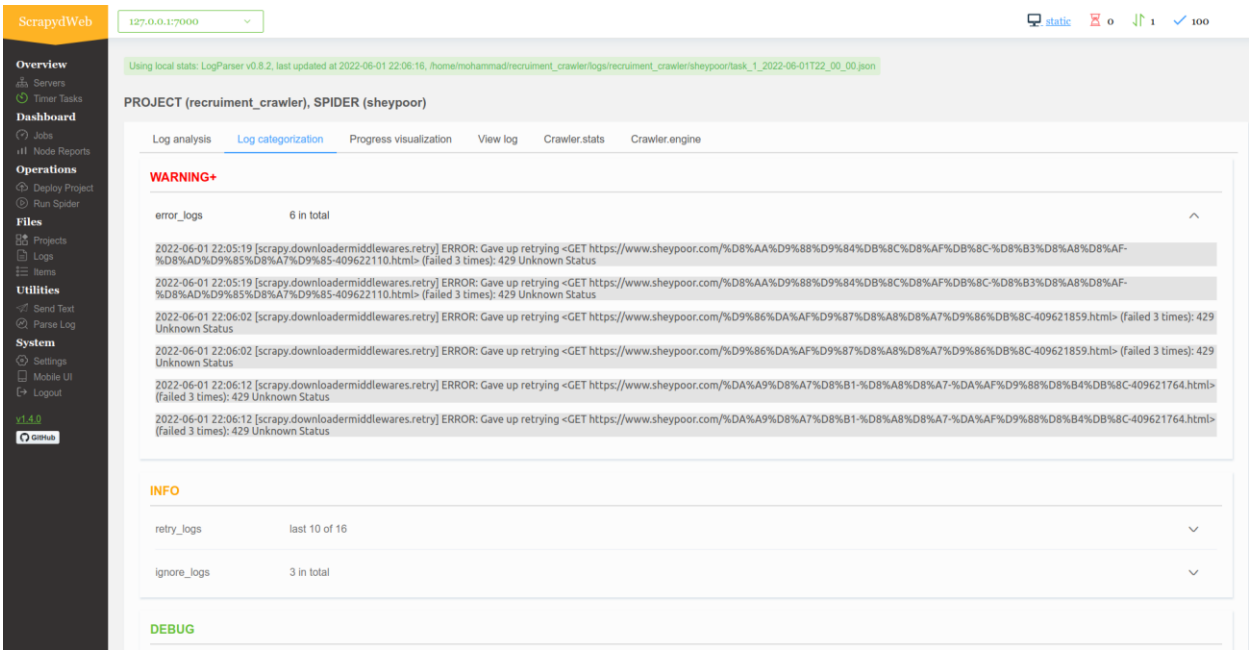
هدف از پیاده سازی این سیستم ها باخبر شدن از جزییات خطا در کمترین زمان ممکن است تا در کمترین زمان ممکن برای رفع آن اقدام کنیم و همچنین کرولرها در هنگام برخورد به خطاهای وب بتوانند زودتر بازیابی شوند. سیستم کرولینگ جدا از سیستم بک اند کار می کند و اطلاعات مستقیما در پایگاه داده وارد می شود یعنی اگر سیستم کرولینگ از کار بیفتد بقیه سیستم ها مانند سیستم بک اند و پایگاه داده به کار خود ادامه می دهند.

