



教育部資安人才培育計畫
Information Security Incubation Program

教育部先進資通安全實務人才培育計畫

113年度新型態資安實務暑期課程

聚類分析技術 - 惡意軟體變體辨識方法研究

Gapa、巴卡阿卡、Jaccard、咪卡巴卡、噶(๑>~<)๑

組別：情資運用及防禦 - B8

組員：鄭宇兩、曹凱翔、吳玗儒、蔡士煒

組員



鄭宇兩
貝坦



吳玥儒
slimu



曹凱翔
l3obo



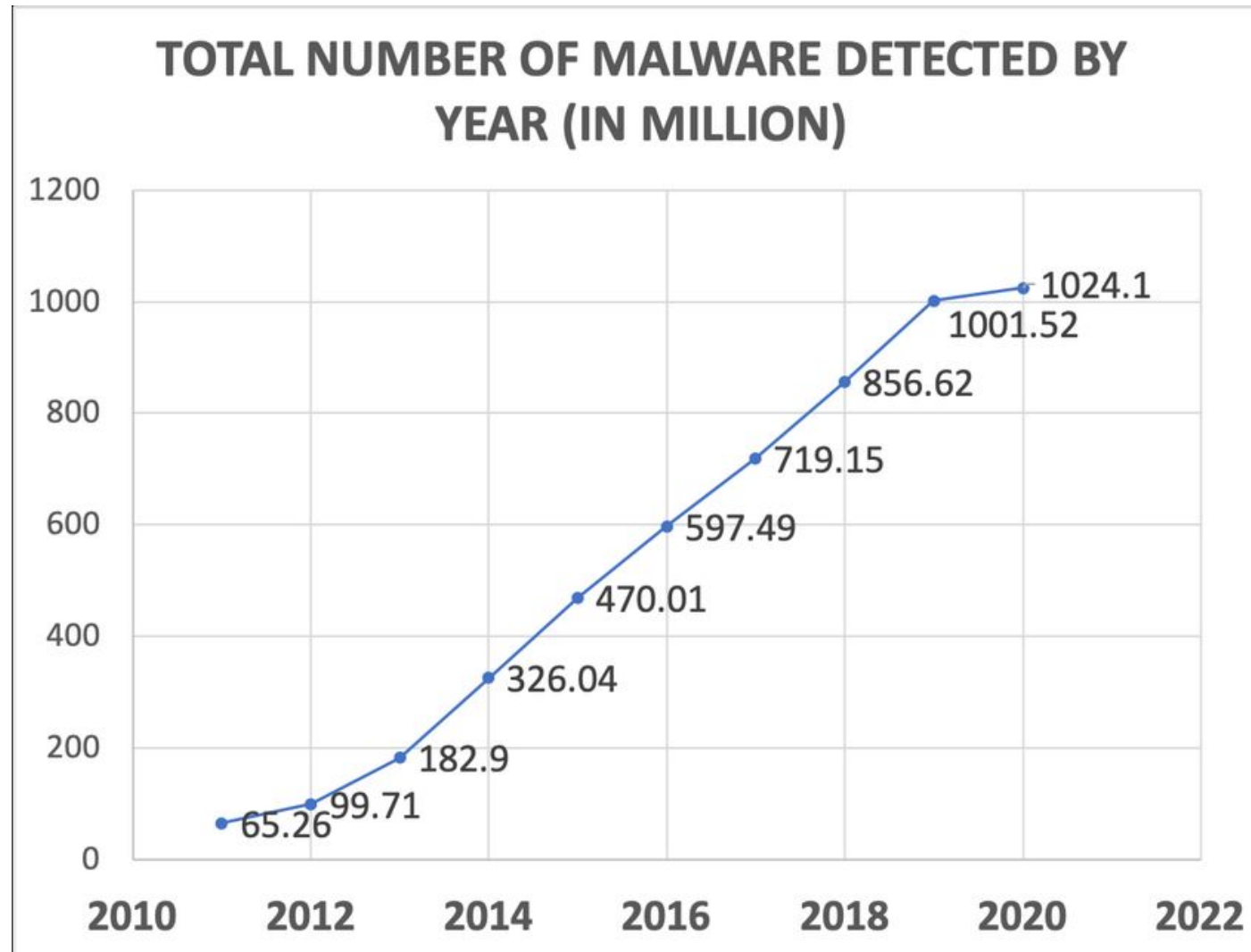
蔡士煒
小蔡

目錄



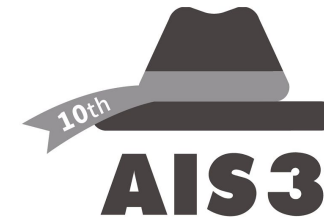
- 動機
- 目標
- 架構
 - 病毒來源
 - CAPA
 - Jaccard
- 方法一：Jaccard
- 方法二：文本向量化
- 方法三：二進位格式、Jaccard
