#### AIL Framework for Analysis of Information Leaks

workshop - A generic analysis information leak open source software



## Alexandre Dulaunoy

alexandre.dulaunoy@circl.lu

Sami Mokaddem

Aurélien Thirion aurelien.thirion@circl.lu

info@circl.lu

@SUNET 20180207

Objectives of the workshop

#### Our objectives of the workshop

- Demonstrate why data-analysis is critical in information security
- Explain challenges and the design of the AIL framework
- Learn how to install and start AIL
- · Learn how to properly feed AIL with custom data
- Learn how to manage current modules
- Learn how to create new modules
- Practical part: Workshop

# Sources of leaks

## Sources of leaks: Paste monitoring

- Example: http://pastebin.com/
  - Easily storing and sharing text online
  - Used by programmers and legitimate users
    - → Source code & information about configurations

## Sources of leaks: Paste monitoring

- Example: http://pastebin.com/
  - Easily storing and sharing text online
  - Used by programmers and legitimate users
    - $\rightarrow$  Source code & information about configurations
- Abused by attackers to store:
  - List of vulnerable/compromised sites
  - Software vulnerabilities (e.g. exploits)
  - Database dumps
    - $\rightarrow$  User data
    - $\rightarrow$  Credentials
    - → Credit card details
  - More and more ...

#### Examples of pastes

```
text 2.02 KB
text 4.41 KB
                                                                                                                                                            KillerGram - Yuffie - Smoke The Big Dick [smkwhr] (Upload
                          - - - - Tool by Y3t1v3t ( u
                                                                                                                                                             Danish the Control of the Control of
                            text 4.57 KB
                                                                                                                                                                 text 2.66 KB

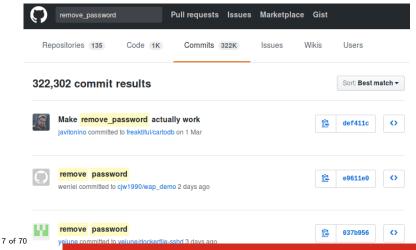
    <item name="%the component to be disabled%" xsi:type="array">

                                    1. #include "wejwyj.h"
                                                                                                                                                                                                <item name="config" xsi:type="array">
                                                                                                                                                                                                             <item name="componentDisabled" xsi:type="boolean">true</item>
                                   3. int zapisz (FILE *plik_
                                                                                                                                                                                          </item>
                                                  int i, j;
                                                                                                                                                                          5. </item>
                                   5. if (obr->KOLOR==0) {
                                                                                                                                                                         7. <?xml version="1.0"?>
                                                                                                                                              10.
    10.
                                                     fprintf (plik wv. "P2
                                                     fprintf (plik wv. "%d
                                                                                                                                                                         9. <page xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespace
                                                     fprintf (plik wv. "%d
                                                                                                                                                                                    /etc/page configuration.xsd">
                                                     for (i=0: i<obr->wvmv
                                                                                                                                                                      10.
                                                                                                                                                                                                <body>
                                                       for (i=0: i<obr->wvmx: i++
                                                                                                                                                                                                             <referenceBlock name="checkout.root">
                                                              fprintf (plik wy, "%d ",
                                                                                                                                                                                                                          <arguments>
                                                                                                                                                                                                                                       <argument name="jsLayout" xsi:type="array">
```

#### Sources of leaks: Others

Mistakes from users

 $\ \ \, \circ \ \, \text{https://github.com/Search?} q = remove\_password \& \textit{type} = \textit{Commits\&ref} = \textit{searchresults} \\$ 



#### Sources of leaks: Others

Mistakes from users

 $\ \ \, \circ \ \, \text{https://github.com/Search?} q = remove\_password \& \textit{type} = \textit{Commits\&ref} = \textit{searchresults} \\$ 



### Why so many leaks?

- Economical interests (e.g. Adversaries promoting services)
- Political motives (e.g. Adversaries showing off)
- Collaboration (e.g. Criminals need to collaborate)
- Operational infrastructure (e.g. malware exfiltrating information on a pastie website)
- Mistakes and Errors

#### Yes!

and we have to deal with this as a CSIRT.

- Contacting companies or organisations who did specific accidental leaks
- Discussing with media about specific case of leaks and how to make it more practical/factual for everyone
- Evaluating the economical market for cyber criminals (e.g. DDoS booters<sup>1</sup> or reselling personal information reality versus media coverage)
- Analysing collateral effects of malware, software vulnerabilities or exfiltration
  - $\rightarrow$  And it's important to detect them automatically.