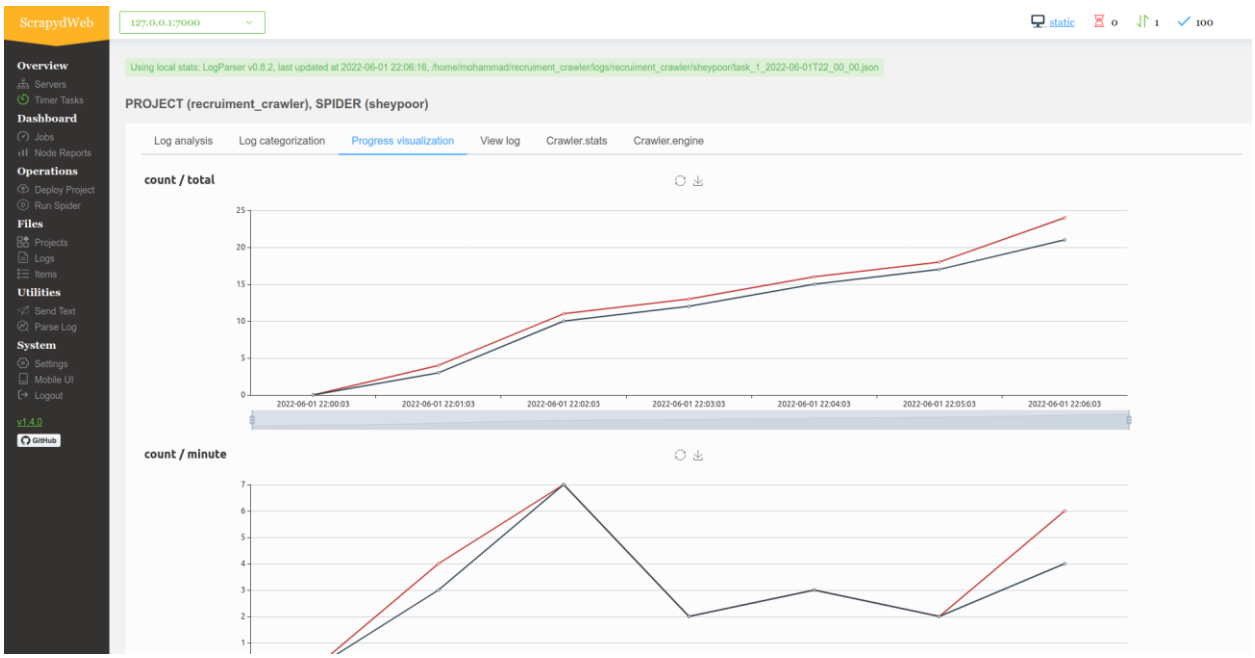
در هنگام پیاده سازی این سیستم تغییر خاصی در معماری نرم افزار اعمال نشده است و صرفا سیستم کرولینگ که یک سیستم جدا در معماری به حساب می آید به یک سیستم مونتورینگ وصل شده است.

ما در پیاده سازی کرولرها به خصوص برای کرولر سایت شیپور به مشکل برخوردیم و سایت شیپور ما را بن کرد. برای دور زدن این مشکل دو نمونه پروکسی سرور tor را در systemd اجرا کردیم تا درخواست ها از طریق یک middlware scrapy به این پروکسی سرورها هدایت شوند.

عکس های زیر تصاویری از سامانه مونیتورینگ می باشد.


ScrapydWeb
127.0.0.1:7000
Overview
Servers
Timer Tasks
Dashboard
Jobs
Node Reports
Operations
Deploy Project
Run Spider
Files
Projects
Logs
Items
Utilities
Send Text
Parse Log
System
Settings
Mobile UI
Logout
v1.4.0
GitHub
Using local stats: LogParser v0.8.2, last updated at 2022-06-01 21:46:59, /home/mohammad/recruiment_crawler/logs/recruiment_crawler/sheypoor/task_1_2022-06-01T21_45_00.json
PROJECT (recruiment_crawler), SPIDER (sheypoor)
Parsed by LogParser 11 secs ago, click to request the latest cached version (FAST)
Realtime version
Log analysis
Log categorization
Progress visualization
View log
Crawler.stats
Crawler.engine
project recruiment_crawler
spider sheypoor
job task_1_2022-06-01T21_45_00
first_log_time 2022-06-01 21:45:02
latest_log_time 2022-06-01 21:46:44
runtime 0:01:42
crawled_pages 6
scraped_items 5
shutdown_reason N/A
finish_reason N/A
log_critical_count 0
log_error_count 0
log_warning_count 0
log_redirect_count 0
log_retry_count 6
log_ignore_count 0
latest_crawl 50 seconds ago
latest_scrape 49 seconds ago
latest_log 18 seconds ago
current_time Thu Jun 02 2022 00:17:01 GMT+0430 (Iran Daylight Time)
latest_item N/A