 - 成果展示
- 總結

動機

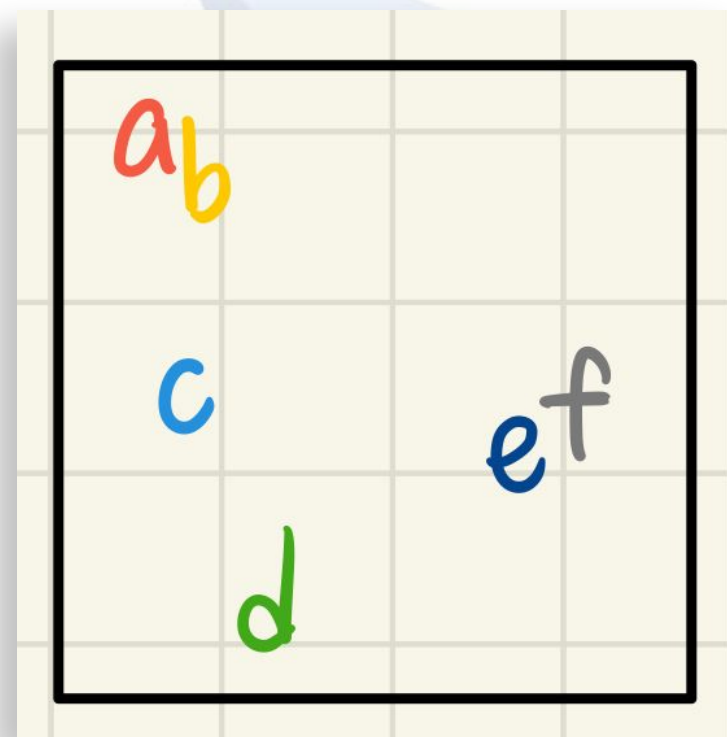


[Total number of malware detected by year \(in million\)](#)

目標



自動解析未知惡意軟體與其他惡意軟體間的 相似度
協助逆向分析人員
更有效識別和分類惡意軟體。

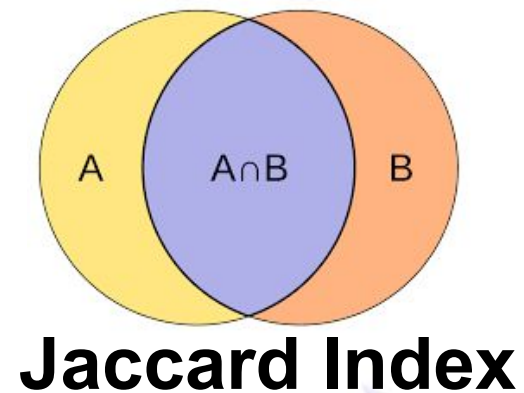


架構



Malware →  CAPA

方法一



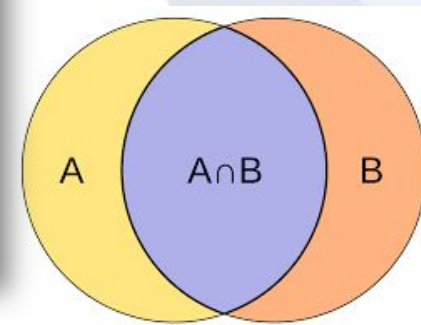
方法二

文本向量化

方法三

```
{  
  "attack_ids": [  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    1,  
    0
```

