

# Type Inference

---

This README file describe how to reproduce our tool to infer type information.

There are four steps in total:

step 1 is collecting the differential results;

step 2 is to parse the differential results to get the type indicator of basic types(bool, string and number) and array type;

step 3 is to get the key-get indicator of object type;

step 4 is to generate the type information of every binding calls.

The following will talk about reproduction in detail.

## Note:

1. for easier reproduction, the provided type inference's code is constructed based on Adobe Acrobat Reader version 2020.013.20074, please check the version before reproduction.

2. in our experience, the type information between different versions of Adobe Reader are usually the same, so we only need to infer type information through one version and use this inferred type information to fuzz different versions of Adobe Reader(so does Foxit Reader)



## Adobe Acrobat Reader DC

Continuous Release | Version 2020.013.20074

Copyright © 1984-2020 Adobe. All rights reserved.

Adobe, the Adobe logo, the Adobe PDF logo, and Acrobat are either registered trademarks or trademarks of Adobe in the United States and/or other countries. All other trademarks are the property of their respective owners.

Portions Copyright IntegrityWare, Inc

Portions copyright Right Hemisphere, Inc.

Portions utilize Microsoft Windows Media Technologies. Copyright (c) 1999-2002, 2006 Microsoft Corporation. All Rights Reserved.

Portions are the result of a cooperative development process by Adobe and Microsoft Corporation.

Copyright 2003-2020 Solid Documents Limited.

Third Party notices, terms and conditions pertaining to third party software can be found at: <http://www.adobe.com/go/thirdparty>.



## 1. diff\_collect

This step will collect the differential results. We use pin.exe to instrument the target PDF Reader to collect execution trace.

### folder structure

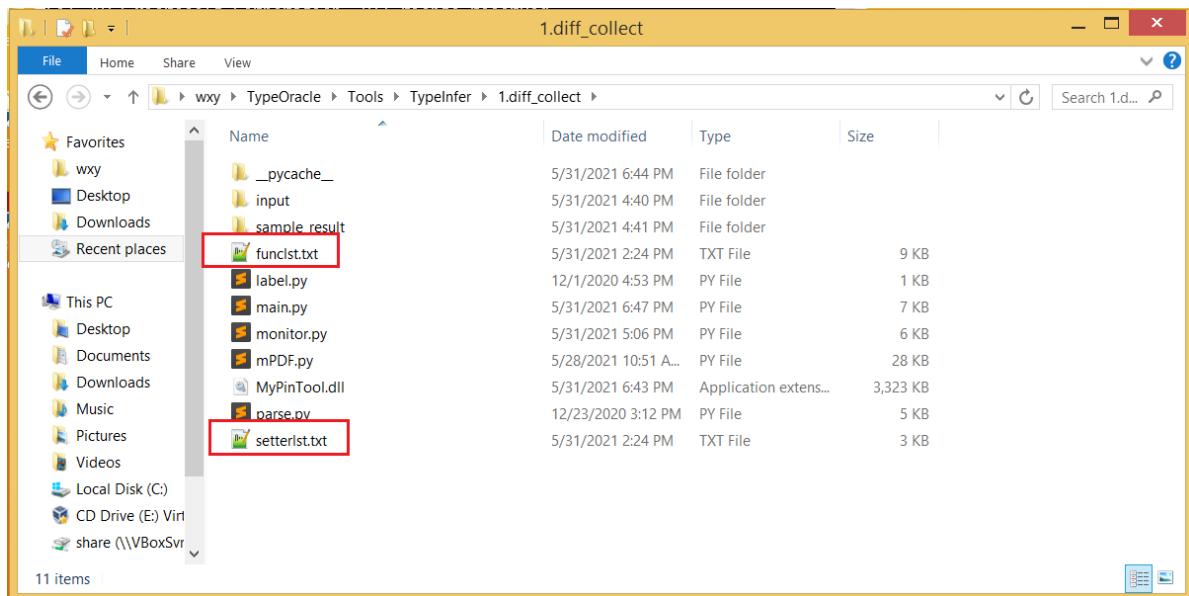
- sample\_input
  - func1st.txt (list of all method names)
  - setter1st.txt (list of all accessors)
- sample\_result (the folder store the sample result of this step)
  - save
- main.py
- monitor.py (to minitor the PDF Reader)
- mPDF.py (to embed javascript code to PDF)
- MyPinTool.dll (the instrumentation file)
- parse.py (parse the results recorded by pin.exe)

## how to reproduce

1. make sure the Page Heap is turned off (execute following command and click yes, for more information about Page Heap, please refer to C:\Users\wxy\TypeOracle\Other\README.pdf)

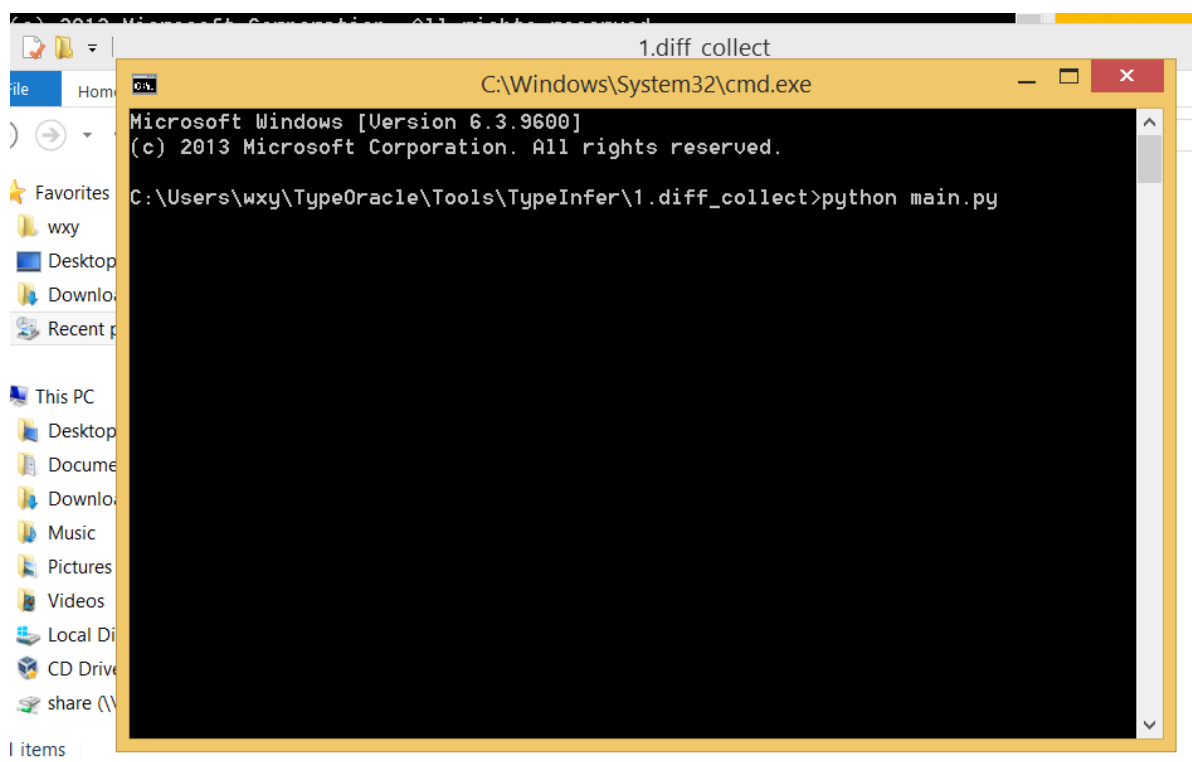
```
"C:\Program Files (x86)\Windows Kits\8.1\Debuggers\x86\gflags.exe" /p /disable  
"C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
```

2. copy `func1st.txt` and `setter1st.txt` from `sample_input` folder to current folder

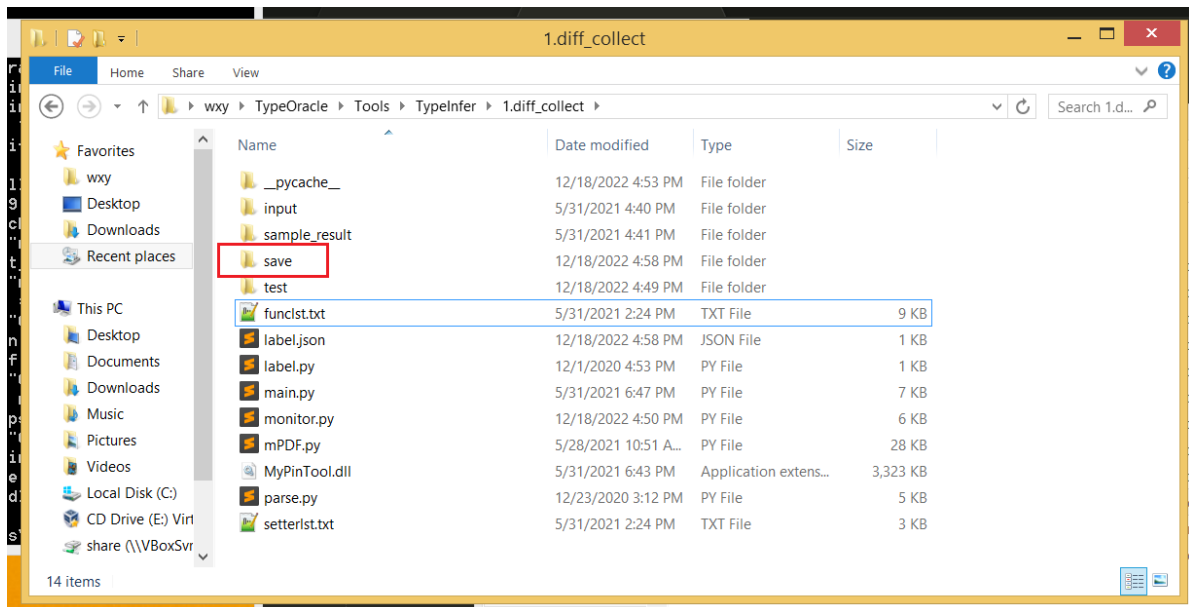


3. execute `main.py` at the current folder to start collect the differential results of every binding call (this step will take more than ten hours)

```
python main.py
```