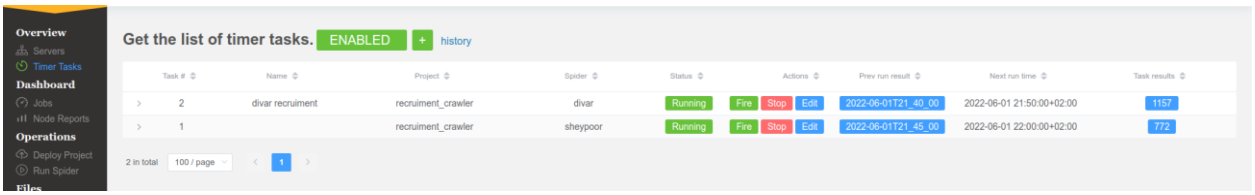

ScrapydWeb
127.0.0.1:7000
static
0
1
100
Using local stats: LogParser v0.8.2, last updated at 2022-06-01 22:06:16, /home/mohammad/recruiment_crawler/logs/recruiment_crawler/sheypoor/task_1_2022-06-01T22_00_00.json
PROJECT (recruiment_crawler), SPIDER (sheypoor)
Log analysis
Log categorization
Progress visualization
View log
Crawler.stats
Crawler.engine
WARNING+
error_logs 6 in total
2022-06-01 22:05:19 [scrapy.downloadermiddlewares.retry] ERROR: Gave up retrying <GET https://www.sheypoor.com/%D8%AA%D9%88%D9%84%DB%8C%D8%AF%DB%8C-%D8%B3%D8%A8%D8%AF-%D8%AD%D9%85%D8%A7%D9%85-409622110.html> (failed 3 times): 429 Unknown Status
2022-06-01 22:05:19 [scrapy.downloadermiddlewares.retry] ERROR: Gave up retrying <GET https://www.sheypoor.com/%D8%AA%D9%88%D9%84%DB%8C%D8%AF%DB%8C-%D8%B3%D8%A8%D8%AF-%D8%AD%D9%85%D8%A7%D9%85-409622110.html> (failed 3 times): 429 Unknown Status
2022-06-01 22:06:02 [scrapy.downloadermiddlewares.retry] ERROR: Gave up retrying <GET https://www.sheypoor.com/%D9%86%DA%AF%D9%87%D8%A8%D8%A7%D9%86%DB%8C-409621859.html> (failed 3 times): 429 Unknown Status
2022-06-01 22:06:02 [scrapy.downloadermiddlewares.retry] ERROR: Gave up retrying <GET https://www.sheypoor.com/%D9%86%DA%AF%D9%87%D8%A8%D8%A7%D9%86%DB%8C-409621859.html> (failed 3 times): 429 Unknown Status
2022-06-01 22:06:12 [scrapy.downloadermiddlewares.retry] ERROR: Gave up retrying <GET https://www.sheypoor.com/%DA%A9%D8%A7%D8%B1-%D8%A8%D8%A7-%DA%AF%D9%88%D8%B4%DB%8C-409621764.html> (failed 3 times): 429 Unknown Status
2022-06-01 22:06:12 [scrapy.downloadermiddlewares.retry] ERROR: Gave up retrying <GET https://www.sheypoor.com/%DA%A9%D8%A7%D8%B1-%D8%A8%D8%A7-%DA%AF%D9%88%D8%B4%DB%8C-409621764.html> (failed 3 times): 429 Unknown Status
INFO
retry_logs last 10 of 16
ignore_logs 3 in total
DEBUG

