Collection of malware dynamic analysis data for implementing behavioral malware detection based on Recurrent neural network

Kamalov Rustem

Gorodetskiy Aleksey

Creed Dmitriy

Goals

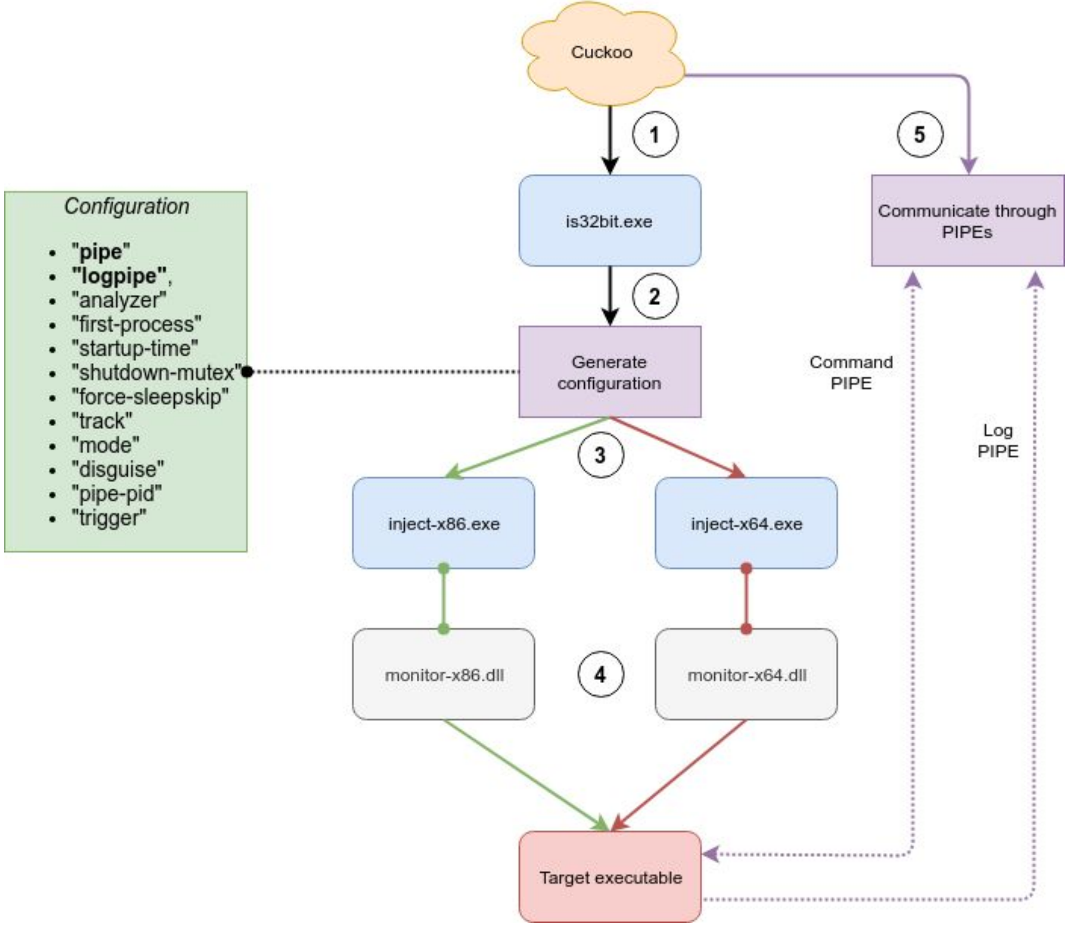
- **Analyze** which dynamic data can be acquired from executables;
- **Investigate** open-source solution(-s) that perform dynamic analysis of malware;
- **Conduct** dynamic analysis data mining from malware;
- **Prepare** data for training in a recurrent neural network;
- **Train** and test RNN and evaluate results;
- (Optional) **Build** a proof of concept application;

Behavioral data

- Assembly code
  - Instruction set
  - Slow to process
- API Calls
  - Variety of calls
  - Arguments
- Side-effects (Memory-, CPU-, Network-, Disk- usage)

How Cuckoo does it?

1. Is target 32 or 64 bit?
2. Generate PIPE and other parameters;
3. Use appropriate **injector** with config as argument;
4. Appropriate **monitor** injected into target EXE;
5. Communicate with monitor through PIPE.



Pipes

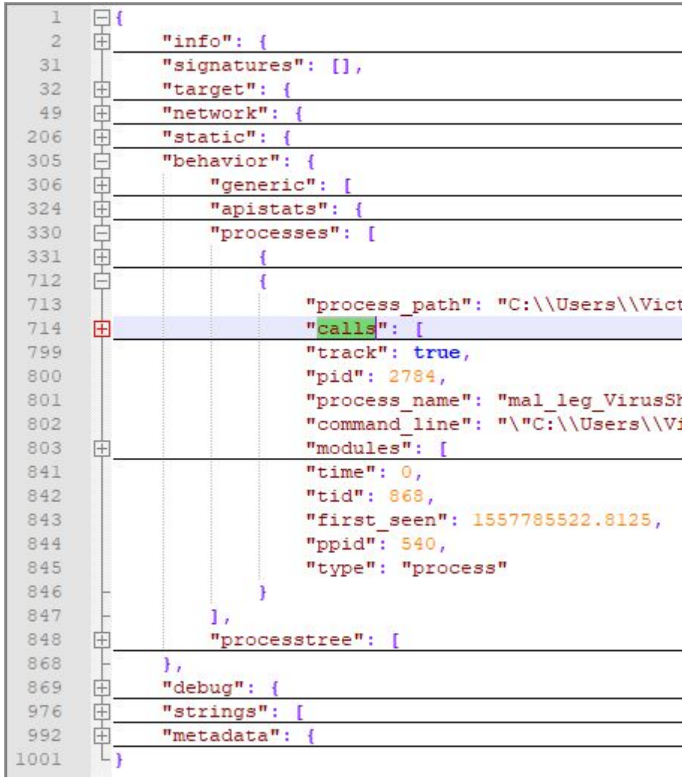
Command pipe:

- bi-directional;
- sends commands;
- commands handled by analyzer.

Logging pipe:

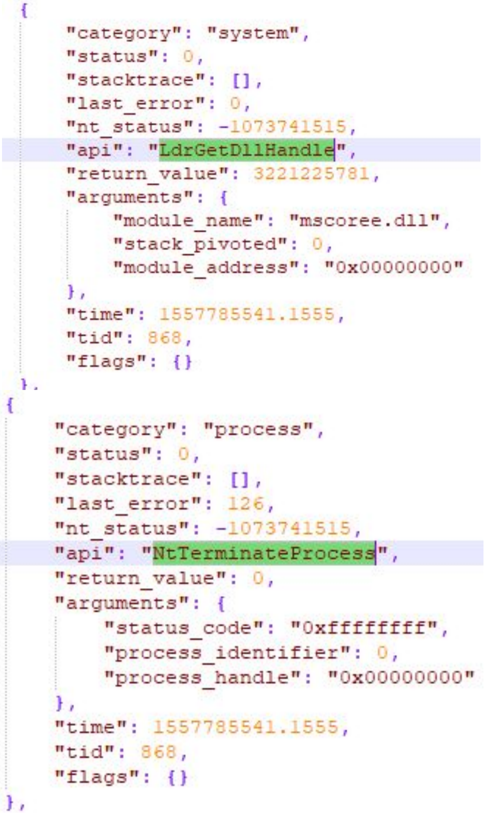
- unidirectional;
- only sends logs from monitor;
- forwarded to the log parser.

JSON Reports



```
{
    "info": {
    "signatures": [],
    "target": {
    "network": {
    "static": {
    "behavior": {
        "generic": [
        "apistats": {
        "processes": [
            {
            {
                "process_path": "C:\\Users\\Vict
                "calls": [
                "track": true,
                "pid": 2784,
                "process_name": "mal_leg_VirusSh
                "command_line": "\"C:\\Users\\Vi
                "modules": [
                "time": 0,
                "tid": 868,
                "first_seen": 1557785522.8125,
                "ppid": 540,
                "type": "process"
            }
        ],
        "processtree": [
    },
    "debug": {
    "strings": [
    "metadata": {
}
```

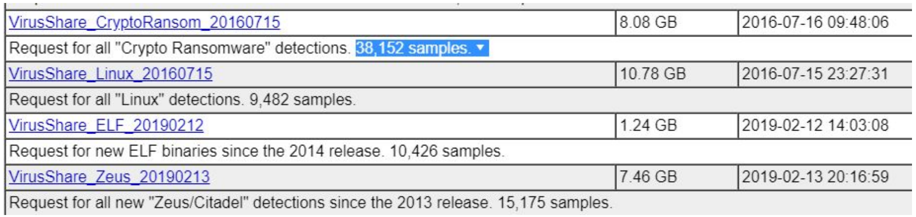
Process behaviour



```
{
    "category": "system",
    "status": 0,
    "stacktrace": [],
    "last_error": 0,
    "nt_status": -1073741515,
    "api": "LdrGetDllHandle",
    "return_value": 3221225781,
    "arguments": {
        "module_name": "mscoree.dll",
        "stack_pivoted": 0,
        "module_address": "0x00000000"
    },
    "time": 1557785541.1555,
    "tid": 868,
    "flags": {}
},
{
    "category": "process",
    "status": 0,
    "stacktrace": [],
    "last_error": 126,
    "nt_status": -1073741515,
    "api": "NtTerminateProcess",
    "return_value": 0,
    "arguments": {
        "status_code": "0xffffffff",
        "process_identifier": 0,
        "process_handle": "0x00000000"
    },
    "time": 1557785541.1555,
    "tid": 868,
    "flags": {}
},
```

