Malicious File Analysis

Version 1.0 | May 2021

Sylvain Hirsch
sylvain@sylvainhirsch.ch
github.com/sylvainhirsch/malware

Introduction

The objective of this document is to provide step-by-step explanations to analyse and triage potential malicious files in order to determine if they are malicious or not and to understand the actions.

Cheat Sheet

A cheat sheet that summarises the process and tools to perform static and dynamic malware analysis is available here.

Malware Analysis Environment

For malware analysis, I recommend and use the two below preconfigured virtual machines.

- Remnux VM – for static malware analysis
- FlareVM – for dynamic malware analysis

WARNING – It is important to only perform malware analysis in a safe environment. If you have never set up a safe malware analysis environment, I recommend reading the chapter 2 "Malware Analysis in Virtual Machines" in the Practical Malware Analysis book.

Additional Information

For additional information, visit github.com/sylvainhirsch/malware or contact me by email at sylvain@sylvainhirsch.ch

Table of Contents

Static Malware Analysis General

The "Static Analysis General" steps can be performed on all types of potentially malicious files. It consists of four steps that permit to assess if the file is malicious, to identify Indicators of Compromises (IOCs) and to determine the file type.

The below sample is used to illustrate the next steps:

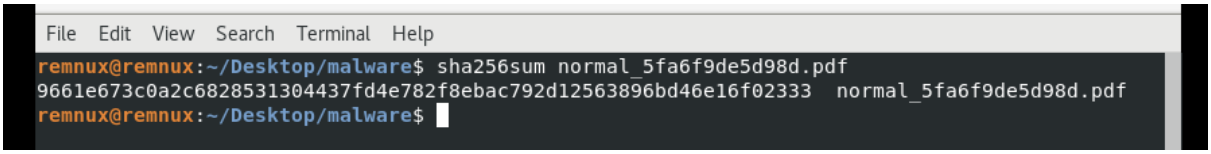
- SHA256: 9661e673c0a2c6828531304437fd4e782f8ebac792d12563896bd46e16f02333

Hash Value

The hash value is a digital fingerprint of a file that is used to uniquely identify it. To avoid collusion, it is recommended to avoid using old cryptographic hash functions (e.g. MD5). SHA256 can be used to compute the file's hash value.

- `certutil.exe -hashfile file SHA256`
  - Windows

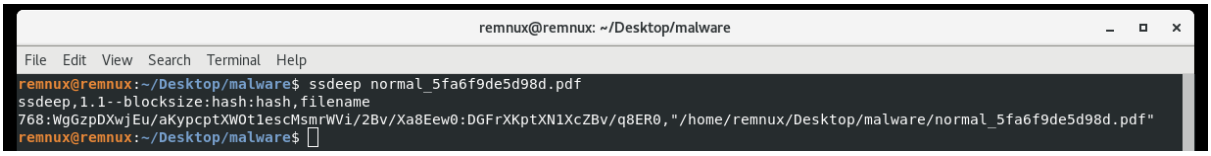
- `sha256sum file`
  - Linux



The fuzzy hash (ssdeep) can be computed in order to identify files with similar or quasi-similar pattern, enabling the identification of similar files.

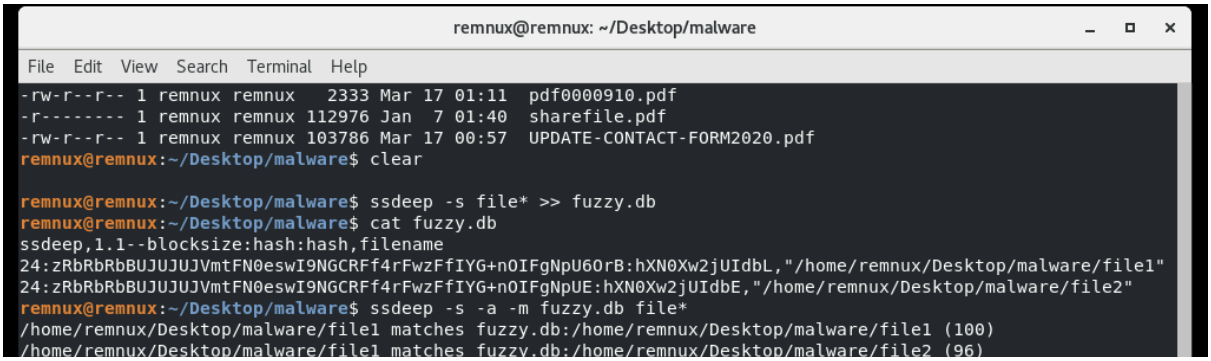
- **`ssdeep.exe file`**
  - Windows

- **`ssdeep file`**
  - Linux



I created two files with almost the same content. I can then use **ssdeep** to compare the two files.

- **ssdeep -s file* >> fuzzy.db**
- **ssdeep -s -a -m fuzzy.db file***



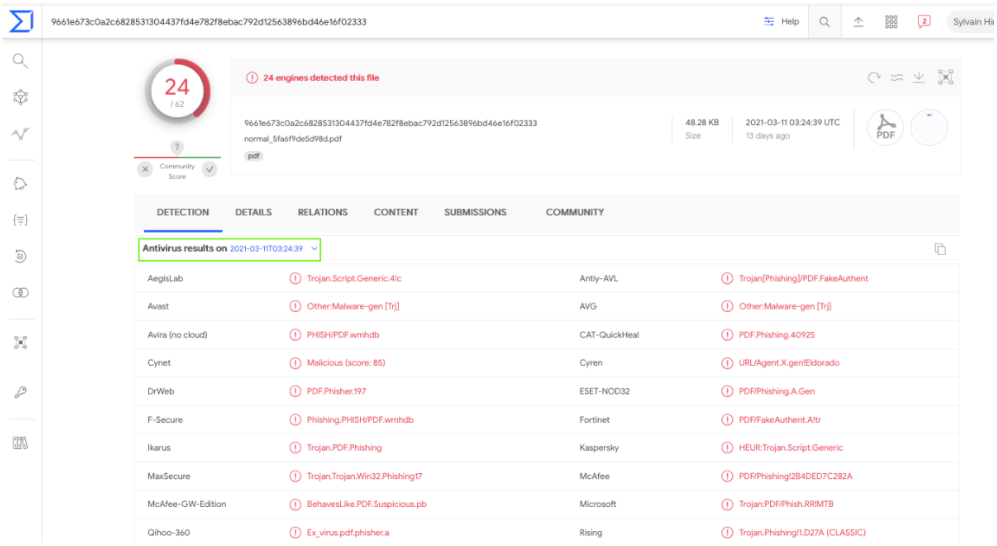
As you can see in the output, the two files are similar to 96%.

Hash Lookup

Open source malware data bases and sandboxes (e.g. VirusTotal, Hybrid Analysis, AlienVault OTX, etc.) can be used to identify if the file is known as malicious, as well as to obtain additional information.

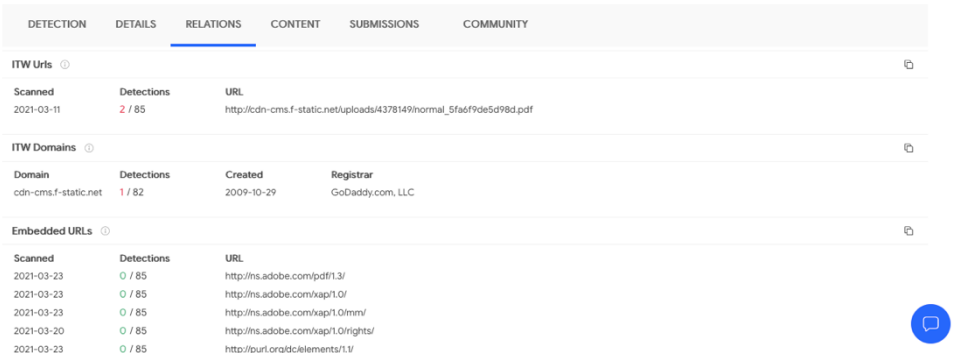
To follow the OpSec best practices, it is recommended to only perform a hash lookup and to not submit the sample to open source data bases or sandboxes.

- Detection Result
  Do not only rely on the hash lookup to determine if a file is malicious. This step only provides some initial result.
  - The "Antivirus results history" (highlighted in green in the below screenshot) can be interesting.



- Relations

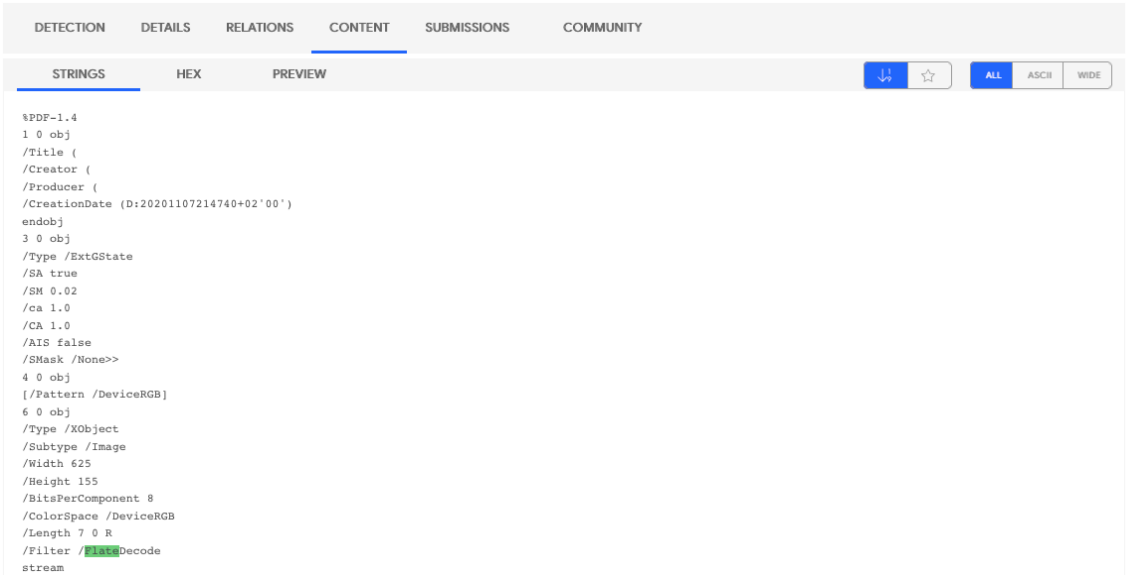
  The relations section enables to identify related IOCs (e.g. C2 server).



- Content

  The content section contains the extracted strings of the file (cf. below steps).

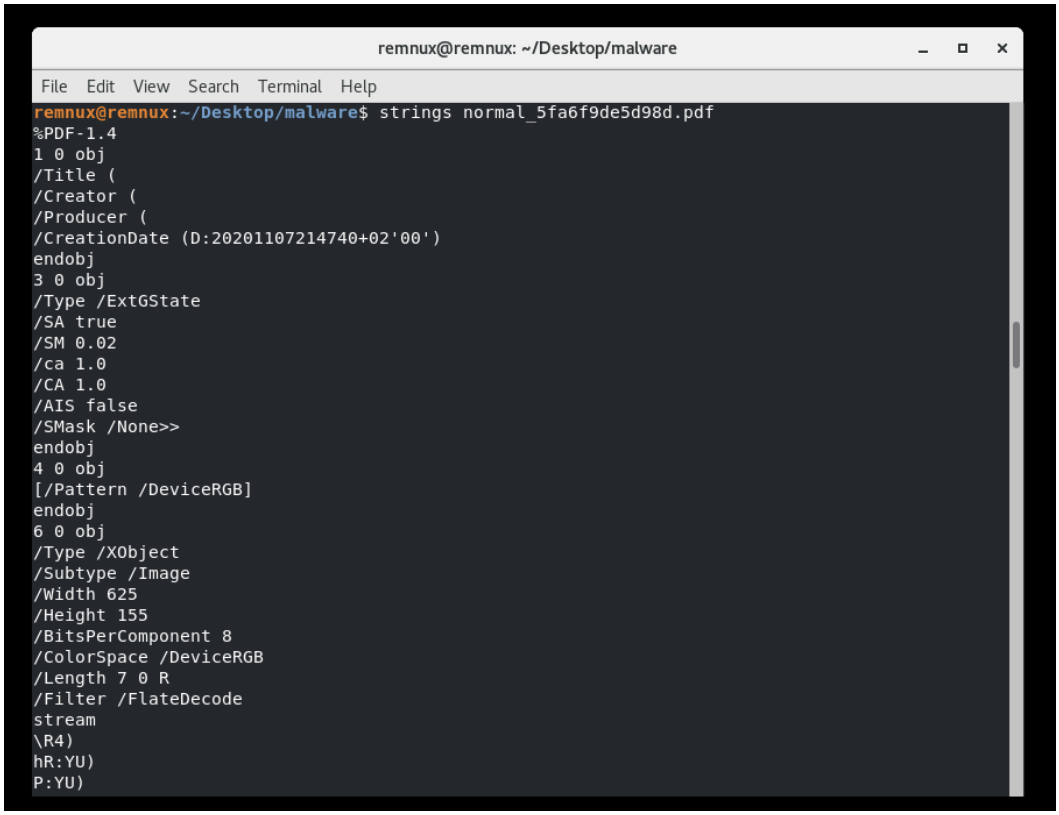
  This step permits to identify interesting elements, as "/Filter /FlateDecode" in this PDF file. For more information about "FlateDecode" refer to the *PDF File Analysis*.



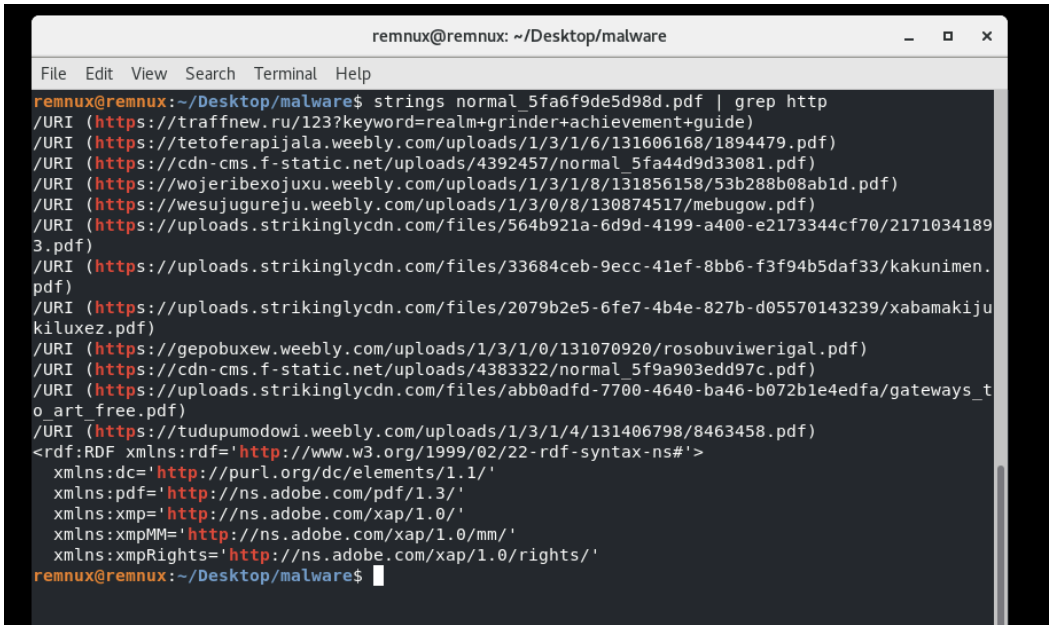
Strings Extraction

The *Strings Extraction* permits to extract all strings from the file and to identify interesting strings/IOCs (e.g. URL, command lines, encoded strings, path, registry keys, etc.).

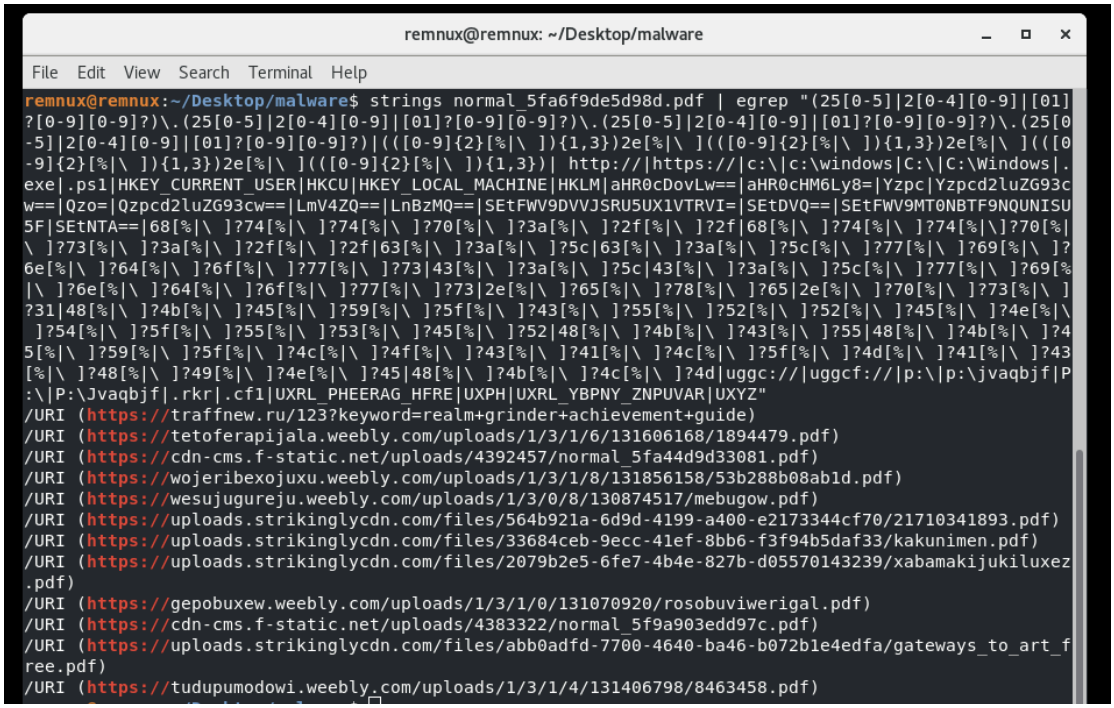
- **`strings file`**
  - **`-n`** - to specify the minimum string length (the default value is 3).



- **`strings file | grep http`**
  - print only strings with http or https.



- **`strings file | egrep Regex`**
  - to find specific pattern (e.g. URL, IP, etc.)



This regex permits to identify IOCs and encoded IOCs that are present in malicious files.

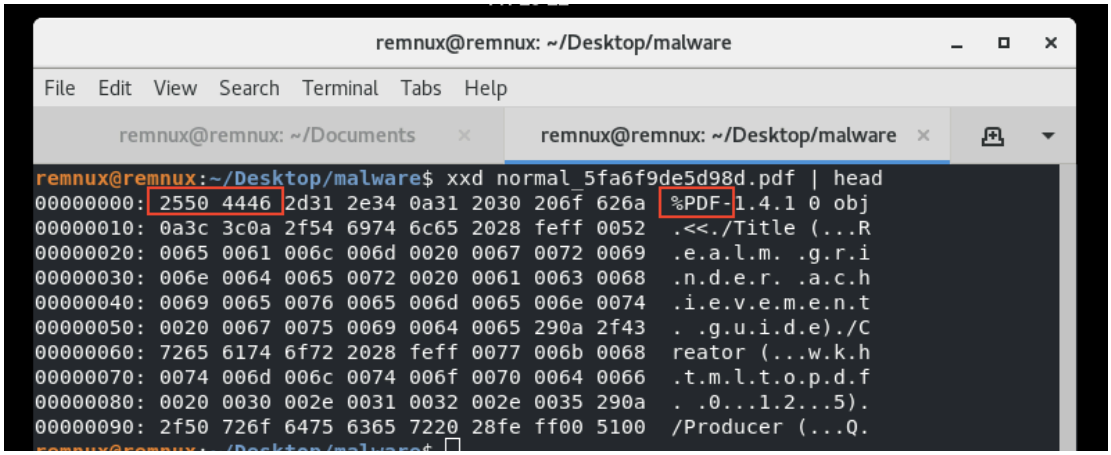
File Signature

The File Signature (aka “magic number”) is a number embedded at or near the beginning of a file that indicates the file format. As the file extension of a file can be renamed, it is important to verify the magic number of the file to determine the correct file type.

The magic number check permits to determine the following steps of the analysis (e.g. pdf, rtf, etc.)

- **`hexdump.exe file | head`**
  - Windows

- **`xxd file | head`**
  - Linux



The below table summarises the main document type's magic numbers. For more magic number values, check this website.

File Type	Hex Magic Value	ASCII Value
PDF	25 50 44 46	%PDF
Microsoft Office OLE2 file (e.g. .doc, .xls, etc.)	D0 CF 11 E0	
Microsoft Office OOXML file (e.g. docx)	50 4B 03 04	PK
Rft	7B 5C 72 74	{ \rft
exe, dll, sys	4D 5A 90 00	MZ
Rar	52 61 72 21	Rar!

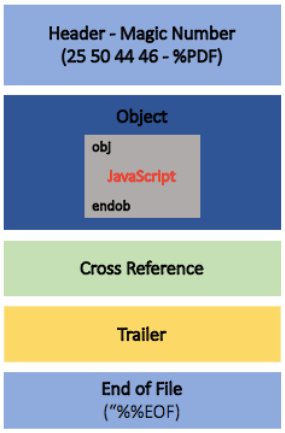
PDF File Analysis

PDF File Format

To analyze malicious PDF it is important to have at least a basic understanding of the pdf file format. Portable Document Format (PDF) is a proprietary file format developed by Adobe. In summary, PDF files are principally composed of objects that contains data or specific functions.

PDF is composed of five parts:

- Header: Magic number %PDF-1.5
- Object(s): Every object has the following format (cf. below section for more information)
  - "1 0 obj" – It starts with two number and "obj"
    - The first number is the numeric id of the object.
    - The second number is the version of the object.
  - "<< … >>" – An object container that stats with "<<" and ends with ">>"
  - "endobj" – It ends by "endobj"
- Cross Reference: not really relevant in the malware context.
- Trailer: not really relevant in the malware context.
- End of File: "%%EOF"

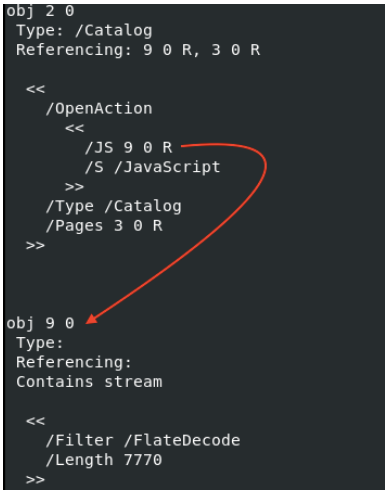


Object

PDF are mainly composed of objects that contain specific data within the document or perform a specific function.