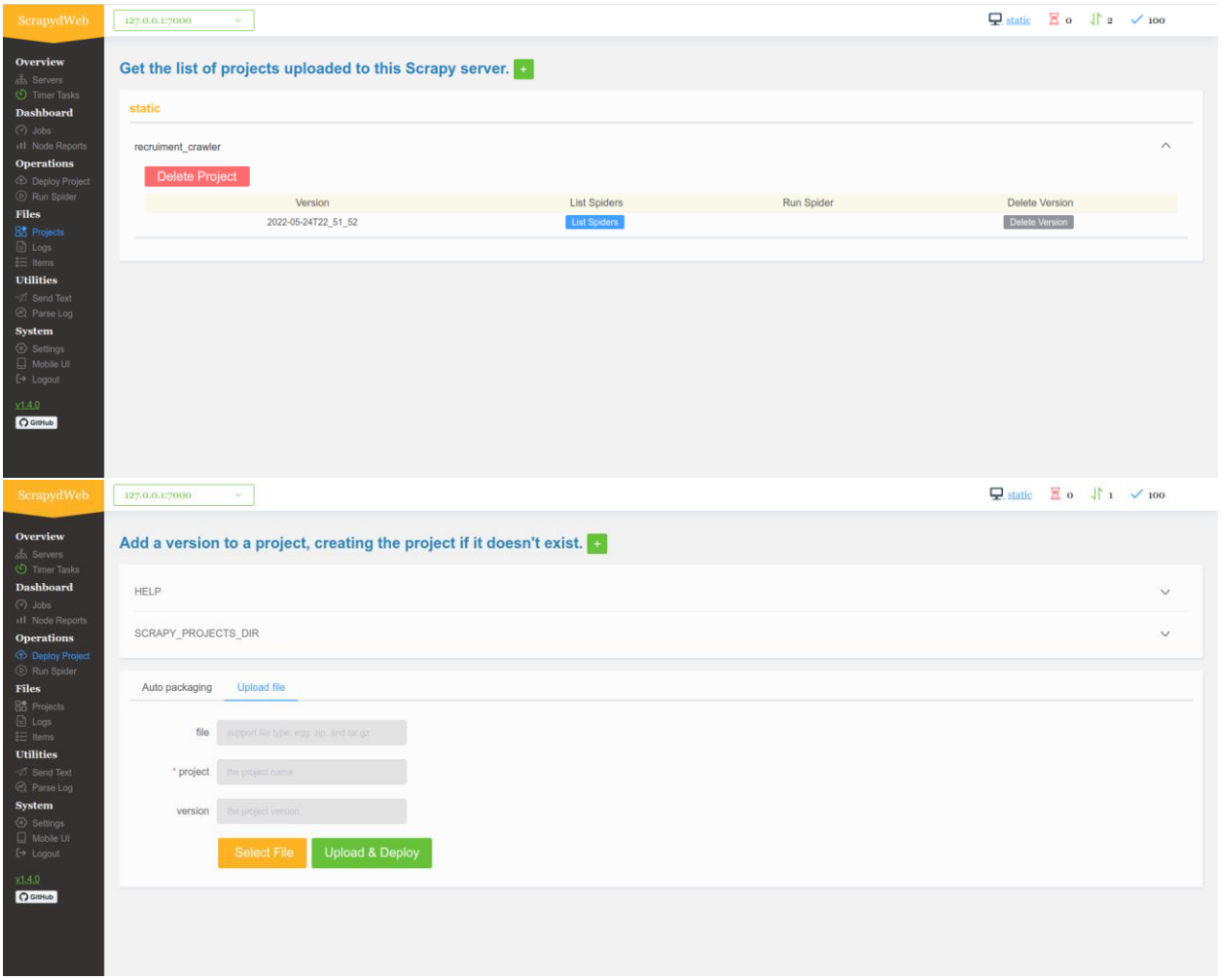


Get the reports of running and finished jobs of all projects after Scrapyd server started. +