→ K-means



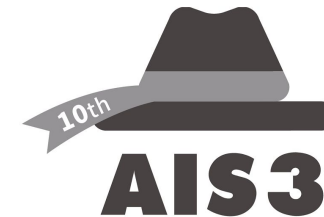
病毒來源



- Malware Bazaar
 - 樣本數量多且有明確分類
 - 可用API大量抓取

Date (UTC)	SHA256 hash	Type	Signature	Tags
2024-08-02 19:05	e9837fc1d609e00844525...	exe	AsyncRAT	AsyncRAT exe RAT
2024-08-02 18:50	82e4fa04f1bbb083ff8657...	elf	Mirai	mirai
2024-08-02 18:50	8184f3e0731d2df9d58c0...	elf	Mirai	mirai
2024-08-02 18:50	ddb07ed1f41dc60aa797...	elf	Mirai	mirai
2024-08-02 18:16	f451f253429a5ea6156e4...	exe		exe signed
2024-08-02 18:16	848260ba966228c4db25...	exe	CobaltStrike	CobaltStrike exe
2024-08-02 18:16	9daffaea889ac9a0a94e7...	exe	Vidar	exe vidar
2024-08-02 18:16	7f08c2afa083a9814989c...	exe	PureLogStealer	exe PureLogStealer

CAPA分析架構



- 基本資料
- MITRE ATT&CK
- MBC(Malware Behavior Catalog Objective)
- Capability

CAPA分析資料那麼多 要選哪些？



基本資料



md5	8	96d8
sha1	5ff4	
sha256	ed01ebfu	0439c6e5babe8e080e41aa
analysis	static	
os	windows	
format	pe	
arch	i386	
path	C:\...	exe

ATT&CK Tactic



ATT&CK Tactic	ATT&CK Technique
DEFENSE EVASION	File and Directory Permissions Modification T1026 Obfuscate Files or Information T1027
DISCOVERY	File and Directory Discovery T1083 Query Registry T1012 System Information Discovery T1082
EXECUTION	Share Files T1129 System Services::Service Execution T1569.002
PERSISTENCE	Create or Modify System Process::Windows T1543.003

MBC Objective



MBC Objective	MBC
CRYPTOGRAPHY	Encrypt Data::AES [C0027.001] Decrypt Data::RC4 [C0027.009] Generate Encryption Key::RC4 KSA [C0028.002] Generate Pseudo-random Sequence [C0021]
DATA	Checksum::CRC32 [C0032.001] Compression Library [C0060] Encrypt Data::XOR [C0026.002]
DEFENSE EVASION	Encrypt Files or Information::Encoding Standard Algorithm [E1027.m02] Obfuscate Files or Information::Encryption Standard Algorithm [E1027.m05]
DISCOVERY	Code Discovery [E1046.001] File and Directory Discovery [E1082] System Information Discovery [E1082]