#### Paste monitoring at CIRCL: Statistics

- Monitored paste sites: 27
  - o pastebin.com
  - ideone.com

o ...

	2016	2017	08.2018
Collected pastes	18,565,124	19,145,300	11,591,987
Incidents	244	266	208

Table: Pastes collected and incident<sup>2</sup> raised by CIRCL

<sup>2</sup>http://www.circl.lu/pub/tr-46

#### Privacy, AIL and GDPR

- Many modules in AIL can process personal data and even special categories of data as defined in GDPR (Art. 9).
- The data controller is often the operator of the AIL framework (limited to the organisation) and has to define legal grounds for processing personal data.
- To help users of AIL framework, a document is available which describe points of AIL in regards to the regulation<sup>3</sup>.

<sup>3</sup>https:

## Potential legal grounds

- Consent of the data subject is in many cases not feasible in practice and often impossible or illogical to obtain (Art. 6(1)(a)).
- Legal obligation (Art. 6(1)(c)) This legal ground applies mostly to CSIRTs, in accordance with the powers and responsibilities set out in CSIRTs mandate and with their constituency, as they may have the legal obligation to collect, analyse and share information leaks without having a prior consent of the data subject.
- Art. 6(1)(f) Legitimate interest Recital 49 explicitly refers to CSIRTs' right to process personal data provided that they have a legitimate interest but not colliding with fundamental rights and freedoms of data subject.

# AIL Framework

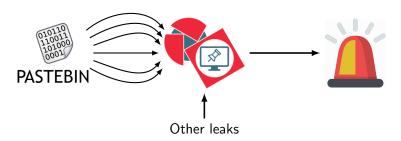
#### From a requirement to a solution: AIL Framework

#### History:

- AlL initially started as an **internship project** (2014) to evaluate the feasibility to automate the analysis of (un)structured information to find leaks.
- In 2018, AIL framework is an open source software in Python. The software is actively used (and maintained) by CIRCL.

# AIL Framework: A framework for Analysis of Information Leaks

"AIL is a modular framework to analyse potential information leaks from unstructured data sources like pastes from Pastebin."



#### AIL Framework: Current capabilities

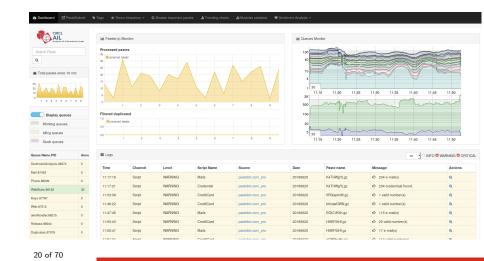
- Extending AIL to add a new analysis module can be done in 50 lines of Python
- The framework **supports multi-processors/cores by default**. Any analysis module can be started multiple times to support faster processing during peak times or bulk import
- Multiple concurrent data input

#### AIL Framework: Current features

- Extracting credit cards numbers, credentials, phone numbers,
   ...
- Extracting and validating potential hostnames
- Keeps track of duplicates
- Submission to threat sharing and incident response platform (MISP and TheHive)
- Full-text indexer to index unstructured information
- Tagging for classification and searches
- Terms, sets and regex tracking and occurences
- Archives, files and raw submission from the UI
- Sentiment/Mood analyser for incoming data
- And many more

# Live demo!

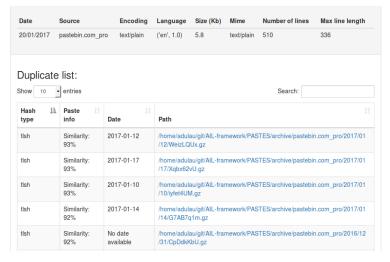
#### Example: Following a notification (0) - Dashboard