4. the result will be stored at `save` folder



## 2. parse\_diffresult

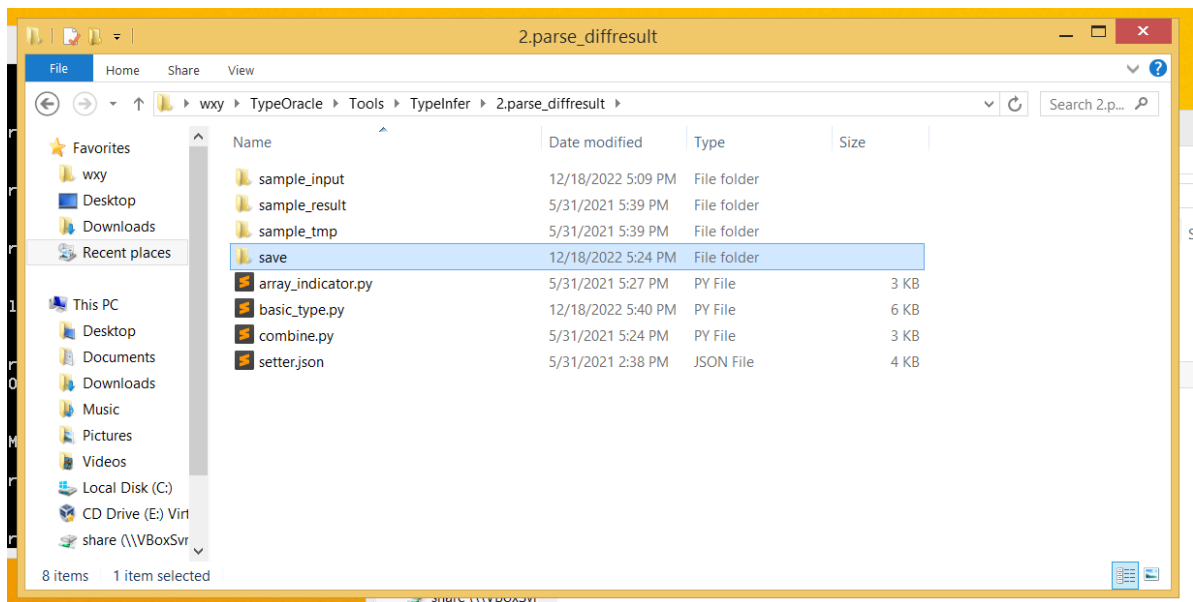
This step is to parse the differential results collected in step 1, and get the type indicator of bool, string, number and array type.

### folder structure

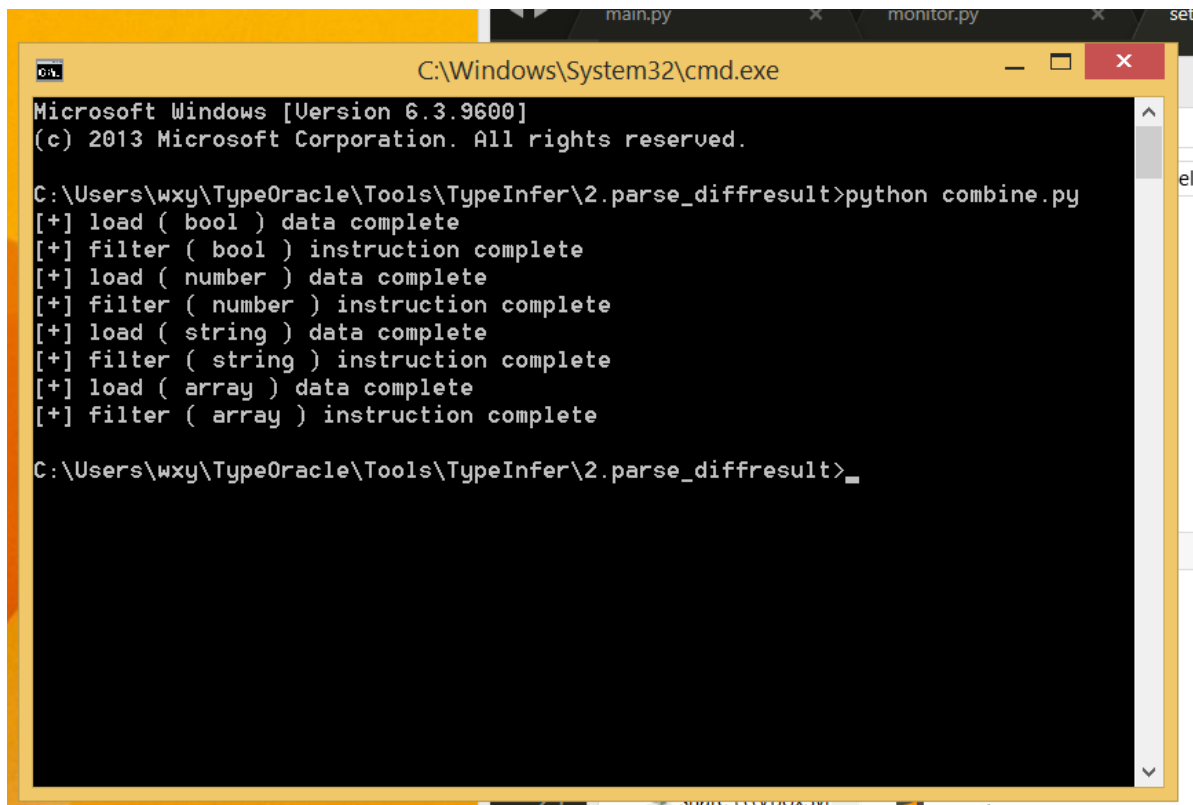
- sample\_input
  - save (differential results collected in step 1)
  - setter.json (list of all accessors and its type)
- sample\_result (the folder store the sample result of this step)
  - basic.txt
  - array.txt
- sample\_tmp (store sample intermediate results)
- array\_indicator.py (get the array indicator)
- basic\_type.py (get basic types' indicator)
- combine.py (combine step 1's result)

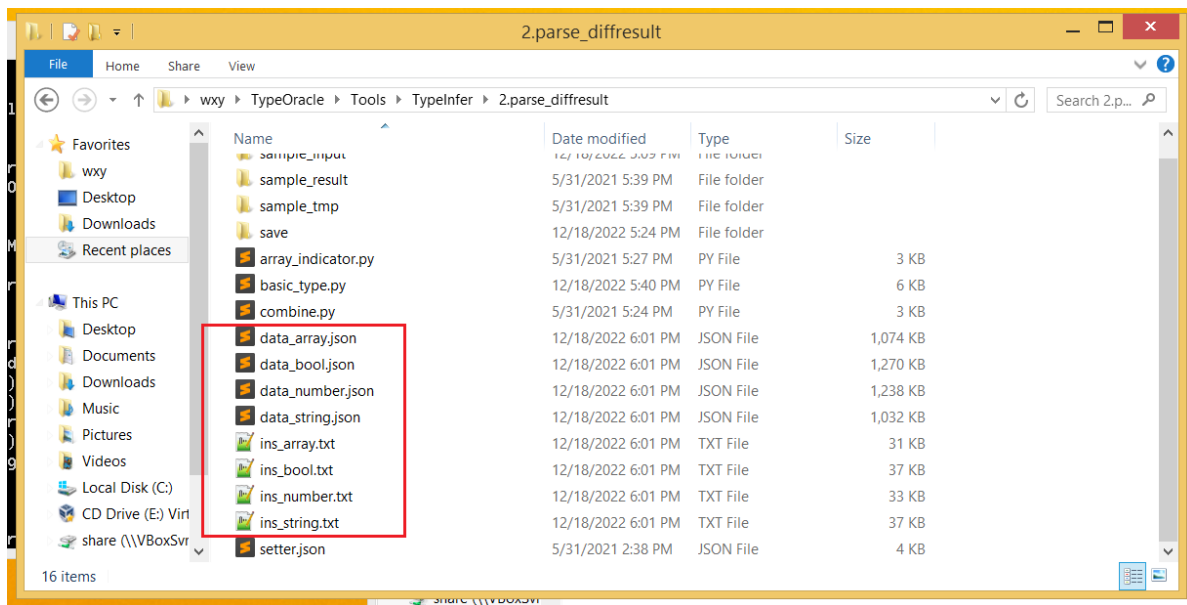
### how to reproduce

1. copy `save` folder in step 1 to this folder (you can also copy `save` from `sample_input` folder), and copy `setter.json` in `sample_input` folder to this folder



2. execute combine.py, this will generate some intermediate result files(data\_x.json ins\_x.json)





3. execute basic.py to get basic types' type indicator, redirect the results to basic.txt

```
python basic_type.py > basic.txt
```

The instructions that show up the most times is the type indicator, from the following snapshot, we can infer following type indicators.

```
type indicator:(h means this instruction belongs to Esript.api module, for other
modules, please refer tp C:\Users\wxy\TypeOracle\Other\label.json)
bool: 0x5be62, 0x5be6b,0x5be9d,0x5bf05,0x5bf0f
number: 0x53d65,0x53d68
string: 0x3ce05,0x3ce34,0x3ce39,0x3ce46
```