Capability



Capability	Namespaces
hash data with CRC32	host-interaction/crc32
encode data using XOR	host-interaction/g/xor
encrypt data using AES (5 matches)	host-interaction/encryption/aes
reference AES constants (5 matches)	host-interaction/encryption/aes
encrypt data using RC4 KSA (3 matches)	host-interaction/encryption/rc4
generate random numbers using the Dr	host-interaction/prng/lcg
extract resource via kernel32 fu	host-interaction/resource
get common file path (3 matches)	host-interaction/file-system
set current directory (3 matches)	host-interaction/system
copy file	host-interaction/system/copy
create directory (2 matches)	host-interaction/system/create
check if file exists (4 matches)	host-interaction/system/exists
get file attributes (6 matches)	host-interaction/system/meta
get file size (2 matches)	host-interaction/file-system/meta
set file attributes	host-interaction/file-system/meta
read file on Windows (3 matches)	host-interaction/file-system/read
write file on Windows (2 matches)	host-interaction/file-system/write
check mutex	host-interaction/mutex

**那又要怎麼確定
CAPA分析結果是正確的
呢？**



確認準確率



- 提供了推測的地方
 - \$ capa.exe malware.exe -vv
- 不信就自己逆看看

確認準確率



```
encode data using XOR (15 matches)
namespace data-manipulation/encoding/xor
author moritz.raabe@mandiant.com
scope basic block
att&ck Defense Evasion::Obfuscated Files or Information [T1027]
mbc Defense Evasion::Obfuscated Files or Information::Encoding
Data::Encode Data::XOR [C0026.002]
```

```
basic block @ 0x402D25 in function 0x402A76
```

確認準確率



- 把自己跟前一位做 xor

```
loc_402D25:  
mov     ebx, [eax-4]  
xor     [eax], ebx
```

```
*v28 ^= *(v28 - 1);
```

確認準確率



- 把自己跟前一位做 xor

```
loc_402D25:  
mov     ebx, [eax-4]  
xor     [eax], ebx
```

```
*v28 ^= *(v28 - 1);
```

Data::Encode Data::XOR [C0026.002]



Jaccard Index介紹

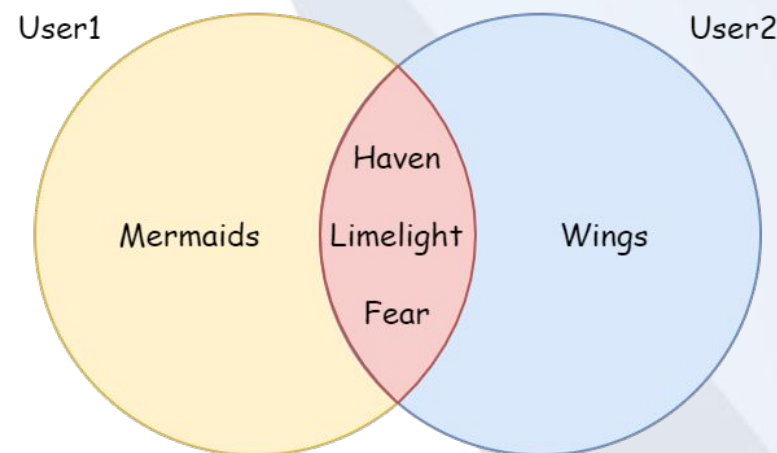


- File and Directory Discovery > 'File' 'and' 'Directory' 'Discovery'
- System and Directory Discovery > 'System' 'and' 'Directory' 'Discovery'

交集 = Information, and, Discovery

聯集 = File, System, and, Directory, Discovery

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{5} = 0.6$$



[用白話文談數學公式 - Jaccard Index](#)