An important point is that an object can refer to another object using the corresponding object id. In the below screenshot the object 2 is referencing to the object 9 (*9 0 R*) and 3 (*3 0 R*).

PDF objects can consist of various types of object. The stream object (one type of pdf objects) contains an unlimited sequence of bytes. Data stream can be modified, compressed or even encoded using filters. When a filter is used, the value following "*/Filter*" indicates the method to decompress or decode the stream.



Data in pdf can be partially or fully encoded in multiples ways (e.g. hexadecimal, octadecimal, additional whitespace, etc.).

The below example shows how data can be encoded in a PDF. The objective of the below encoding is to complicate the identification of the *FlateDecode* filter:

- Filter [/F#6c#61#74e#44e#63#6fde/#41#53#43II#38#35#44#65#63#6fd#65]
- FlateDecode --> 46 6c 61 74 65 44 65 63 6f 64 65

Interesting Elements

The below actions and elements provide a good insight into the PDF behaviors:

- /OpenAction or /AA: an action is automatically performed when the pdf is opened
- /JavasScript or /JS: the pdf contains a JavaScript code
- /Name: refers to name instead of the keyboard
- /EmbeddedFiles: indicates a dictionary with the embedded files
- /URI : an URL is accessed
- /SubmitForm: data can be transferred to a website.
- /Launch: a program is automatically spawned when the pdf is opened

It is worth being careful when a PDF contains:

- /OpenAction or /AA with /JavasScript or /JS
- /URI or /SubmitForm with /Launch

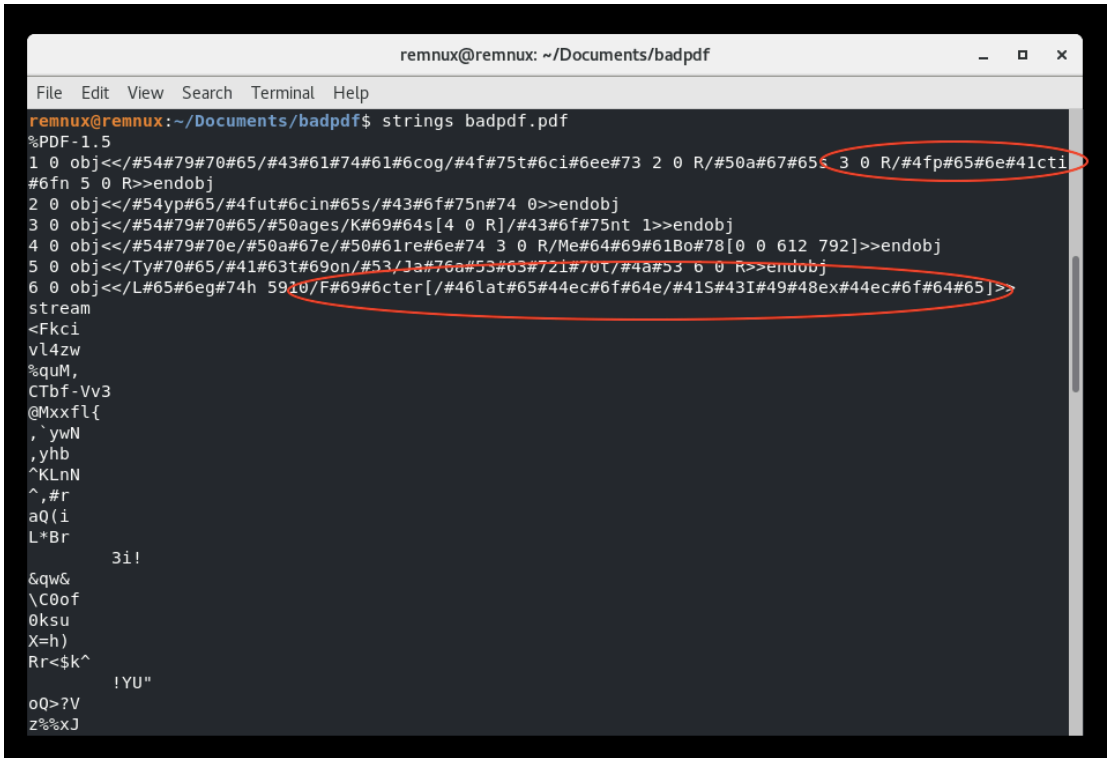
The below sample is used to illustrate the next steps:

- SHA256: e6889f6e98f554a3fd6e4371e9a8323abbdfbced2031b8863ba84a3050c56bea

Strings

Using **`strings`** command, as explained in the "Static Analysis General" section, interesting elements can be revealed.

- **`strings badpdf.pdf`**



The PDF contains some obfuscation techniques. With a bit of experience and as demonstrated later, it can be identified that the pdf contains some OpenAction and some encoded data with *FlateDecode* filter.

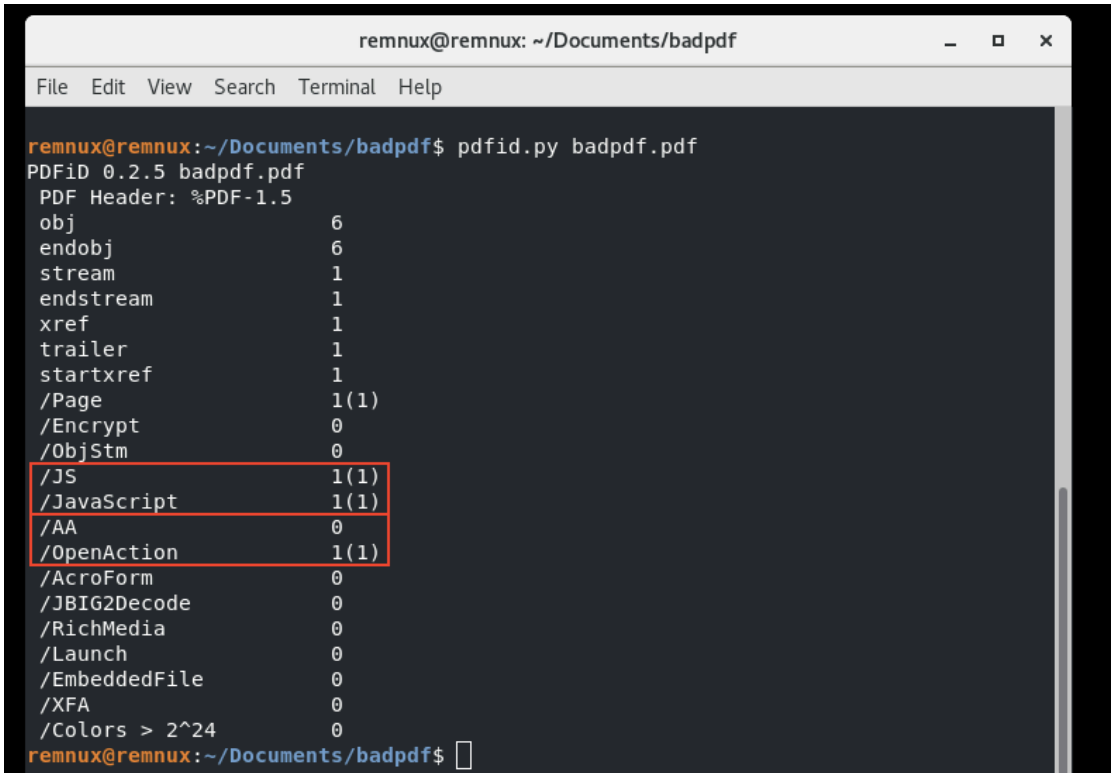
The *OpenAction* and *FlateDecode* words are encoded in hexadecimal:

- OpenAction
  - #4fp#65#6e#41cti#6fn
- FlateDecode
  - /#46lat#65#44ec#6f#64e/#41S#43I#49#48ex#44ec#6f#64#65]

Pdfid

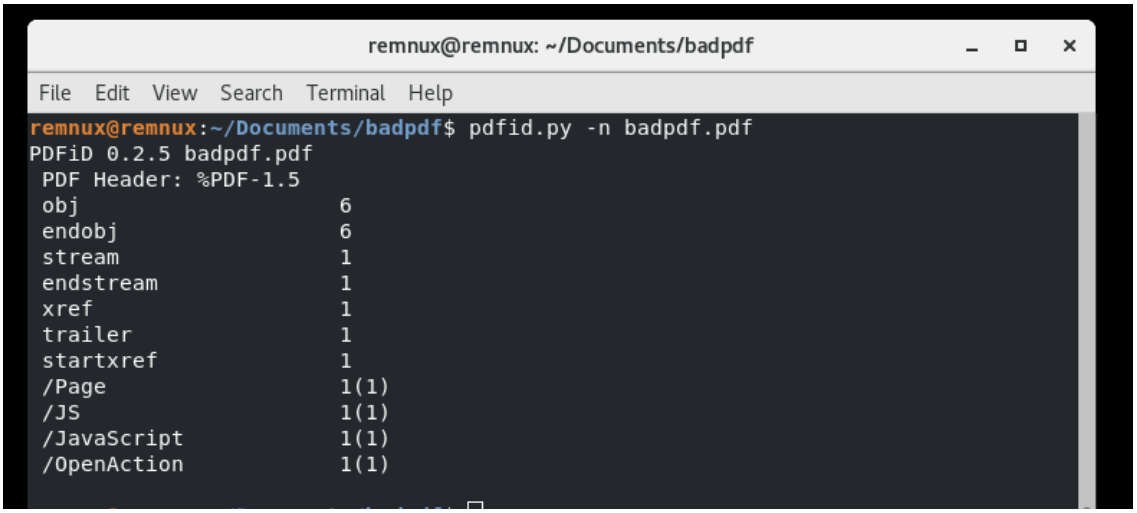
Pdfid (Didier Stevens) is a utility that scans PDF documents to look for certain PDF keywords, allowing to identify for example JavaScript or OpenAction. It also provides statistics about the elements within the document.

- **pdfid.py badpdf.pdf**



This PDF contains an *OpenAction* and a *JavaScript* code.

- **pdfid.py -n badpdf.pdf**
  - **-n** argument can be used to only output the present elements



Pdf-parser

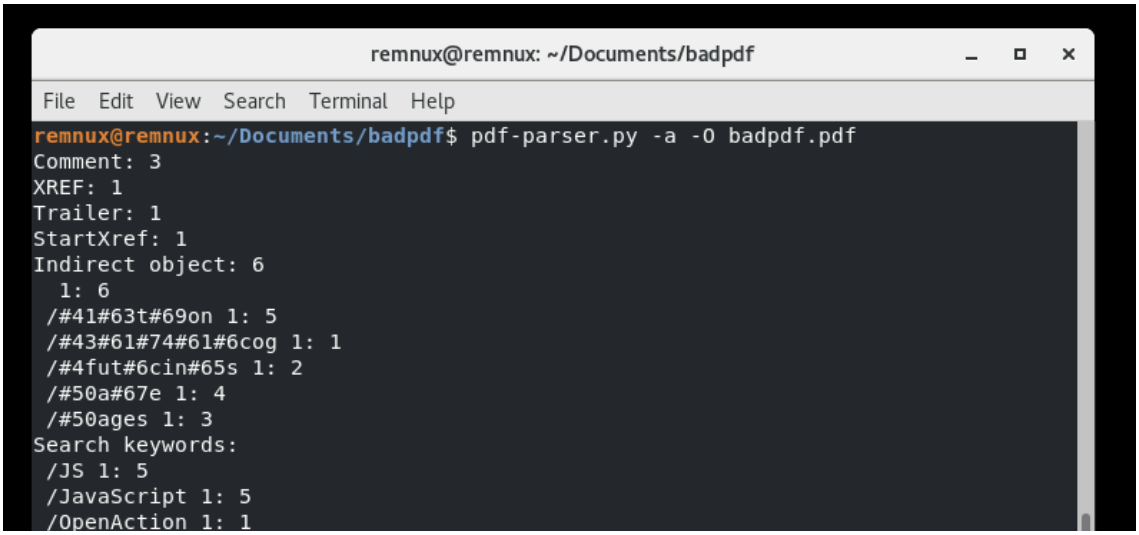
Pdf-parser (Didier Stevens) is a tool that parses a PDF document to identify the fundamental elements in the PDF.

There are some useful arguments for pdf-parser.

- **-s** or **--search** to search for a stream of /ObjStm objects
- **-o** or **--object** to search for an object id
- **-f** or **--filter** to pass the stream through filters to decode it
- **-w** or **--raw** to output the raw output
- **-d** or **--dump** to dump a stream

Using the pdf-parser **-a** and **-O** arguments provides the equivalent information obtained with pdfid.

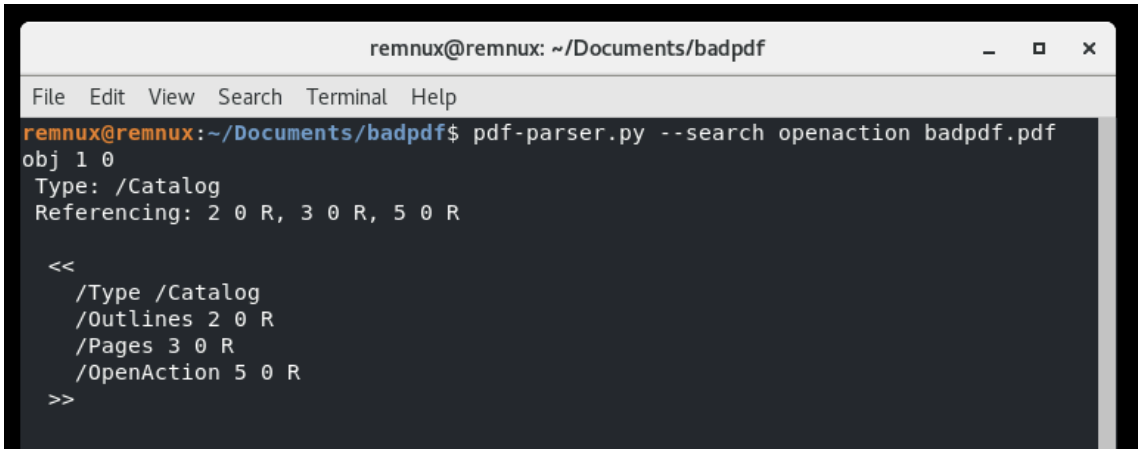
- **pdf-parser.py -a -O badpdf.pdf**
  - **-a** - displays the stats for the file
  - **-O** - extracts and parses the objects inside stream objects.



The *OpenAction* is located in the object number 1 and the JavaScript is in the object number 5.

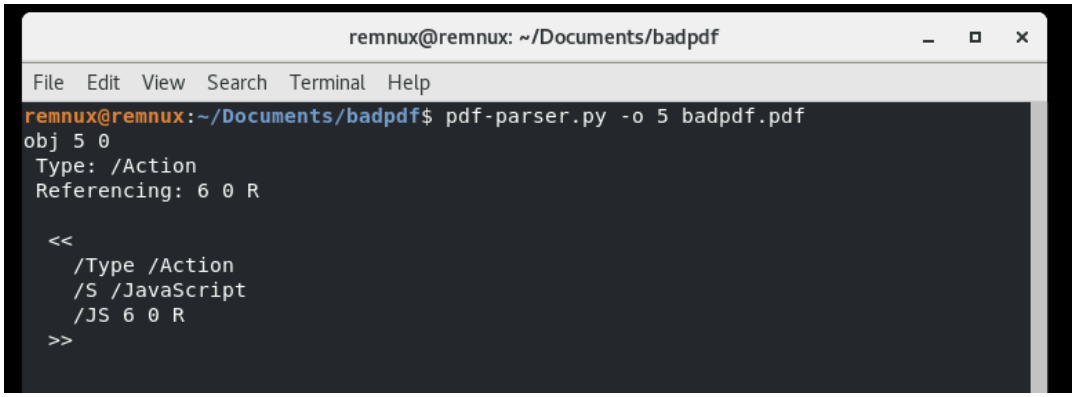
The *OpenAction* can also be found with the below command. In the below output you can see that the object number 1 contains an *OpenAction* and will spawn the content of the object 5.

- **pdf-parser.py --search openaction badpdf.pdf**
  - **--search javascript**
  - **--search JS**
  - **--search AA**

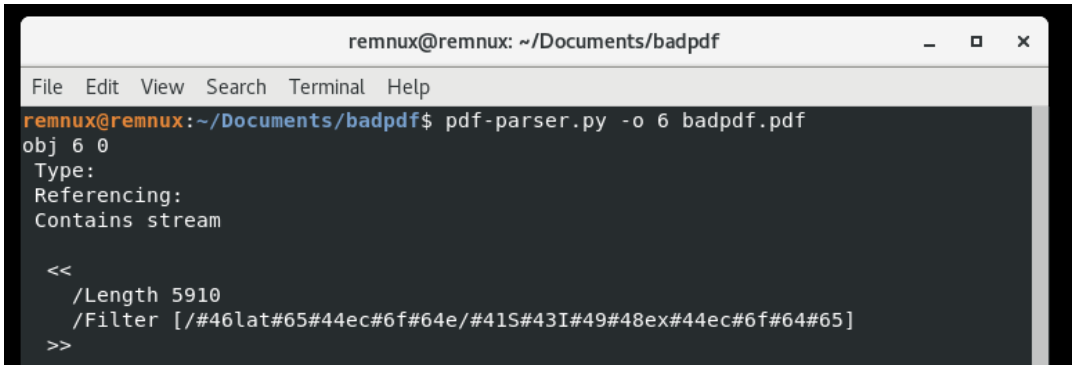


The content of object number 5 can be analysed with the below command. The object 5 refers to the object 6.

- **pdf-parser.py --object 5 badpdf.pdf**



- **pdf-parser.py --object 6 badpdf.pdf**



The data contained in the object 6 is encoded with the FlateDecode filter. The name FlateDecode is partially encoded with hexadecimal values. You can use CyberChef to decode it.

f	l	a	t	e	d	e	c	o	d	e
46	6c	61	74	65	44	65	63	6f	64	65

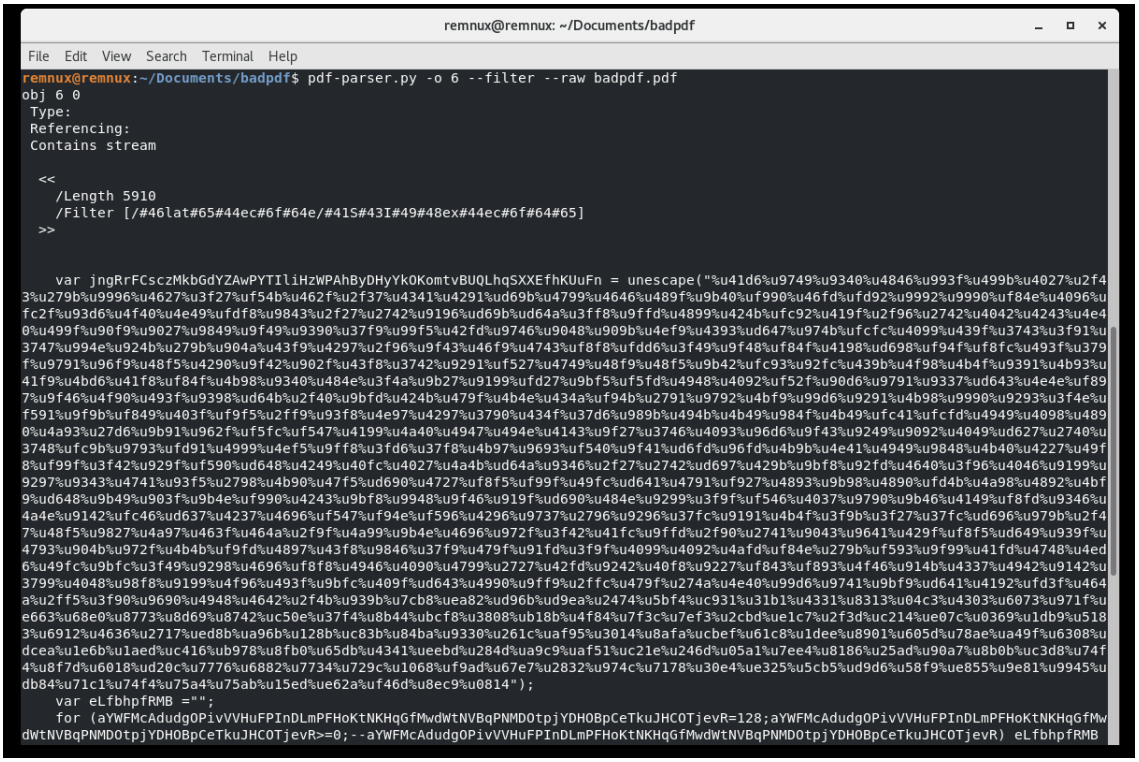
In non-encoded malicious pdf, you will see the below output.

```
obj 9 0
 Type:
 Referencing:
 Contains stream

  <<
    /Filter /FlateDecode
    /Length 7770
  >>
```

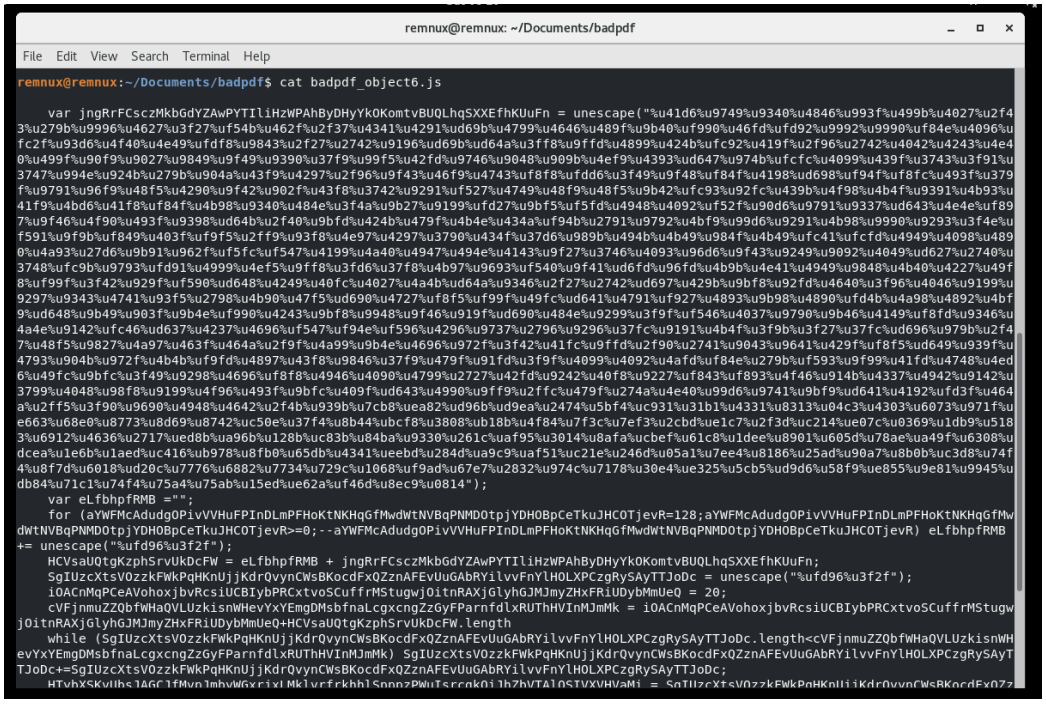
To decode the content of the object 6, use the arguments **--filter** and **--raw**.