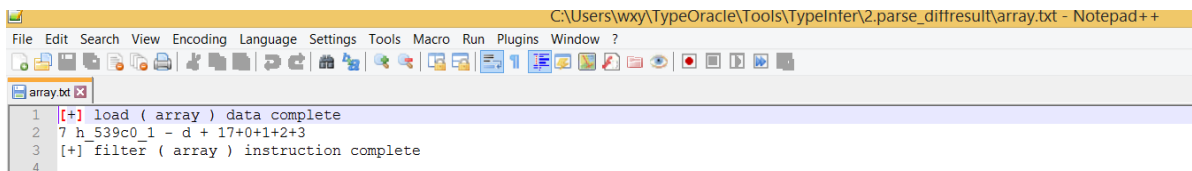
```
File Edit Search View Encoding Language Settings Tools Ma
basic.txt
1 bool:
2 200
3
4 62
5 h_5be62_0:- 1 + 0
6 h_5be6b_1:- 1 + 0
7 h_5be73_0:- 1 + 0
8 h_5be73_1:- 1 + 0
9 h_5be9d_0:- 1 + 0
10 h_5bf05_1:- 1 + 0
11 h_5bf0f_1:- 1 + 0
12 10
13 h_ffff9_1:- 1 + 0
14 h_ffffc_1:- 1 + 0
15 h_b00d1_0:- 1 + 0
16
17
18 number:
19 45
20 h_53d65_1:- 17 + b3
21 h_53d68_1:- 17 + b3
22
23
24 string:
25 194
26 h_3ce05_1:- d4 + 174
27 h_3ce34_1:- d4 + 174
28 h_3ce39_1:- d + 17
29 h_3ce46_1:- d4 + 174
30 16
31 h_3da69_0:- 1a + 2e
32 h_3dac8_1:- d + 17
33 h_3dacd_0:- d + 17
34 h_3dad5_0:- 10 + 1a
35 h_3dae0_0:- d + 17
36 14
37 h_2c3e_0:- d + 17
```

4. execute array\_indicator.py, redirect the results to array.txt

```
python array_indicator.py > array.txt
```

we can get the following array type indicator

```
type indicator:
array: 0x539c0
```



### 3. generic\_object

An object contains key-value pairs, and one can access the values by referring to the keys. The keys can not be obtained by blind guessing. Observe that the binding calls generally get the desired key of an object from the data segment of the program. Hence the strings read from the data segment during the binding call execution are considered as possible keys.

The script engines of Adobe Reader and Foxit Reader both have two important APIs for handling keys in objects, similar to `hasKeyInObject()` and `getValueByKeyInObject()`. The former is to check whether the key exists, and the latter is to get the value from the object according to the key. This is our main idea to get key-get indicator.

### folder structure

- stage1
  - sample\_result (the folder store the sample result of this step)
  - sample\_tmp (the folder store sample intermediate results)
  - checkstring.py (script file executed in IDA pro)
  - cmd.txt (some commands to start Adobe Reader with pin.exe)
  - get\_string.py (get the string in execution strace)
  - label.py
  - makepdf.py (generate PDFs with special javascript codes)
  - mPDF.py (to embed javascript code to PDF)
  - MyPinTool.dll (the instrumentation file)
- stage2
  - sample\_result (the folder store the sample result of this step)
  - sample\_tmp (the folder store sample intermediate results)
  - cmd.txt (some commands to start Adobe Reader with pin.exe)
  - idapush.py (script file executed in IDA pro)
  - label.py
  - makepdf.py (generate PDFs with special javascript codes)
  - mPDF.py (to embed javascript code to PDF)
  - MyPinTool.dll (the instrumentation file)



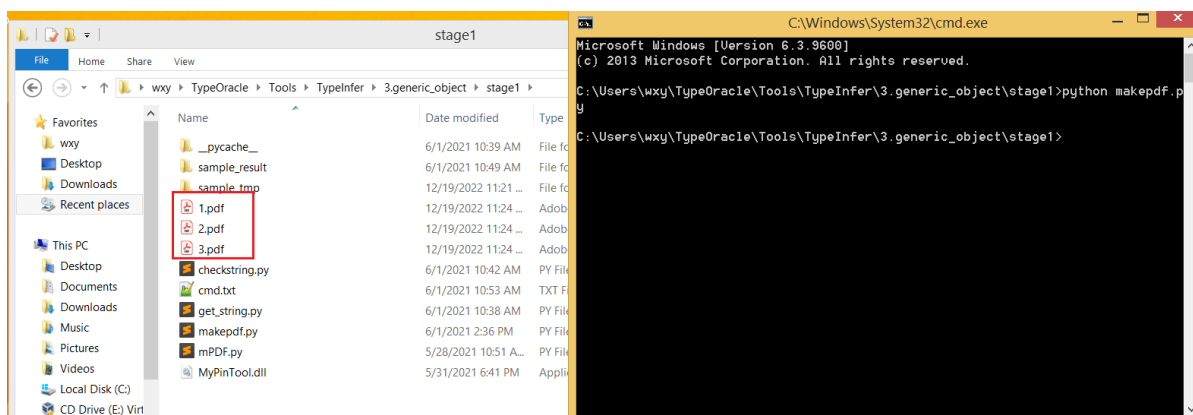
- parse.py
- parse2.py
- record.py (script file executed in windbg)
- EScript.api.idb (the result database of IDA pro to parse EScript.api)

## how to reproduce

### stage 1

1. execute `makepdf.py` to generate three test PDFs

```
python makepdf.py
```



2. execute the commands in `cmd.txt`, remember to close Adobe Reader when the execution of every command is finished(execution of each command will take about 5 minutes), this will generate three files: `1.out`, `2.out` and `3.out`

commands in `cmd.txt`:

```
C:\Users\wxy\Desktop\pin-3.17\pin.exe -t MyPinTool.dll -o 1.out -- "C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
```

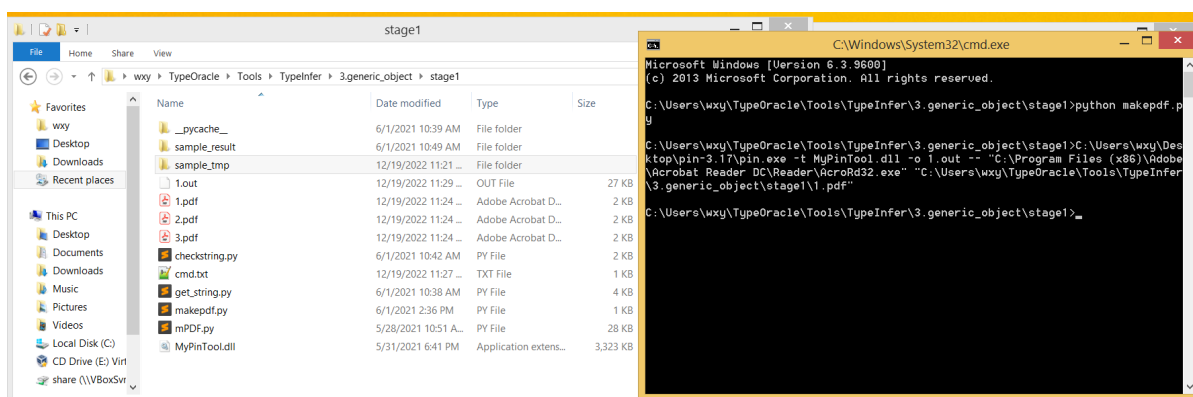
```
"C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage1\1.pdf"
```

```
C:\Users\wxy\Desktop\pin-3.17\pin.exe -t MyPinTool.dll -o 2.out -- "C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
```

```
"C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage1\2.pdf"
```

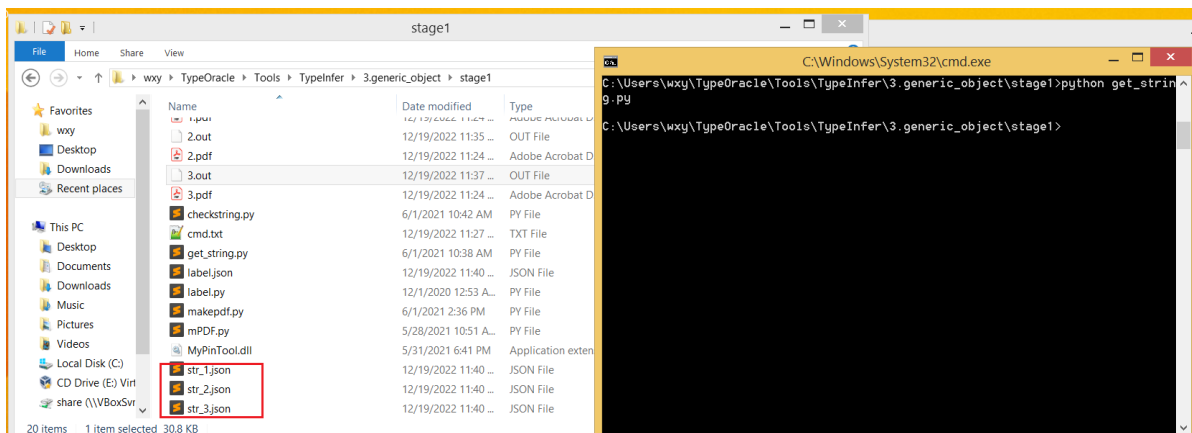
```
C:\Users\wxy\Desktop\pin-3.17\pin.exe -t MyPinTool.dll -o 3.out -- "C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
```

```
"C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage1\3.pdf"
```