# Example: Following a notification (1) - Searching



## Example: Following a notification (2) - Metadata



# Example: Following a notification (3) - Browsing content

#### Content:

```
http://members2.mofosnetwork.com/access/login/
somosextremos:buddy1990
brazzers_glenn:cocklick
hrazzers61:hraves01
http://members.naughtyamerica.com/index.php?m=login
gernblanston:3unc2352
Janhuss141200:310575
igetalliwant:1377zeph
pwilks89:mon22key
Bman1551:hockey
MoFos IKnowThatGirl PublicPickUps
http://members2.mofos.com
Chrismagg40884:loganm40
brando1:zzbrando1
aacoen:1q2w3e4r
1rstunkle23:mv8self
Bra77ers
http://ma.brazzers.com
qcjensen:qcj21pva
skycsc17:rbcdnd
                                 >| Get Daily Update Fresh Porn Password Here |<
                                              http://www.erg.1o/4mF1
```

# Example: Following a notification (3) - Browsing content

#### Content:

```
Over 50000+ custom hacked xxx passwords by us! Thousands of free xxx passwords to the hottest paysites!
>| Get Fresh New Premium XXX Site Password Here |<
    http://www.erg.io/4mF1
http://ddfnetwork.com/home.html
eu172936:hCSBaKh
UecwB6zs:159X0$!r#6K78FuU
http://pornxn.stiffia.com/user/login
feldwWek8939:RObluJ8XtB
dabudka:17891789
braiits:braiits1
http://members.pornstarplatinum.com/sbloqin/loqin.php/
gigiriveracom:xxxjay
jayx123:xxxjay69
http://members.vividceleb.com/
Rufio99:fairhaven
ScH1FRv1:102091
Chaos84:HOLE5244
Riptor705.bl
```

# Setting up the framework

# Setting up AIL-Framework from source or virtual machine

# Setting up AlL-Framework from source 1 git clone https://github.com/CIRCL/AIL-framework.git 2 cd AIL-framework 3 ./installing\_deps.sh 4 cd var/www/ 5 ./update\_thirdparty.sh

#### Using the virtual machine:

- Download https://www.circl.lu/assets/files/ ail-training/AIL\_v@4986352.ova
- 2. Start virtualbox
- 3. File  $\rightarrow$  import appliance  $\rightarrow$  select AIL\_June.ova
- 4. (for now) Prevent the automatic launch and git pull the changes  $\frac{26}{20}$  of 70

AIL ecosystem - Challenges and design

#### AIL ecosystem: Technologies used

Programing language: Full python3

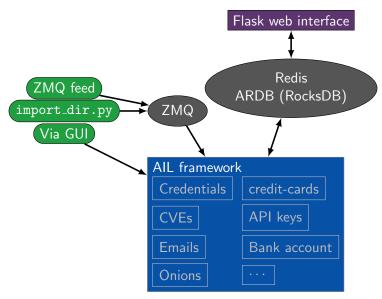
Databases: Redis and ARDB

Server: Flask

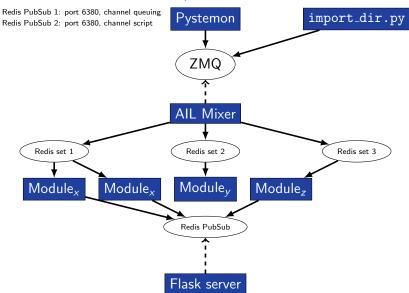
Data message passing: ZMQ, Redis list and Redis

Publisher/Subscriber

#### AIL global architecture 1/2



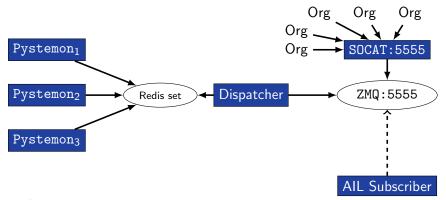
#### AIL global architecture 2/2



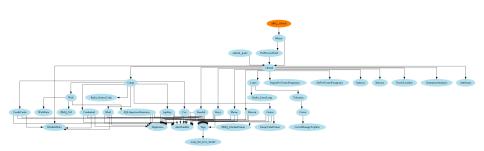
## Data feeder: Gathering pastes with pystemon

#### Pystemon global architecture

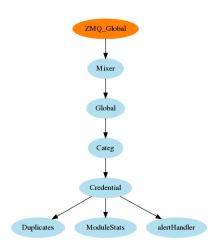
Redis PubSub 1: port 6380, channel queuing Redis PubSub 2: port 6380, channel script



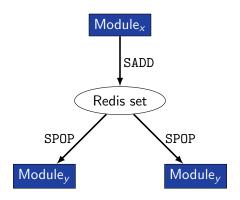
# AIL global architecture: Data streaming between module



# AIL global architecture: Data streaming between module (Credential example)



## Message consuming



- ightarrow No message lost nor double processing
- $\rightarrow$  Multiprocessing!