- **pdf-parser.py --object 6 --filter --raw badpdf.pdf**
  - **--filter** enables to decode the content of the object. It supports the following filters: FlateDecode, ASCIIHexDecode, ASCII85Decode, LZWDecode and RunLengthDecode
  - **--raw** outputs the raw data



Using the **`--dump`** argument, the object can be dumped to be analyzed.

- **`pdf-parser.py --object 6 --filter --raw --dump=badpdf_object6.js badpdf.pdf`**



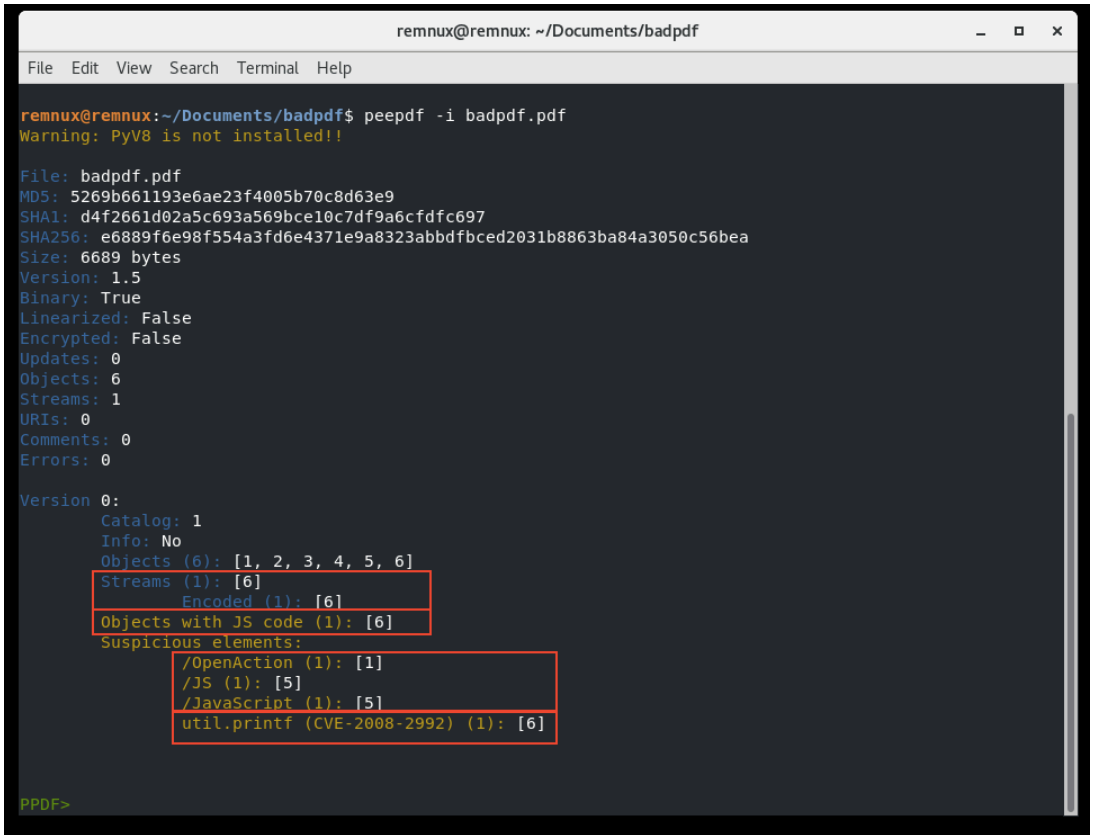
Now the dumped JavaScript can be analyzed. Refer to the JavaScript section for more information.

Peepdf

Peepdf is a great tool that analyses pdf document and provides a summary of its content, as well as suspicious elements identified.

A summary of the content of the pdf can be obtained using the below command.

- **peepdf -i badpdf.pdf**
  - **-i** provides an interactive console



In the above output, it can be observed that the pdf contains:

- *OpenAction* in the object number 1
- *JavaScript* in the object number 5
- Some encoded streams in the object 6
- A CVE (CVE-2008-2992) exploitation in the object 6

This tool is great as it summarizes the content of the pdf using a single command.

The available commands can be displayed using:

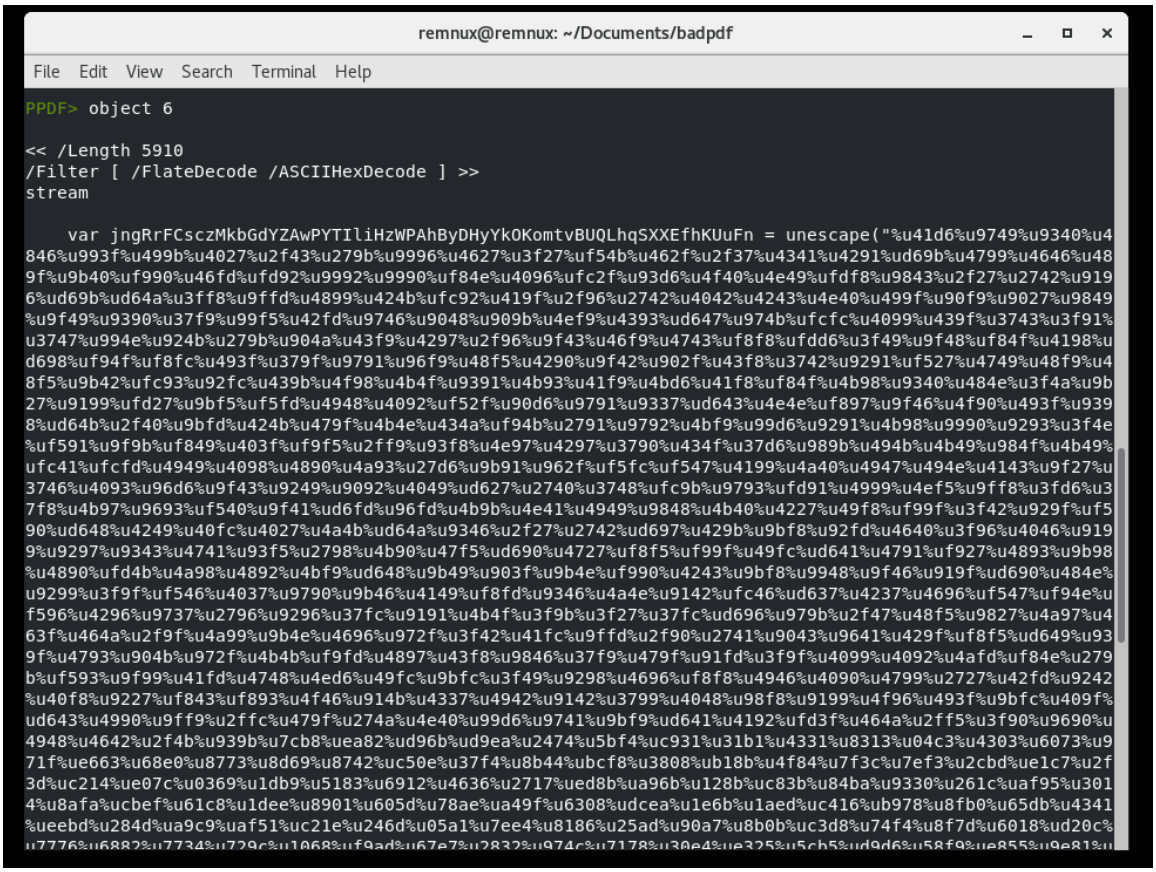
- **help**

```
PPDF> help

Documented commands (type help <topic>):
========================================
bytes           exit        js_jjdecode       open          search
changelog       extract     js_join           quit          set
create          filters     js_unescape       rawobject     show
decode          hash        js_vars           rawstream     stream
decrypt         help        log               references    tree
embed           info        malformed_output  replace       vtcheck
encode          js_analyse  metadata          reset         xor
encode_strings  js_beautify modify            save          xor_search
encrypt         js_code     object            save_version
errors          js_eval     offsets           sctest
```

The content of an object can be viewed using **object** following by the object id. As you can see, the object was decoded automatically.

- **object 6**



The object 6 can be extracted using the below command.

- **object 6 > badpadf_object6.js**

The JavaScript can be analyzed using the **--js_analysis**. Refer to the JavaScript section for more information.

Microsoft Office OLE2 File

Microsoft Office OLE2 document can be seen as a mini *FAT* filesystem in a file. It is divided into fixed length sections (usually 512 bytes). The first sector contains the header, and the other sectors contain data in streams or other structures (e.g. Macro or embedded object). Each stream has a name (cf. oledump output below, e.g. *'Macros/VBA/ThisDocument'*) and has specific properties (e.g. author, timestamps, etc.)

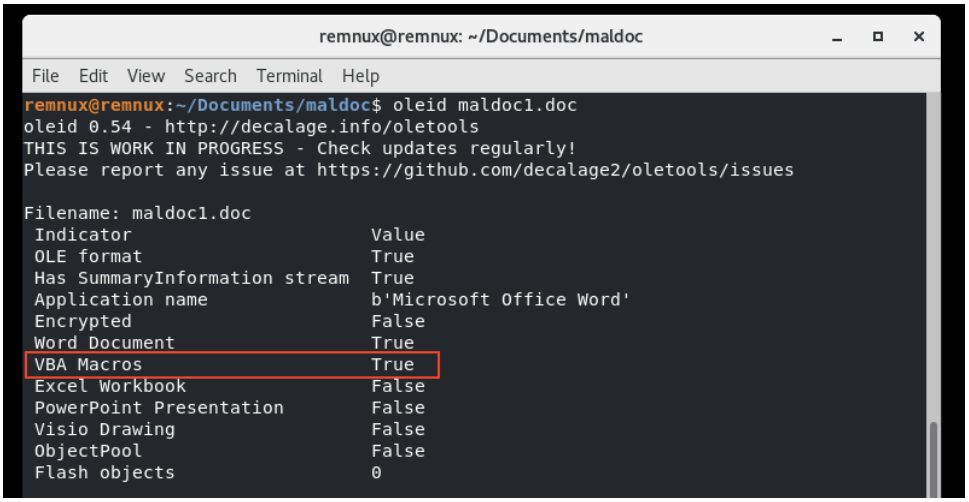
The below sample is used to illustrate the next steps:

- SHA256: 8b92c23b29422131acc150fa1ebac67e1b0b0f8cfc1b727805b842a88de447de

Oleid

Oleid is a script to parse OLE documents to identify specific malicious characteristics such as VBA macros or embedded objects.

- **`oleid maldoc1.pdf`**

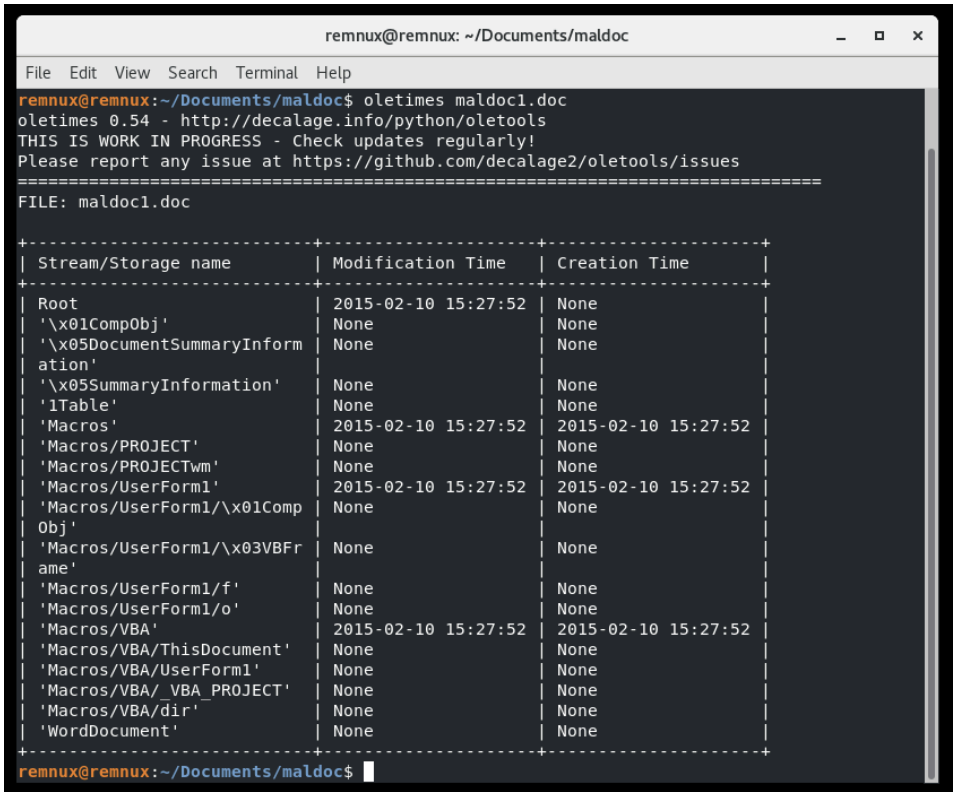


The output indicates that a VBA Marco (*VBA Marcos True*) is present in the document, however it does not contain any object (*Flash objects 0*).

Oletimes

Oletimes is a script to parse OLE documents that provides the creation and modification timestamps of streams and objects.

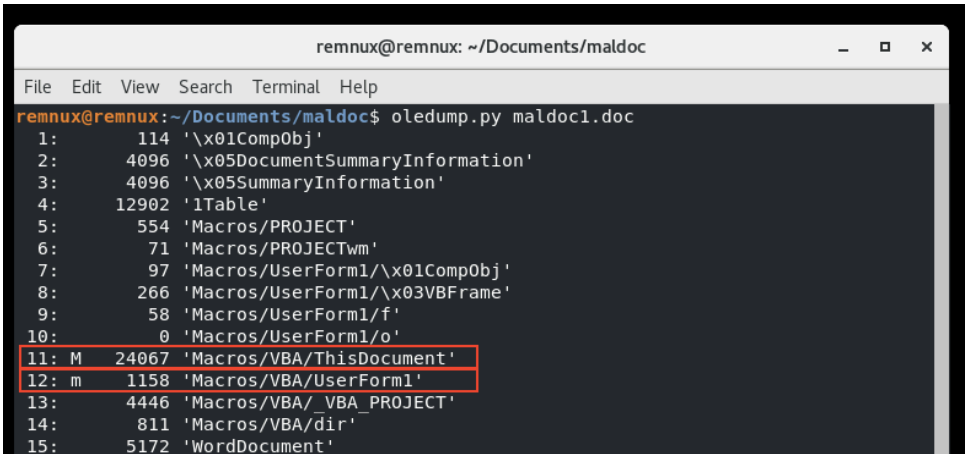
- **`oletimes maldoc1.doc`**



Oledump

Oledump (Didier Stevens) is a script that parses OLE documents to identify macros and objects and extract them. The letter 'M' indicates a macro and the letter 'O' an embedded object.

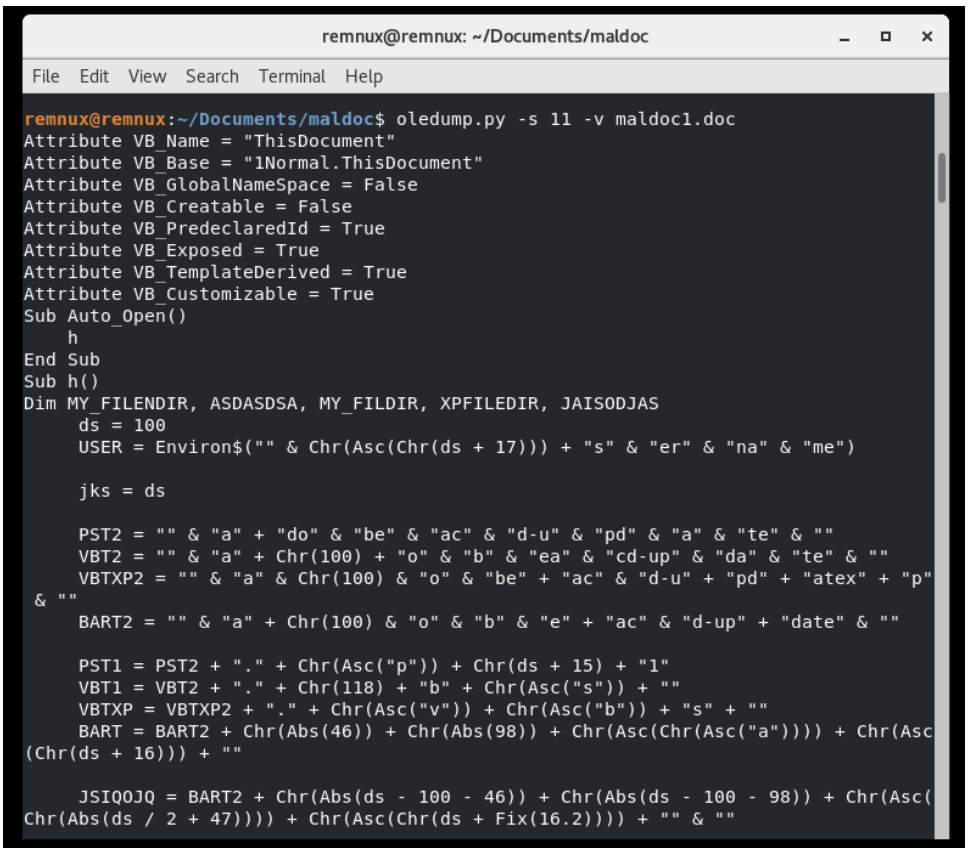
- **`oledump maldoc1.doc`**



The output indicates that the streams 11 and 12 contain a macro.

The content of the stream 11 can be analyzed with the below command.

- **oledump -s 11 -v maldoc1.doc**



As you can see below, the macro is obfuscated and partially encoded. Using the following command, it can be extracted to be further analysed (~~cf. olevba~~).