3. execute `get_string.py` to get `str_x.json`

```
python get_string.py
```

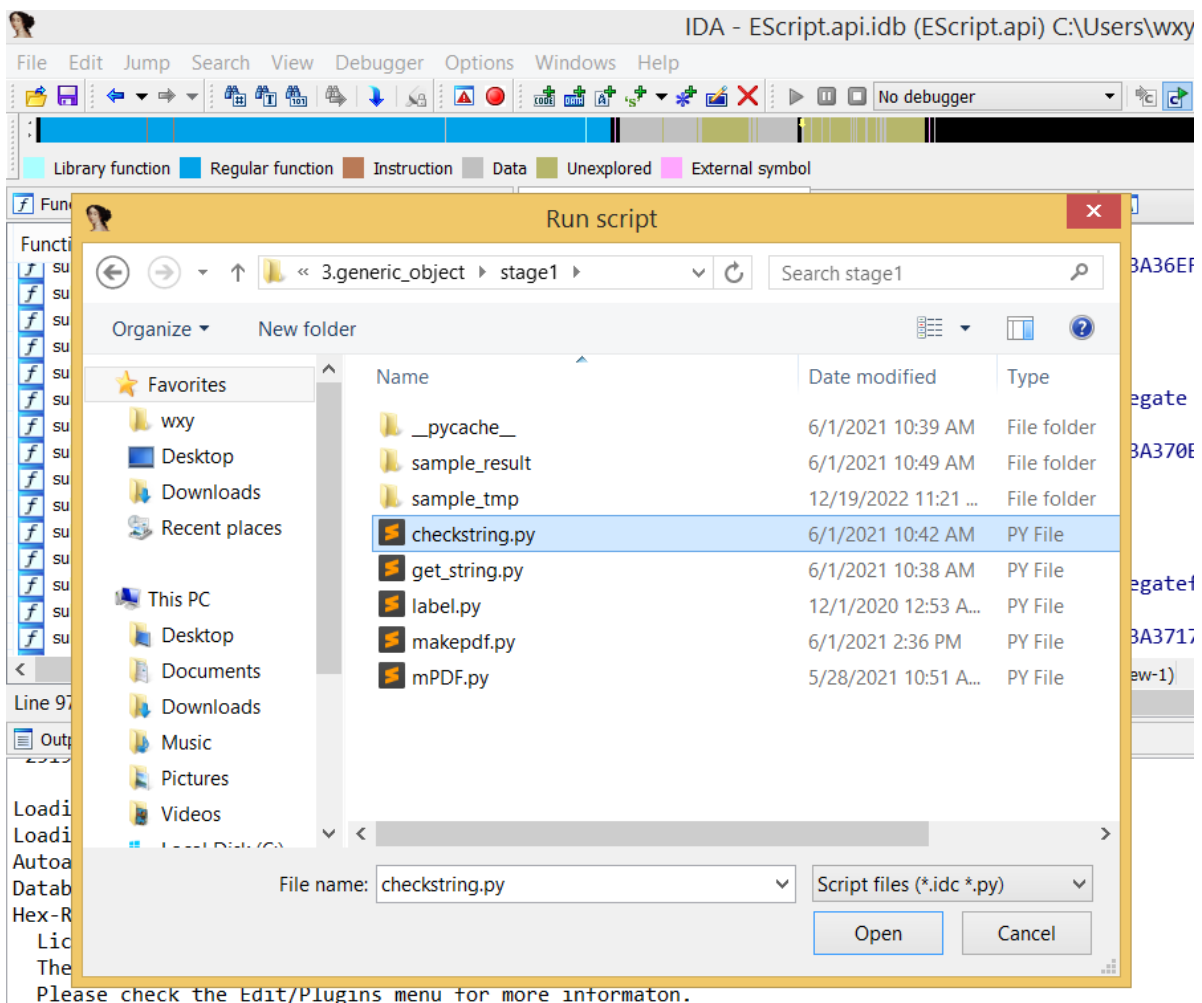


4. open `EScript.api.idb` with IDA pro, and execute scripts `checkstring.py`

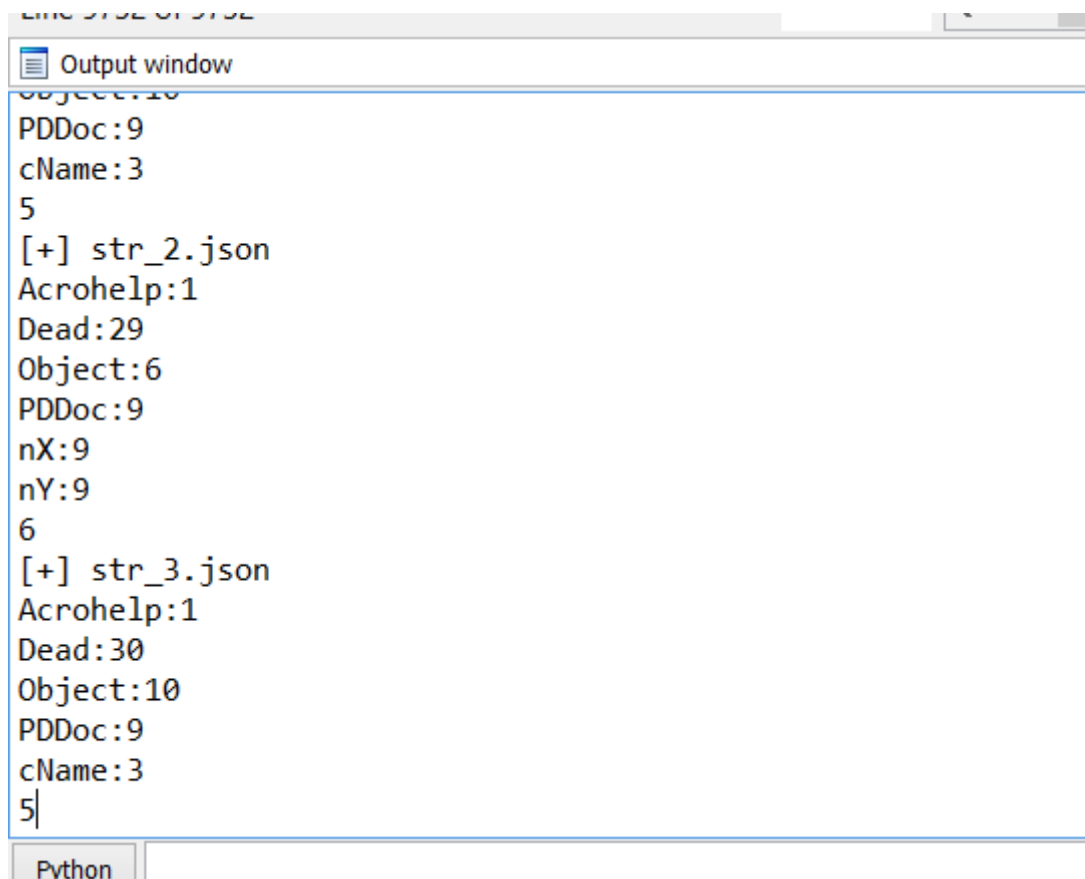
This is IDA pro



Click file->Script file (file button is in the upper left corner)



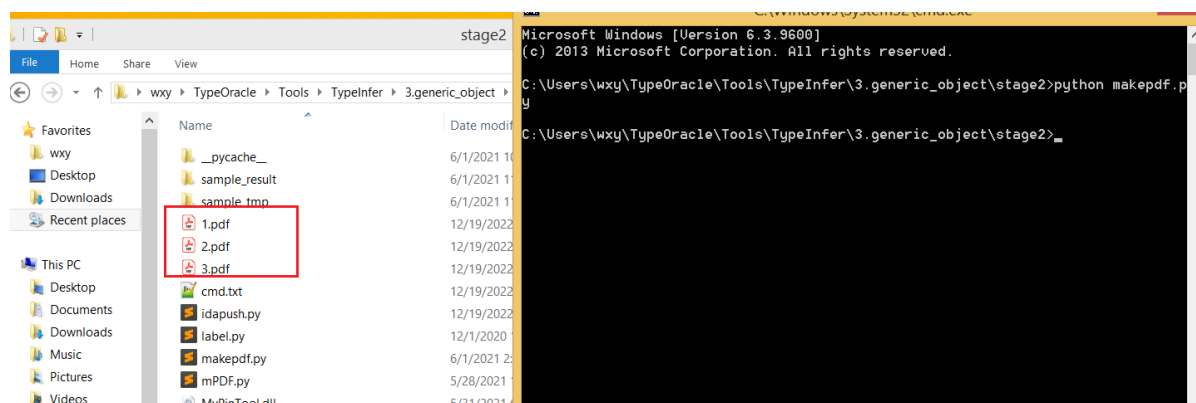
The result will output through output windows



## stage2

1. execute `makepdf.py` to generate three test PDFs (this step use the key inferred in stage1)

```
python makepdf.py
```