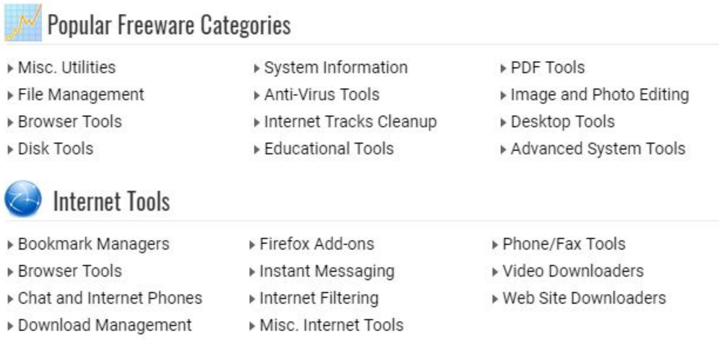
API calls of process

Collecting data



VirusShare_CryptoRansom_20160715	8.08 GB	2016-07-16 09:48:06
Request for all "Crypto Ransomware" detections. 38,152 samples. ▾		
VirusShare_Linux_20160715	10.78 GB	2016-07-15 23:27:31
Request for all "Linux" detections. 9,482 samples.		
VirusShare_ELF_20190212	1.24 GB	2019-02-12 14:03:08
Request for new ELF binaries since the 2014 release. 10,426 samples.		
VirusShare_Zeus_20190213	7.46 GB	2019-02-13 20:16:59
Request for all new "Zeus/Citadel" detections since the 2013 release. 15,175 samples.		

virusshare.com



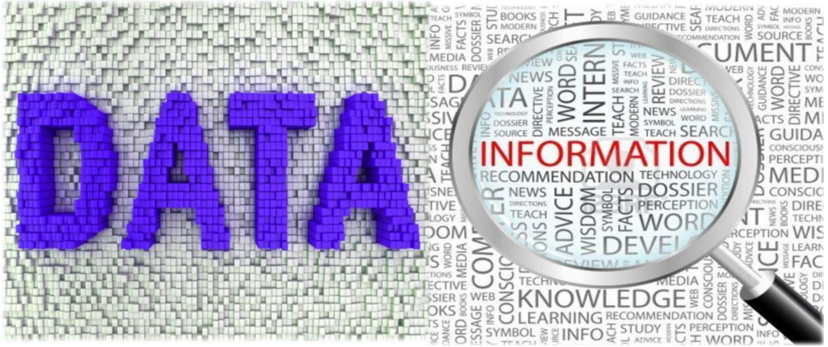
Popular Freeware Categories

- Misc. Utilities
- File Management
- Browser Tools
- Disk Tools
- System Information
- Anti-Virus Tools
- Internet Tracks Cleanup
- Educational Tools
- PDF Tools
- Image and Photo Editing
- Desktop Tools
- Advanced System Tools

Internet Tools

- Bookmark Managers
- Browser Tools
- Chat and Internet Phones
- Download Management
- Firefox Add-ons
- Instant Messaging
- Internet Filtering
- Misc. Internet Tools
- Phone/Fax Tools
- Video Downloaders
- Web Site Downloaders

www.snapfiles.com



Data

3000 - malicious files

2000 - legal files

3,5 min - time for a report

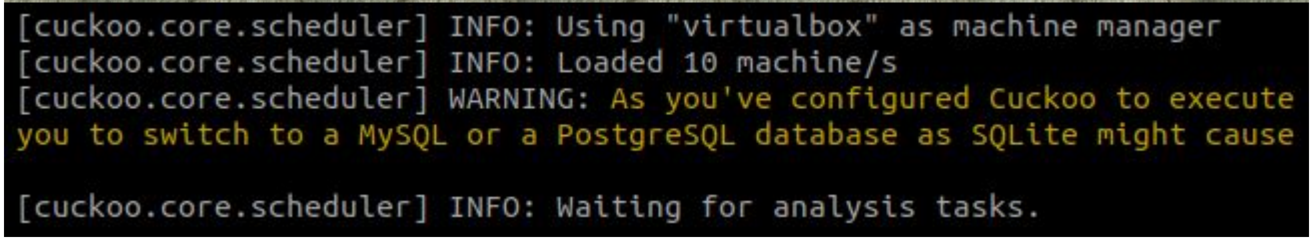
5000 * 3,5 = 17500 min =

= 12 days

Solution

10

linked clones



```
[cuckoo.core.scheduler] INFO: Using "virtualbox" as machine manager
[cuckoo.core.scheduler] INFO: Loaded 10 machine/s
[cuckoo.core.scheduler] WARNING: As you've configured Cuckoo to execute
you to switch to a MySQL or a PostgreSQL database as SQLite might cause

[cuckoo.core.scheduler] INFO: Waiting for analysis tasks.
```

Fun (no)

- Submitting
- RAM
- No responding
- Unknown errors

Submitting

Cuckoo CLI

- crashed
- no automatic way to continue

Cuckoo Web

- more than 50 files make Chrome crash
- more than 50 files make Firefox show error

API

Submitting one by one

- No response - **sandbox crashing**
- Attempt to continue - **different sorting**
- Crash after submitting - **empty reports**

Cuckoo sandbox phases

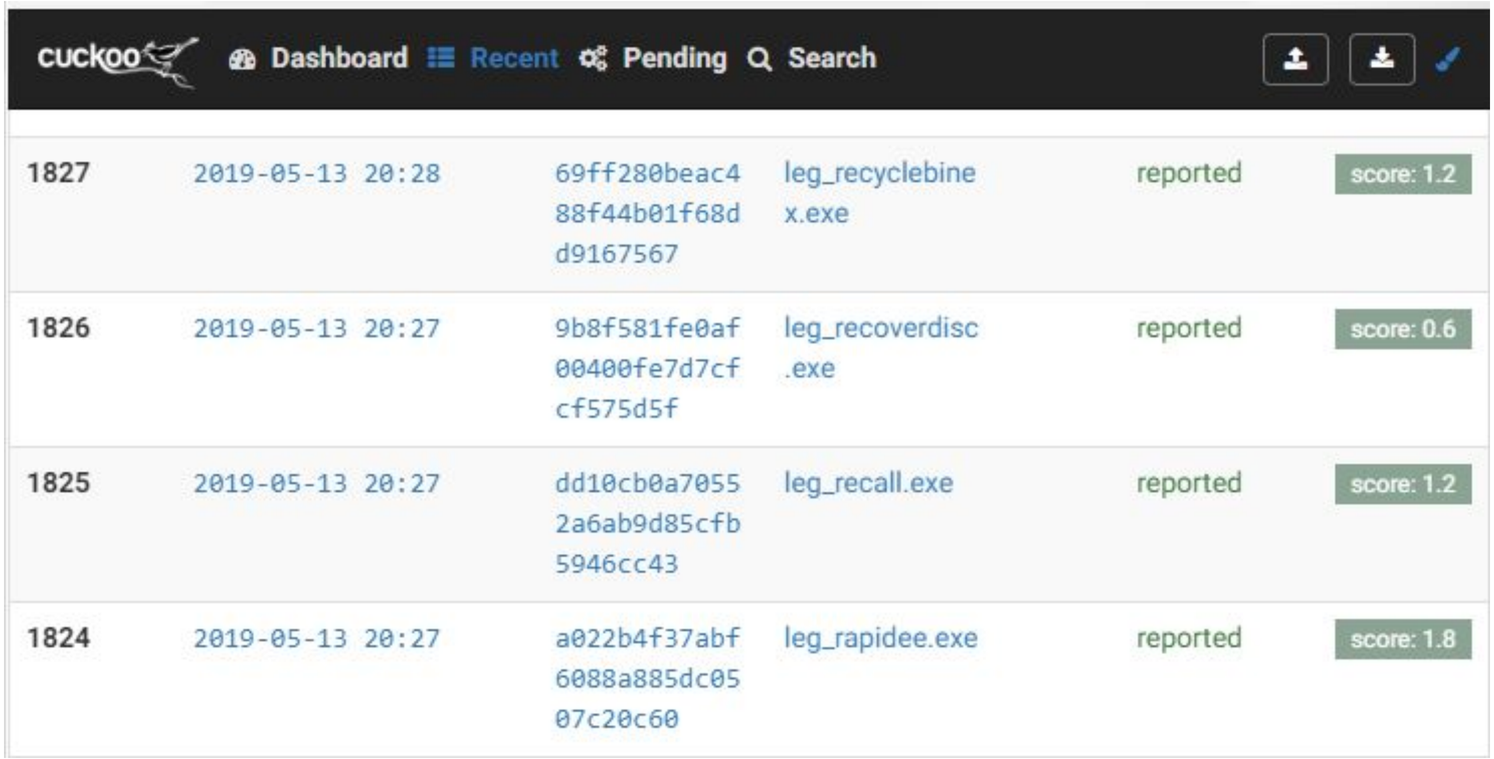
- Analysis
- Execution
- Processing
- Reporting

The road so far

- gathered all analysis logs and data
- went throw all errors
- created reports by multiple processes
- gathered data that we need