- **oledump -s 11 -v maldoc1.doc >maldoc1.vba**

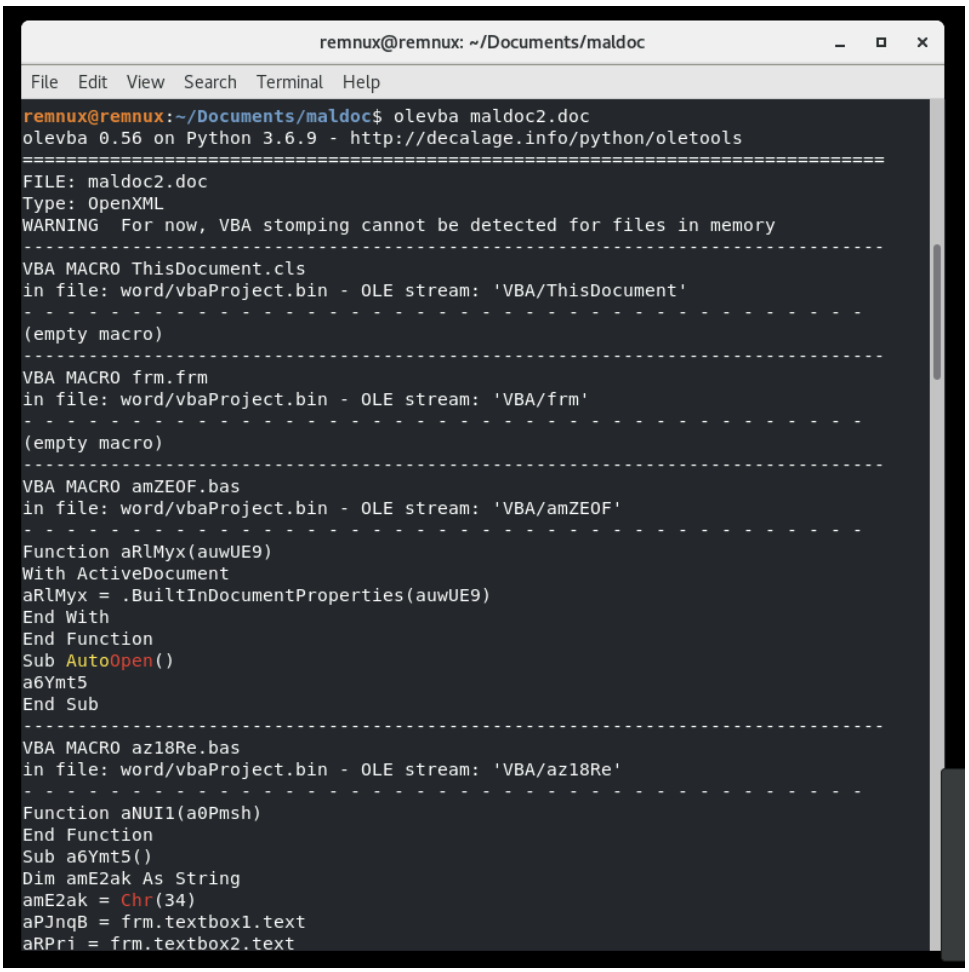
Olevba

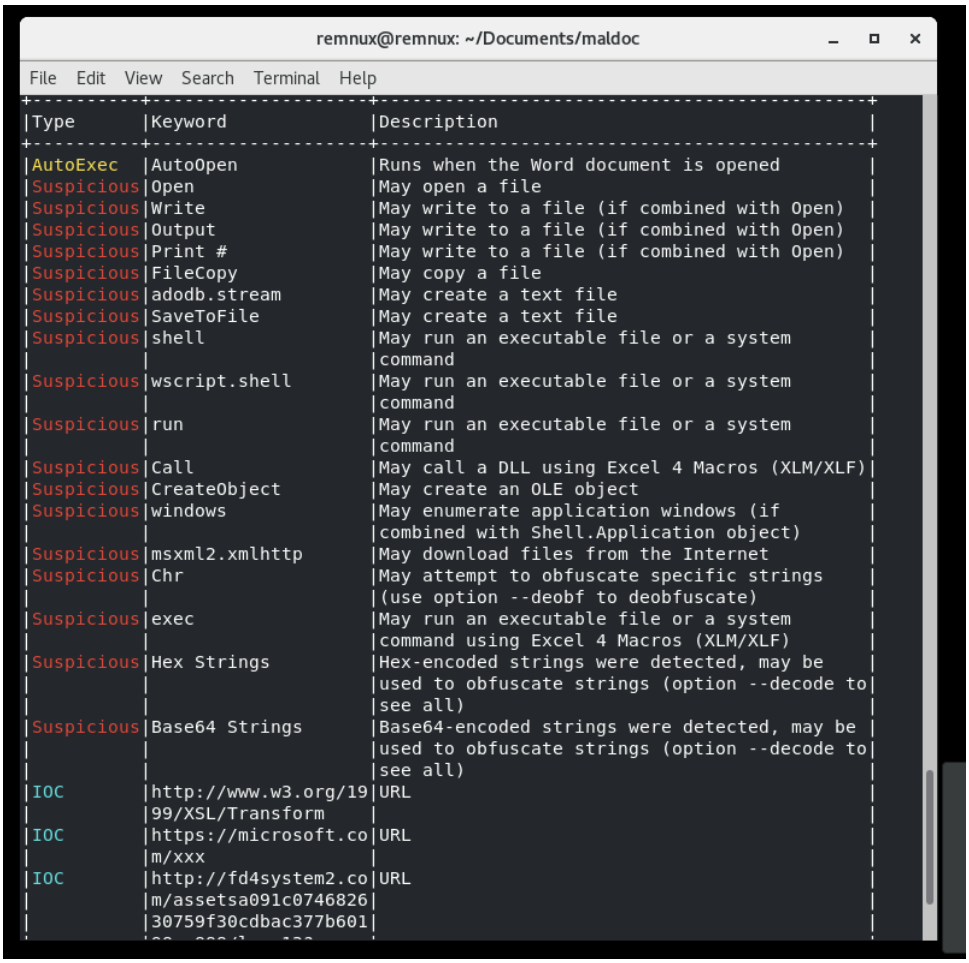
olevba is a tool that parses Mircosoft Office OLE and OpenXML files and extracts, deobfuscates and analyses malicious VBA macros.

The below sample is used to illustrate the next steps:

- SHA256: 0e865e28928cd37b6783a508d36bd905c816d9271412740bfe5b2b6bfaaeec26

- **`olevba maldoc2.doc`**
  - **`--deobf`** – attempts to desobfuscate VBA expressions
  - **`-a`** – displays only analysis results, not the macro source code
  - **`--decode`** – displays all the obfuscated strings with their decoded content (Hex, Base64, StrReverse, Dridex, VBA).





Type	Keyword	Description
AutoExec	AutoOpen	Runs when the Word document is opened
Suspicious	Open	May open a file
Suspicious	Write	May write to a file (if combined with Open)
Suspicious	Output	May write to a file (if combined with Open)
Suspicious	Print #	May write to a file (if combined with Open)
Suspicious	FileCopy	May copy a file
Suspicious	adodb.stream	May create a text file
Suspicious	SaveToFile	May create a text file
Suspicious	shell	May run an executable file or a system command
Suspicious	wscript.shell	May run an executable file or a system command
Suspicious	run	May run an executable file or a system command
Suspicious	Call	May call a DLL using Excel 4 Macros (XLM/XLF)
Suspicious	CreateObject	May create an OLE object
Suspicious	windows	May enumerate application windows (if combined with Shell.Application object)
Suspicious	msxml2.xmlhttp	May download files from the Internet
Suspicious	Chr	May attempt to obfuscate specific strings (use option --deobf to deobfuscate)
Suspicious	exec	May run an executable file or a system command using Excel 4 Macros (XLM/XLF)
Suspicious	Hex Strings	Hex-encoded strings were detected, may be used to obfuscate strings (option --decode to see all)
Suspicious	Base64 Strings	Base64-encoded strings were detected, may be used to obfuscate strings (option --decode to see all)
IOC	http://www.w3.org/1999/XSL/Transform	URL
IOC	https://microsoft.com/xxx	URL
IOC	http://fd4system2.com/assetsa091c0746826/30759f30cdbac377b601	URL

The above table summarizes the content of the VBA macro in the office document. It permits to identify IOCs and others potentially malicious content.

mraptor

mraptor (MacroRaptor) is a script that identifies malicious VBA macros using a generic heuristic methodology.

"In a nutshell, mraptor detects keywords corresponding to the three following types of behavior that are present in clear text in almost any macro malware:

- A: Trigger of auto-execution
- W: Writing to the file system or memory
- X: Execution of a file or any payload outside the VBA context

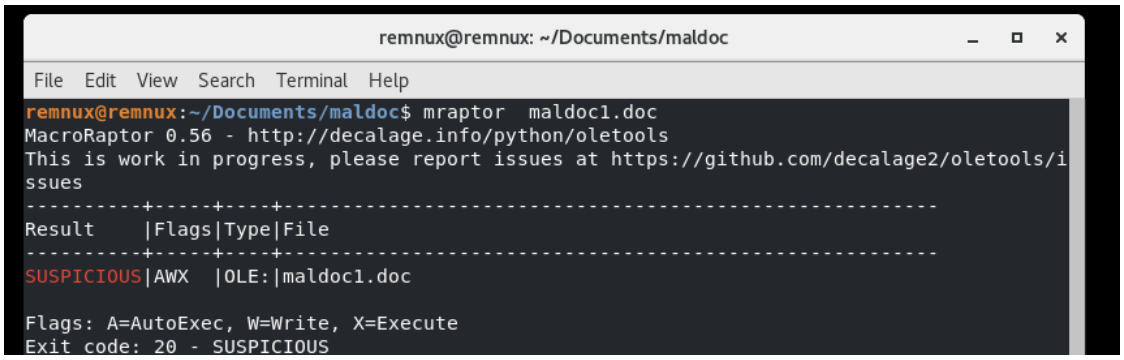
mraptor considers that a macro is suspicious when A and (W or X) is true."[1]

An exit code is returned based on the analysis result:

- 0: No Macro
- 1: Not MS Office
- 2: Macro OK
- 10: ERROR
- 20: SUSPICIOUS

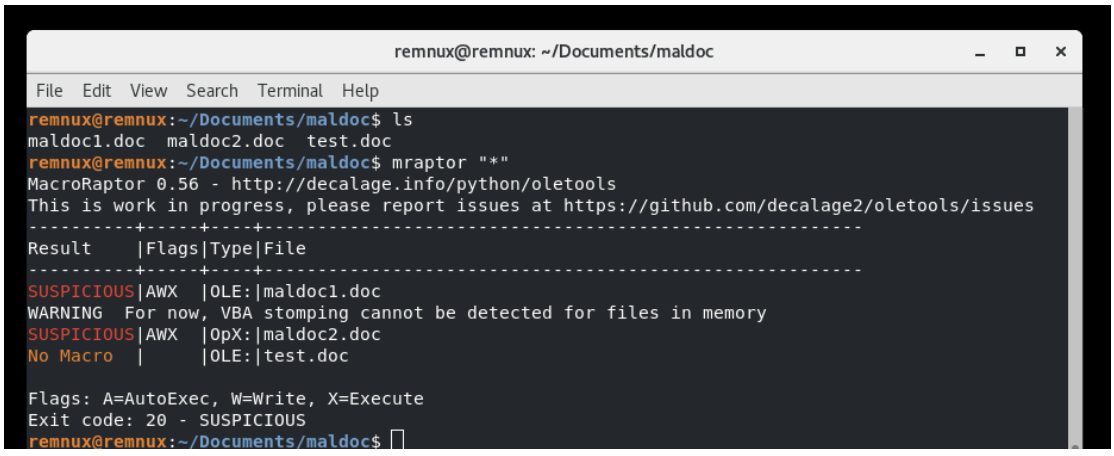
mraptor permits to quickly identify if a Microsoft Office document contains a malicious macro.

- **`mraptor file.doc`**



Mraptor is an excellent tool to scan several OLE files at a time.

- **`mraptor "directory_name/*"`**



[1] https://github.com/decalage2/oletools/wiki/mraptor

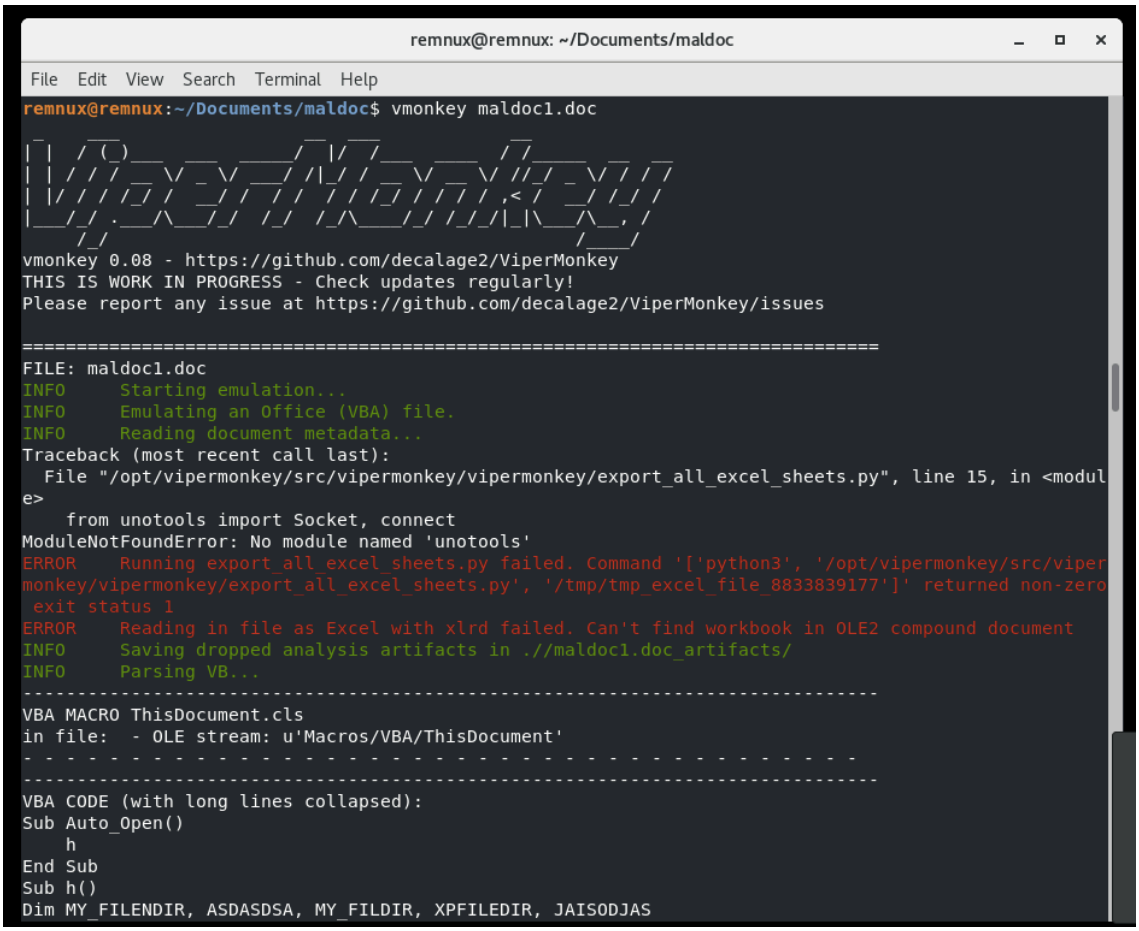
The *maldoc* directory contains three OLE files. As you can see in the above output from mraptor, two files are detected as malicious and *test.doc* does not contain any macro.

ViperMonkey

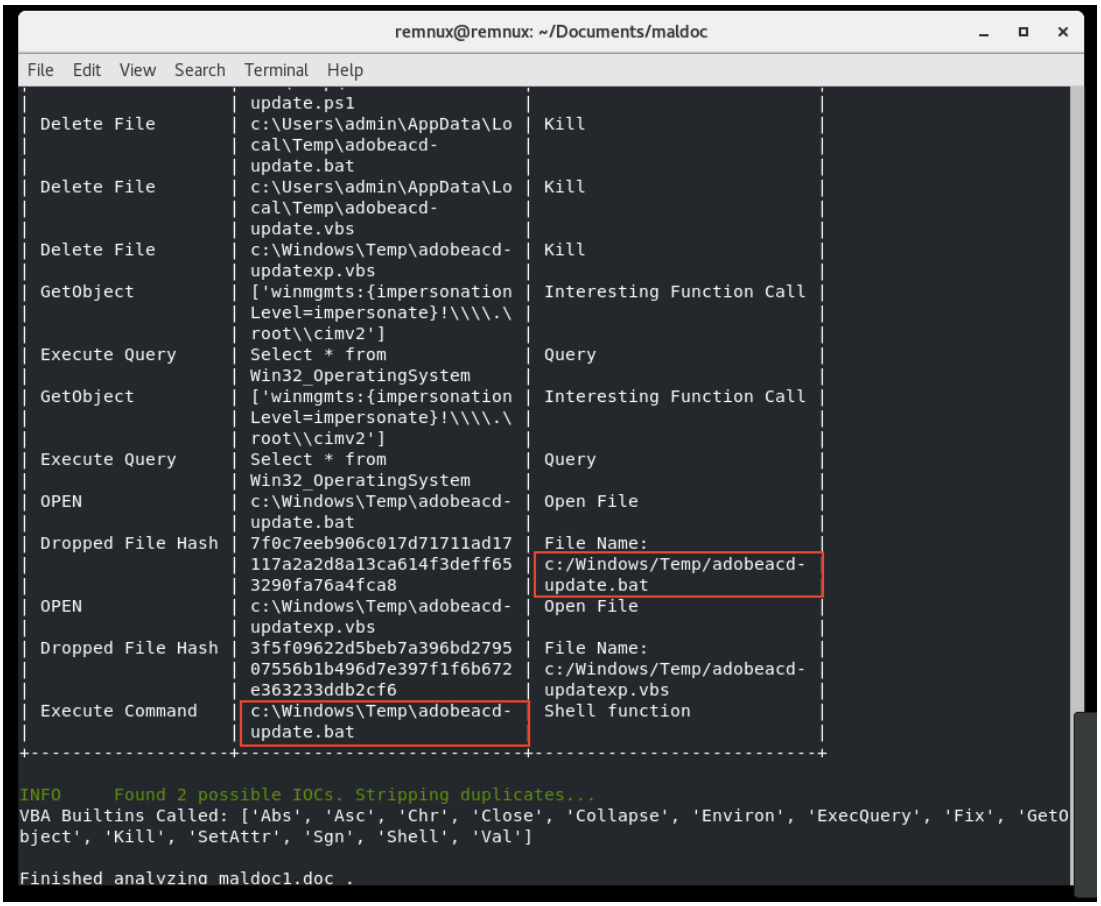
ViperMonkey "a VBA Emulation engine written in Python, designed to analyze and deobfuscate malicious VBA Macros contained in Microsoft Office files."[2]

As you can see in the ViperMonkey output (second screenshot), the IOCs found during the analysis are displayed in a table at the end of the analysis. It is a great tool to emulate the actions of the VBA macro(s), however it does not work with every malicious documents.

- **vmonkey maldoc1.doc**
  - **-p** – creates a log file



[2] https://github.com/decalage2/ViperMonkey



LOffice

Loffice (Lazy Office Analyzer) is a script that utilizes WinAppDbg to extract URL, VBA script and JavaScript form OLE files. Loffice uses various exit-modes which determine if execution is to be aborted.

- url: exit when the first URL is found
- proc: exit if a new process is to be created
- thread : before resuming a suspended thread (RunPE style)
- none: do not interrupt execution, URL and file information will still be printed.

WARNING: use loffice only in a safe malware analysis environment (e.g. properly configured Windows Flare VM) because depending on the chosen exit-modes, the system can be compromised. Before performing this analysis, I recommend reading the Dynamic Analysis Section.

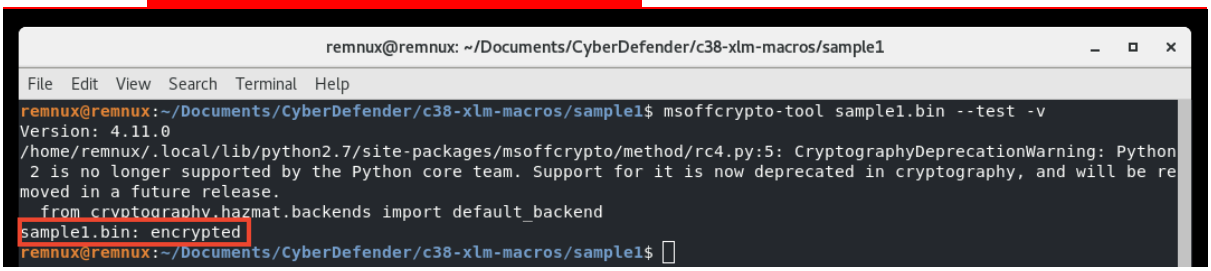
To use loffice, the correct parameters have to be selected.

- **python loffice.py type_file exit-on path_to_file -p path_to_office**
  - type of file to analyse: word, excel, power, script
  - exit-on: url, proc, thread, none
  - path_to_type_file: C:\ Program Files\Microsft Office\root\Office15

msoffcrypto-crack.py

python msoffcrypto-crack.py sample1.bin

- fb5ed444ddc37d748639f624397cff2a

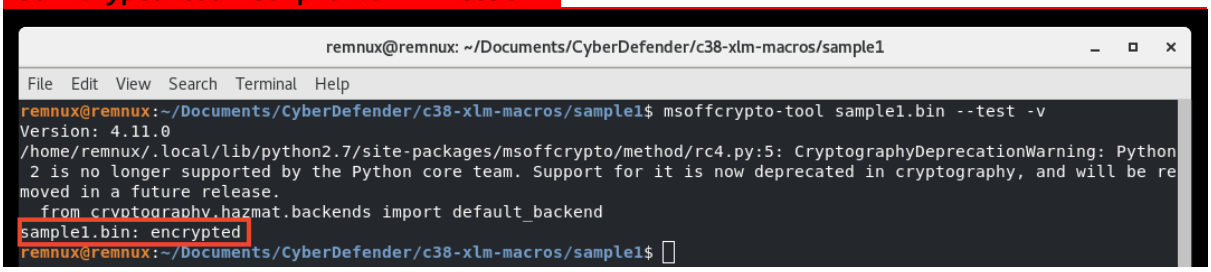


msoffcrypto-tool

msoffcrypto-tool - https://pypi.org/project/msoffcrypto-tool/

- Test if the file is encrypted or not (exit code 0 or 1 is returned):
- fb5ed444ddc37d748639f624397cff2a

msoffcrypto-tool sample1.bin --test -v



msoffcrypto-tool sample1.bin sample1.unenc -p VelvetSweatshop

Microsoft Office Open XML File

Microsoft Office Open XML (OOXML) document is the new Microsoft Word document format introduced in 2007. Docx file is a collection of XML files (metadata files and documents) contained in a zipped archive.

The below sample is used to illustrate the next steps:

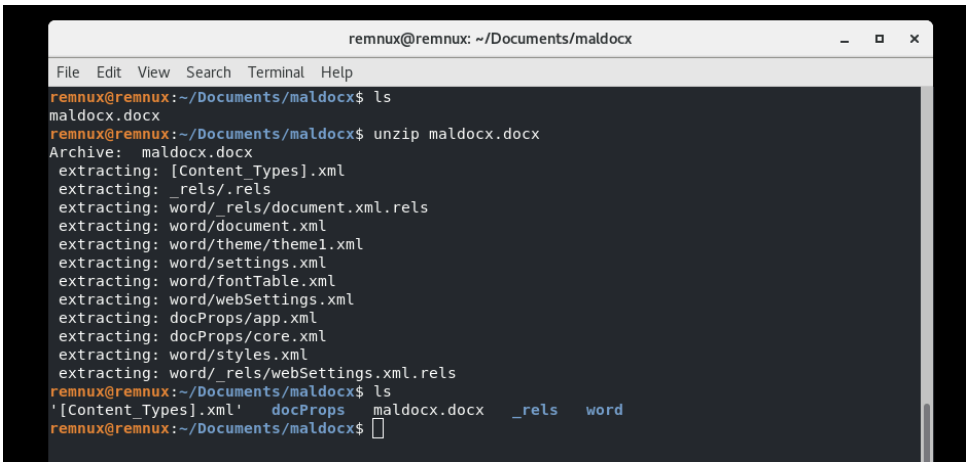
- SHA256: efe907b594cf491d0c8f013560d87baf1bea20ac3bfa7bb35d6a51b83b4ee357

Example 1

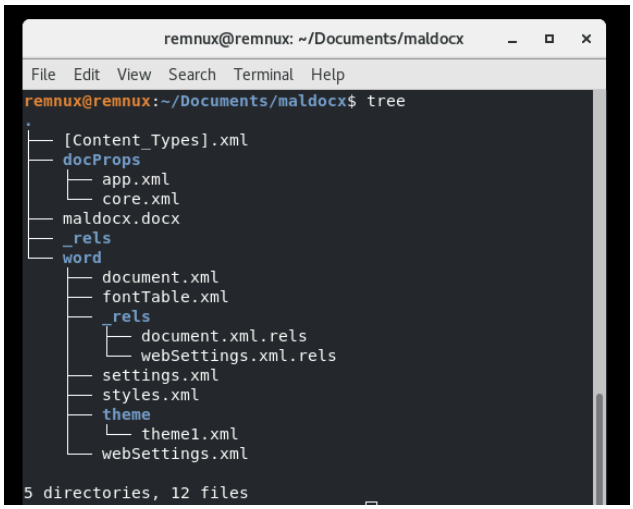
Unzip

OOXML documents are zip file containers that contains data. By unzipping the OOXML file, data in the document can be analyzed.