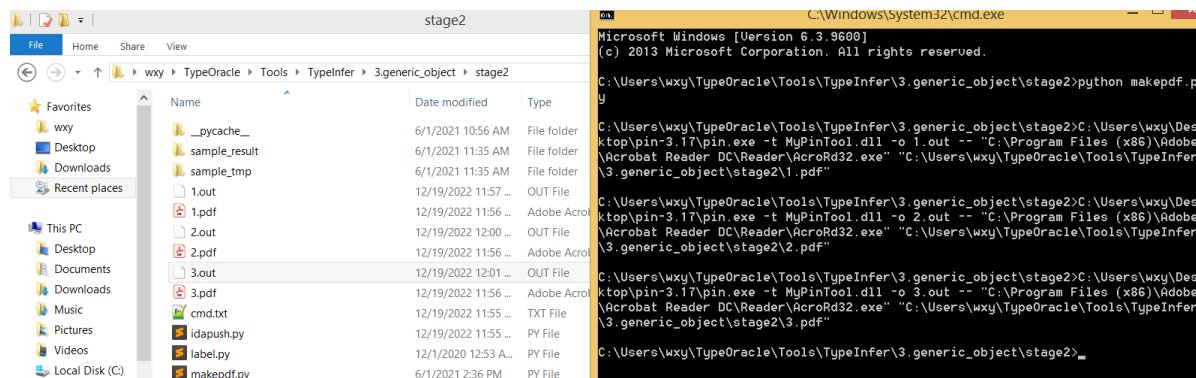


2. execute the following three commands in `cmd.txt`, remember to close Adobe Reader when the execution of every command is finished(execution of each command will take about 5 minutes), this will generate three files: `1.out`, `2.out` and `3.out`

```

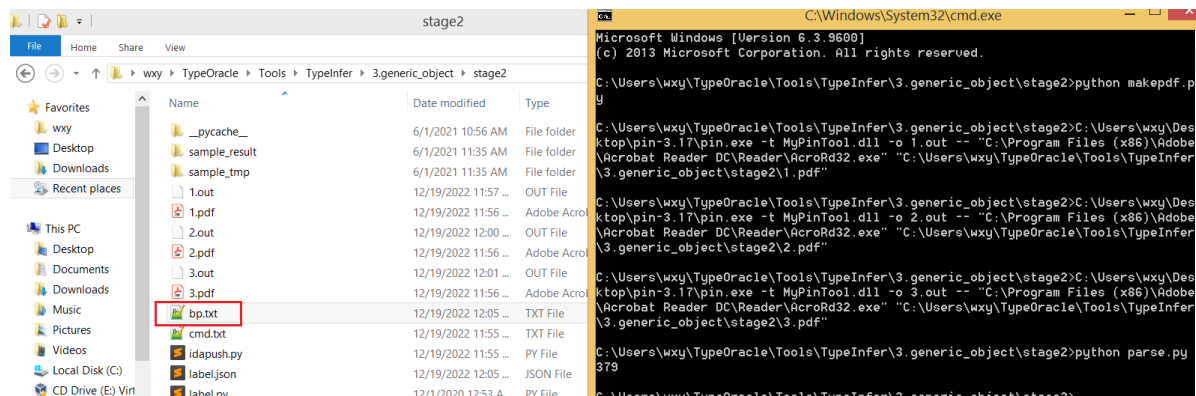
C:\Users\wxy\Desktop\pin-3.17\pin.exe -t MyPinTool.dll -o 1.out -- "C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
"C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage2\1.pdf"
C:\Users\wxy\Desktop\pin-3.17\pin.exe -t MyPinTool.dll -o 2.out -- "C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
"C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage2\2.pdf"
C:\Users\wxy\Desktop\pin-3.17\pin.exe -t MyPinTool.dll -o 3.out -- "C:\Program Files (x86)\Adobe\Acrobat Reader DC\Reader\AcroRd32.exe"
"C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage2\3.pdf"

```

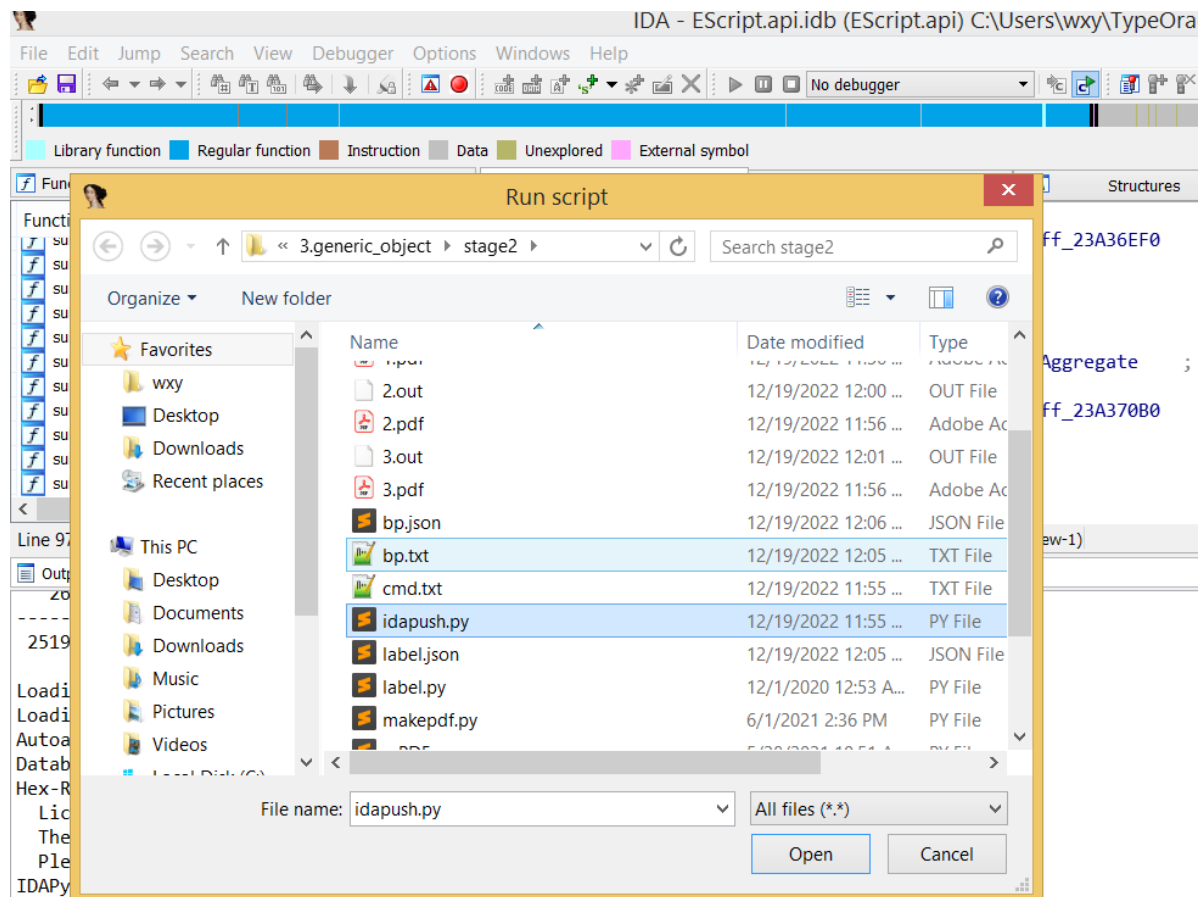


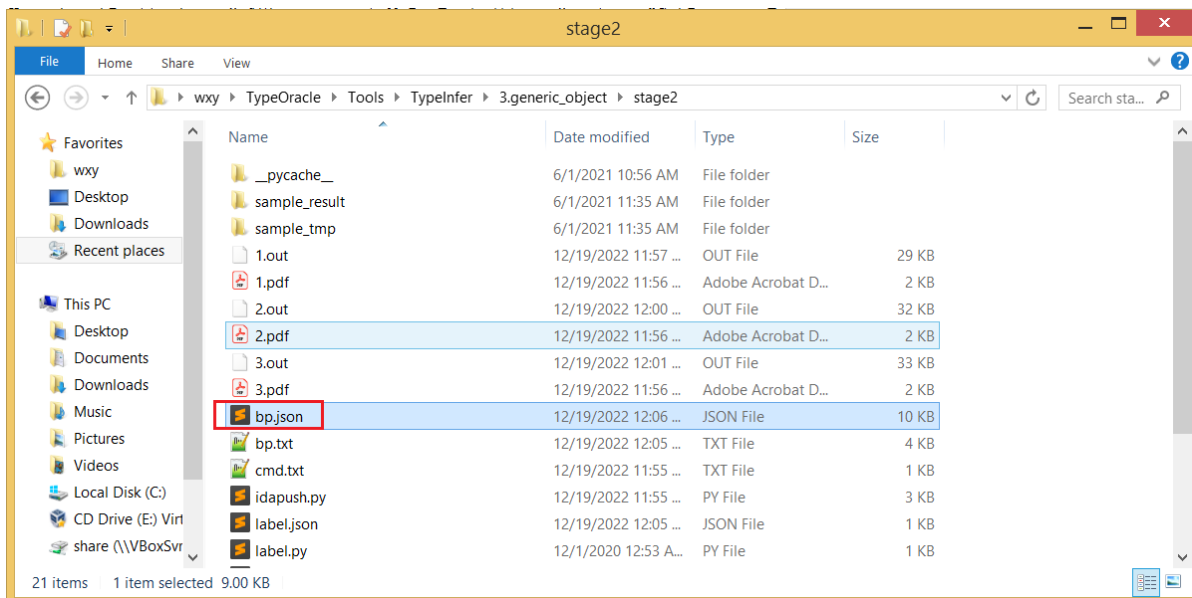
3. execute `parse.py` to get `bp.txt`

`python parse.py`



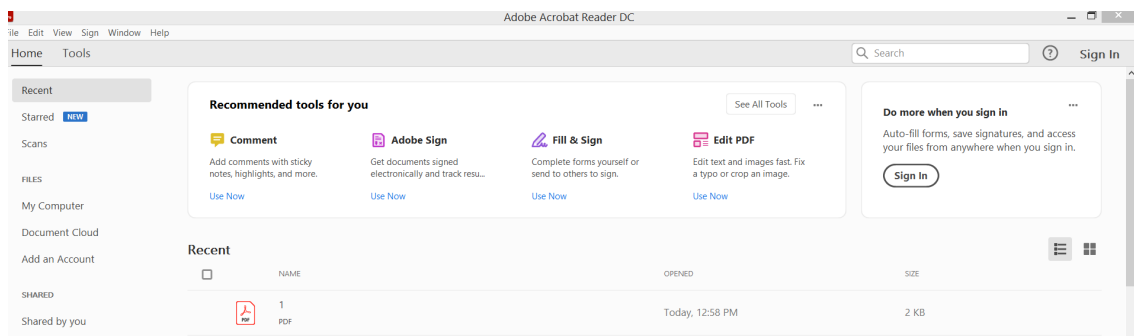
4. open `Escript.api.idb` with IDA pro, and execute scripts `idapush.py`