Legitimate files

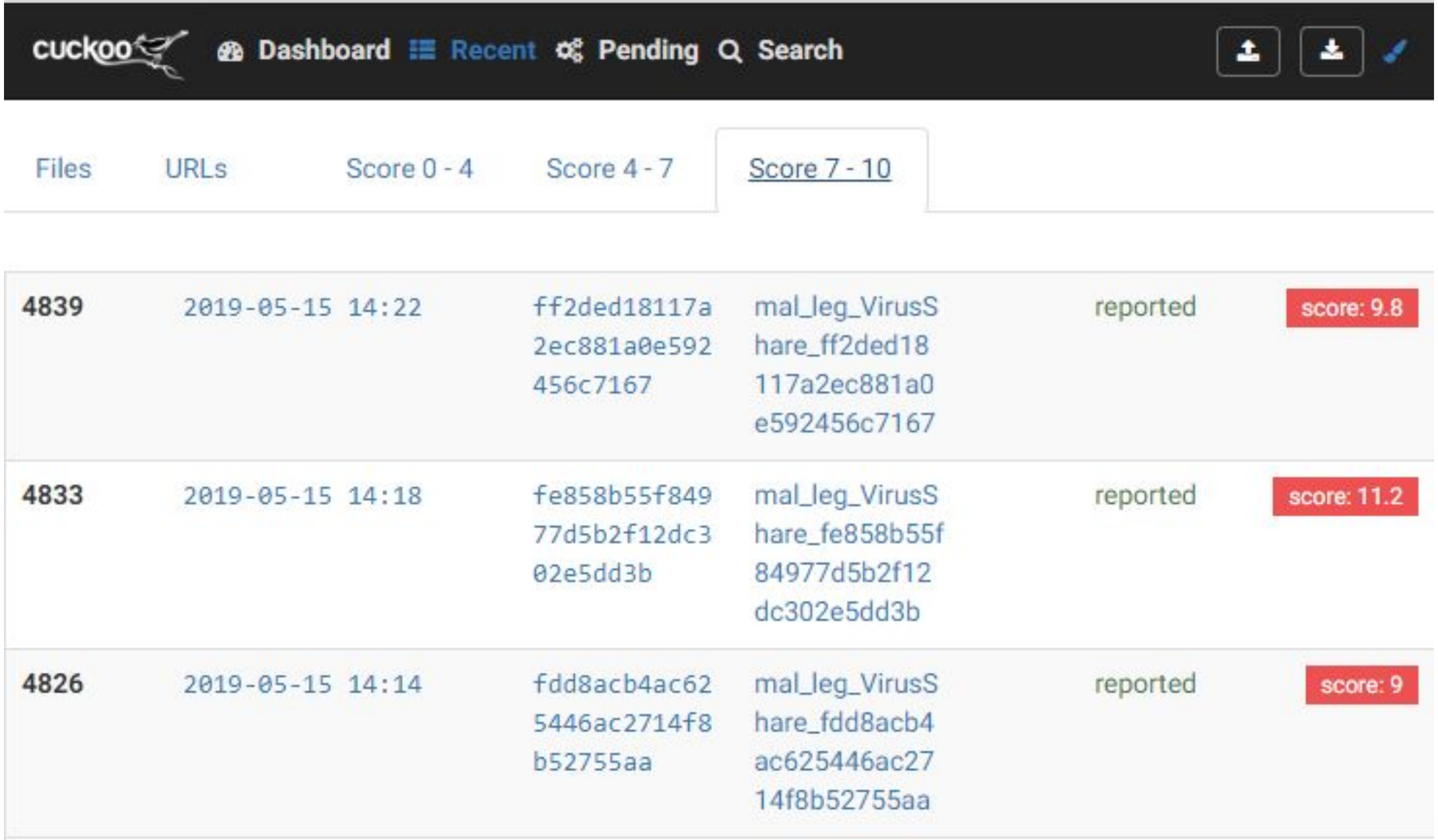
Legitimate files: **2196**



1827	2019-05-13 20:28	69ff280beac488f44b01f68dd9167567	leg_recyclebinex.exe	reported	score: 1.2
1826	2019-05-13 20:27	9b8f581fe0af00400fe7d7cfcf575d5f	leg_recoverdisc.exe	reported	score: 0.6
1825	2019-05-13 20:27	dd10cb0a70552a6ab9d85cfb5946cc43	leg_recall.exe	reported	score: 1.2
1824	2019-05-13 20:27	a022b4f37abf6088a885dc0507c20c60	leg_rapidee.exe	reported	score: 1.8

Malwares

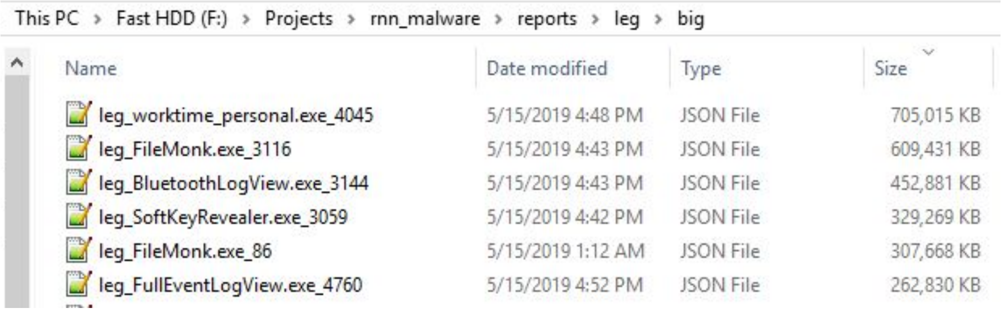
Malicious files: **2884**



4839	2019-05-15 14:22	ff2ded18117a2ec881a0e592456c7167	mal_leg_VirusShare_ff2ded18117a2ec881a0e592456c7167	reported	score: 9.8
4833	2019-05-15 14:18	fe858b55f84977d5b2f12dc302e5dd3b	mal_leg_VirusShare_fe858b55f84977d5b2f12dc302e5dd3b	reported	score: 11.2
4826	2019-05-15 14:14	fdd8acb4ac625446ac2714f8b52755aa	mal_leg_VirusShare_fdd8acb4ac625446ac2714f8b52755aa	reported	score: 9

Remove big reports

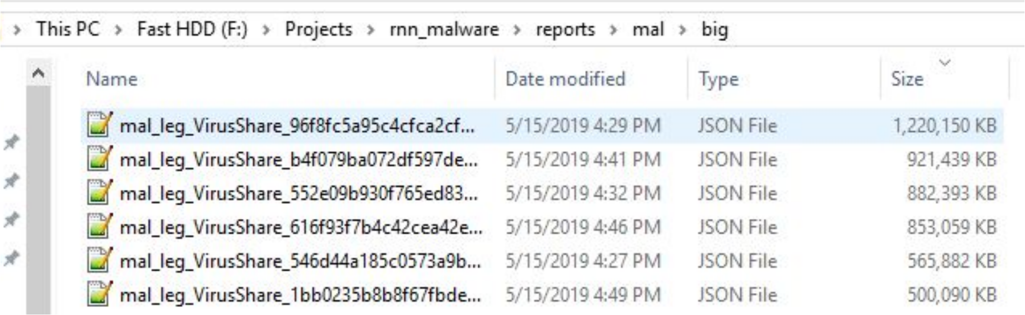
-10.5 Gb,
total 6.74 Gb



This PC › Fast HDD (F:) › Projects › rnn_malware › reports › leg › big

Name	Date modified	Type	Size
leg_worktime_personal.exe_4045	5/15/2019 4:48 PM	JSON File	705,015 KB
leg_FileMonk.exe_3116	5/15/2019 4:43 PM	JSON File	609,431 KB
leg_BluetoothLogView.exe_3144	5/15/2019 4:43 PM	JSON File	452,881 KB
leg_SoftKeyRevealer.exe_3059	5/15/2019 4:42 PM	JSON File	329,269 KB
leg_FileMonk.exe_86	5/15/2019 1:12 AM	JSON File	307,668 KB
leg_FullEventLogView.exe_4760	5/15/2019 4:52 PM	JSON File	262,830 KB

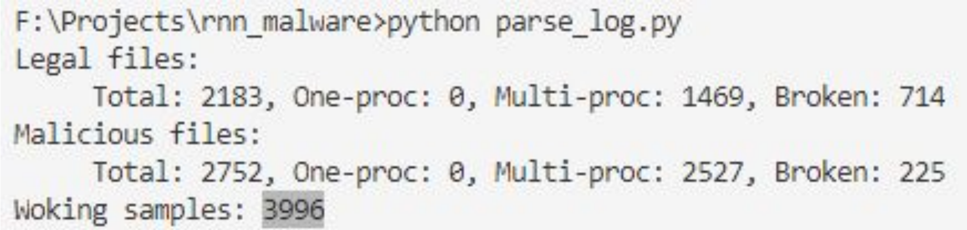
-**32.1** Gb,
total 16.1 Gb



› This PC › Fast HDD (F:) › Projects › rnn_malware › reports › mal › big

Name	Date modified	Type	Size
mal_leg_VirusShare_96f8fc5a95c4cfca2cf...	5/15/2019 4:29 PM	JSON File	1,220,150 KB
mal_leg_VirusShare_b4f079ba072df597de...	5/15/2019 4:41 PM	JSON File	921,439 KB
mal_leg_VirusShare_552e09b930f765ed83...	5/15/2019 4:32 PM	JSON File	882,393 KB
mal_leg_VirusShare_616f93f7b4c42cea42e...	5/15/2019 4:46 PM	JSON File	853,059 KB
mal_leg_VirusShare_546d44a185c0573a9b...	5/15/2019 4:27 PM	JSON File	565,882 KB
mal_leg_VirusShare_1bb0235b8b8f67fbde...	5/15/2019 4:49 PM	JSON File	500,090 KB

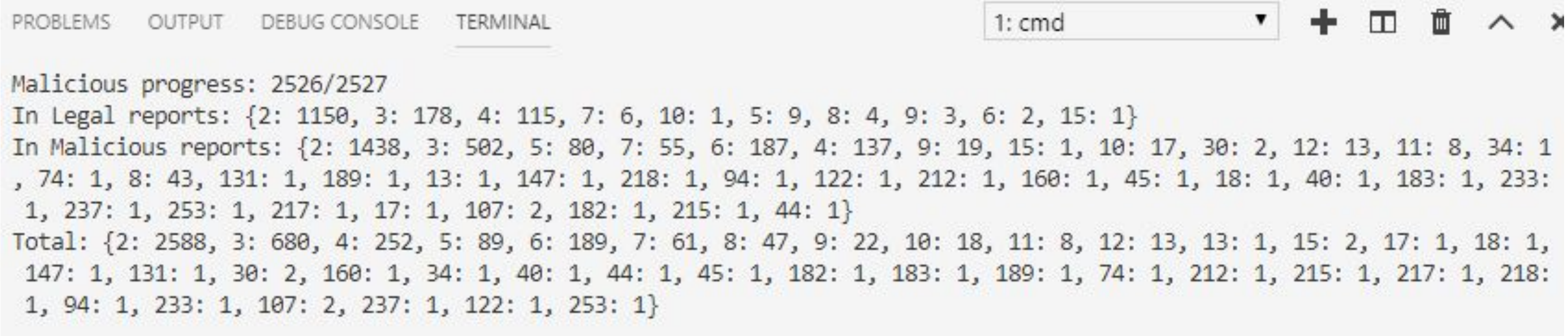
Empty calls



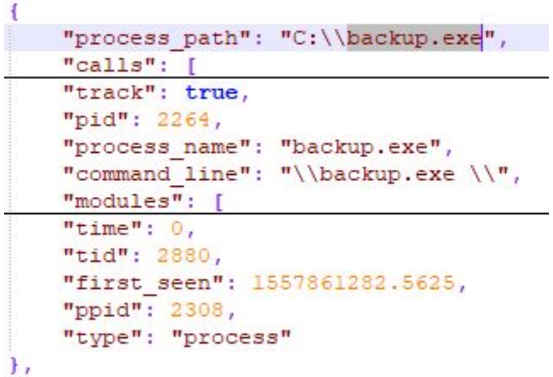
```
F:\Projects\rnn_malware>python parse_log.py
Legal files:
     Total: 2183, One-proc: 0, Multi-proc: 1469, Broken: 714
Malicious files:
     Total: 2752, One-proc: 0, Multi-proc: 2527, Broken: 225
Woking samples: 3996
```