- **`unzip maldocx.docx`**



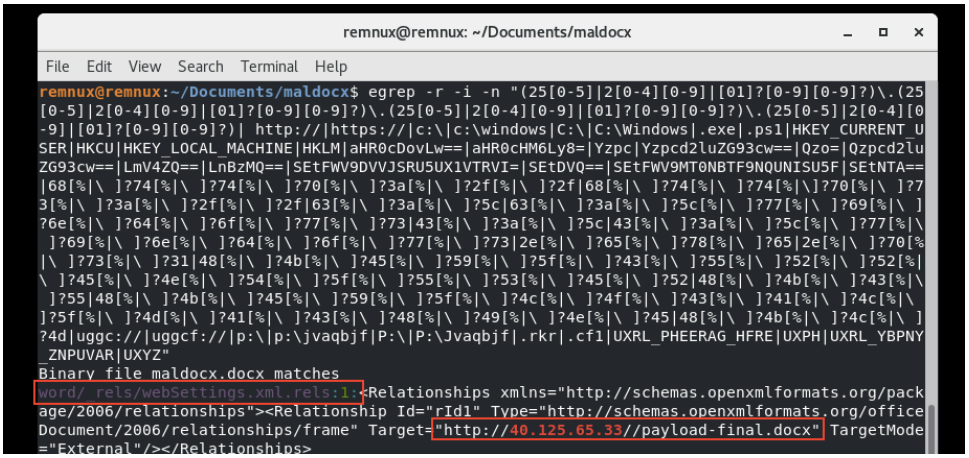
The structure of the zipped file can be viewed using the **`tree`** command.



egrep

The potentially malicious content can be identified by manually reviewing the content of the xml files. I created a regex that permits to identify malicious encoded or not URL, IP, command lines, etc.

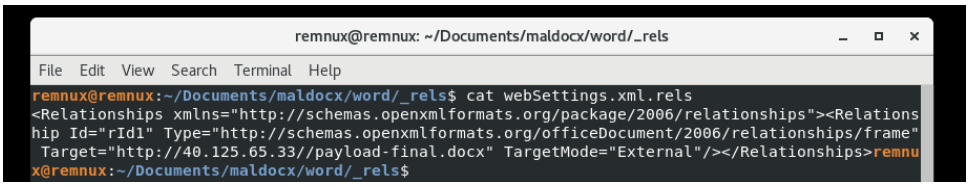
- **egrep -r -i -n Regex ***



The above screen displays the URLs that downloads the second stage of the malware.

For the demonstration, the content of *webSettings.xml.rels* was outputted and displays the same result.

- **cat webSettings.xml.rels**



Example 2

The below sample is used to illustrate the next steps:

- SHA256: 6ae5583ec767b7ed16aaa45068a1239a827a6dae95387d5d147c58c7c5f71457

Unzip

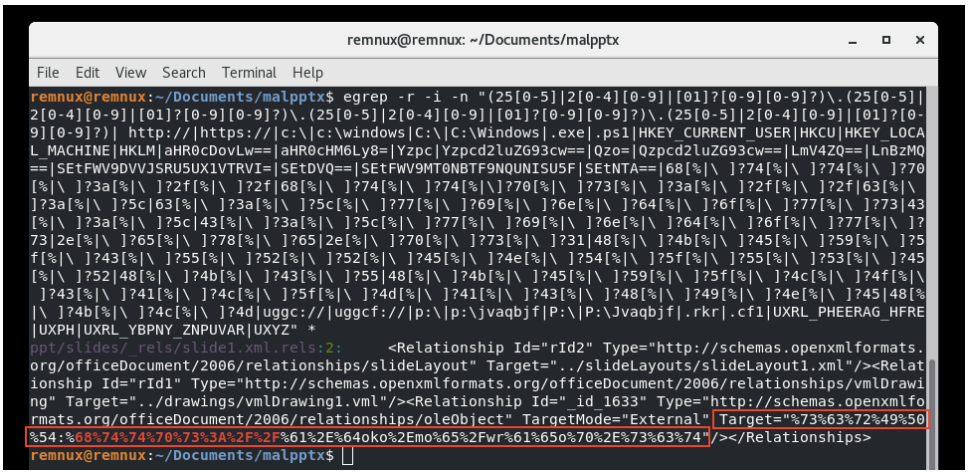
The ppsx file is unzipped.

- **unzip SCAN.ppsx**

egrep

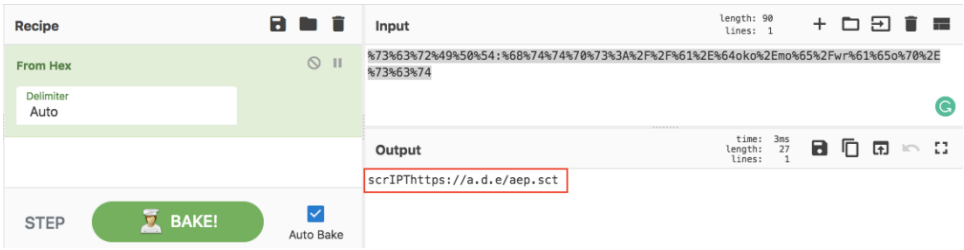
The regex is run to identify IOCs.

- **egrep -r -i -n Regex ***



An encoded hexadecimal string was detected.

Using for example CyberChef the encoded hexadecimal strings can be decoded.



Olevba

Olevba can also be used to analyse Microsoft Office OpenXML (cf. Olevba in the Microsoft Office OLE section).

XLMMacroDeobfuscator

https://github.com/DissectMalware/XLMMacroDeobfuscator

xlmdeobfuscator --file sample1.unenc

xlmdeobfuscator --file sample1.bin -p VelvetSweatshop

- b5d469a07709b5ca6fee934b1e5e8e38
- https://hatching.io/blog/excel-xlm-extraction/

Anti-Sandboxing

IF(GET.WORKSPACE(19),,CLOSE(TRUE))
IF(GET.WORKSPACE(42),,CLOSE(TRUE))
RETURN()

(It should be noted that GET.WORKSPACE(19) returns true if and only if a mouse is present and GET.WORKSPACE(42) returns true if and only if sound can be played on the computer).

Rich Text Format - RTF

Rich Text Format (RTF) is a file format developed by Microsoft to standardize data transfer between different word processing software and operating systems.

RTF files do not contain macros like office documents, however it can contain linked and embedded objects. The data related to this embedded object is stored in a hexadecimal format in “objdata” RTF sub-destination.

Example 1

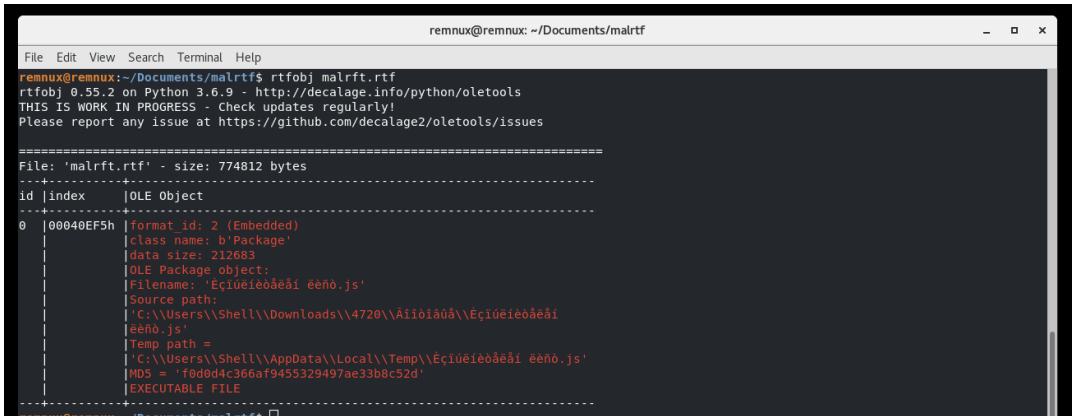
The below sample is used to illustrate the next steps:

- SHA256: fe201e51084b54a1a855ea4e765448d305157064824e026249d6b8d7326f1b73

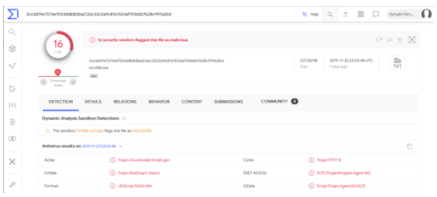
Rtfobj

rtfobj is a script that automatically extract embedded objects from rtf files.

- **`rtfobj malrft.rtf`**



The output displays that an executable file (md5: f0d0d4c366af9455329497ae33b8c52d) is embedded in the rtf file. By submitting the hash value on VirusTotal, it can be determined that the object is known as malicious.



The object can be extracted with the following command:

- **`rtfobj -s 0 malrft.rtf`**
  - **`-s`** – to save a specific object.

It is important to know that rtfobj does not work for every malicious files, this is why it is important to be able to manually extract embedded objects. With rtfdump, I will demonstrate how we can extract the same executable manually.

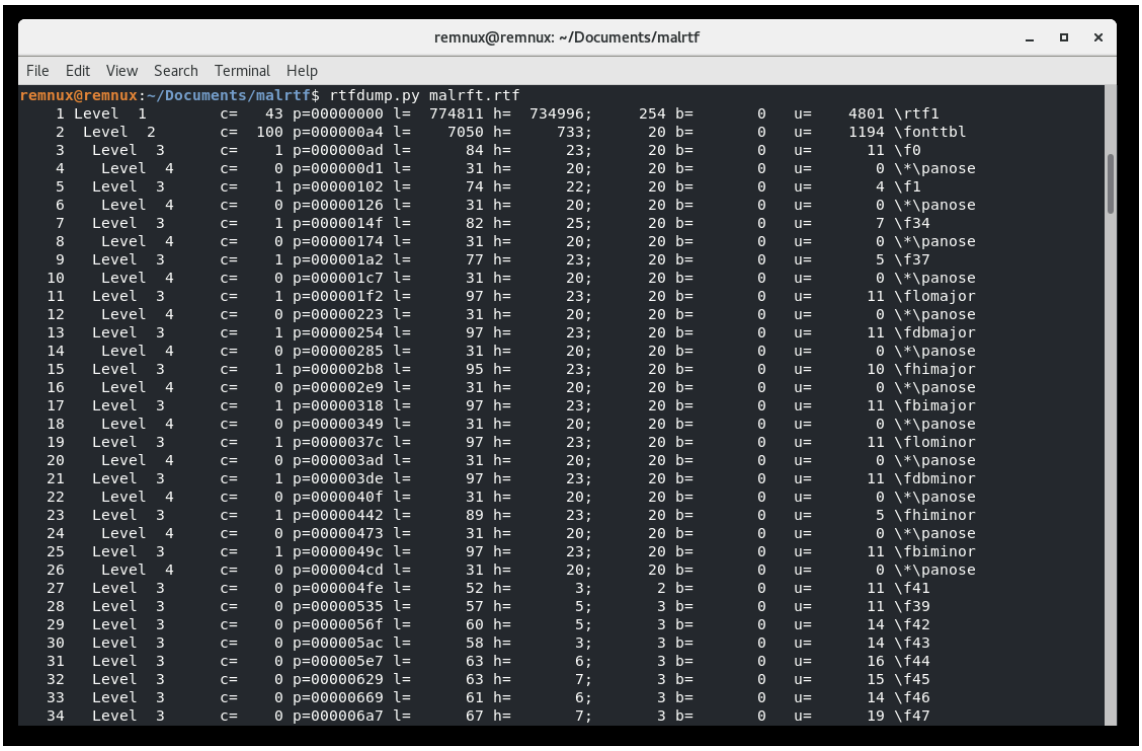
Rtfdump

-d, --dump perform dump
-x, --hexdump perform hex dump
-a, --asciidump perform ascii dump
-A, --asciidumprle perform ascii dump with RLE
-H, --hexdecode decode hexadecimal data; append 0 in case of uneven number of hexadecimal digits

Rtfdump (Didier Stevens) is a script that helps to identify and extract objects embedded in RTF files.

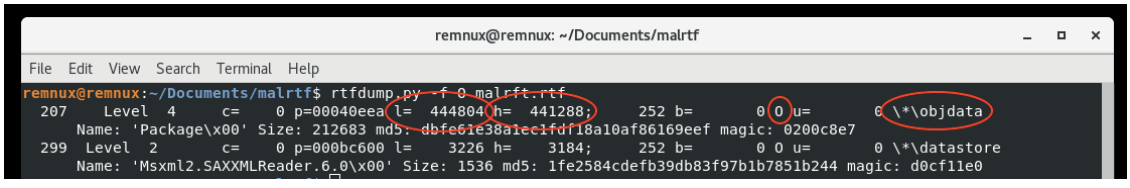
Rftdump permits to dump all items from the rtf files.

- **rtfdump.py malrtf.rtf**



There are 299 items in the above output. It is possible to filter out for the items that contain an object using **-f O**

- **rtfdump.py -f O malrtf.rtf**



This rtf file contains two objects (“O”)

- “l=444804” is the number of characters in the object
- “h=441288” is the number of hexadecimal characters in the object

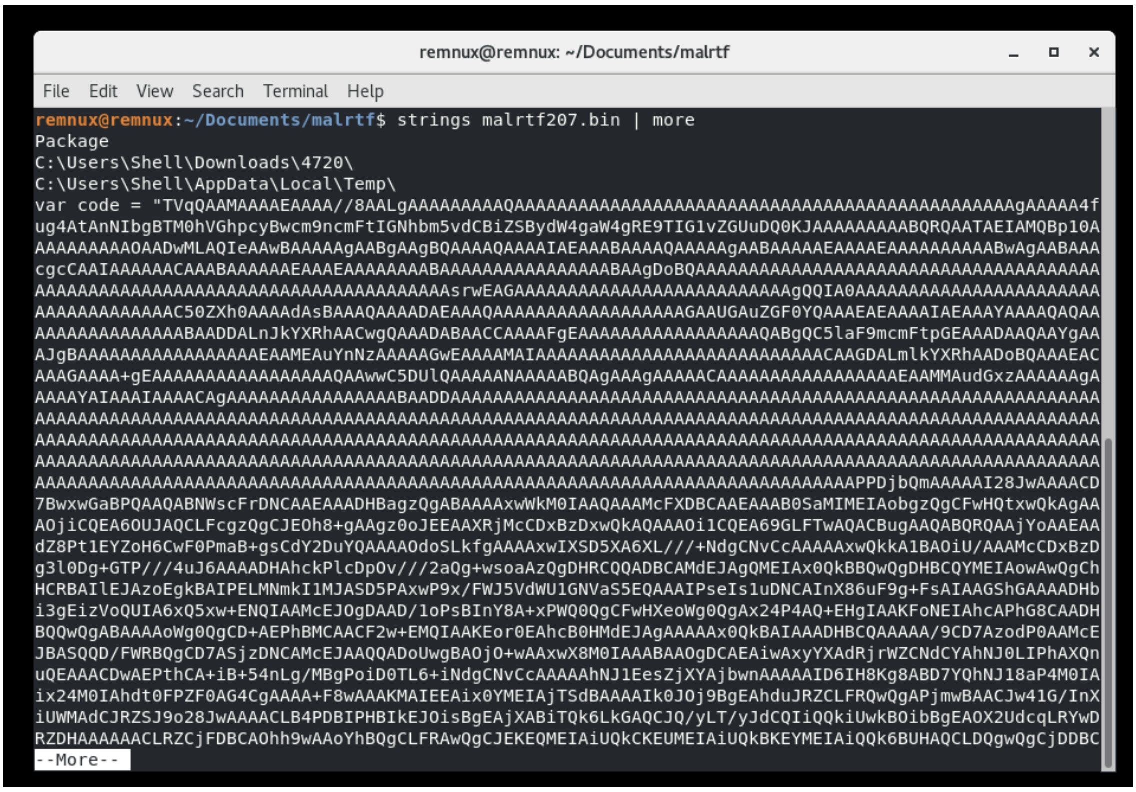
NB: having a “l” value significantly higher than the “h” value can be a sign of obfuscation.

The item 207 contains some data "\objdata". It is possible to dump and decode the content of the data with the below command:

- **rtfdump.py -s 207 -H -d malrft.rtf >malrtf207.bin**

The content of the dump file "malrtf207.bin" can be checked with the **strings** command.

- **`strings malrtf207.bin | more`**



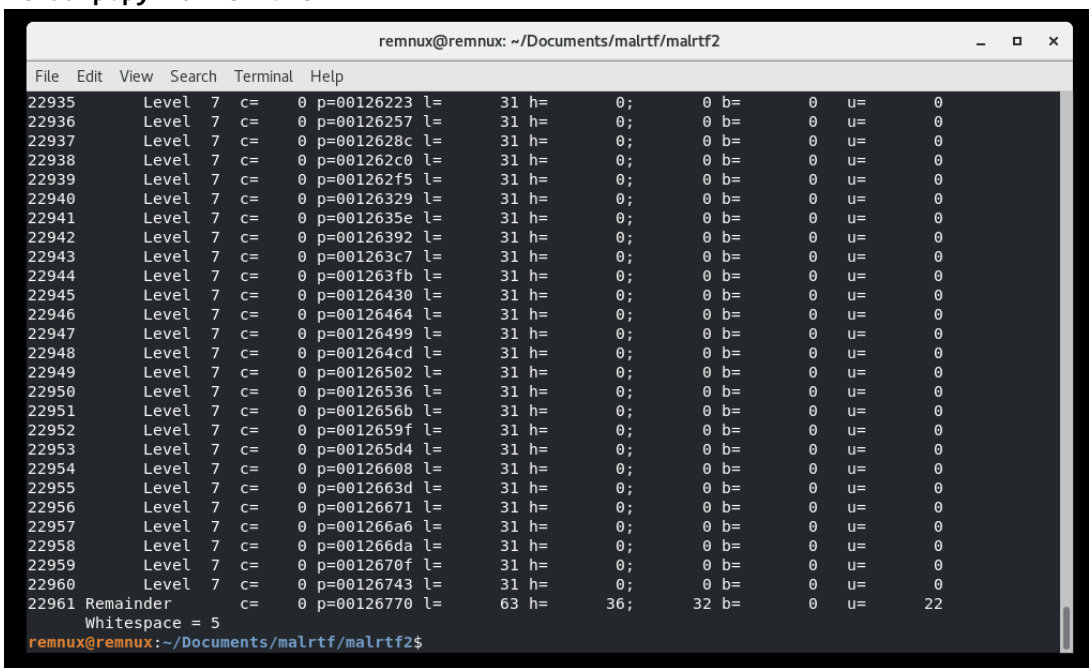
The binfile can be manually decoded to determine the actions and identify IOCs. If it is too hard to desobfuscate, the file can be analyzed using dynamic malware analysis.

Example 2

The below sample is used to illustrate the next steps:

- SHA256: 5627020b060387afb16f9d2de2b3bbd6ea870fe91d5658caabca8209227ad5f1

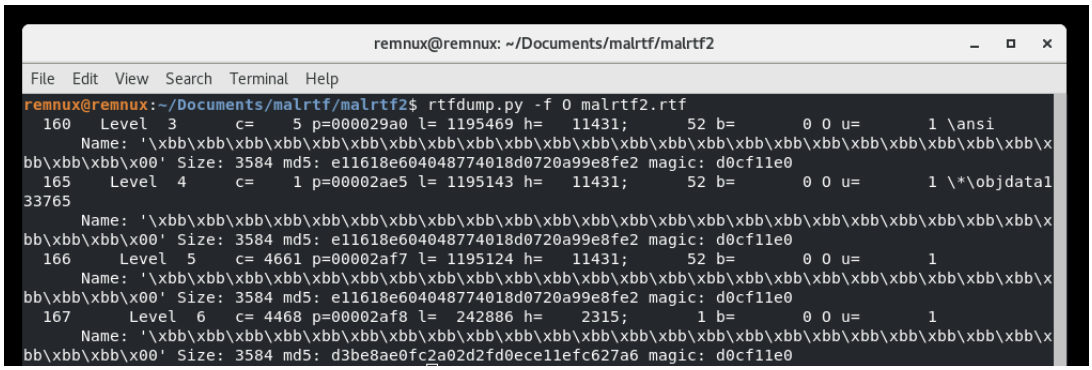
- **`rtfdump.py malrtf2.rtf`**



There are more than 22'000 thousandsitems. It is a clear sign of obfuscation.

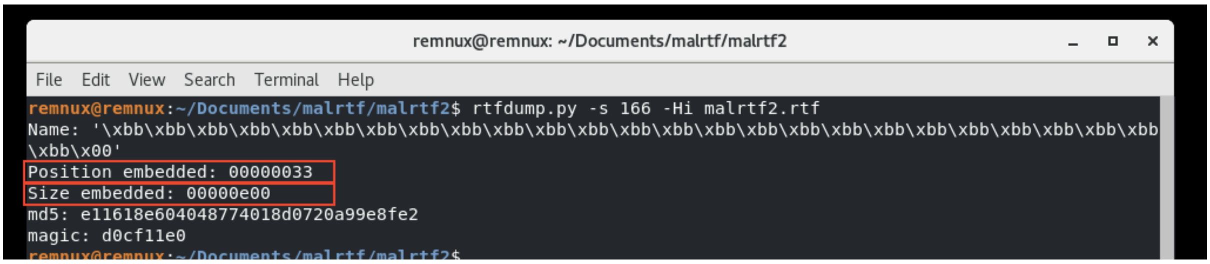
The below command can be used to filter out the items with embedded objects.