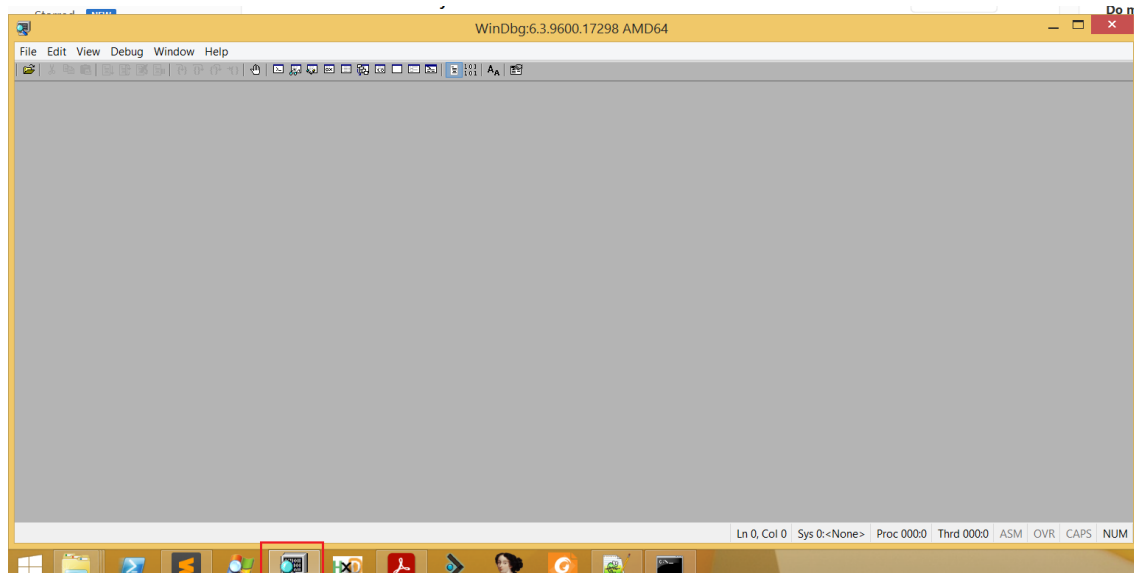


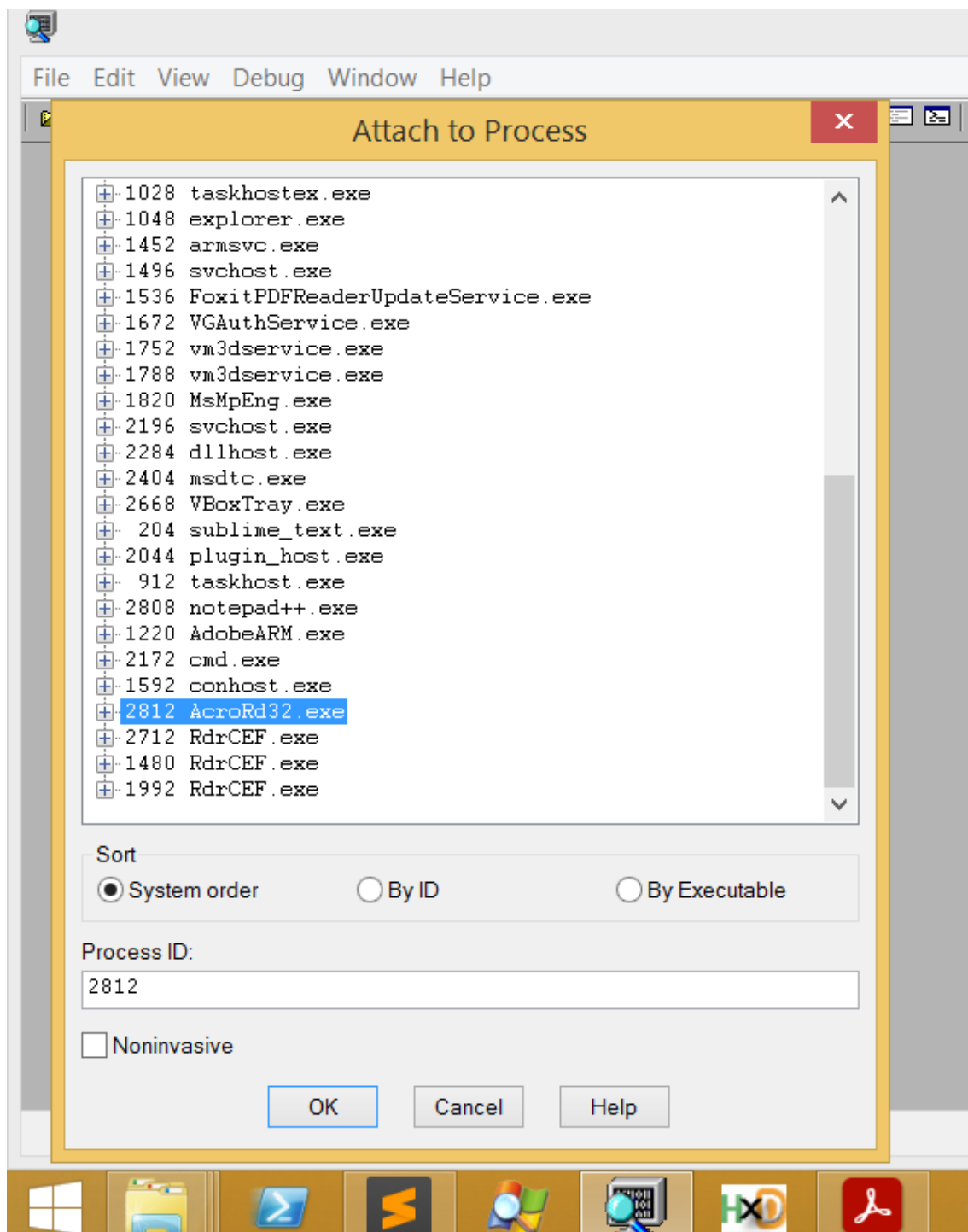
5. 打开1.pdf并关闭(载入EScript模块), attach windbg并运行cmd.txt中的指令载入pykd脚本, 然后打开1.pdf,将结果result.json重命名为1.json,打开2.pdf,重命名为2.json,打开3.pdf,重命名为3.json

5. open 1.pdf with Adobe Reader (this step is to load EScript.api module, 1.pdf will automatically close itself)



use windbg to attach to Adobe Reader





execute the following command to load record.py

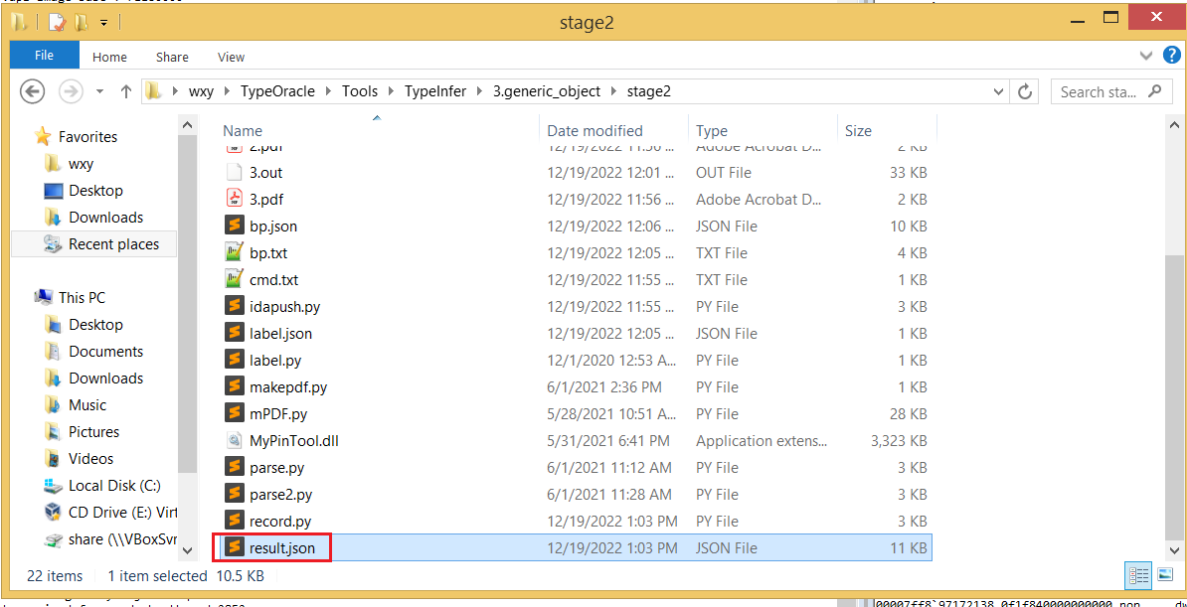
```
.load pykd;!py -g
C:\Users\wxy\TypeOracle\Tools\TypeInfer\adobe\3.generic_object\stage2\record.py
```

```
0:009> .load pykd;!py -g C:\Users\wxy\TypeOracle\Tools\TypeInfer\3.generic_object\stage2\record.py
[*] load 377 breakpoint
[*] EScrip.api image base : 71180000

*BUSY* Debuggee is running...
```