Fetch remaining reports

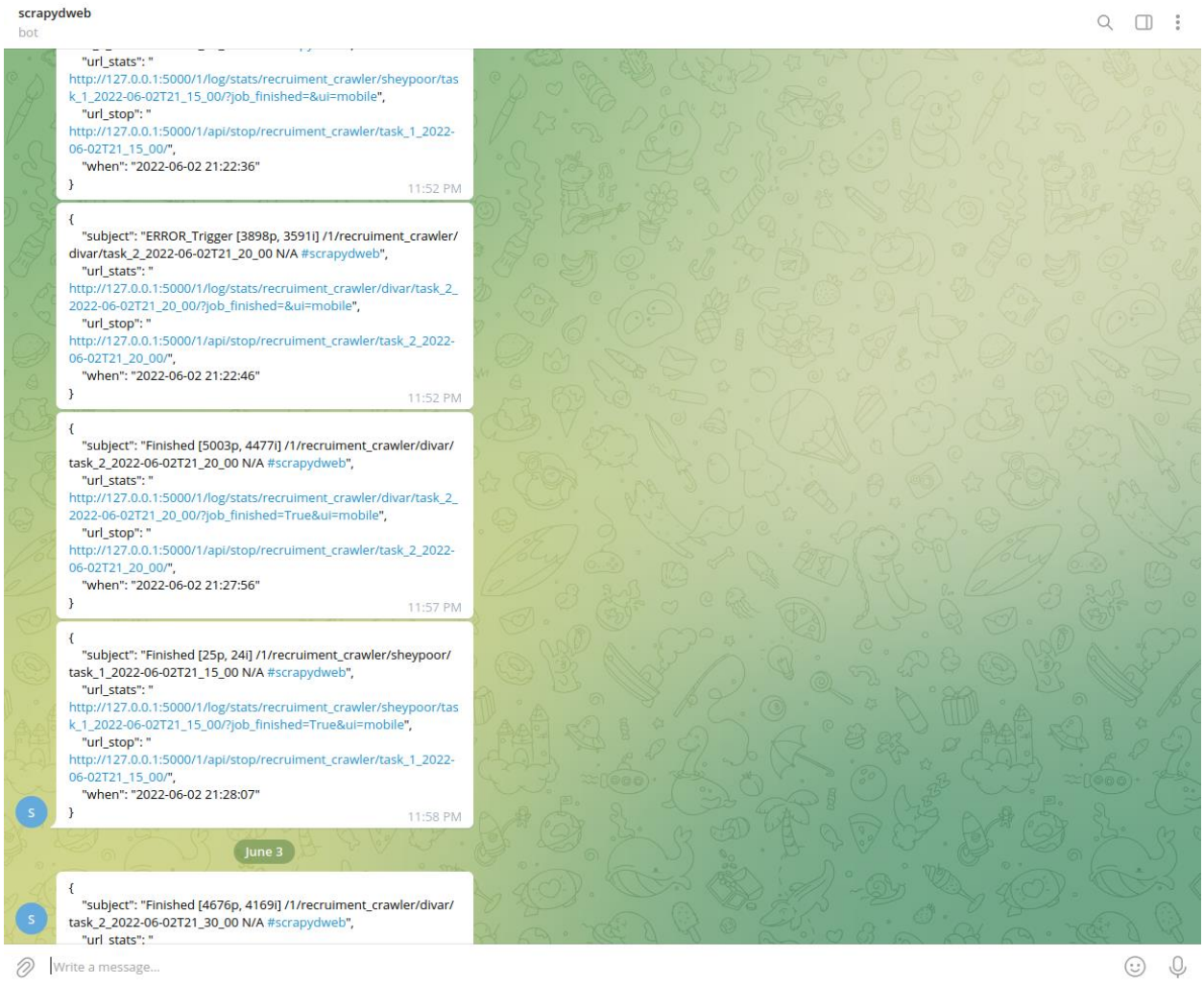
Index	Stats	Spider	Pages	Items	Runtime	Reason	Critical	Error	Warning	Redirect	Retry	Ignore	latest_item
1	Stats	sheypoor	6	5	0:02:11	N/A	0	0	0	0	8	0	N/A
100	Stats	divar	4990	4477	0:03:18	finished	0	242	0	0	583	213	N/A
99	Stats	sheypoor	25	24	0:03:33	finished	0	0	0	0	0	0	N/A
98	Stats	divar	4788	4265	0:03:08	finished	0	247	0	0	571	207	N/A
97	Stats	divar	4585	4077	0:03:13	finished	0	235	0	0	584	209	N/A
96	Stats	sheypoor	25	24	0:04:22	finished	0	0	0	0	6	0	N/A
95	Stats	divar	5507	4962	0:04:19	finished	0	238	0	0	563	198	N/A
94	Stats	sheypoor	25	24	0:04:03	finished	0	0	0	0	4	0	N/A