- **`rtfdump.py -f O malrtf2.rtf`**

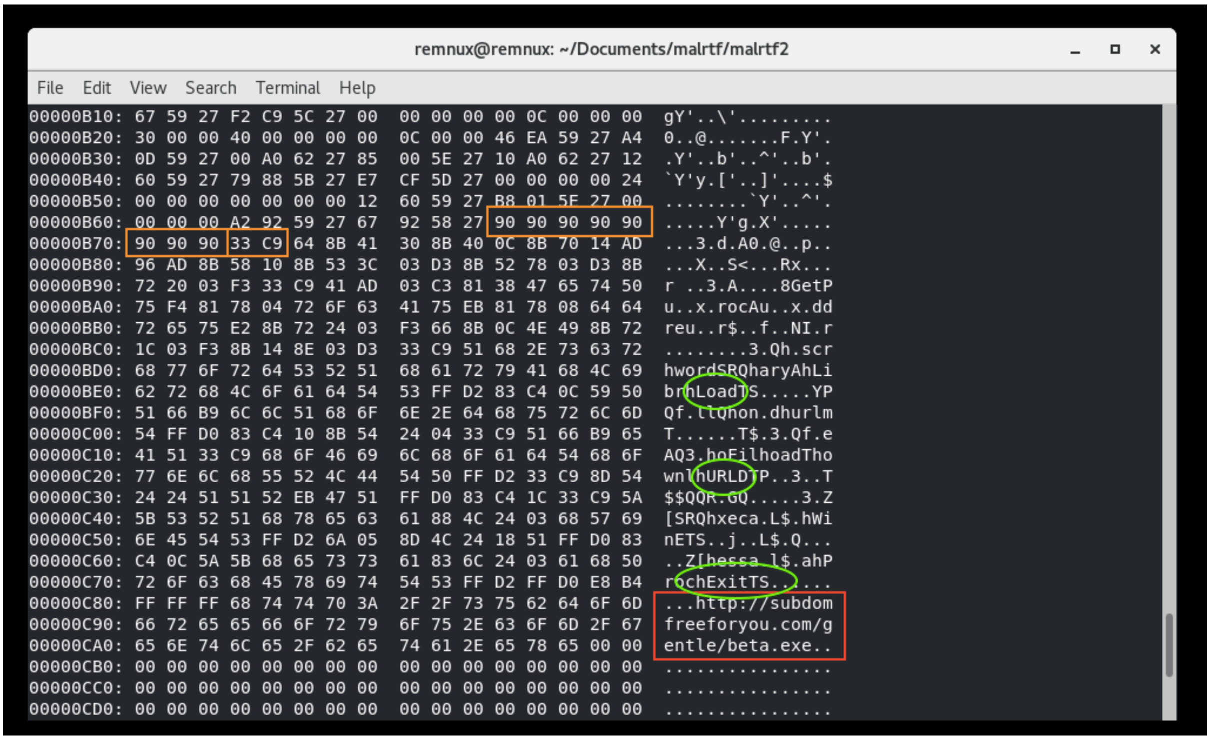


Additional information can be extracted for the object 166

- **`rtfdump.py -s 166 -Hi malrtf2.rtf`**



- **`rtfdump.py -s 166 -H -c "0x33:0xe00" malrtf2.rtf`**

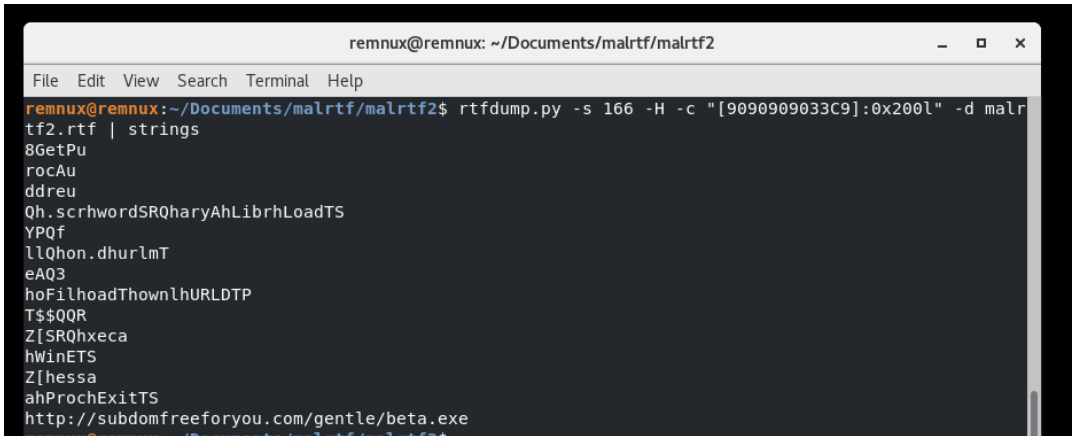


The output of this command is interesting.

- In red, a URL is present "hXXp://subdomfreeforyou[.]com/gntle/beta.exe"
- In green, before the URL, there are lots of bytes, but interesting words can be identified: "Load", "URL" and "Exit".
- Before, highlighted in orange, no-operation (a NOP-sled) sequence is present (\x90), which is a good indicator of an exploit for a shell code.

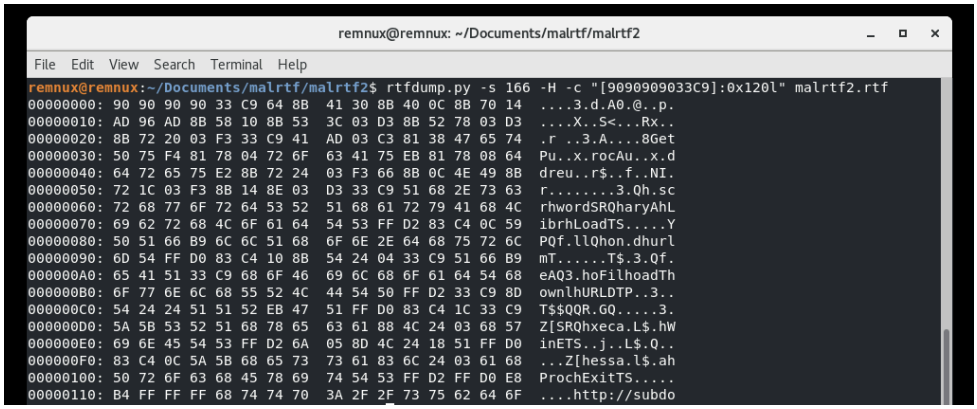
The following command can be used to decode and extract the strings.

- **rtfdump.py -s 166 -H -c "[9090909033C9]:0x200l" -d malrtf2.rtf | strings**



Using rtfdump, we will look for "*90 90 90 90 90 33 C9*" specific sequence.

- **rtfdump.py -s 166 -H -c "[9090909033C9]:0x200l" malrtf2.rtf**
  - -c "[9090909033C9]:0x200l"
    - [9090909033C9] – strings
    - 0x120l – length of the strings



The shellcode can be extracted with the below command to be further analyzed.

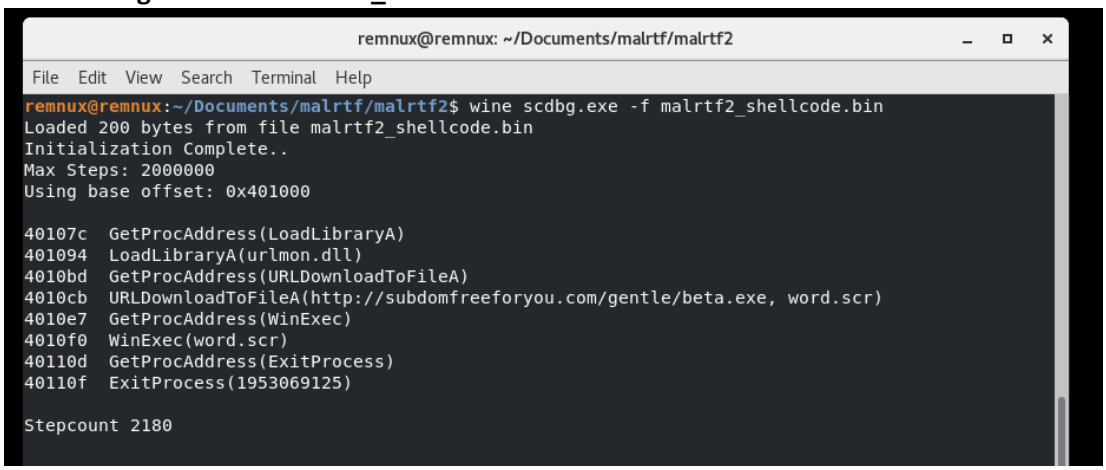
- **rtfdump.py -s 166 -H -c "[9090909033C9]:0x200l" -d malrtf2.rtf > malrtf2_shellcode.bin**

SCBD - ShellCode

Scdbg.exe is a program that runs a shellcode and emulates 32bit processor, memory and Windows API calls. It displays the Windows API call performed by the shellcode.

The emulation of shellcode is an easy way to identify the action performed by the shellcode.

- **`wine scdbg.exe -f malrtf2_shellcode.bin`**



Using scdbg, the actions performed by the shellcode can be analyzed. Basically, the above shellcode uses “URLDownloadToFileA” to download a file and writes this file in “word.scr”. Then, using “WinExec”, the “word.scr” file is executed.

JavaScript

Obfuscation techniques

- Formatting: Modify the code to make it more difficult to read. A code beautification program can be used to parse the JS code in order to reformat it.
- Extraneous Code: add extra lines of code that are not usewithin the code. To solve this obfuscation technique, looks for variables and codes that are only used once and delete them.
- Data Obfuscation: use operations to make data unreadable or confusing. Replace the readable values.
- Substitution: uses random name for variable. Read the code and rename the variable.

Example 1

The below sample is used to illustrate the next steps:

- SHA256: e6889f6e98f554a3fd6e4371e9a8323abbdfbced2031b8863ba84a3050c56bea

Example 2

I used the JavaScript (object 6) that we extracted from the analysed pdf in the pdf section (cf. pdf section).

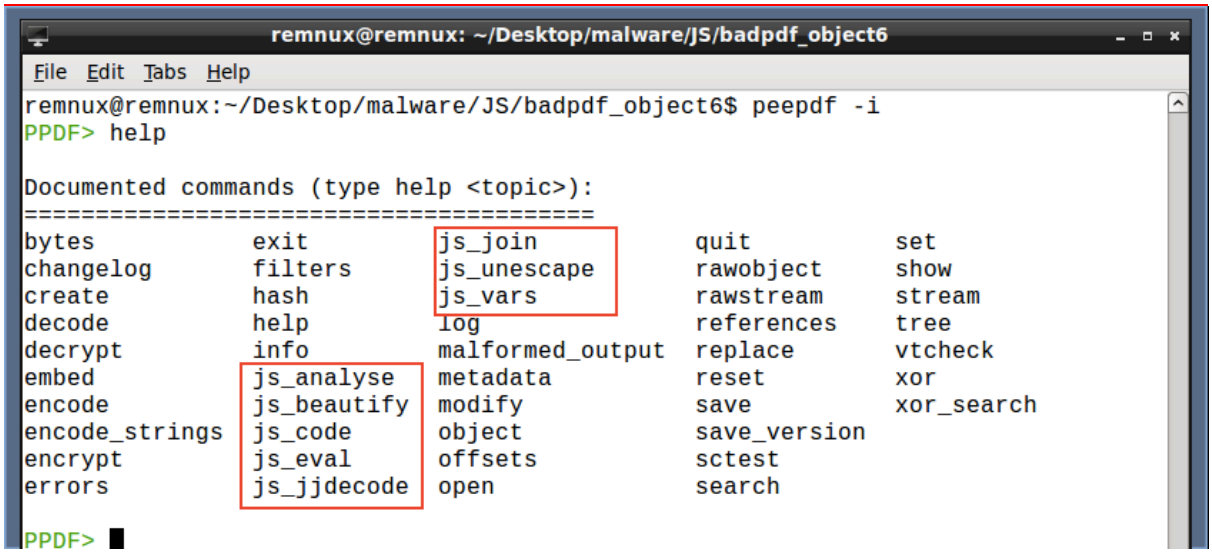
- PDF – SHA256: e6889f6e98f554a3fd6e4371e9a8323abbdfbced2031b8863ba84a3050c56bea

Peepdf – js_analyze

Peepdf is a tool that is used for the analysis of pdf cf. pdf section, moreover it has also the capability to analysis JavaScript.

Peepdf has several js analysis tool.

- **peepdf -i**



https://eternal-todo.com/blog/extract-streams-shellcode-peepdf

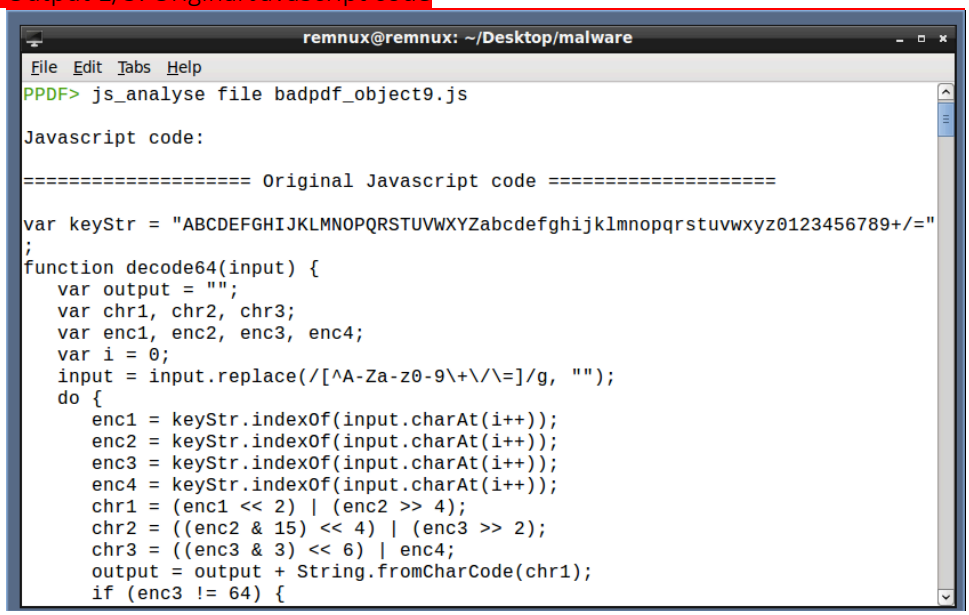
Peepdf – js_analyze

The analysed JavaScript was extracted from the following malicious pdf file:

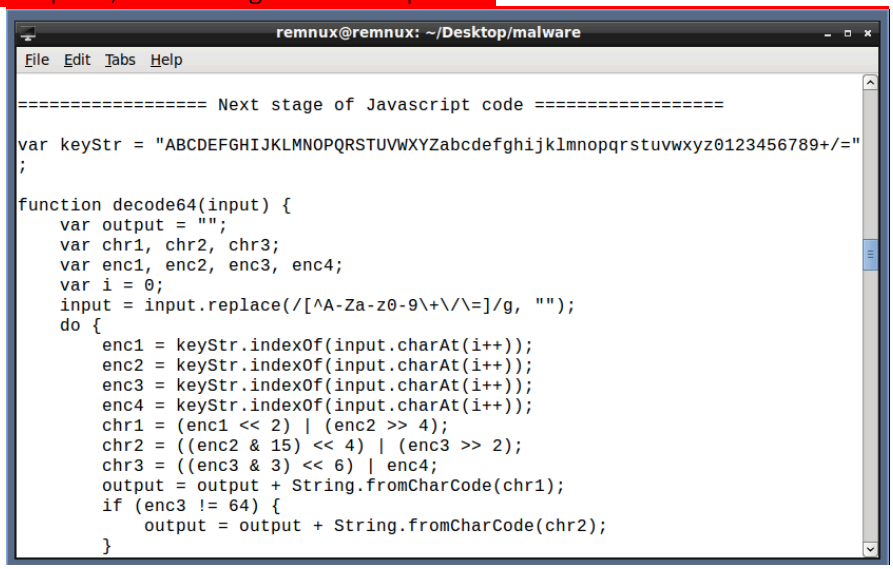
- SHA256: 6f33cdf096a8744a1ad0ae15d67784405d6182b529b8ccd1b332377fbd4169f7

- **peepdf -i**
  - Peepdf interactive mode

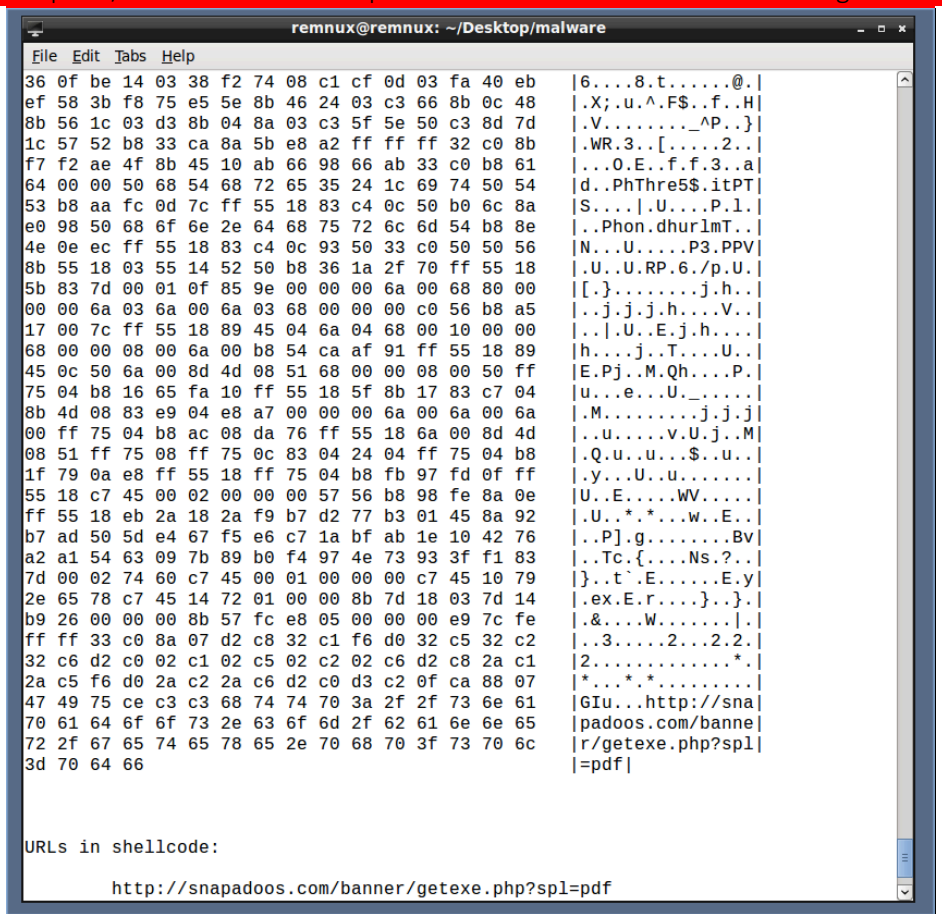
- **js_analyse file badpdf_object9.js**
  - Analyse the PDF file
  - Output 1/3: Original JavaScript code



- Output 2/3: Next stage of JavaScript code



- Output 3/3: decode the JS and provide the URL used for the second stage of the attack



Js_analysis manages to extract one URL http://snapadoos[.]com/banner/getexe.php?spl=pdf known as malicious on VirusTotal.

Peepdf – js_analyze

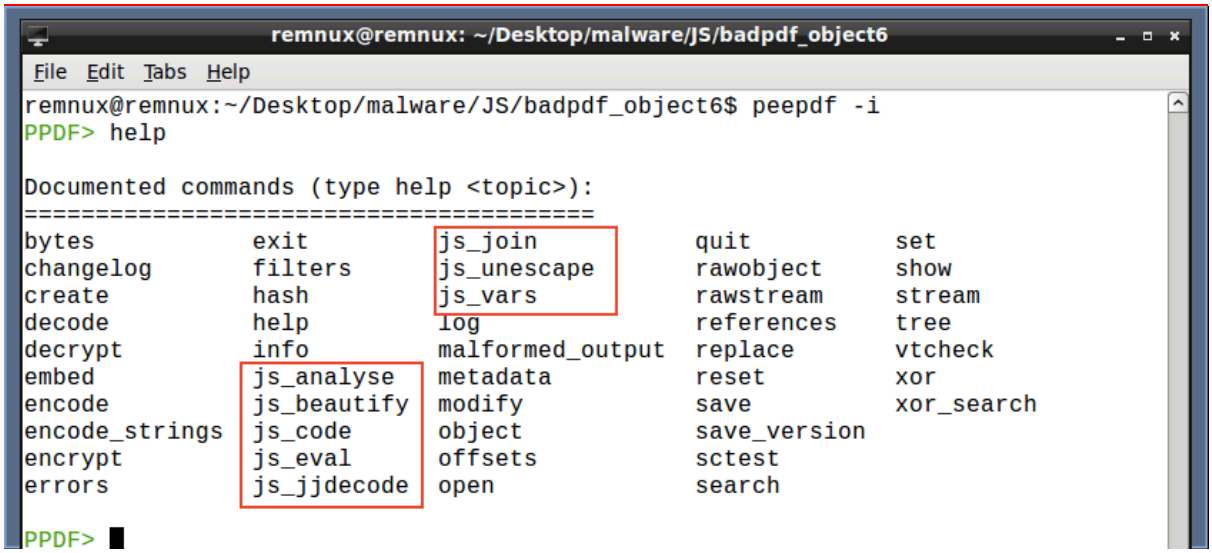
Peepdf is a tool that is used for the analysis of pdf (cf. pdf section), moreover it has also the capability to analyze JavaScript.

The analyzed JavaScript was extracted from the following malicious pdf file:

- SHA256: 6f33cdf096a8744a1ad0ae15d67784405d6182b529b8ccd1b332377fbd4169f7

Peepdf has several js analysis tools.