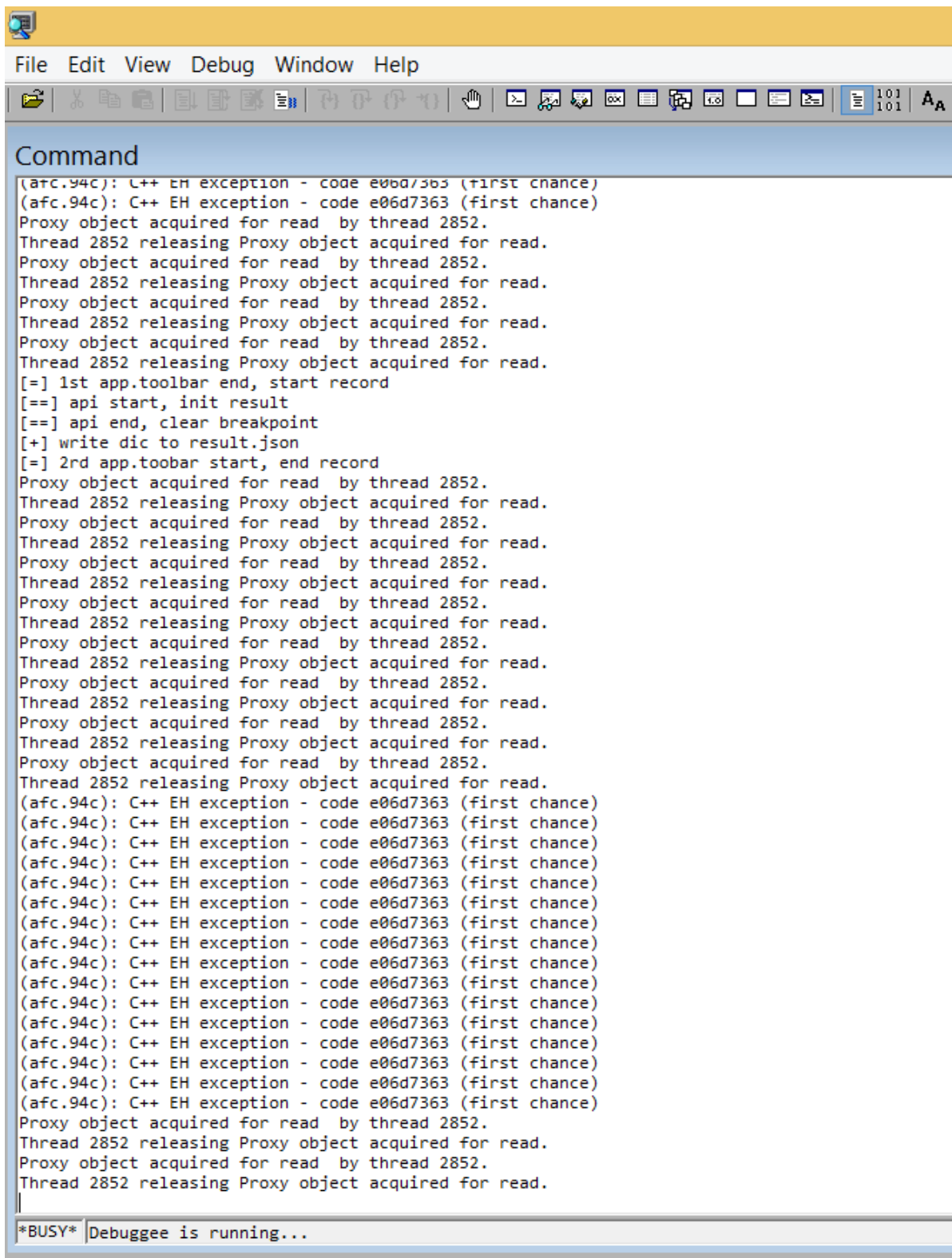
open 1.pdf with Adobe Reader, this will generate result.json, rename it to 1.json

```
Command
0:009> .load pykd;ipy -g C:\Users\wxy\TypeOracle\Tools\TypeInfer\3.generic_object\stage2\record.py
[*] load 377 breakpoint
[*] EScript.api image base : 71180000
[=] 1st app.toolbar end, start record
[==] api start, init result
[==] api end, clear breakpoint
[+] write dic to result.json
[=] 2rd app.toobar start, end record
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
*BUSY* Debuggee is running...
```

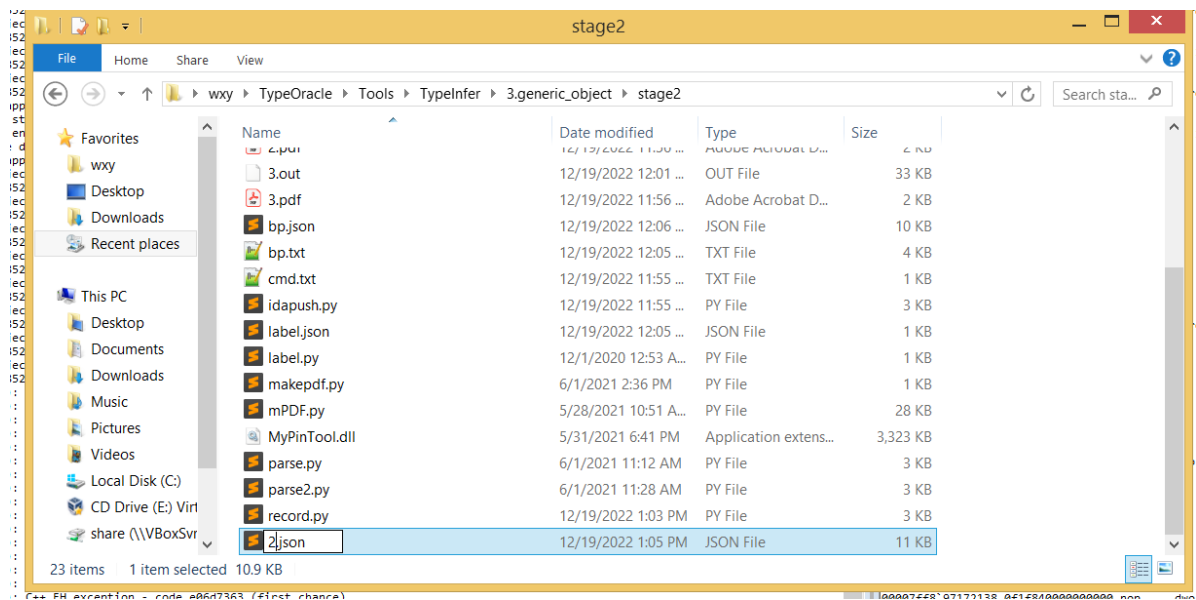


open 2.pdf with Adobe Reader, this will generate result.json, rename it to 2.json







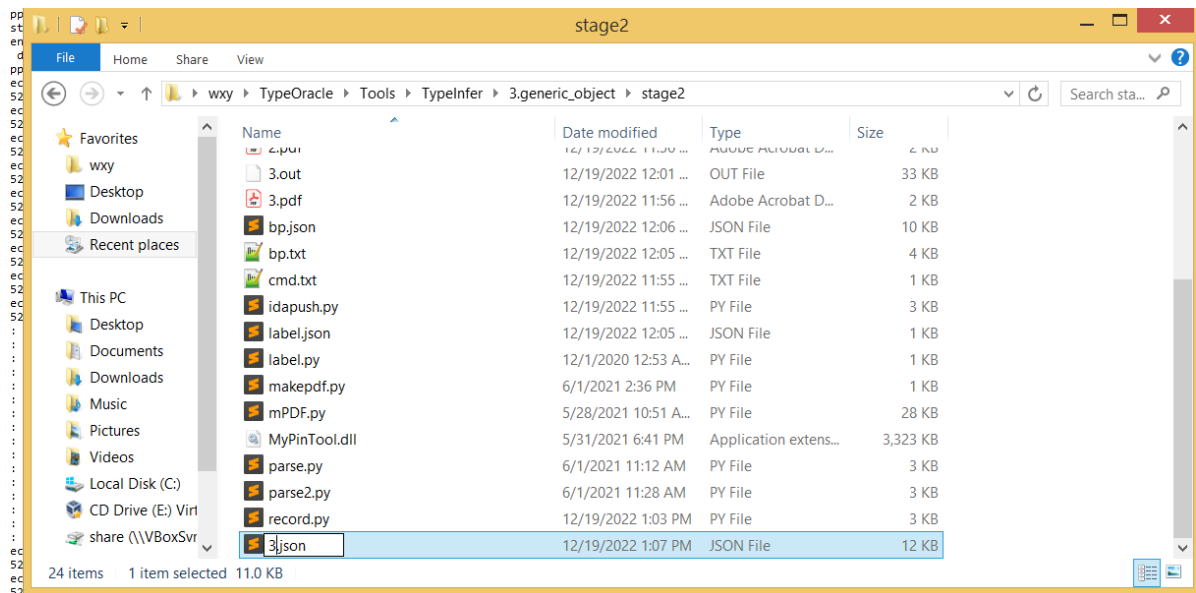


open 3.pdf with Adobe Reader, this will generate result.json, rename it to 3.json

## Command

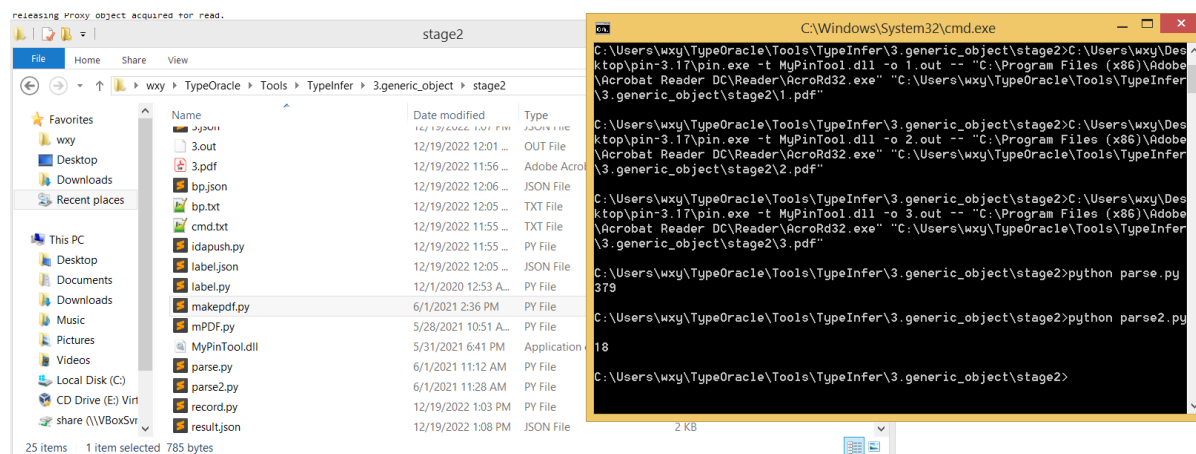
```
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
[=] 1st app.toolbar end, start record
[==] api start, init result
[==] api end, clear breakpoint
[+] write dic to result.json
[=] 2nd app.toobar start, end record
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
(afc.94c): C++ EH exception - code e06d7363 (first chance)
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
Proxy object acquired for read by thread 2852.
Thread 2852 releasing Proxy object acquired for read.
```

```
*BUSY* Debuggee is running...
```



6. execute parse2.py to get result.json, it contains the key-get indicators

```
python parse2.py
```



## 4.arg\_probe

This step uses the type indicator extracted previously to obtain type information.