方法一

提取CAPA報告特徵



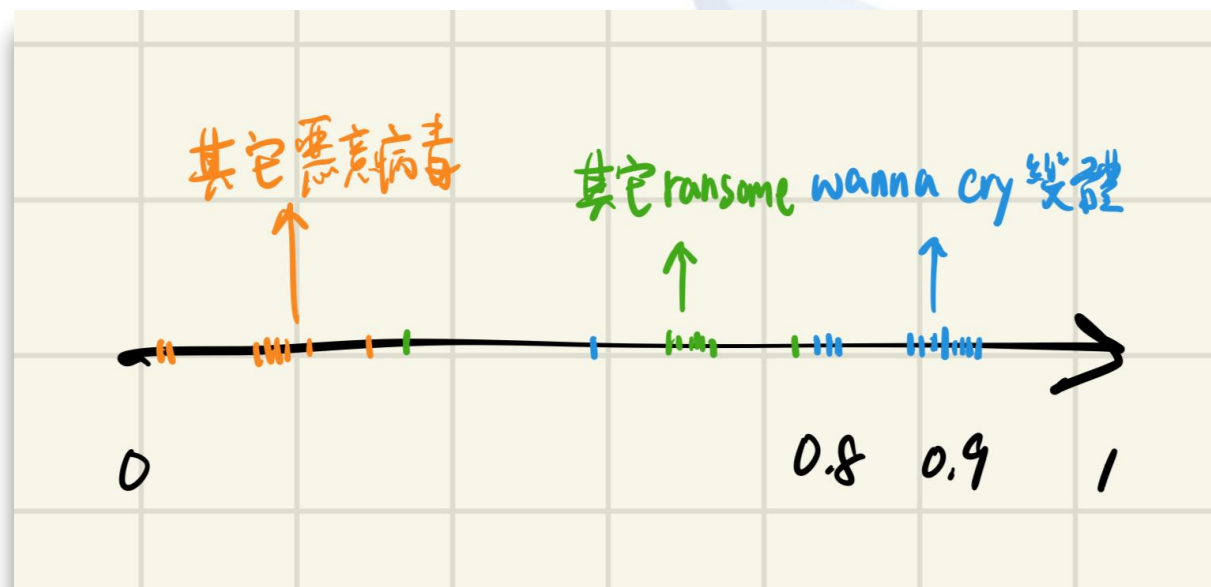
- 從編號下手
- 開一個陣列，紀錄有無該編號

```
wannacry_sample = ['T1001', 'T1003', 'T1005', 'C0011', 'C0004', 'C0002']  
wannacry_vriant = ['T1005', 'C0011', 'C0004']
```