#### Web crawler

- Web crawler is used to crawl regular website as well as .onion addresses
- Splash (scriptable browser) is rending the pages (including javascript) and produce screenshots (HAR archive too)

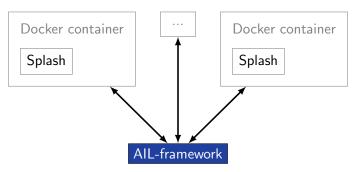


Figure: Architecture of AIL and its hidden services crawler

# Starting the framework

### Running your own instance from source

```
Make sure that ZMQ_Global \rightarrow address = tcp://crf.circl.lu:5556,tcp://127.0.0.1:5556 in bin/package/config.cfg
```

```
Accessing the environment and starting AIL

1  # Activate the virtualenv
2  . ./AILENV/bin/activate

3  # Launch the system
5 cd bin/
6  ./LAUNCH -1

7  8  # Will also start the web interface
```

### Running your own instance using the virtual machine

#### Login and passwords:

```
Web interface (default network settings):
    http://192.168.56.51:7000/
Shell/SSH:
    ail/Password1234
```

Feeding the framework

### Feeding AIL

There are differents way to feed AIL with data:

- 1. Be a trusted partner with CIRCL and ask to get access to our feed info@circl.lu
- 2. Setup pystemon and use the custom feeder
  - o pystemon will collect pastes for you
- Feed your own data using the import\_dir.py script
- Feed your own file/text using the UI (/PasteSubmit/)

### Feeding AIL

There are differents way to feed AIL with data:

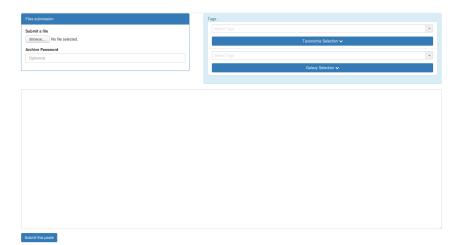
- 1. CIRCL trusted partners can ask to access our feed info@circl.lu
- 2. Setup pystemon and use the custom feeder
  - o pystemon will collect pastes for you
- Feed your own file/text using the UI (/PasteSubmit/)
- 4. Feed your own data using the import\_dir.py script

### Plug-in AIL to the CIRCL feed

You can freely access the CIRCL feed during this workshop!

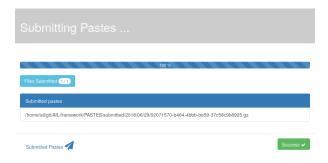
- In the file bin/package/config.cfg,
- Set ZMQ\_Global->address to tcp://crf.circl.lu:5556

# Via the UI (1)



43 of 70

# Via the UI (2)



# Feeding AIL with your own data - import\_dir.py (1)

/!\ 2 requirements:

- 1. Data to be fed must have the path hierarchy as the following:
  - 1.1 year/month/day/(textfile/gzfile)
  - 1.2 This is due to the inner representation of paste in AIL
- 2. Each file to be fed must be of a raisonable size:
  - $2.1 \sim 3$  Mb is already large
  - 2.2 This is because some modules are doing regex matching
  - 2.3 If you want to feed a large file, better split it in multiple ones

# Feeding AIL with your own data - import\_dir.py (2)

- 1. Check your local configuration bin/package/config.cfg
  - In the file bin/package/config.cfg,
  - Add 127.0.0.1:5556 in ZMQ\_Global
  - (should already be set by default)

# Feeding AIL with your own data - import\_dir.py (2)

- 1. Check your local configuration bin/package/config.cfg
  - In the file bin/package/config.cfg,
  - Add 127.0.0.1:5556 in ZMQ\_Global
  - (should already be set by default)
- 2. Launch import\_dir.py with de directory you want to import
  - o import\_dir.py -d dir\_path

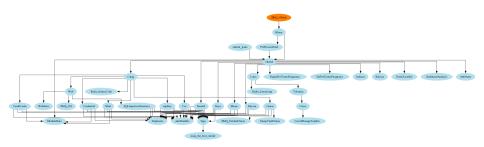
# Feeding AIL with your own data - import\_dir.py (2)