~**4k** reports in total

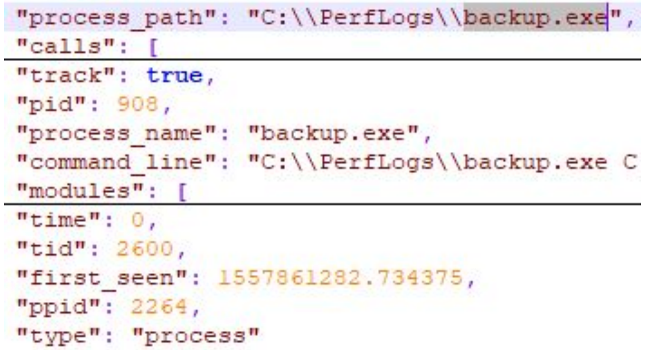
Spawning processes



```
Malicious progress: 2526/2527
In Legal reports: {2: 1150, 3: 178, 4: 115, 7: 6, 10: 1, 5: 9, 8: 4, 9: 3, 6: 2, 15: 1}
In Malicious reports: {2: 1438, 3: 502, 5: 80, 7: 55, 6: 187, 4: 137, 9: 19, 15: 1, 10: 17, 30: 2, 12: 13, 11: 8, 34: 1
, 74: 1, 8: 43, 131: 1, 189: 1, 13: 1, 147: 1, 218: 1, 94: 1, 122: 1, 212: 1, 160: 1, 45: 1, 18: 1, 40: 1, 183: 1, 233:
 1, 237: 1, 253: 1, 217: 1, 17: 1, 107: 2, 182: 1, 215: 1, 44: 1}
Total: {2: 2588, 3: 680, 4: 252, 5: 89, 6: 189, 7: 61, 8: 47, 9: 22, 10: 18, 11: 8, 12: 13, 13: 1, 15: 2, 17: 1, 18: 1,
 147: 1, 131: 1, 30: 2, 160: 1, 34: 1, 40: 1, 44: 1, 45: 1, 182: 1, 183: 1, 189: 1, 74: 1, 212: 1, 215: 1, 217: 1, 218:
 1, 94: 1, 233: 1, 107: 2, 237: 1, 122: 1, 253: 1}
```

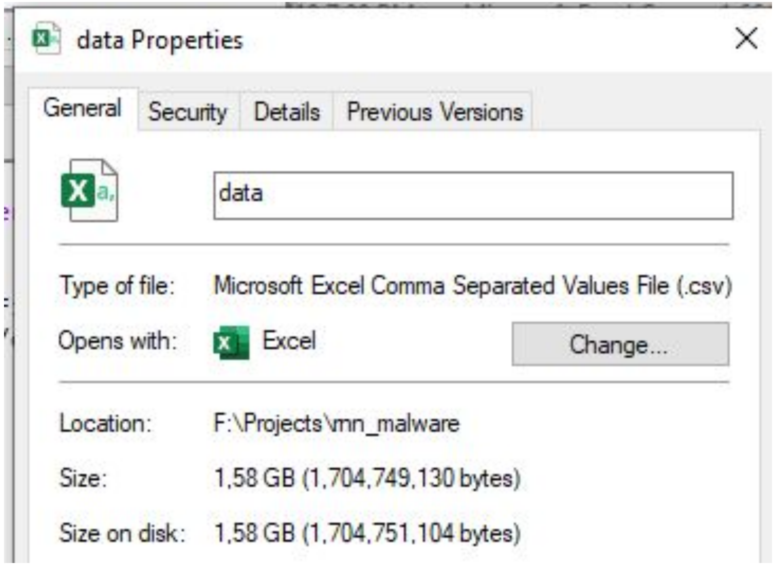


```
{
    "process_path": "C:\\backup.exe",
    "calls": [
    "track": true,
    "pid": 2264,
    "process_name": "backup.exe",
    "command_line": "\\backup.exe \\",
    "modules": [
    "time": 0,
    "tid": 2880,
    "first_seen": 1557861282.5625,
    "ppid": 2308,
    "type": "process"
},
```



```
"process_path": "C:\\PerfLogs\\backup.exe",
"calls": [
"track": true,
"pid": 908,
"process_name": "backup.exe",
"command_line": "C:\\PerfLogs\\backup.exe C
"modules": [
"time": 0,
"tid": 2600,
"first_seen": 1557861282.734375,
"ppid": 2264,
"type": "process"
```

Dataset ver. 1



data

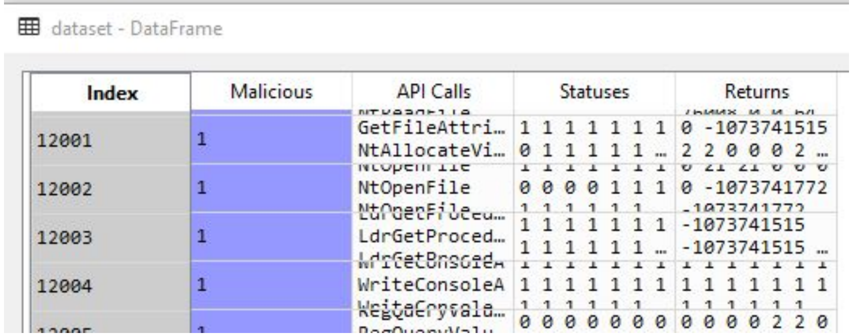
Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel

Location: F:\Projects\rnn_malware

Size: 1,58 GB (1,704,749,130 bytes)

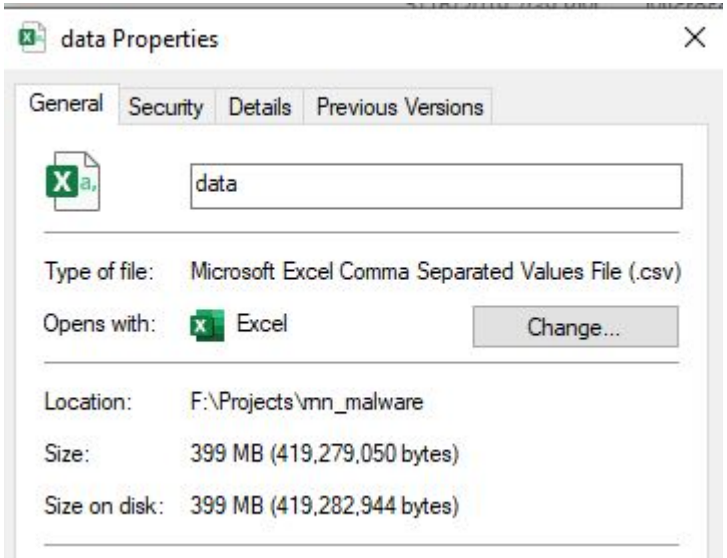
Size on disk: 1,58 GB (1,704,751,104 bytes)



Index	Malicious	API Calls	Statuses	Returns
12001	1	GetFileAttri… NtAllocateVi…	1 1 1 1 1 1 1 0 1 1 1 1 1 …	0 -1073741515 2 2 0 0 0 2 …
12002	1	NtOpenFile	0 0 0 0 1 1 1	0 -1073741772
12003	1	LdrGetProced…	1 1 1 1 1 1 1 1 1 1 1 1 1 …	-1073741515 -1073741515 …
12004	1	WriteConsoleA	1 1 1 1 1 1 1	1 1 1 1 1 1 1
12005	1	RegQueryValu…	0 0 0 0 0 0 0	0 0 0 0 2 2 0

There were also status and return values

Dataset ver.2



data Properties

data

Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel

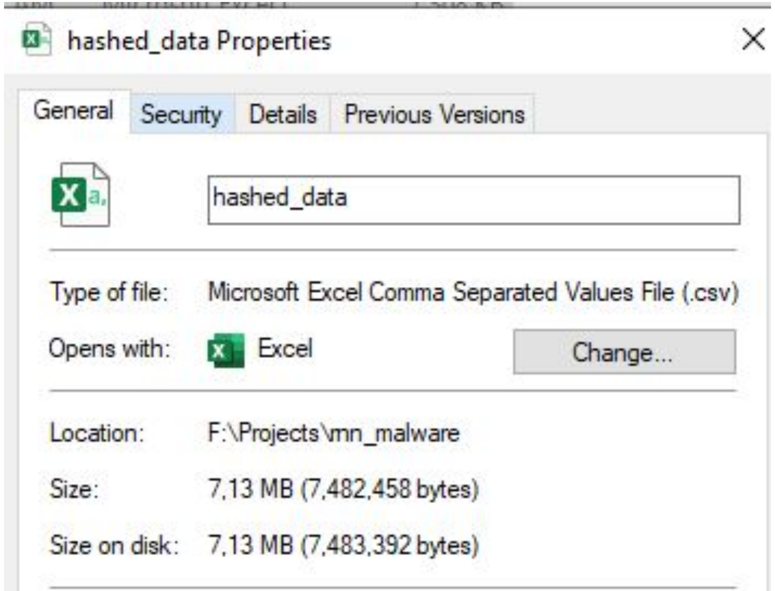
Location: F:\Projects\rnn_malware

Size: 399 MB (419,279,050 bytes)

Size on disk: 399 MB (419,282,944 bytes)

Only first 150 API calls

Dataset ver. 3



hashed_data Properties

hashed_data

Type of file: Microsoft Excel Comma Separated Values File (.csv)

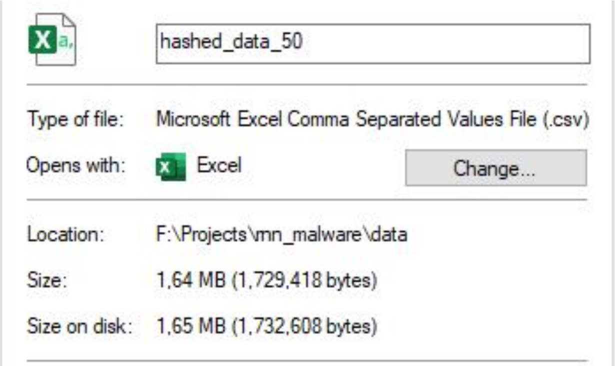
Opens with: Excel

Location: F:\Projects\rnn_malware

Size: 7,13 MB (7,482,458 bytes)