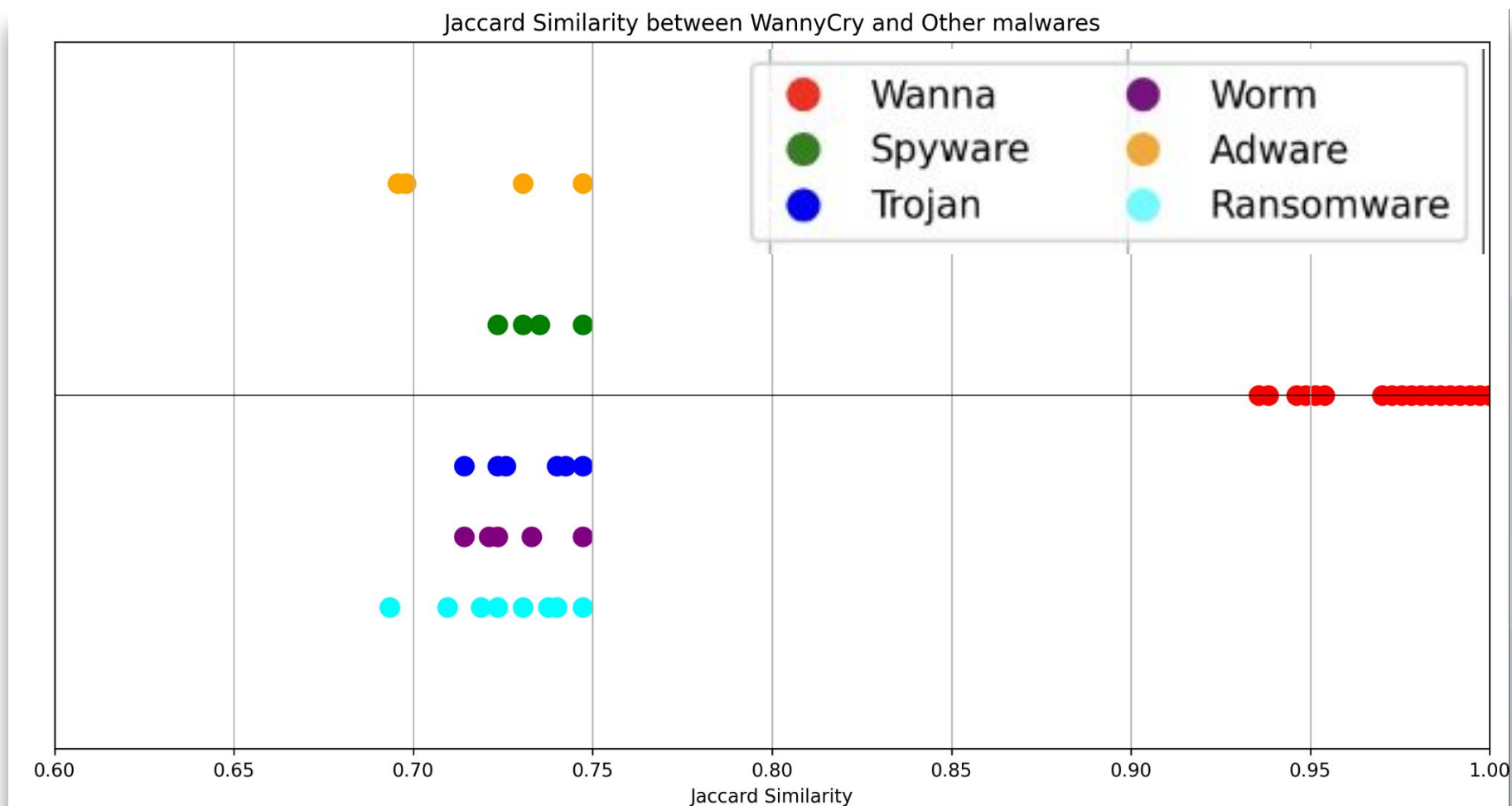
Jaccard Index比較類別



- WannaCry變體
- 其他勒索軟體
- 其他惡意程式



實驗分析



方法二

使用技術



- **K-means**
- **CountVectorizer**
- **SentenceTransformer**
- **PCA**

AIS3
10th

MBC Objective	MBC Behavior
ANTI-STATIC ANALYSIS	Executable Code Obfuscation::Argument Obfuscation [B0032.020] Executable Code Obfuscation::Stack Strings [B0032.017]
COLLECTION	Keylogging::Polling [F0002.002]
DATA	Encode Data::XOR [C0026.002]
DEFENSE EVASION	Obfuscated Files or Information::Encoding-Standard Algorithm [E1027.m02]
DISCOVERY	System Information Discovery [E1082]
FILE SYSTEM	Writes File [C0052]
MEMORY	Allocate Memory [C0007]
PROCESS	Allocate Thread Local Storage [C0040] Check Mutex [C0043] Create Mutex [C0042] Set Thread Local Storage Value [C0041] Terminate Process [C0018]

```
all_attack_ids = ['T1001', 'T1003', 'T1005', 'T1006', 'T1007', 'T1008', 'T1009', 'T1010', 'T1011', 'T1012', 'T1013', 'T1014', 'T1015', 'T1016', 'T1017', 'T1018', 'T1019', 'T1020', 'T1021', 'T1022', 'T1023', 'T1024', 'T1025', 'T1026', 'T1027', 'T1029', 'T1030', 'T1031', 'T1032', 'T1033', 'T1034', 'T1035', 'T1036', 'T1037', 'T1038', 'T1039', 'T1040', 'T1041', 'T1042', 'T1043', 'T1044', 'T1045', 'T1046', 'T1047', 'T1048', 'T1049', 'T1050', 'T1051', 'T1052', 'T1053', 'T1054', 'T1055', 'T1056', 'T1057', 'T1058', 'T1059', 'T1060', 'T1061', 'T1062', 'T1063', 'T1064', 'T1065', 'T1066', 'T1067', 'T1068', 'T1069', 'T1070', 'T1071', 'T1072', 'T1073', 'T1074', 'T1075', 'T1076', 'T1077', 'T1078', 'T1079', 'T1080', 'T1082', 'T1083', 'T1084', 'T1085', 'T1086', 'T1087', 'T1088', 'T1089', 'T1090', 'T1091', 'T1092', 'T1093', 'T1094', 'T1095', 'T1096', 'T1097', 'T1098', 'T1099', 'T1100', 'T1101', 'T1102', 'T1103', 'T1104', 'T1105', 'T1106', 'T1107', 'T1108', 'T1109', 'T1110', 'T1111', 'T1112', 'T1113', 'T1114', 'T1115', 'T1116', 'T1117', 'T1118', 'T1119', 'T1120', 'T1121', 'T1122', 'T1123', 'T1124', 'T1125', 'T1126', 'T1127', 'T1128', 'T1129', 'T1130', 'T1131', 'T1132', 