- **peepdf -i**



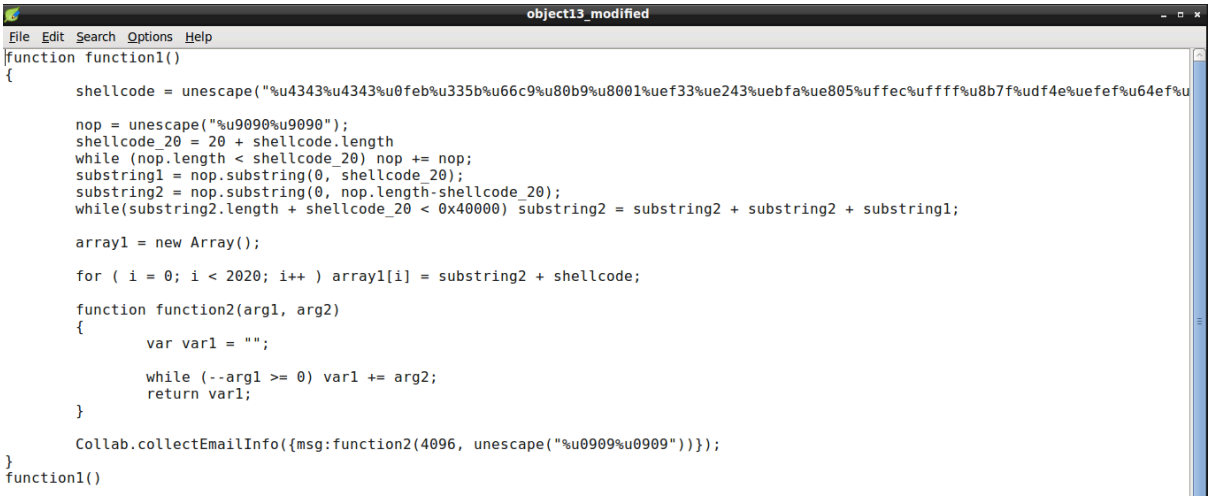
Js_analysis permits to analyze and decrypt the JavaScript automatically.

- **js_analyse file object13**

Peepdf

- **peepdf -i**
- **js_beautify file object13**
  - js_beautify can be used to structure the JS code correctly to be more readable.

The JS code contains some long variable names to make the analysis more difficult. Rename them to make the reading of the code easier variables.

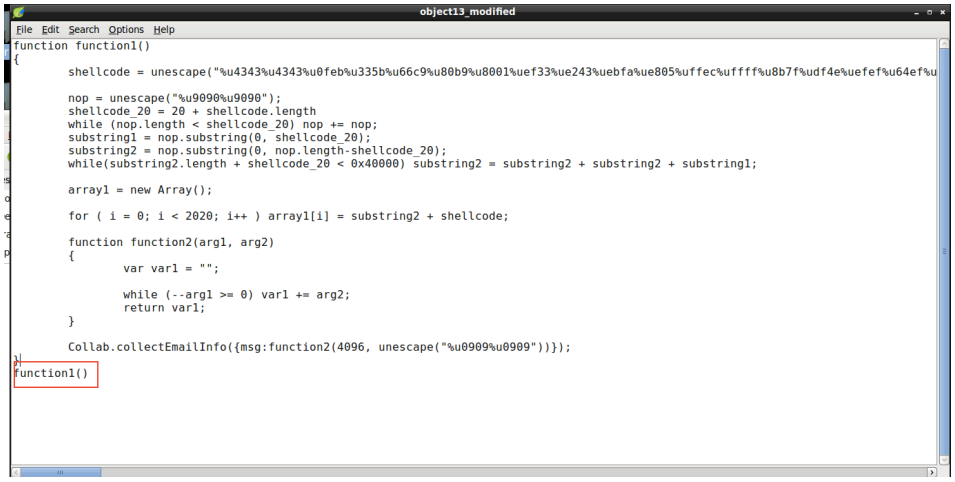


Js-didier

Js-didier is a modified version of the Spidermonkey tool.

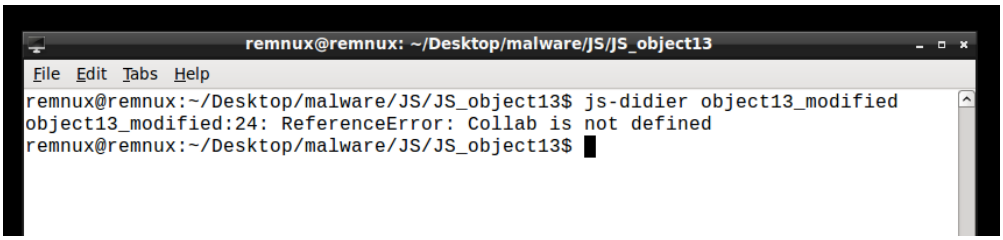
- **js-didier object13**
  - To decode the JavaScript

In this situation, nothing is decoded because the function containing the shellcode is never called and then run. Consequently, add *function1()* at the end of the script to run the code.



Even after calling the function, nothing was run. This is because js-didier did not identify any “eval” or “document.write()” command.

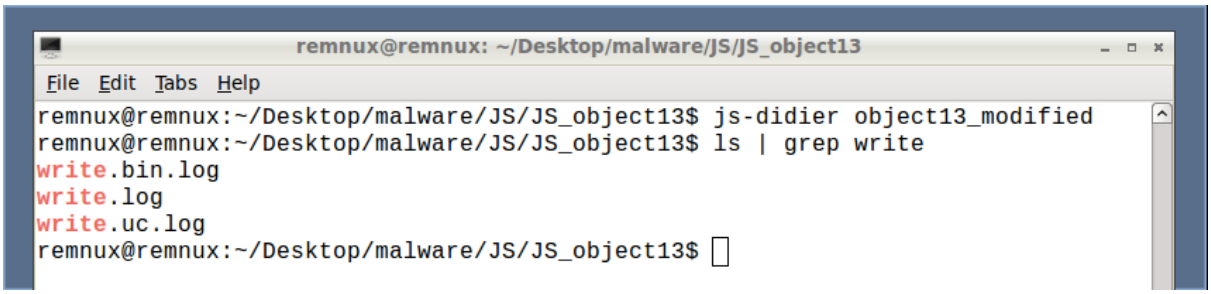
- **js-didier object13**



Add the following code after the shellcode: “document.write(shellcode); return;” to run it and return it output.

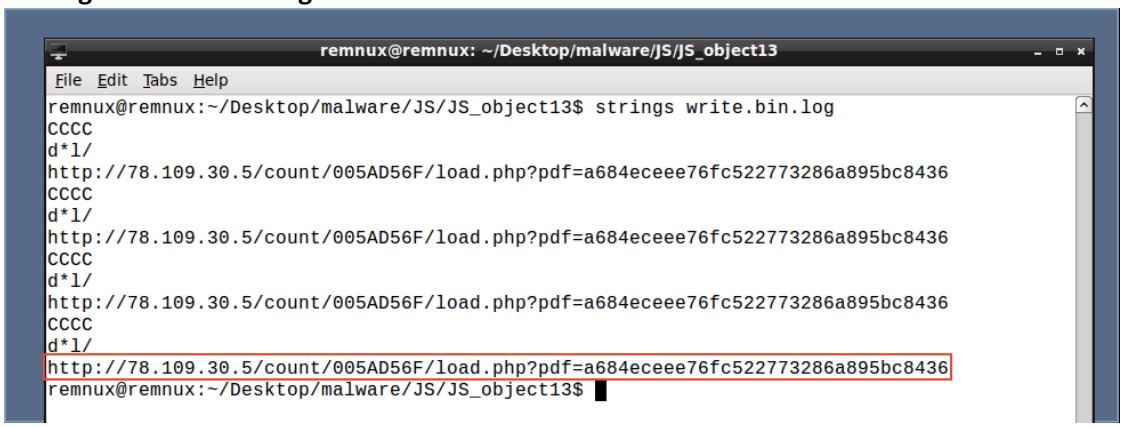
Run again js-didier object13, and finally the script runs successfully. It created tree output files:

- Write.bin.log: binary representation of the document
- Write.log: ASCII version
- Write.uc.log: Unicode version



The content of the file can be outputted.

- **`strings write.bin.log`**



http://78.109.30[.]5/count/005AD56F/load.php?pdf=a684eceee76fc522773286a895bc8436

JSDetox - Mozilla's JavaScript engine for debugging malicious JS.

- http://www.relentless-coding.com/projects/jsdetox/screencasts

https://github.com/svent/jsdetox

ShellCode

A shellcode is a small piece of hexadecimal code used as a payload by malware to perform various actions (e.g. download the second stage of the attack, contact the C2 sever, etc.).

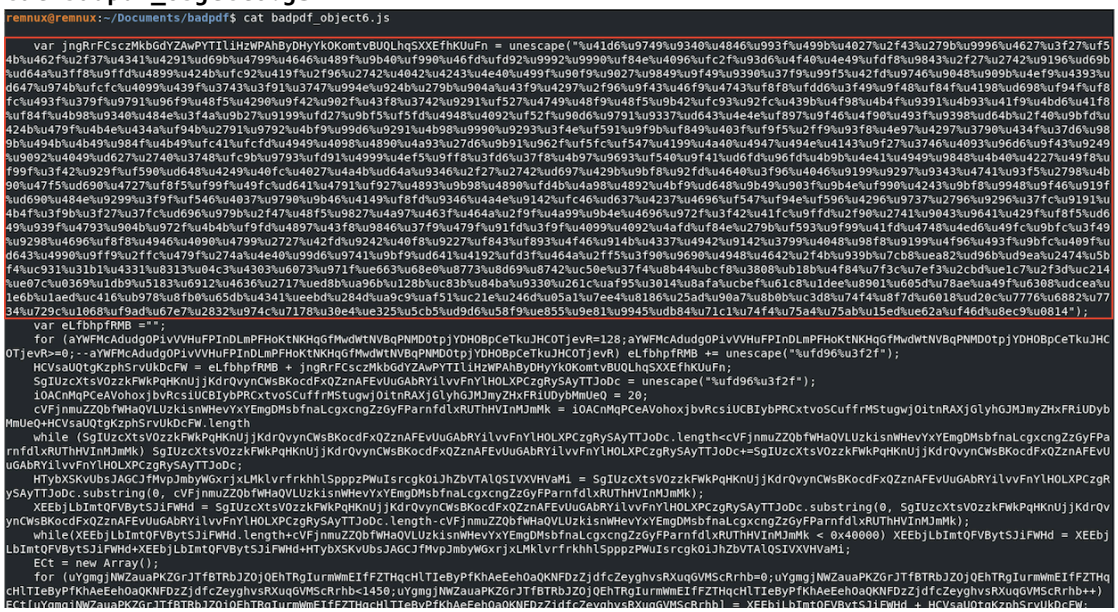
Shellcode can be stored/encoded in various formats:

- hexademical (68 74 74 70)
- percentage unicode (%u7468%u7074)
- backslash (\x68\x74\x74\x70)
- backslash unicode (\u7468\u7074)
- byte array ([0x68, 0x74, 0x74, 0x70])

The below JavaScript was extracted from the below pdf (cf. pdf section):

- SHA256: e6889f6e98f554a3fd6e4371e9a8323abbdfbced2031b8863ba84a3050c56bea

- **`cat badpdf_object6.js`**



It can be identified that the JavaScript code contains a *percentage Unicode* ("*%u*") encoded shellcode. To analyze the shellcode, it must be extracted before being converted in hexadecimal values.

https://resources.infosecinstitute.com/topic/hacking-pdf-part-2/

CVE - CVE-2008-2992

To decode the *percentage Unicode* shellcode, I recommend using the below code written by the authors of Practical Malware Analysis book. I pasted the shellcode in payload strings at the line 30.

Shellcode_unescape_decode.py

home > remnux > Documents > badpdf > Shellcode_unescape_decode.py

```
1
2   #Code Created by Practical Malware Analysis
3   #https://nostarch.com/malware
4   def decU16(inbuff):
5       """Manually perform JavaScript's unescape() function. """
6       i=0
7       outArr = [ ]
8       while i < len(inbuff):
9           if inbuff[i] == '"':
10              i += 1
11          elif inbuff[i] == '%':
12              if ((i+6) <= len(inbuff)) and (inbuff[i+1] == 'u'): #it's a 2-byte "unicode" value
13                  currchar = int(inbuff[i+2:i+4], 16)
14                  nextchar = int(inbuff[i+4:i+6], 16)
15                  #switch order for little-endian
16                  outArr.append(chr(nextchar))
17                  outArr.append(chr(currchar))
18                  i += 6
19              elif (i+3) <= len(inbuff):
20                  #it's just a single byte
21                  currchar = int(inbuff[i+1:i+3], 16)
22                  outArr.append(chr(currchar))
23                  i += 3
24              else:
25                  # nothing to change
26                  outArr.append(inbuff[i])
27                  i += 1
28      return ''.join(outArr)
29
30  payload = "%u41d6%u9749%u9340%u4846%u993f%u499b%u4027%u2f43%u279b%u9996%u4627%u3f27%uf54b%u462f%u2f37%u434
31
32  outFile = file('badpdf_object6_shellcode_hex.bin', 'wb')
33
34  outFile.write(decU16(payload))
```

The script can be run using the below command line:

- **`python shellcode_unescape_decode.py`**

The script will create a file named after the *outFile* variable, here “badpdf_object6_shellcode_hex.bin”, with the shellcode in hexadecimal.

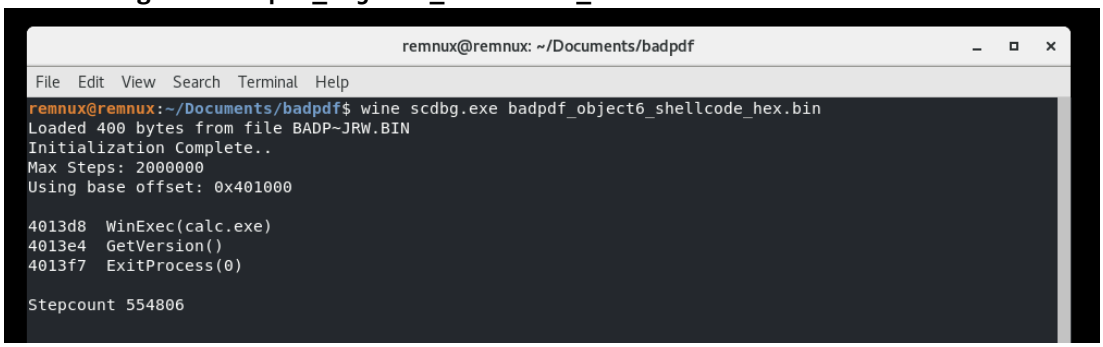
The shellcode code can then be run in a safe environment (dynamic malware analysis) or emulated with scdbg.exe.

SCBD - ShellCode

Scdbg.exe is a program that runs a shellcode and emulates 32bit processor, memory and Windows API calls. It displays the Windows API call performed by the shellcode.

The emulation of shellcode is an easy way to identify the action performed by the shellcode.

- **wine scdbg.exe badpdf_object6_shellcode_hex.bin**



Using scdbg, the actions performed by the shellcode can be analyzed. In the above example, the shellcode spawned "calc.exe" using "WinExec" and obtained the version of "calc.exe" before closing the process.

I haven't talked about the exploit itself, it's located at the end of the javascript code (like stated by Exploit search, "util.printf – found in stream: 6"). It's exploiting a buffer overflow on printf function to execute arbitrary code (here, our heap-sprayed shellcode)

https://www.appservgrid.com/paw92/index.php/2018/11/13/payload-in-pdf-ls-blog/

shellcode2exe

shellcode2exe converts a shellcode (shellcode.bin) into an executable (file.exe). It can be useful to create the executable binary to debug the shellcode or to run it and analyze it using dynamic malware analysis techniques.

- **shellcode2exe 32 file.bin file.exe**

https://tho-le.medium.com/techniques-and-tools-for-shellcode-analysis-9a49a1e15b2f

https://www.youtube.com/watch?v=9Ld53EkVJus

Disassembler

Shellcode can be analyzed with a dissembler to understand its logic without requiring its execution. IDA or Ghidra could be used to disassemble the malware.

I will demo this in a later version of this document. In the meantime, if you are interested, you can check the Practical Malware Analysis book.

Dynamic Malware Analysis

Dynamic Malware Analysis permits to analyze a malware during its execution. It is important to only execute malware in a safe and controlled environment. To set up your malware environment, I recommend reading the *Chapter 2 – Malware Analysis in Virtual Machines* in the Practical Malware Analysis book.

FLARE VM is a Windows based virtual machine for malware analysis and incident response, which was created by FireEye. It contains most of the tools to perform malware analysis. FLARE VM was used to analyze the below example.

It is important to mention that some malware are designed to detect test/analysis environments. Consequently, the malware will not execute. If you do not detect any malicious activity during a dynamic malware analysis, do not always consider that the sample is not malicious.

The below sample is used to illustrate the next steps:

- SHA256: d8cd7e914884d8cf9f715e69826e6a871dc73697c2005535abd578ff2bc5e4c3

Process

Before starting the analysis, take a snapshot of your virtual machine to save a clean state.

Moreover, before spawning or opening the malicious file:

- Take a snapshot of the registry with Regshot to save the state of the registry.
- Launch *Process Explorer*, *FakeDNS* and *Process Monitor* to monitor the host and network activities.

Process Monitor

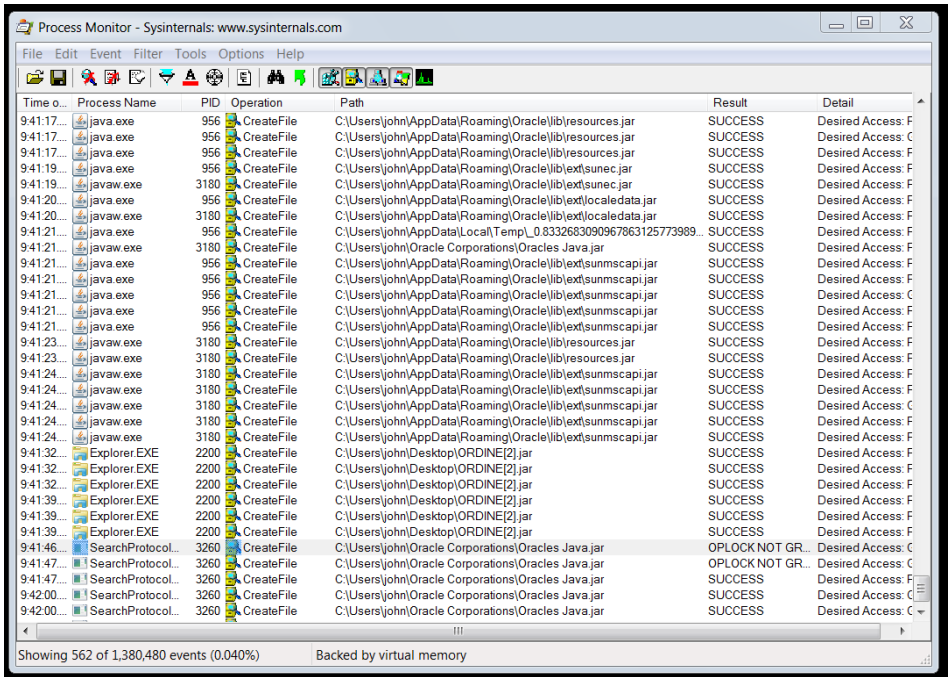
Process Monitor monitors and records file system, registry, processes, and other activities done on the system. It has a basic configuration by default (for example it excludes *procmon.exe*). It is possible to use additional filters to filter out and identify activities linked to the analyzed file (e.g. *Operation* for RegSetValue and WriteFile, etc.).

Before launching the malware, go to *Edit*, click on *Clear Display* and start monitoring events by clicking on *Capture Event* under *File.* After the malware execution, stop the capture of the event.

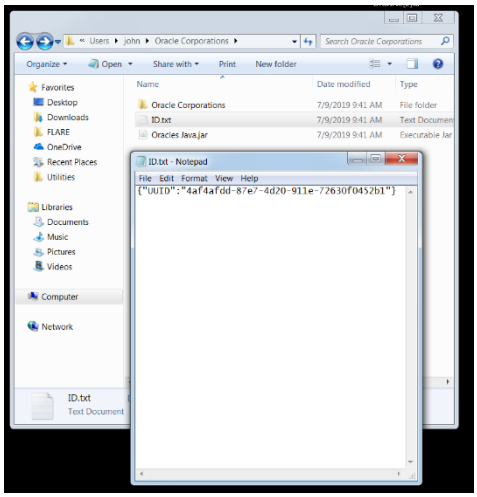
After running the malware, filters can be added:

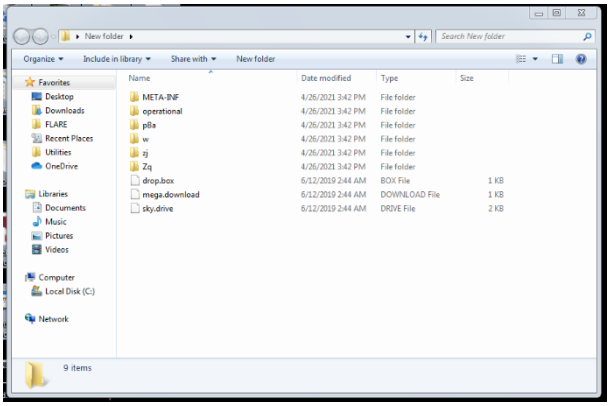
1. On the *Filter* on menu, click on *Filter*
2. Create a new filter under *Operation* based on the *CreateFile* for example
3. Click on *Add* and *OK*

Using the filter *CreateFile*, it can be observed that the sample malware creates lots of files and folders on the system. One of the interesting files is “Oracles Java.jar”.



“Oracles Java.jar” is in the “Oracle Corporations” folder that contains the jar file and an ID file (“ID.txt”). The “Oracles Java.jar” contains three files that are typical of Adwind RAT (cf. below), which is a cross-platform, multifunctional malware program also known as AlienSpy, Frutas, Unrecom, Sockrat, JSocket and jRat.[3] Adwind malware can” be analyzed using “AdWindDecryptor.py ”. The text file contains the unique ID of the infected system. This is used by the C2 server to identify the infected system (cf. on the right screenshot).