- 1. Check your local configuration bin/package/config.cfg
  - In the file bin/package/config.cfg,
  - Add 127.0.0.1:5556 in ZMQ\_Global
  - (should already be set by default)
- 2. Launch import\_dir.py with de directory you want to import
  - o import\_dir.py -d dir\_path
- 3. Watch your data being feed to AIL

# Creating new features

# Developping new features: Plug-in a module in the system

Choose where to put your module in the data flow:



Then, modify bin/package/modules.cfg accordingly

# Writing your own modules - /bin/template.py

```
import time
   from pubsublogger import publisher
   from Helper import Process
   if __name__ == '__main__':
       # Port of the redis instance used by pubsublogger
6
       publisher.port = 6380
 7
       # Script is the default channel used for the modules.
       publisher.channel = 'Script'
       # Section name in bin/packages/modules.cfg
10
       config_section = '<section name>'
11
       # Setup the I/O queues
12
       p = Process(config_section)
13
       # Sent to the logging a description of the module
14
       publisher.info("<description of the module>")
15
       # Endless loop getting messages from the input queue
16
       while True:
17
           # Get one message from the input queue
18
           message = p.get_from_set()
19
           if message is None:
20
               publisher.debug("{} queue is empty, waiting".format(config_section))
21
               time.sleep(1)
22
               continue
23
           # Do something with the message from the queue
24
           something has been done = do something(message)
25
    49 of 70
```

### AIL - Add your own web interface

- 1. Launch var/www/create\_new\_web\_module.py
- 2. Enter the module's name
- 3. A template and flask skeleton has been created for your new webpage in var/www/modules/
- 4. You can start **coding** server-side in:
  - var/www/modules/your\_module\_name/Flask\_your\_module\_name.py
- 5. You can start **coding** client-side in:
  - var/www/modules/your\_module\_name/templates/your\_module\_name.html

var/www/modules/your\_module\_name/templates/header\_your\_module\_name.html

# Case study: Push alert to MISP

#### Push alert to MISP







Goal: push tags to MISP.

#### Push alert to MISP





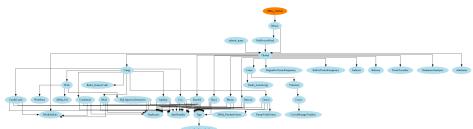


- 1. Use infoleak taxonomy<sup>4</sup>
- 2. Add your own tags
- 3. Create an event on a paste

<sup>4</sup>https://www.misp-project.org/taxonomies.html

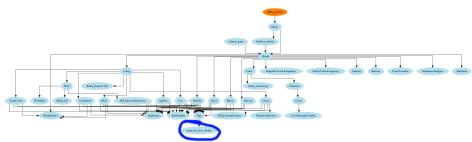
# Case study: Finding the best place in the system

#### Best place to put it?



# Case study: Finding the best place in the system

#### Best place to put it?



### Case study: Updating Flask server.py

#### Flask server.py

```
[...]
               INITIAL tags auto export
  r_serv_db = redis.StrictRedis(
      host=cfg.get("ARDB DB", "host"),
      port=cfg.getint("ARDB_DB", "port"),
      db=cfg.getint("ARDB_DB", "db"),
6
      decode responses=True)
8 infoleak_tags = taxonomies.get('infoleak').machinetags()
9 infoleak_automatic_tags = []
10 for tag in taxonomies.get('infoleak').machinetags():
      if tag.split('=')[0][:] == 'infoleak:automatic-detection':
11
12
          r_serv_db.sadd('list_export_tags', tag)
13
14 r_serv_db.sadd('list_export_tags', 'infoleak:submission="manual"')
15 r_serv_db.sadd('list_export_tags', '<your_tag>')
16
```

### Auto Push Tags









#### Create an event



#### Duplicate list:



Hash type ↓1		Paste info	Date	Path			
['tlsh']		Similarity: [59]%	2018-05-30	/home/aurelien/git/python3/AlL-framework/PASTES/archive/pastebin.com_pro/2018/05/30/ePtpckUe.gz			

Showing 1 to 1 of 1 entries

#### Content:

[Raw content]

powershell -noP -sta -w 1 -enc JABHAFIATwBVAFAAUABVAEwAaQBDAHKAUwBFAFQAVABJAG4ARwBzACAAPQAgAFsacgBFAEYAXQAUAEEAUwBTAGUADQBCAGwAeQAuAEcAZQBBAFQAeQBwAGUAKAANAF