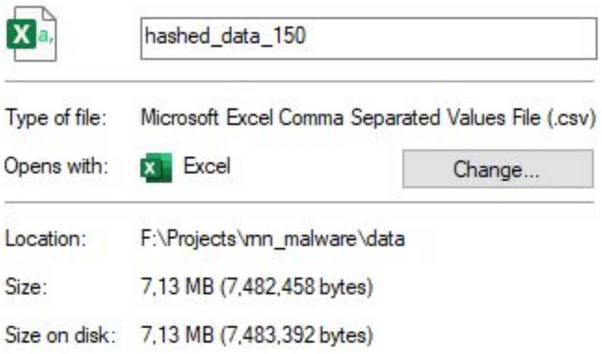
Size on disk: 7,13 MB (7,483,392 bytes)

Divide and conquer



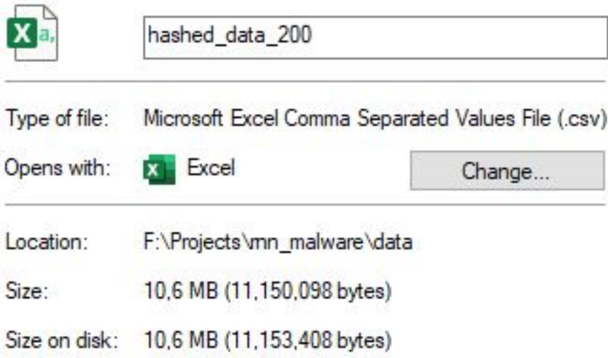
hashed_data_50

Type of file:	Microsoft Excel Comma Separated Values File (.csv)
Opens with:	Excel
Location:	F:\Projects\mn_malware\data
Size:	1,64 MB (1,729,418 bytes)
Size on disk:	1,65 MB (1,732,608 bytes)



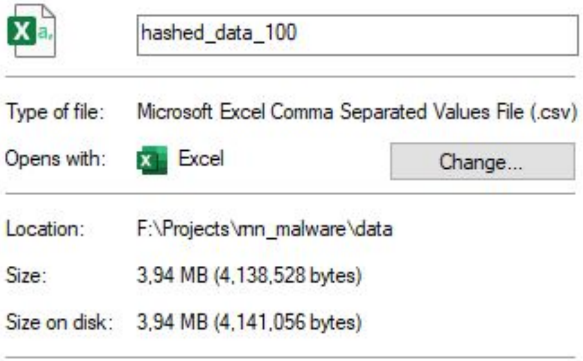
hashed_data_150

Type of file:	Microsoft Excel Comma Separated Values File (.csv)
Opens with:	Excel
Location:	F:\Projects\mn_malware\data
Size:	7,13 MB (7,482,458 bytes)
Size on disk:	7,13 MB (7,483,392 bytes)



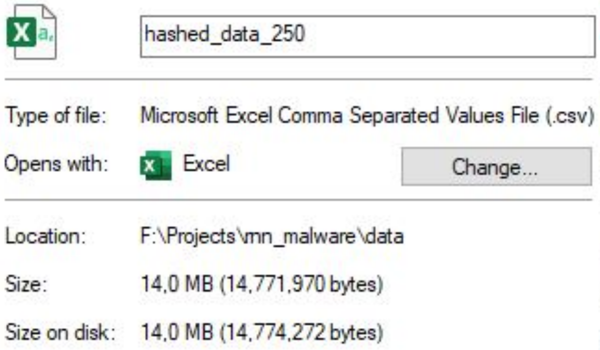
hashed_data_200

Type of file:	Microsoft Excel Comma Separated Values File (.csv)
Opens with:	Excel
Location:	F:\Projects\mn_malware\data
Size:	10,6 MB (11,150,098 bytes)
Size on disk:	10,6 MB (11,153,408 bytes)



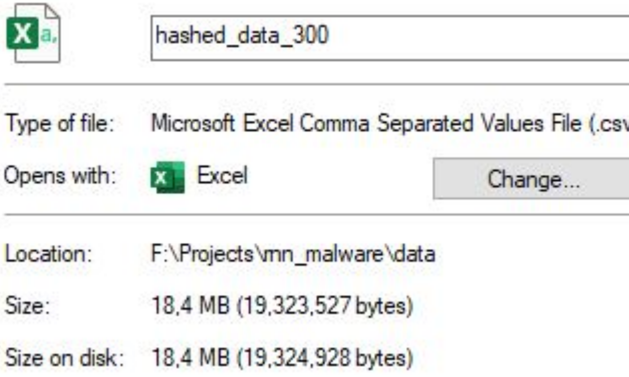
hashed_data_100

Type of file:	Microsoft Excel Comma Separated Values File (.csv)
Opens with:	Excel
Location:	F:\Projects\mn_malware\data
Size:	3,94 MB (4,138,528 bytes)
Size on disk:	3,94 MB (4,141,056 bytes)



hashed_data_250

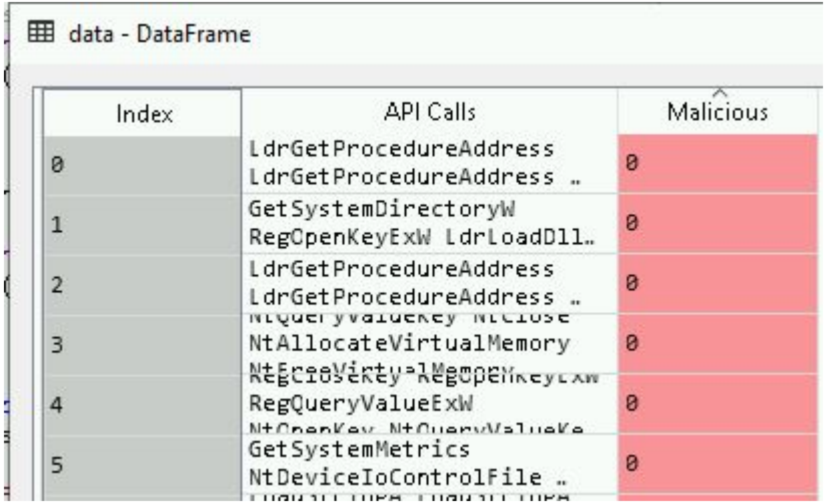
Type of file:	Microsoft Excel Comma Separated Values File (.csv)
Opens with:	Excel
Location:	F:\Projects\mn_malware\data
Size:	14,0 MB (14,771,970 bytes)
Size on disk:	14,0 MB (14,774,272 bytes)



hashed_data_300

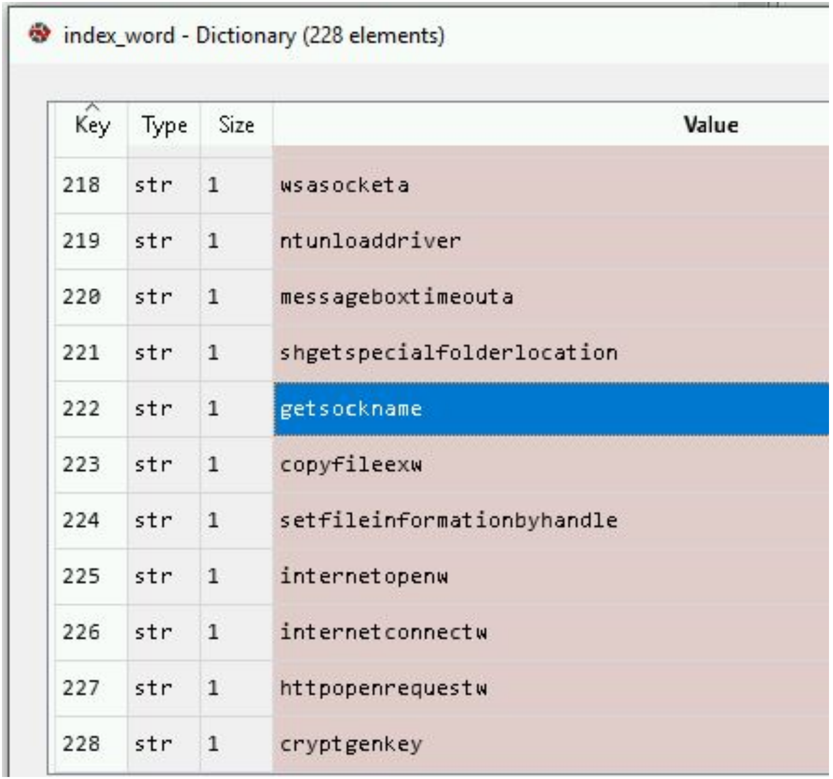
Type of file:	Microsoft Excel Comma Separated Values File (.csv
Opens with:	Excel
Location:	F:\Projects\mn_malware\data
Size:	18,4 MB (19,323,527 bytes)
Size on disk:	18,4 MB (19,324,928 bytes)