[3] https://www.kaspersky.com/resource-center/threats/adwind

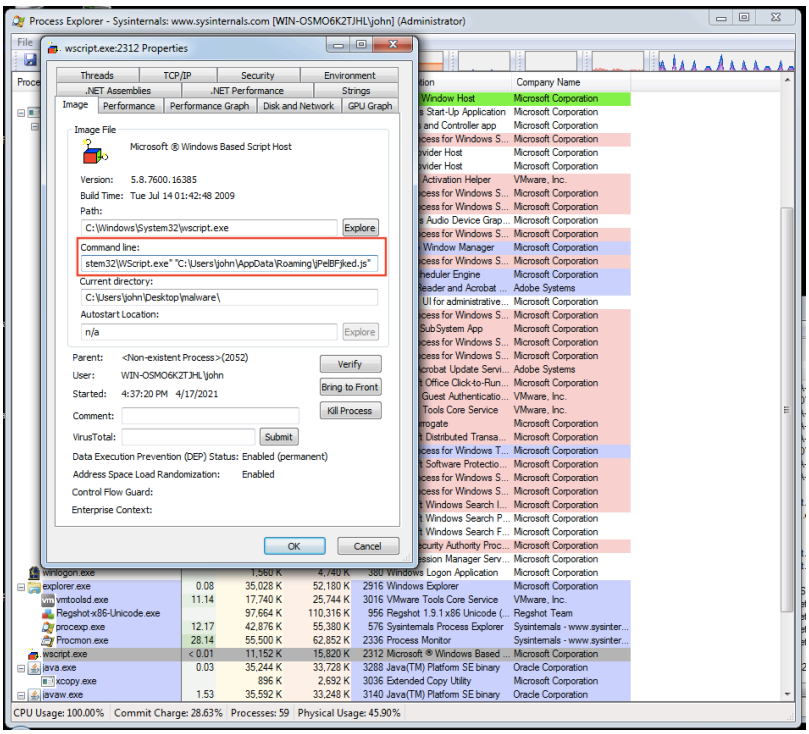
Process Explorer

Process Explorer (*procexp.exe*) is a Sysinternal tool that monitors and displays all the processes running on a Windows system, including the hierarchical parent relationship, the loaded DLL files, the processes spawned, suspended and killed, and others activities.

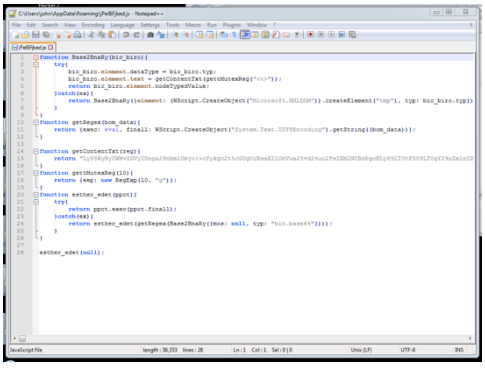
When performing a dynamic malware analysis, it is worth monitoring the activities with *Process Explorer* as it records all processes spawned, stopped, or killed.

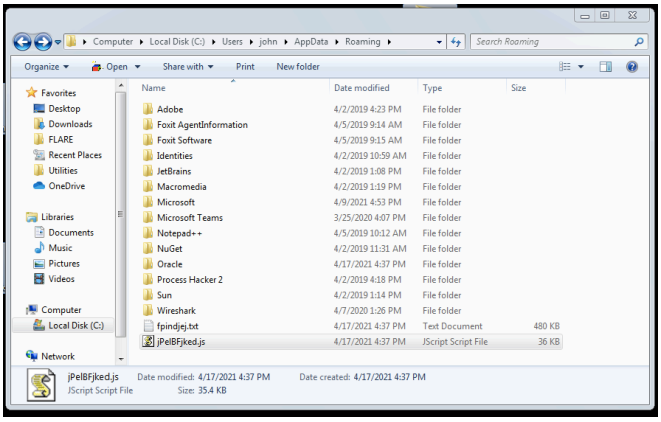
After launching our sample, we can easily identify that the malware starts lots of executables: *javaw.exe*, *wscript.exe*, *killtask.exe* and other processes. By clicking on a process (e.g. *wscript.exe* in the below screenshot) and selecting the *Properties*, you can see under *Image* the *Command Line* run.

In the below example, the malicious file spawned wscript.exe, which launched "jPelBFjked.js"

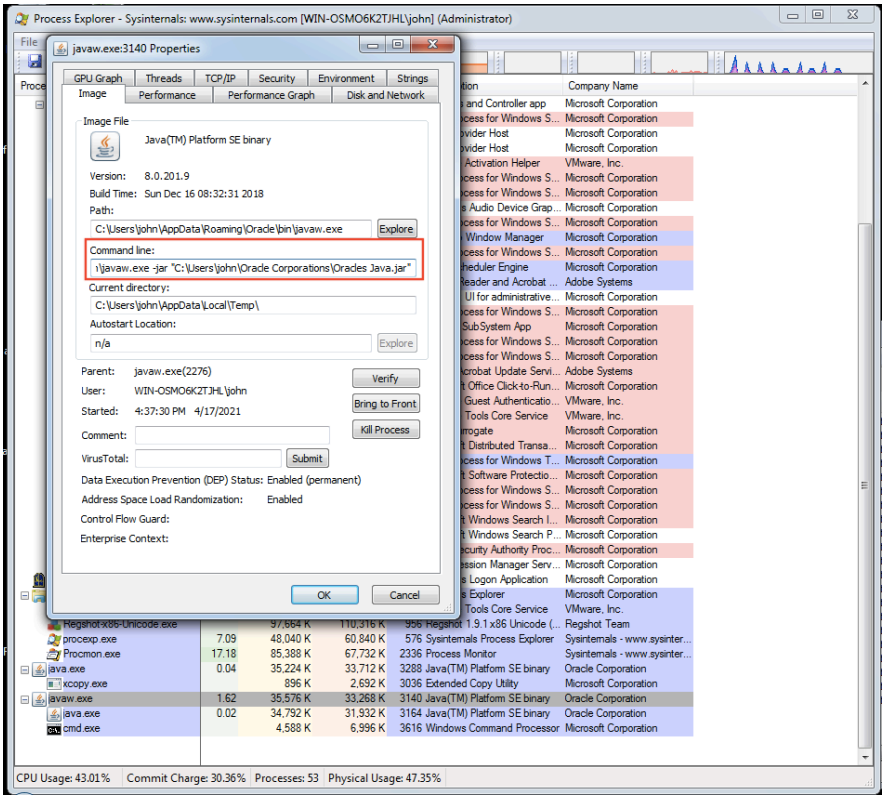


The JavaScript file can then be checked. The JavaScript strings in the JavaScript is highly encrypted (AdWind malware).

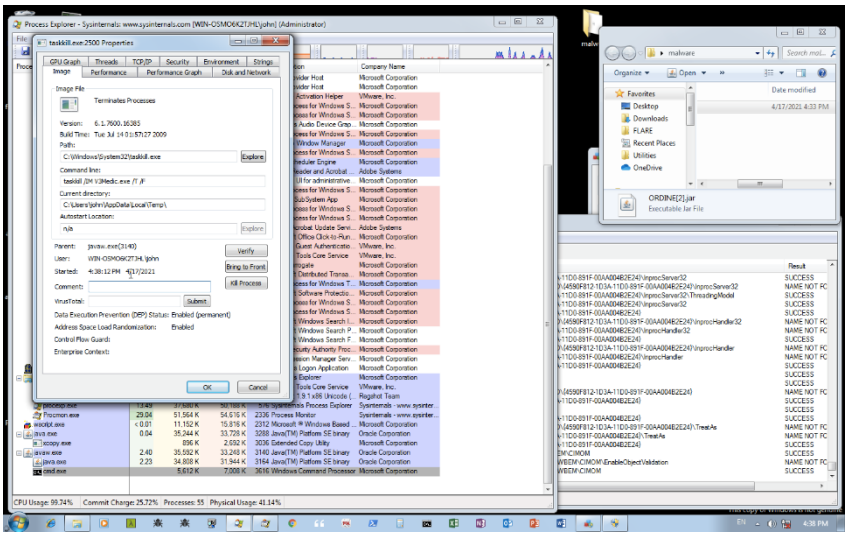




It also spawns *javaw.exe* which is going to run the *Oracle Java.jar* file previously identified with Process Monitor.



Using *taskill.exe*, this malware identifies and kills lots of analysis programs like V3Medic.exe (cf. below screenshot) and Process Explorer, so it is normal if you see Process Explorer exiting without any action from your side.



At the end of the analysis, kill all the processes created by the malicious file, by right clicking on the process and clicking on *Kill Process Tree*.

FakeNet-NG

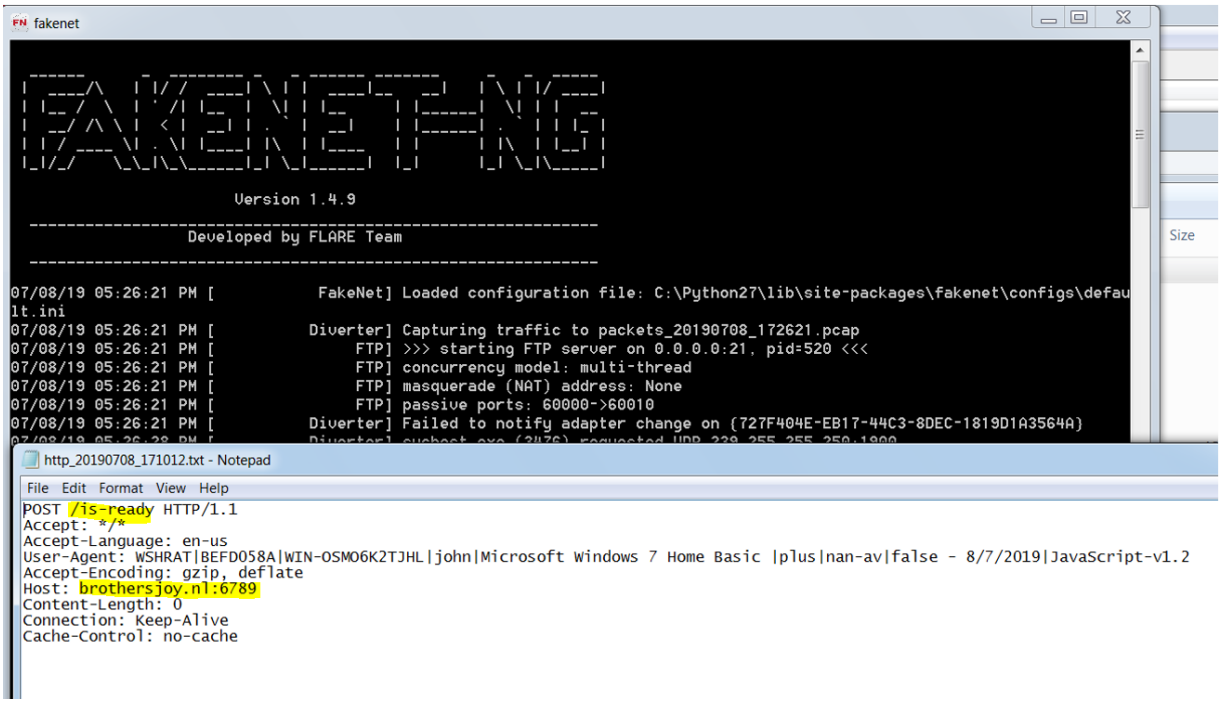
FakeNet-NG simulates legitimate network services (e.g. HTTP, HTTPS, FTP, ICR, DNS, SMPT, etc.) and intercepts and redirects all the network traffic.

FakeNet-NG can be extremely useful to identify network IOCs (e.g. downloading of the second stage, C2 server, etc.).

To launch FakeNet, simply start the executable. At the end of the analysis, the program can be ended with “ctrl + C” and the logs are available in the *fakenet_logs* folder.

In the below image, FAKENET provides the information concerning the domain contacted by the malware:

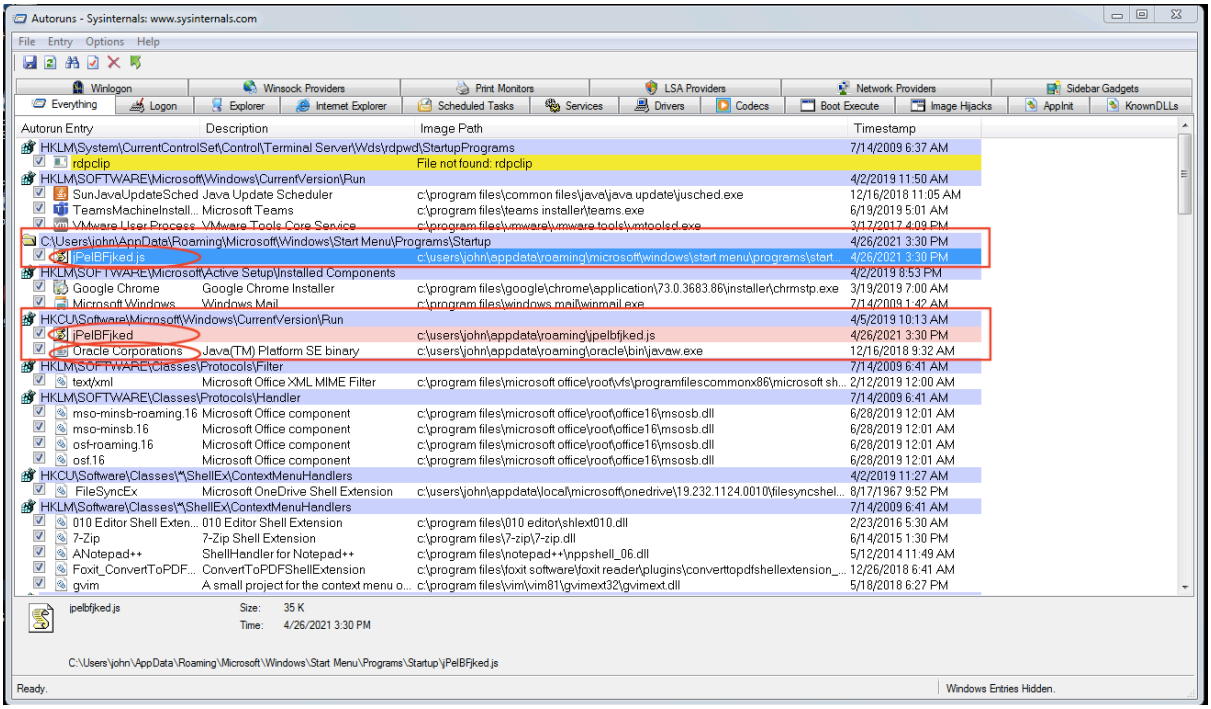
- brothersjoy[.]nl:6789/is-ready



Autoruns

Autoruns enumerates all the programs that automatically start when a Windows system is switched on. It can be useful to launch autoruns after the execution of the malware to identify persistence mechanisms added by the malware.

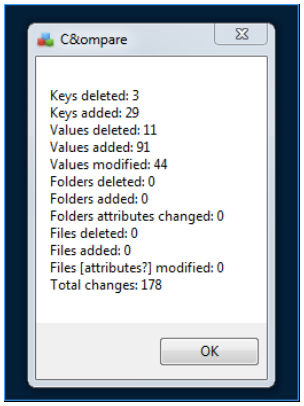
The timestamp column is one of the easiest ways to identify the new autoruns locations (cf. below screenshot). It can be identified that the previously identified JavaScript and jar file is added as a persistence mechanism.

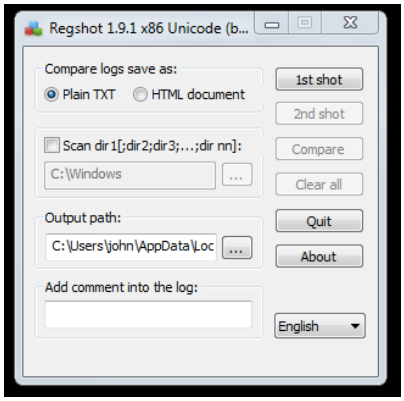


Regshot

Regshot compares the changes in the Registry at two different point in time. Take a screenshot of the Registry before (*1st shot*) and after (*2nd shot*) the execution of the malware.

Then, review the differences between the two screenshots to identify registry changes made by the malware.

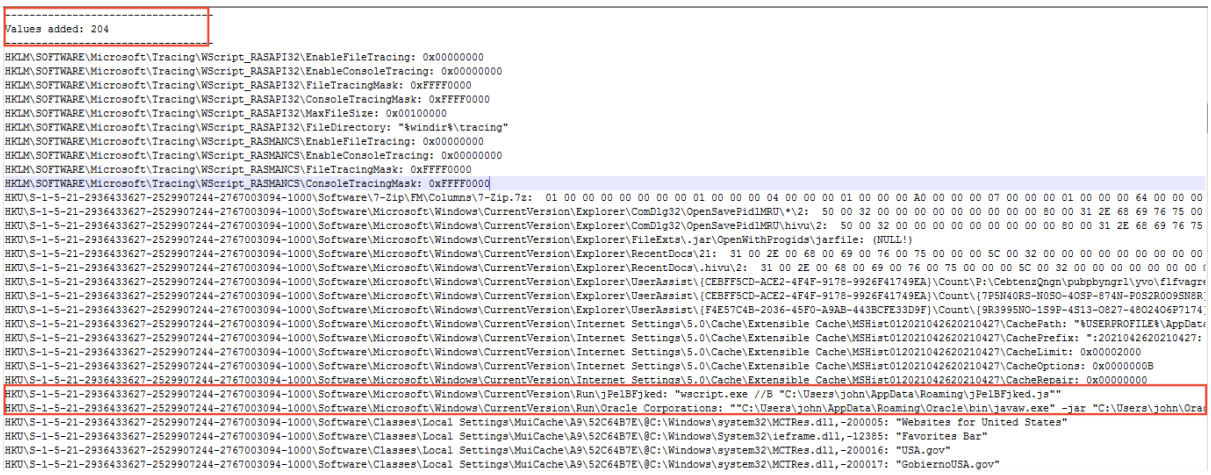




As it can be seen in the below screenshot, using Regshot, the malicious JavaScript and Jar file can also be identified as a persistence mechanism. The two below persistence keys were added:

- HKU\S-1-5-21-2936433627-2529907244-2767003094-1000\Software\Microsoft\Windows\CurrentVersion\Run\jPelBFjked: "wscript.exe //B "C:\Users\john\AppData\Roaming**jPelBFjked.js**""

- HKU\S-1-5-21-2936433627-2529907244-2767003094-1000\Software\Microsoft\Windows\CurrentVersion\Run\Oracle Corporations: ""C:\Users\john\AppData\Roaming\Oracle\bin\javaw.exe" -jar "C:\Users\john\Oracle Corporations**Oracles Java.jar**""



Appendix

Regex

I developed the below regex in order to identify the below strings and encoded strings.

Strings	Base64	Hex	Rot13
http://	aHR0cDovLw==	**68 74 74 703a 2f 2f**	uggc://
https://	aHR0cHM6Ly8=	68 74 74 70 73 **3a 2f 2f**	uggcf://
c:\	Yzpc	**63 3a 5c**	p:\
c:\windows	Yzpcd2luZG93cw==	63 3a 5c 77 69 6e 64 6f 77 73	p:\jvaqbjf
C:\	Qzo=	**43 3a 5c**	P:\
C:\Windows	Qzpcd2luZG93cw==	43 3a 5c 77 69 6e 64 6f 77 73	P:\Jvaqbjf
.exe	LmV4ZQ==	2e 65 78 65	.rkr
.ps1	LnBzMQ==	2e 70 73 31	.cf1
HKEY_CURRENT_USER	SEtFWV9DVVJSRU5UX1VTRVI=	**48 4b 45 59 5f** 43 55 52 52 45 4e 54 5f 55 53 45 52	**UXRL**_PHEERAG_HFRE
HKCU	SEtDVQ==	**48 4b 43 55**	**UXPH**
HKEY_LOCAL_MACHINE	SEtFWV9MT0NBTF9NQUNISU5F	**48 4b 45 59 5f** 4c 4f 43 41 4c 5f 4d 41 43 48 49 4e 45	**UXRL**_YBPNY_ZNPUVAR
HKLM	SEtNTA==	**48 4b 4c 4d**	**UXYZ**

"(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)|(([0-9]{2}[%|\ ]){1,3})2e[%|\ ](([0-9]{2}[%|\ ]){1,3})2e[%|\ ](([0-9]{2}[%|\ ]){1,3})2e[%|\ ](([0-9]{2}[%|\ ]){1,3})|http://|https://|c:\|c:\windows|C:\|C:\Windows|.exe|.ps1|HKEY_CURRENT_USER|HKCU|HKEY_LOCAL_MACHINE|HKLM|aHR0cDovLw==|aHR0cHM6Ly8=|Yzpc|Yzpcd2luZG93cw==|Qzo=|Qzpcd2luZG93cw==|LmV4ZQ==|LnBzMQ==|SEtFWV9DVVJSRU5UX1VTRVI=|SEtDVQ==|SEtFWV9MT0NBTF9NQUNISU5F|SEtNTA==|68[%|\ ]?74[%|\ ]?74[%|\ ]?70[%|\ ]?3a[%|\ ]?2f[%|\ ]?2f|68[%|\ ]?74[%|\ ]?74[%|\]?70[%|\ ]?73[%|\ ]?3a[%|\ ]?2f[%|\ ]?2f|63[%|\ ]?3a[%|\ ]?5c|63[%|\ ]?3a[%|\ ]?5c[%|\ ]?77[%|\ ]?69[%|\ ]?6e[%|\ ]?64[%|\ ]?6f[%|\ ]?77[%|\ ]?73|43[%|\ ]?3a[%|\ ]?5c|43[%|\ ]?3a[%|\ ]?5c[%|\ ]?77[%|\ ]?69[%|\ ]?6e[%|\ ]?64[%|\ ]?6f[%|\ ]?77[%|\ ]?73|2e[%|\ ]?65[%|\ ]?78[%|\ ]?65|2e[%|\ ]?70[%|\ ]?73[%|\ ]?31|48[%|\ ]?4b[%|\ ]?45[%|\ ]?59[%|\ ]?5f[%|\ ]?43[%|\ ]?55[%|\ ]?52[%|\ ]?52[%|\ ]?45[%|\ ]?4e[%|\ ]?54[%|\ ]?5f[%|\ ]?55[%|\ ]?53[%|\ ]?45[%|\ ]?52|48[%|\ ]?4b[%|\ ]?43[%|\ ]?55|48[%|\ ]?4b[%|\ ]?45[%|\ ]?59[%|\ ]?5f[%|\ ]?4c[%|\ ]?4f[%|\ ]?43[%|\ ]?41[%|\ ]?4c[%|\ ]?5f[%|\ ]?4d[%|\ ]?41[%|\ ]?43[%|\ ]?48[%|\ ]?49[%|\ ]?4e[%|\ ]?45|48[%|\ ]?4b[%|\ ]?4c[%|\ ]?4d|uggc://|uggcf://|p:\|p:\jvaqbjf|P:\|P:\Jvaqbjf|.rkr|.cf1|UXRL_PHEERAG_HFRE|UXPH|UXRL_YBPNY_ZNPUVAR|UXYZ"