#### Create an event



W 1 -enc JABHAFIATWBVAFAAUABVAEWABOBDAHKAUWBFAFOAVABJAG4ARWBZACAAPOAGAFSACGBFAEYAXOAUAEEAUWBTAGUADOBCAGWAEGAZOBOAFOAEOBWAGUAKAANAFMAGOBZAHOAZOBtAC4ATOBhAG4AYOBNAGUAHOBIAG4AdaA

# Practical part

### Practical part: Pick your choice

- 1. Update support of docker/ansible
- 2. Graph database on Credential.py
  - o Top used passwords, most compromised user, ...
- 3. Webpage scrapper
  - Download html from URL found in pastes
  - Re-inject html as paste in AIL
- 4. Improvement of Phone.py
  - Way to much false positive as of now. Exploring new ways to validate phone numbers could be interesting
- 5. Your custom feature

# Contribution rules



# Glimpse of contributed features

- Docker
- Ansible
- Email alerting
- SQL injection detection
- Phone number detection

• Feel free to fork the code, play with it, make some patches or add additional analysis modules.

- Feel free to fork the code, play with it, make some patches or add additional analysis modules.
- Feel free to make a pull request for your contribution

- Feel free to fork the code, play with it, make some patches or add additional analysis modules.
- Feel free to make a pull request for your contribution
- That's it!



#### Final words

- Building AIL helped us to find additional leaks which cannot be found using manual analysis and improve the time to detect duplicate/recycled leaks.
  - $\rightarrow$  Therefore quicker response time to assist and/or inform proactively affected constituents.

# Annexes

# Managing the framework

# Managing AIL: Old fashion way

#### Access the script screen

1 screen -r Script

#### Table: GNU screen shortcuts

Shortcut	Action					
C-a d	detach screen					
C-a c	Create new window					
C-a n	next window screen					
C-a p	previous window screen					

# Managing your modules: Using the helper

creen(1	ModuleInformation)									🤝 En 🕴 🖂 🖭	· •(ii) 00:24 😃
Action <a></a> <a> <a>&lt;</a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a></a>	Queue name Attributes BrowschardingPaste Catog Catog Credital Creditards DonClassifier Indoar	PID 31731 31952 31766 31822 31783 31755 31876 31744 31784 31932 31888 31894 31941 31775 31818 31922		\$ Tue	R Tine 0.00:01 0.00:01 0.00:00 0.00:00 0.00:00 0.00:00 0.00:00 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01 0.00:01	Processed claimst Gribnog Gribnog PPDBLED Hebbert He		CPU x 3.100 (c. 600 (c	1.50% 1.43% 1.64% 1.63% 1.60% 1.64% 1.57% 0.43% 1.57% 0.438 1.59% 1.64% 1.49% 1.57%	Ang CFUN 3.66% 3.66% 3.66% 3.70% 3.7	
Action <k> <k> <k> <k></k></k></k></k>			Idling Q dle Time :00:00 :00:00 :00:01	ueues Last paste hash nnDewerkX yCkUXRtp rhn2f3Yt		Action	Queue Curve CurveHanageTopSets Cve DumpValldOnton Duplicates Onton PreProcessFeed RegexForTermsFrequency SentimentAnalysis SetforTermsFrequency	State Stuck or idle, Not running by Stuck or idle, Not running by Stuck or idle, Stuck or idle, Not running by	default restarting di default restarting di restarting di default restarting di restarting di	isabled isabled isabled isabled isabled	
						Tine 00:23:2 00:23:2 00:23:2 00:23:2 00:23:2 00:23:1	9 SentimentAnalysi 9 RegexForTermsFre 9 Curve 9 SetForTermsFrequ	31725 s 31961 equency 31852 31837 sency 31864	*invalid pid i *id pid in MOI Cleared invali	id pid in MODULE_TYPE_Dupli in MODULE_TYPE_SentimentAma XULE_TYPE_RegexForTernsFreq di pid in MODULE_TYPE_Curve MODULE_TYPE_SetForTermsFreq module_info	lysis uency

0:24 0% bash [1 ModuleInformation] 2-\$ Mixer 3% Global 4% Duplicates 5% Attributes 6% Lines 7% DomClassifier 8% Categ 9% Tokenize 10% CreditCards 11% Onion 12% Mail 13% Heb 14% Creden