API Calls



data - DataFrame

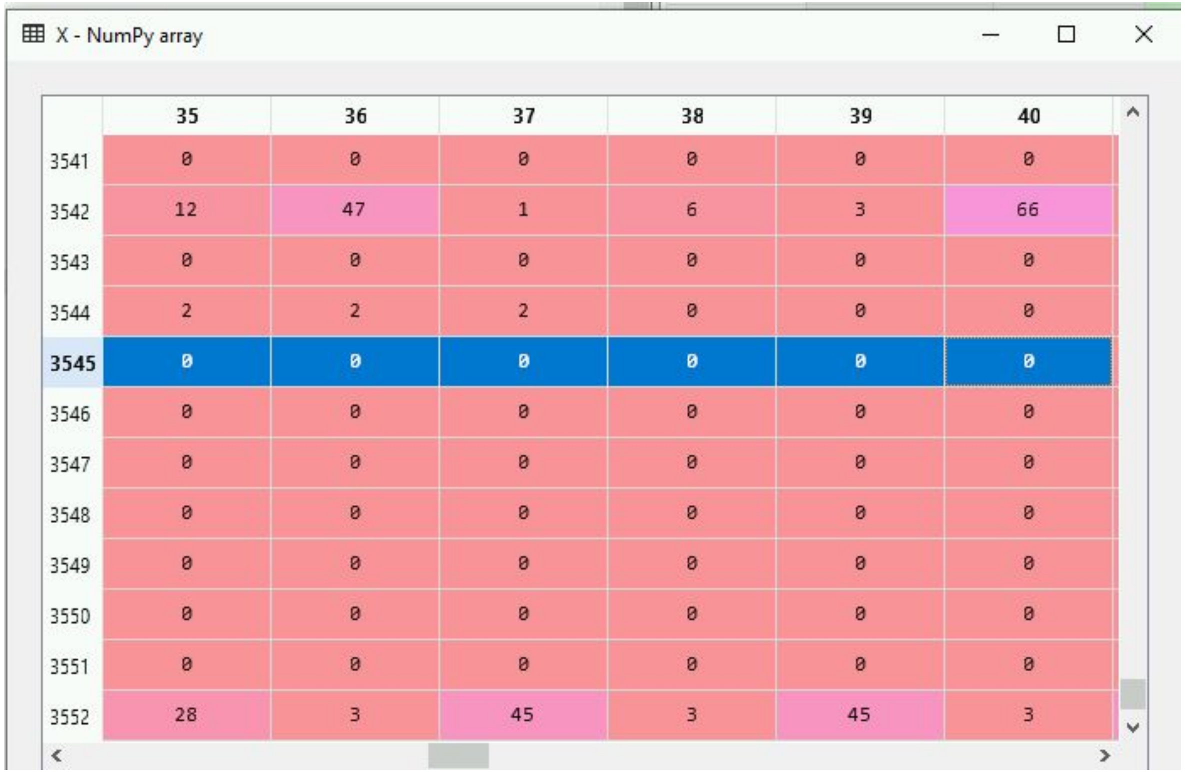
Index	API Calls	Malicious
0	LdrGetProcedureAddress LdrGetProcedureAddress …	0
1	GetSystemDirectoryW RegOpenKeyExW LdrLoadDll…	0
2	LdrGetProcedureAddress LdrGetProcedureAddress …	0
3	NtQueryValueKey NtClose NtAllocateVirtualMemory NtFreeVirtualMemory	0
4	RegCloseKey RegOpenKeyExW RegQueryValueExW NtOpenKey NtQueryValueKe	0
5	GetSystemMetrics NtDeviceIoControlFile …	0



index_word - Dictionary (228 elements)

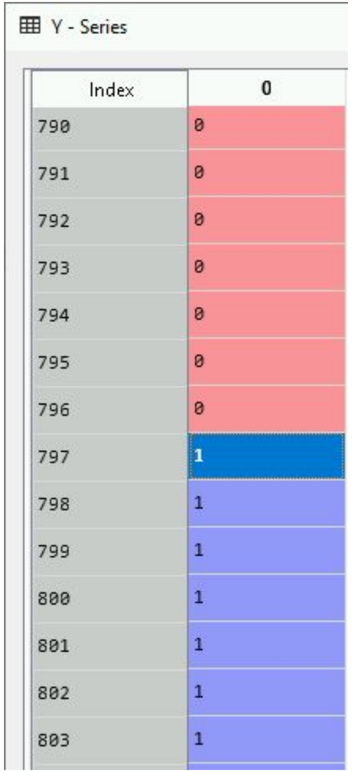
Key	Type	Size	Value
218	str	1	wsasocketa
219	str	1	ntunloaddriver
220	str	1	messageboxtimeouta
221	str	1	shgetspecialfolderlocation
222	str	1	getsockname
223	str	1	copyfileexw
224	str	1	setfileinformationbyhandle
225	str	1	internetopenw
226	str	1	internetconnectw
227	str	1	httpopenrequestw
228	str	1	cryptgenkey

Encoded API



X - NumPy array

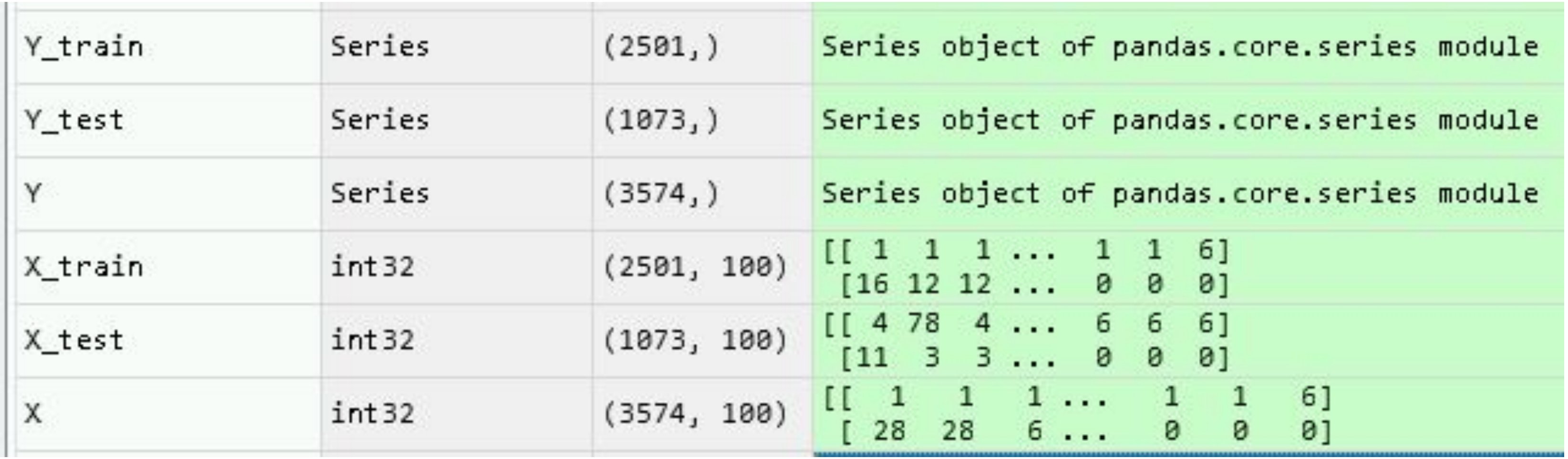
	35	36	37	38	39	40
3541	0	0	0	0	0	0
3542	12	47	1	6	3	66
3543	0	0	0	0	0	0
3544	2	2	2	0	0	0
3545	0	0	0	0	0	0
3546	0	0	0	0	0	0
3547	0	0	0	0	0	0
3548	0	0	0	0	0	0
3549	0	0	0	0	0	0
3550	0	0	0	0	0	0
3551	0	0	0	0	0	0
3552	28	3	45	3	45	3



Y - Series

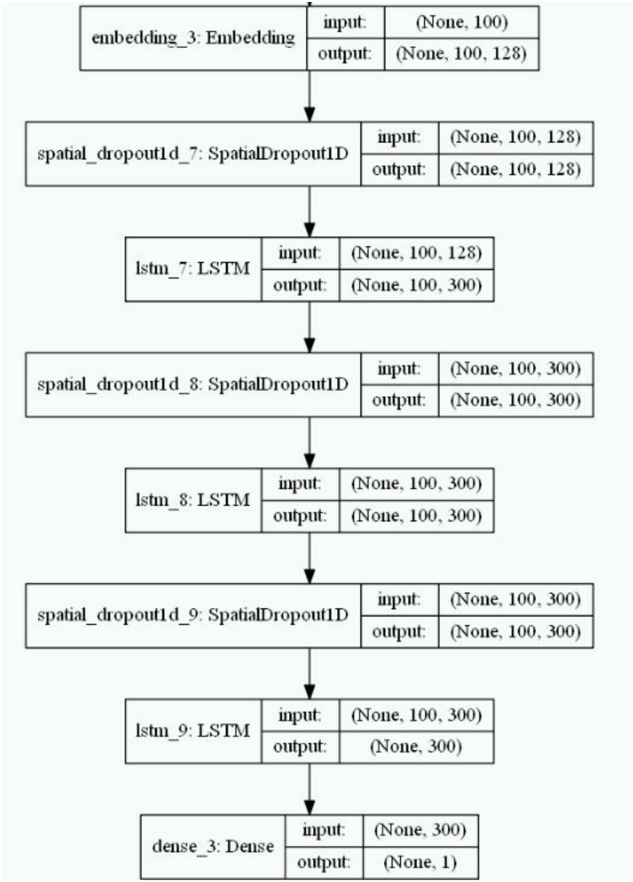
Index	0
790	0
791	0
792	0
793	0
794	0
795	0
796	0
797	1
798	1
799	1
800	1
801	1
802	1
803	1

Train test split



Y_train	Series	(2501,)	Series object of pandas.core.series module
Y_test	Series	(1073,)	Series object of pandas.core.series module
Y	Series	(3574,)	Series object of pandas.core.series module
X_train	int32	(2501, 100)	[[1 1 1 ... 1 1 6] [16 12 12 ... 0 0 0]
X_test	int32	(1073, 100)	[[4 78 4 ... 6 6 6] [11 3 3 ... 0 0 0]
X	int32	(3574, 100)	[[1 1 1 ... 1 1 6] [28 28 6 ... 0 0 0]