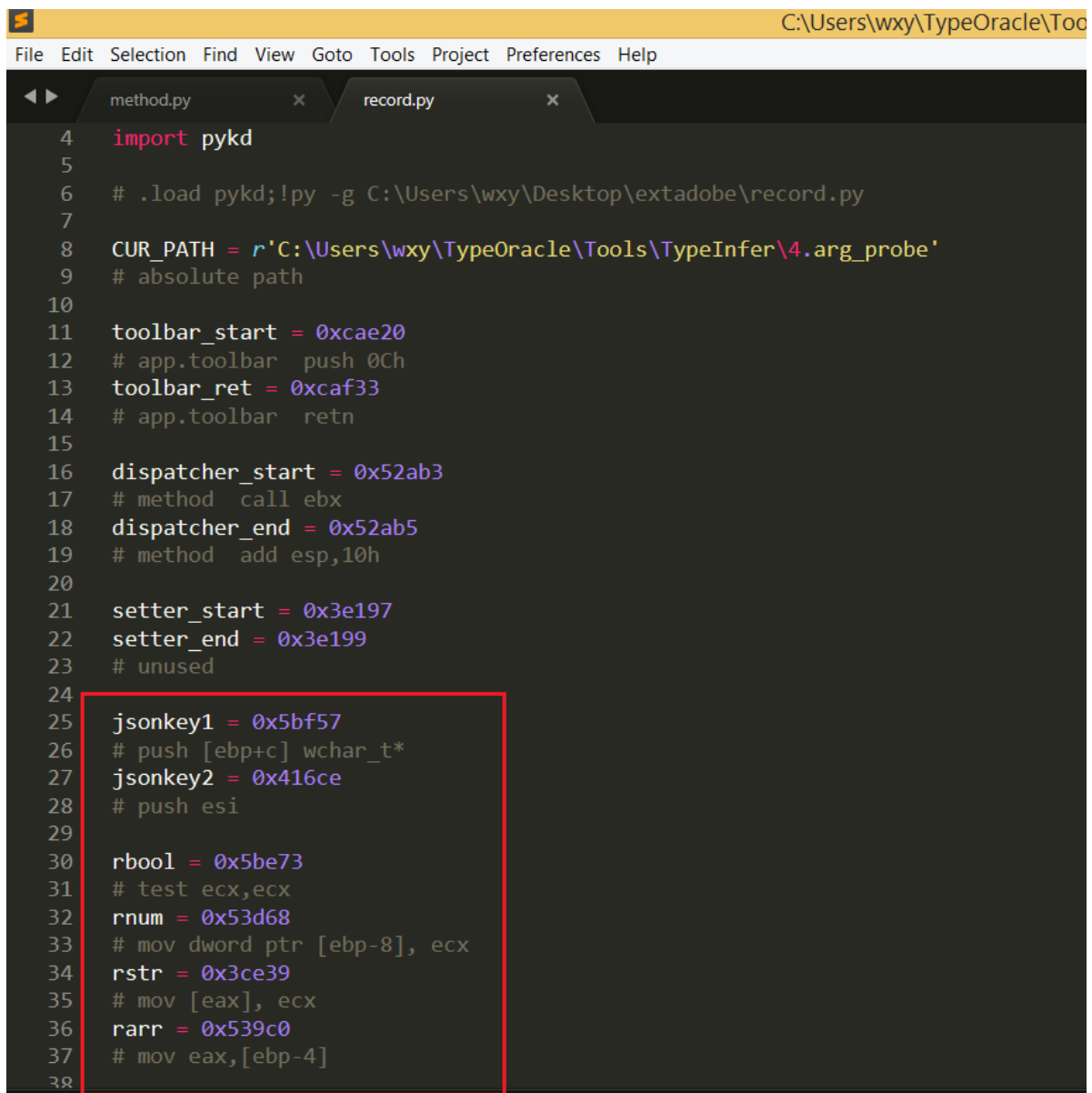
## folder structure

- sample\_result (the folder store the sample result of this step)
- sample\_tmp (the folder store sample intermediate results)
- 1.json-9.json
- func1st.txt (list of all methods)

- init.pdf (PDF file to init Adobe Reader)
- jstemplate.js
- method.py
- monitor.py (monitor the execution of PDF Reader)
- mPDF.py (generate PDF file)
- parse.py
- pattern.py (generate javascript code for one binding call)
- record.py (script executed in windbg)
- runtime.py
- s1.py (combine the results)
- setter1st.txt (list of all accessors)

## how to reproduce

1. fill the type indicator inferred before in `record.py`

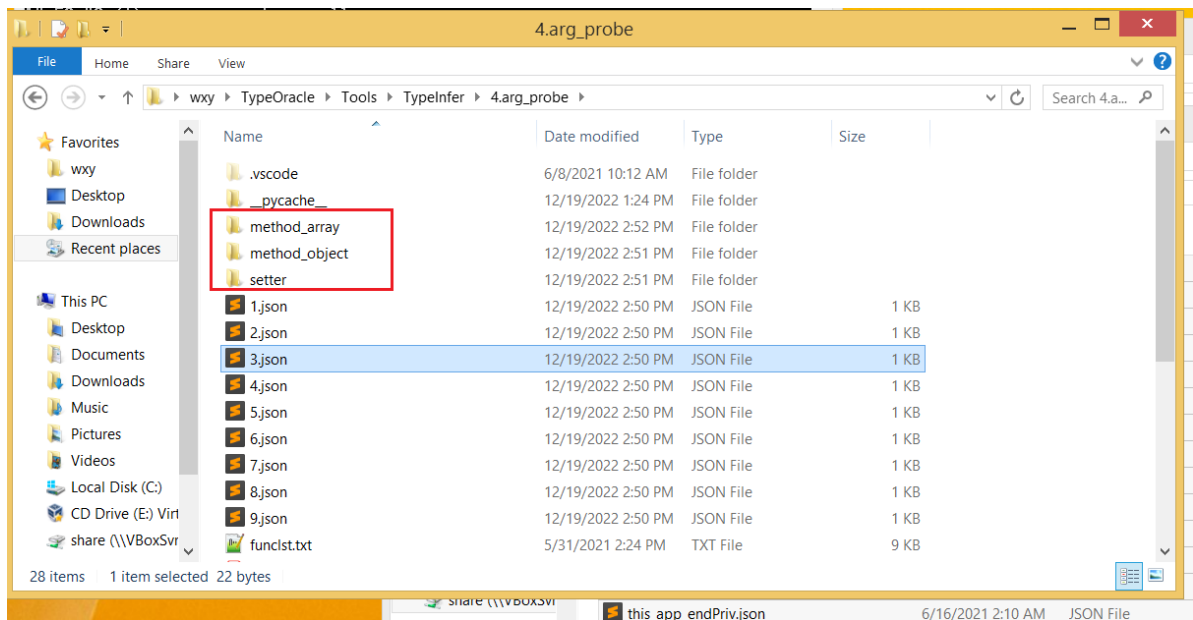


```
4 import pykd
5
6 # .load pykd;!py -g C:\Users\wxy\Desktop\extadobe\record.py
7
8 CUR_PATH = r'C:\Users\wxy\TypeOracle\Tools\TypeInfer\4.arg_probe'
9 # absolute path
10
11 toolbar_start = 0xcae20
12 # app.toolbar push 0Ch
13 toolbar_ret = 0xcaf33
14 # app.toolbar retn
15
16 dispatcher_start = 0x52ab3
17 # method call ebx
18 dispatcher_end = 0x52ab5
19 # method add esp,10h
20
21 setter_start = 0x3e197
22 setter_end = 0x3e199
23 # unused
24
25 jsonkey1 = 0x5bf57
26 # push [ebp+c] wchar_t*
27 jsonkey2 = 0x416ce
28 # push esi
29
30 rbool = 0x5be73
31 # test ecx,ecx
32 rnum = 0x53d68
33 # mov dword ptr [ebp-8], ecx
34 rstr = 0x3ce39
35 # mov [eax], ecx
36 rarr = 0x539c0
37 # mov eax,[ebp-4]
38
```

2. execute `method.py` (it will take about five hours), this will generate three folders:

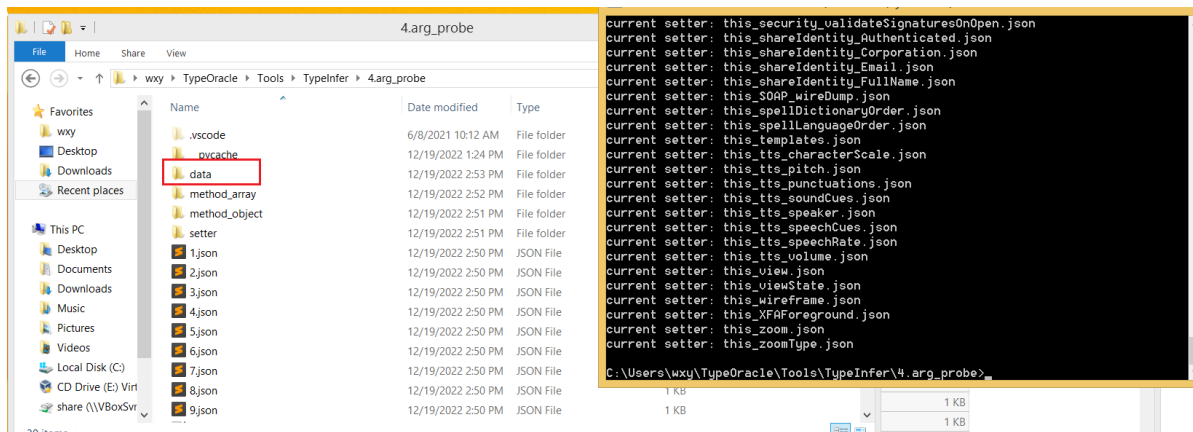
`method_array`, `method_object`, `setter`

```
python method.py
```



3. execute s1.py to combine all results, the type information is store in `data` folder

```
python s1.py
```



## illustration of type information

api: api name

apitype: 0-method 1-accessor

root: root object's index

info:

23x generic object

22x array

req\_type: required argument

opt\_type: optional argument

5x value: multiple types of parameters are allowed

0 bool

1 number

3 string

2 built-in object

5 any type