'T1133', 'T1134', 'T1135', 'T1136', 'T1137', 'T1138', 'T1139', 'T1140', 'T1141', 'T1142', 'T1143', 'T1144', 'T1145', 'T1146', 'T1147', 'T1148', 'T1149', 'T1150', 'T1151', 'T1152', 'T1153', 'T1154', 'T1155', 'T1156', 'T1157', 'T1158', 'T1159', 'T1160', 'T1161', 'T1162', 'T1163', 'T1164', 'T1165', 'T1166', 'T1167', 'T1168', 'T1169', 'T1170', 'T1171', 'T1172', 'T1173', 'T1174', 'T1175', 'T1176', 'T1177', 'T1178', 'T1179', 'T1180', 'T1181', 'T1182', 'T1183', 'T1184', 'T1185', 'T1186', 'T1187', 'T1188', 'T1189', 'T1190', 'T1191', 'T1192', 'T1193', 'T1194', 'T1195', 'T1196', 'T1197', 'T1198', 'T1199', 'T1200', 'T1201', 'T1202', 'T1203', 'T1204', 'T1205', 'T1206', 'T1207', 'T1208', 'T1209', 'T1210', 'T1211', 'T1212', 'T1213', 'T1214', 'T1215', 'T1216', 'T1217', 'T1218', 'T1219', 'T1220', 'T1221', 'T1222', 'T1223', 'T1224', 'T1225', 'T1226', 'T1227', 'T1228', 'T1229', 'T1230', 'T1231', 'T1232', 'T1233', 'T1234', 'T1235', 'T1236', 'T1237', 'T1238', 'T1239', 'T1240', 'T1241', 'T1242', 'T1243', 'T1244', 'T1245', 'T1246', 'T1247', 'T1248', 'T1249', 'T1250', 'T1251', 'T1252', 'T1253', 'T1254', 'T1255', 'T1256', 'T1257', 'T1258', 'T1259', 'T1260', 'T1261', 'T1262', 'T1263', 'T1264', 'T1265', 'T1266', 'T1267', 'T1268', 'T1269', 'T1270', 'T1271', 'T1272', 'T1273', 'T1274', 'T1275', 'T1276', 'T1277', 'T1278', 'T1279', 'T1280', 'T1281', 'T1282', 'T1283', 'T1284', 'T1285', 'T1286', 'T1287', 'T1288', 'T1289', 'T1290', 'T1291', 'T1292', 'T1293', 'T1294', 'T1295', 'T1296', 'T1297', 'T1298', 'T1299', 'T1300', 'T1301', 'T1302', 'T1303', 'T1304', 'T1305', 'T1306', 'T1307', 'T1308', 'T1309', 'T1310', 'T1311', 'T1312', 'T1313', 'T1314', 'T1315', 'T1316', 'T1317', 'T1318', 'T1319', 'T1320', 'T1321', 'T1322', 'T1323', 'T1324', 'T1325', 'T1326', 'T1327', 'T1328', 'T1329', 'T1330', 'T1331', 'T1332', 'T1333', 'T1334', 'T1335', 'T1336', 'T1337', 'T1338', 'T1339', 'T1340', 'T1341', 'T1342', 'T1343', 'T1344', 'T1345', 'T1346', 'T1347', 'T1348', 'T1349', 'T1350', 'T1351', 