RNN Architecture

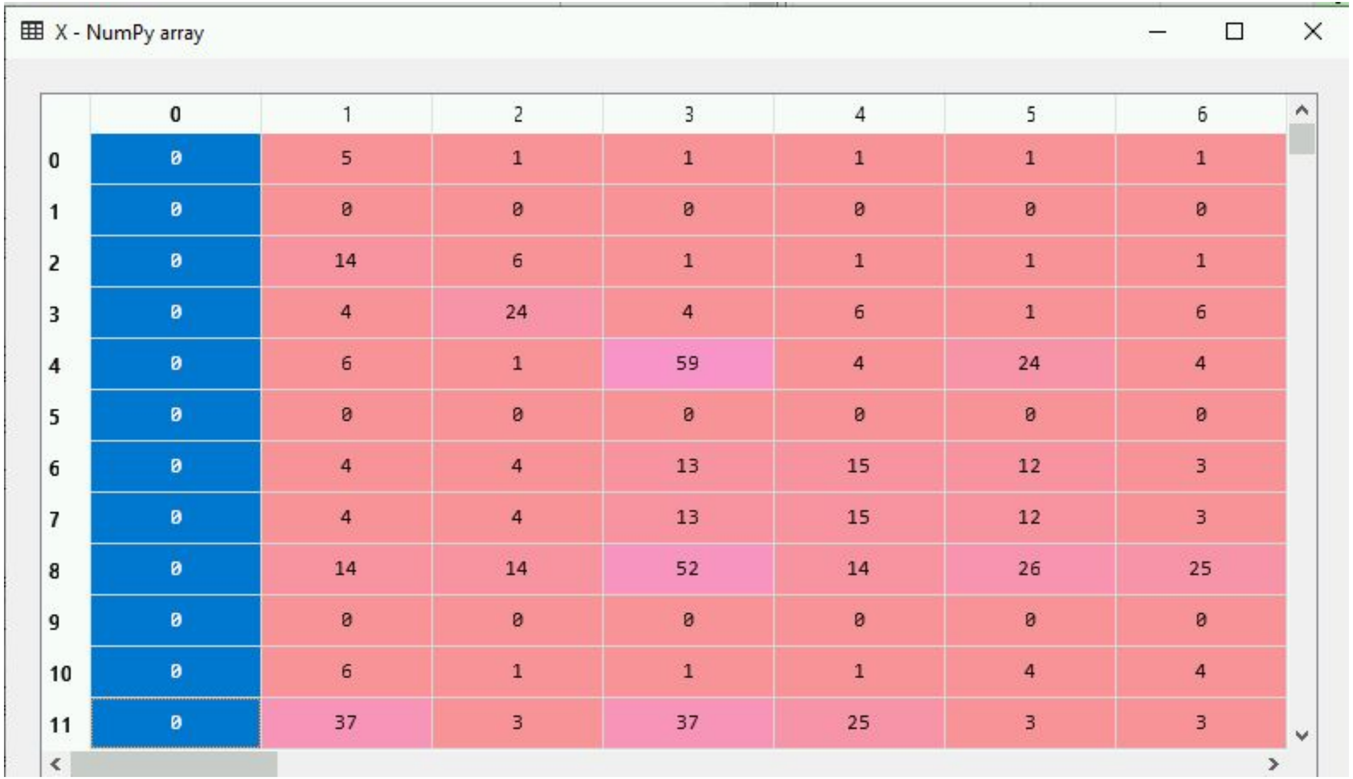


embedding_3: Embedding	input:	(None, 100)
	output:	(None, 100, 128)
spatial_dropout1d_7: SpatialDropout1D	input:	(None, 100, 128)
	output:	(None, 100, 128)
lstm_7: LSTM	input:	(None, 100, 128)
	output:	(None, 100, 300)
spatial_dropout1d_8: SpatialDropout1D	input:	(None, 100, 300)
	output:	(None, 100, 300)
lstm_8: LSTM	input:	(None, 100, 300)
	output:	(None, 100, 300)
spatial_dropout1d_9: SpatialDropout1D	input:	(None, 100, 300)
	output:	(None, 100, 300)
lstm_9: LSTM	input:	(None, 100, 300)
	output:	(None, 300)
dense_3: Dense	input:	(None, 300)
	output:	(None, 1)

Build and train RNN

APIs	LSTM layers	Accuracy	FP	FN	Epochs	Leg/Mal samples
50	150/150/150	**89.1**%	4.7%	6.2%	500	587/1533
100	300/300/300	88.1%	6.1%	5.8%	600	797/1790
150	150/150/150	49% *	34.7% *	16.2%	400	979/2178
200	250/250/250	88.2%	6.6%	5.2%	500	1111/2490
250						1194/2686
300						1273/3002

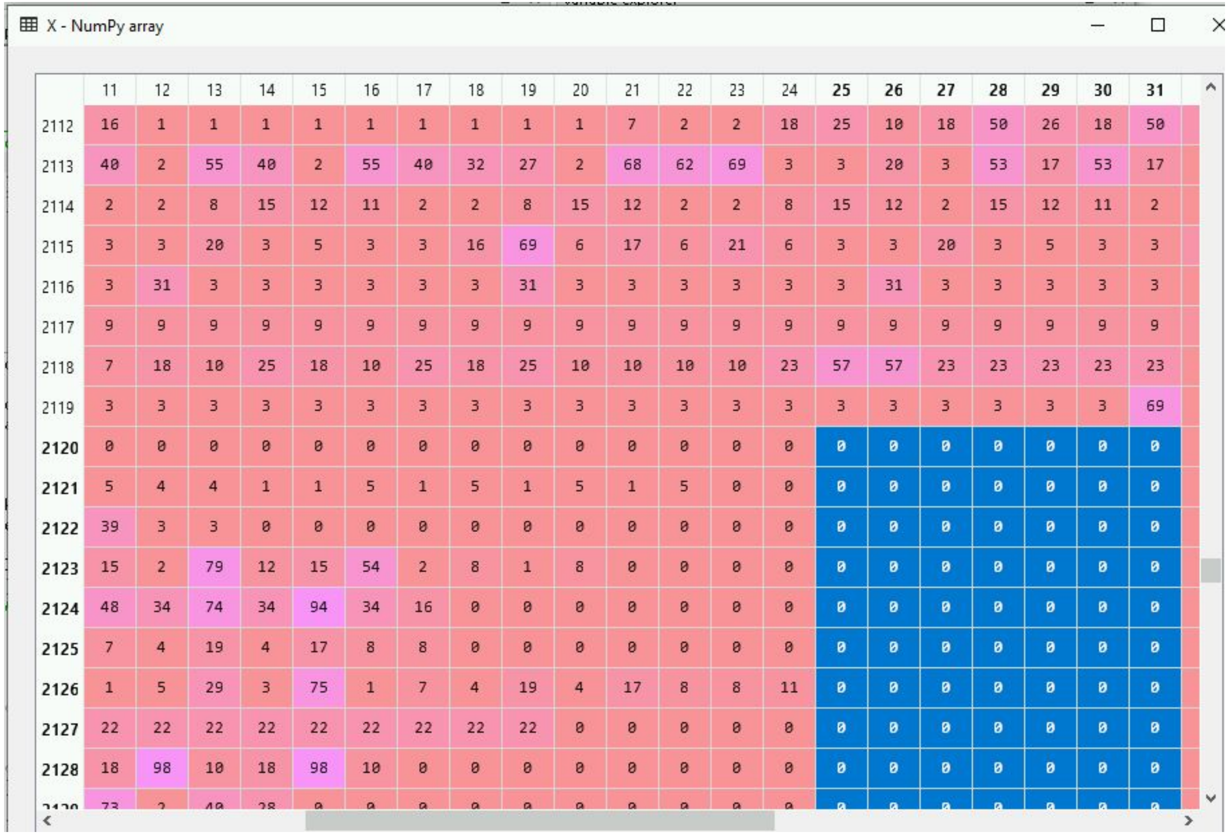
Tolerate 0s. 1st attempt



X - NumPy array

	0	1	2	3	4	5	6
0	0	5	1	1	1	1	1
1	0	0	0	0	0	0	0
2	0	14	6	1	1	1	1
3	0	4	24	4	6	1	6
4	0	6	1	59	4	24	4
5	0	0	0	0	0	0	0
6	0	4	4	13	15	12	3
7	0	4	4	13	15	12	3
8	0	14	14	52	14	26	25
9	0	0	0	0	0	0	0
10	0	6	1	1	1	4	4
11	0	37	3	37	25	3	3

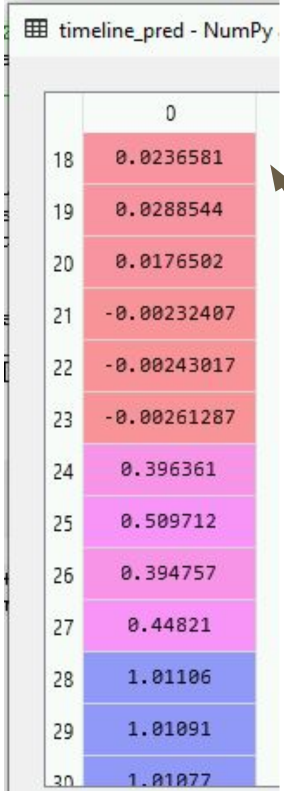
Tolerate 0s. 2nd attempt



X - NumPy array

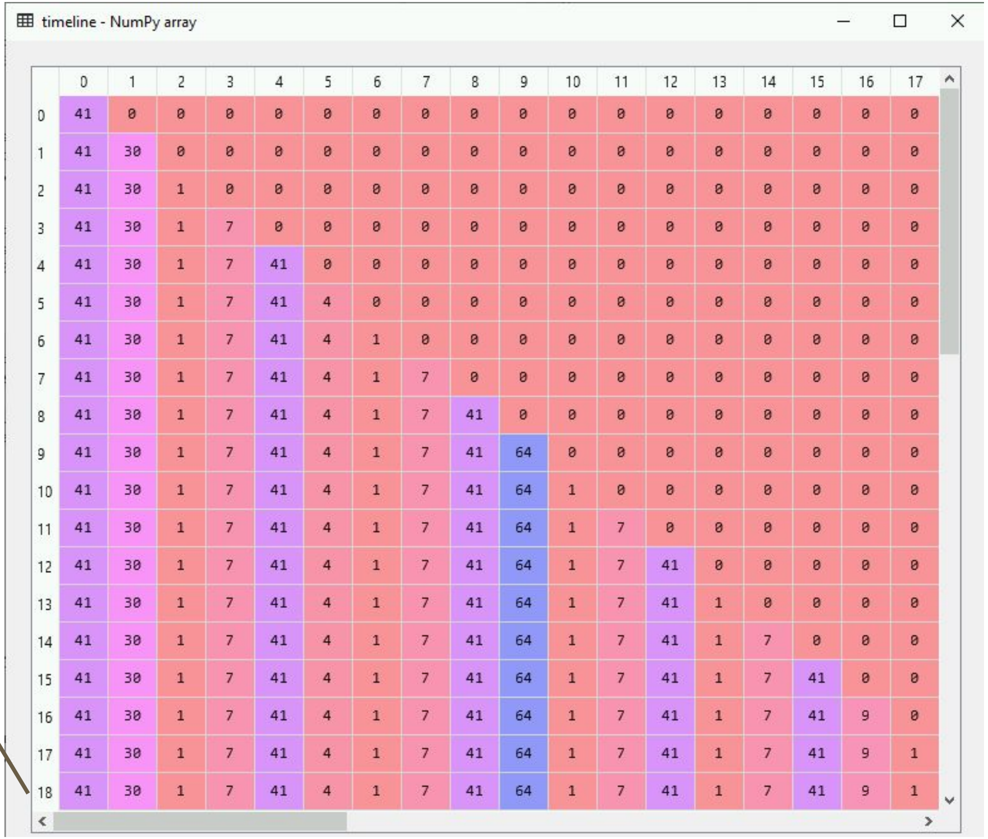
	11	12	13	14	15	16	17	18	19	20	21	22	23	24	**25**	**26**	**27**	**28**	**29**	**30**	**31**
2112	16	1	1	1	1	1	1	1	1	1	7	2	2	18	25	10	18	50	26	18	50
2113	40	2	55	40	2	55	40	32	27	2	68	62	69	3	3	20	3	53	17	53	17
2114	2	2	8	15	12	11	2	2	8	15	12	2	2	8	15	12	2	15	12	11	2
2115	3	3	20	3	5	3	3	16	69	6	17	6	21	6	3	3	20	3	5	3	3
2116	3	31	3	3	3	3	3	3	31	3	3	3	3	3	3	31	3	3	3	3	3
2117	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
2118	7	18	10	25	18	10	25	18	25	10	10	10	10	23	57	57	23	23	23	23	23
2119	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	69
2120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2121	5	4	4	1	1	5	1	5	1	5	1	5	0	0	0	0	0	0	0	0	0
2122	39	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2123	15	2	79	12	15	54	2	8	1	8	0	0	0	0	0	0	0	0	0	0	0
2124	48	34	74	34	94	34	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2125	7	4	19	4	17	8	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2126	1	5	29	3	75	1	7	4	19	4	17	8	8	11	0	0	0	0	0	0	0
2127	22	22	22	22	22	22	22	22	22	0	0	0	0	0	0	0	0	0	0	0	0
2128	18	98	10	18	98	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2129	73	2	40	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Timeline generation - Window-Sliding