ScrapydWeb
127.0.0.1:7000
static 0 2 100
Overview
Servers
Timer Tasks
Dashboard
Jobs
Node Reports
Operations
Deploy Project
Run Spider
Files
Projects
Logs
Items
Utilities
Send Text
Parse Log
System
Settings
Mobile UI
Logout
v1.4.0
GitHub
Get the list of projects uploaded to this Scrapy server.
static
recruiment_crawler
Delete Project
Version
List Spiders
Run Spider
Delete Version
2022-05-24T22_51_52
List Spiders
Delete Version
ScrapydWeb
127.0.0.1:7000
static 0 1 100
Add a version to a project, creating the project if it doesn't exist.
HELP
SCRAPY_PROJECTS_DIR
Auto packaging
Upload file
file
support file type: egg, zip, and tar.gz
* project
the project name
version
the project version
Select File
Upload & Deploy

triggers


```
{
  "ON_JOB_RUNNING_INTERVAL": 300,
  "ON_JOB_FINISHED": True,
  "CRITICAL": {
    "LOG_CRITICAL_THRESHOLD": 3,
    "LOG_CRITICAL_TRIGGER_STOP": False,
    "LOG_CRITICAL_TRIGGER_FORCESTOP": False
  },
  "ERROR": {
    "LOG_ERROR_THRESHOLD": 50,
    "LOG_ERROR_TRIGGER_STOP": False,
    "LOG_ERROR_TRIGGER_FORCESTOP": False
  },
  "WARNING": {
    "LOG_WARNING_THRESHOLD": 0,
    "LOG_WARNING_TRIGGER_STOP": False,
    "LOG_WARNING_TRIGGER_FORCESTOP": False
  },
  "REDIRECT": {
    "LOG_REDIRECT_THRESHOLD": 0,
    "LOG_REDIRECT_TRIGGER_STOP": False,
    "LOG_REDIRECT_TRIGGER_FORCESTOP": False
  },
  "RETRY": {
    "LOG_RETRY_THRESHOLD": 0,
    "LOG_RETRY_TRIGGER_STOP": False,
    "LOG_RETRY_TRIGGER_FORCESTOP": False
  },
  "IGNORE": {
    "LOG_IGNORE_THRESHOLD": 10,
    "LOG_IGNORE_TRIGGER_STOP": False,
    "LOG_IGNORE_TRIGGER_FORCESTOP": False
  }
}
```


scrapydweb
bot
"url_stats": "
http://127.0.0.1:5000/1/log/stats/recruiment_crawler/sheypoor/task_1_2022-06-02T21_15_00/?job_finished=&ui=mobile",
"url_stop": "
http://127.0.0.1:5000/1/api/stop/recruiment_crawler/task_1_2022-06-02T21_15_00/",
"when": "2022-06-02 21:22:36"
}
11:52 PM
{
"subject": "ERROR_Trigger [3898p, 3591i] /1/recruiment_crawler/divar/task_2_2022-06-02T21_20_00 N/A #scrapydweb",
"url_stats": "
http://127.0.0.1:5000/1/log/stats/recruiment_crawler/divar/task_2_2022-06-02T21_20_00/?job_finished=&ui=mobile",
"url_stop": "
http://127.0.0.1:5000/1/api/stop/recruiment_crawler/task_2_2022-06-02T21_20_00/",
"when": "2022-06-02 21:22:46"
}
11:52 PM
{
"subject": "Finished [5003p, 4477i] /1/recruiment_crawler/divar/task_2_2022-06-02T21_20_00 N/A #scrapydweb",
"url_stats": "
http://127.0.0.1:5000/1/log/stats/recruiment_crawler/divar/task_2_2022-06-02T21_20_00/?job_finished=True&ui=mobile",
"url_stop": "
http://127.0.0.1:5000/1/api/stop/recruiment_crawler/task_2_2022-06-02T21_20_00/",
"when": "2022-06-02 21:27:56"
}
11:57 PM
{
"subject": "Finished [25p, 24i] /1/recruiment_crawler/sheypoor/task_1_2022-06-02T21_15_00 N/A #scrapydweb",
"url_stats": "
http://127.0.0.1:5000/1/log/stats/recruiment_crawler/sheypoor/task_1_2022-06-02T21_15_00/?job_finished=True&ui=mobile",
"url_stop": "
http://127.0.0.1:5000/1/api/stop/recruiment_crawler/task_1_2022-06-02T21_15_00/",
"when": "2022-06-02 21:28:07"
}
11:58 PM
S
June 3
S
{
"subject": "Finished [4676p, 4169i] /1/recruiment_crawler/divar/task_2_2022-06-02T21_30_00 N/A #scrapydweb",
"url_stats": "
Write a message...