timeline_pred - NumPy

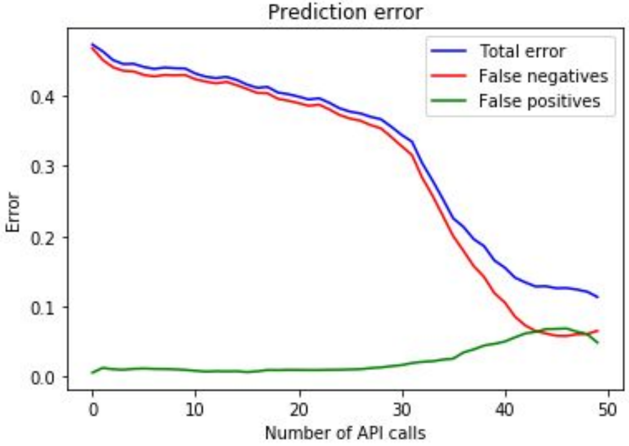
	0
18	0.0236581
19	0.0288544
20	0.0176502
21	-0.00232407
22	-0.00243017
23	-0.00261287
24	0.396361
25	0.509712
26	0.394757
27	0.44821
28	1.01106
29	1.01091
30	1.01077

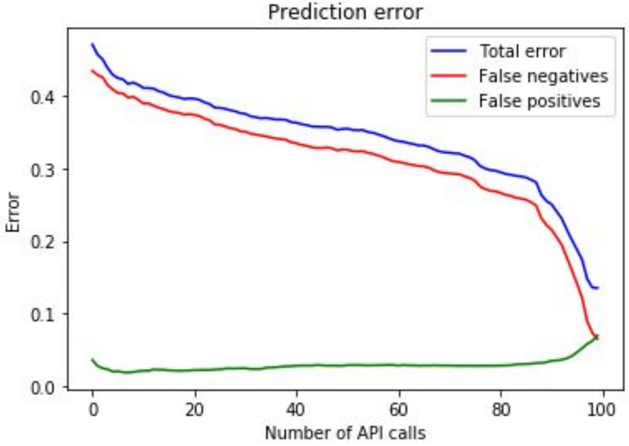


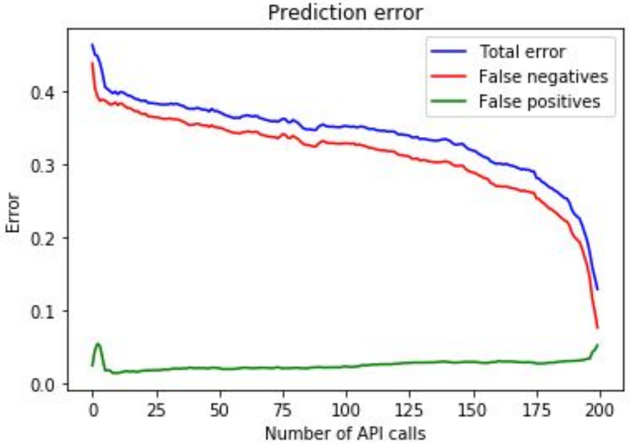
timeline - NumPy array

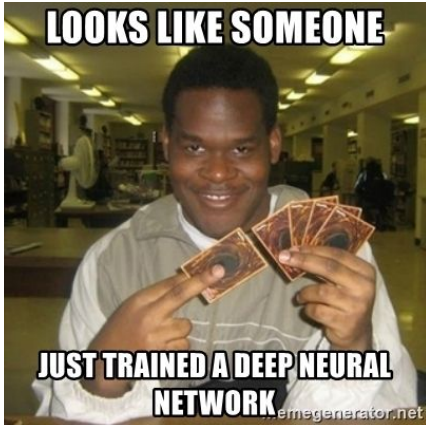
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	41	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	41	30	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	41	30	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	41	30	1	7	41	0	0	0	0	0	0	0	0	0	0	0	0	0
5	41	30	1	7	41	4	0	0	0	0	0	0	0	0	0	0	0	0
6	41	30	1	7	41	4	1	0	0	0	0	0	0	0	0	0	0	0
7	41	30	1	7	41	4	1	7	0	0	0	0	0	0	0	0	0	0
8	41	30	1	7	41	4	1	7	41	0	0	0	0	0	0	0	0	0
9	41	30	1	7	41	4	1	7	41	64	0	0	0	0	0	0	0	0
10	41	30	1	7	41	4	1	7	41	64	1	0	0	0	0	0	0	0
11	41	30	1	7	41	4	1	7	41	64	1	7	0	0	0	0	0	0
12	41	30	1	7	41	4	1	7	41	64	1	7	41	0	0	0	0	0
13	41	30	1	7	41	4	1	7	41	64	1	7	41	1	0	0	0	0
14	41	30	1	7	41	4	1	7	41	64	1	7	41	1	7	0	0	0
15	41	30	1	7	41	4	1	7	41	64	1	7	41	1	7	41	0	0
16	41	30	1	7	41	4	1	7	41	64	1	7	41	1	7	41	9	0
17	41	30	1	7	41	4	1	7	41	64	1	7	41	1	7	41	9	1
18	41	30	1	7	41	4	1	7	41	64	1	7	41	1	7	41	9	1

Error slope

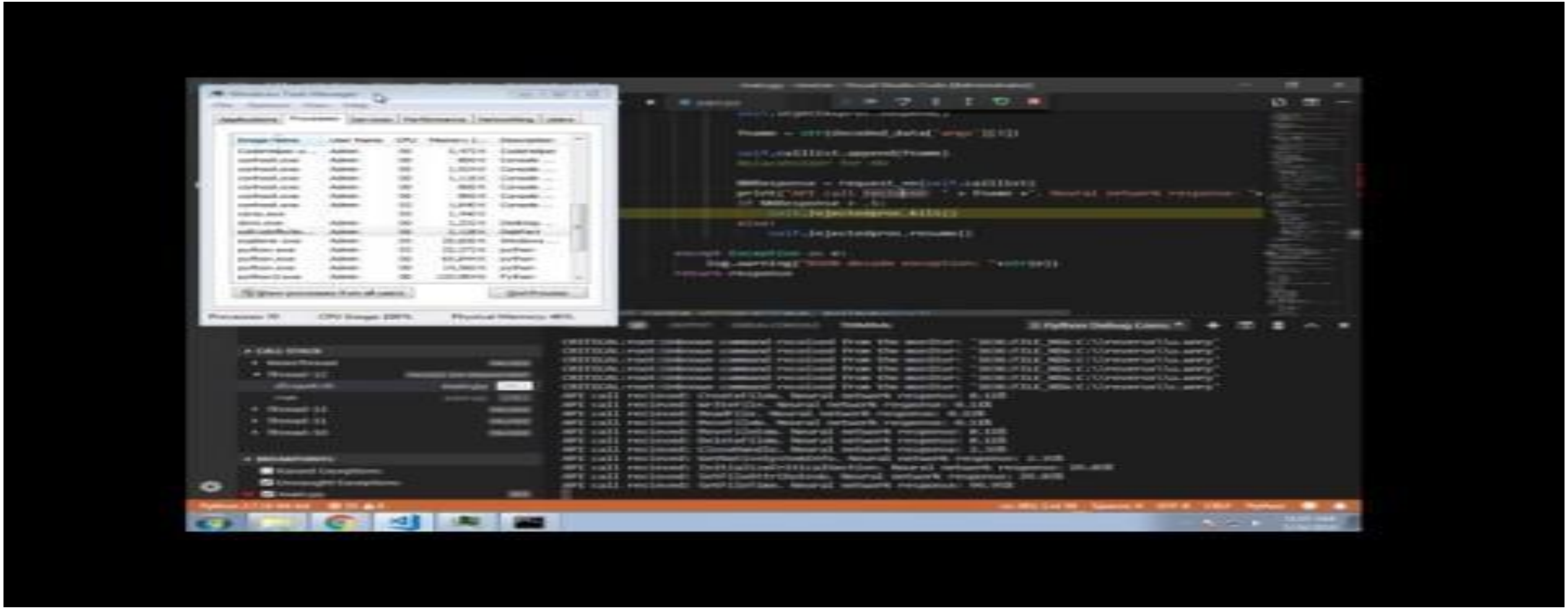

Prediction error
Total error
False negatives
False positives
Error
0.0
0.1
0.2
0.3
0.4
0
10
20
30
40
50
Number of API calls


Prediction error
Total error
False negatives
False positives
Error
0.0
0.1
0.2
0.3
0.4
0
20
40
60
80
100
Number of API calls


Prediction error
Total error
False negatives
False positives
Error
0.0
0.1
0.2
0.3
0.4
0
25
50
75
100
125
150
175
200
Number of API calls


LOOKS LIKE SOMEONE
JUST TRAINED A DEEP NEURAL NETWORK
memegenerator.net

Proof of Concept



Further work

- Delete viruses from legal files
- Investigate the dependency of error upon number of API calls
- Which API functions trigger RNN?
- Expand dataset with API calls after first N

-QuestionS?
imgflip.com