'T1352', 'T1353', 'T1354', 'T1355', 'T1356', 'T1357', 'T1358', 'T1359', 'T1360', 'T1361', 'T1362', 'T1363', 'T1364', 'T1365', 'T1366', 'T1367', 'T1368', 'T1369', 'T1370', 'T1371', 'T1372', 'T1373', 'T1374', 'T1375', 'T1376', 'T1377', 'T1378', 'T1379', 'T1380', 'T1381', 'T1382', 'T1383', 'T1384', 'T1385', 'T1386', 'T1387', 'T1388', 'T1389', 'T1390', 'T1391', 'T1392', 'T1393', 'T1394', 'T1395', 'T1396', 'T1397', 'T1398', 'T1399', 'T1400', 'T1401', 'T1402', 'T1403', 'T1404', 'T1405', 'T1406', 'T1407', 'T1408', 'T1409', 'T1410', 'T1411', 'T1412', 'T1413', 'T1414', 'T1415', 'T1416', 'T1417', 'T1418', 'T1419', 'T1420', 'T1421', 'T1422', 'T1423', 'T1424', 'T1425', 'T1426', 'T1427', 'T1428', 'T1429', 'T1430', 'T1431', 'T1432', 'T1433', 'T1434', 'T1435', 'T1436', 'T1437', 'T1438', 'T1439', 'T1440', 'T1441', 'T1442', 'T1443', 'T1444', 'T1445', 'T1446', 'T1447', 'T1448', 'T1449', 'T1450', 'T1451', 'T1452', 'T1453', 'T1454', 'T1455', 'T1456', 'T1457', 'T1458', 'T1459', 'T1460', 'T1461', 'T1462', 'T1463', 'T1464', 'T1465', 'T1466', 'T1467', 'T1468', 'T1469', 'T1470', 'T1471', 'T1472', 'T1473', 'T1474', 'T1475', 'T1476', 'T1477', 'T1478', 'T1479', 'T1480', 'T1481', 'T1482', 'T1483', 'T1484', 'T1485', 'T1486', 'T1487', 'T1488', 'T1489', 'T1490', 'T1491', 'T1492', 'T1493', 'T1494', 'T1495', 'T1496', 'T1497', 'T1498', 'T1499', 'T1500', 'T1501', 'T1502', 'T1503', 'T1504', 'T1505', 'T1506', 'T1507', 'T1508', 'T1509', 'T1510', 'T1511', 'T1512', 'T1513', 'T1514', 'T1515', 'T1516', 'T1517', 'T1518', 'T1519', 'T1520', 'T1521', 'T1522', 'T1523', 'T1524', 'T1525', 'T1526', 'T1527', 'T1528', 'T1529', 'T1530', 'T1531', 'T1532', 'T1533', 'T1534', 'T1535', 'T1536', 'T1537', 'T1538', 'T1539', 'T1540', 'T1541', 'T1542', 'T1543', 'T1544', 'T1545', 'T1546', 'T1547', 'T1548', 'T1549', 'T1550', 'T1551', 'T1552', 'T1553', 'T1554', 'T1555', 'T1556', 'T1557', 'T1558', 'T1559', 'T1560', 'T1561', 'T1562', 'T1563', 'T1564', 'T1565', 'T1566', 'T1567', 'T1568', 'T1569', 'T1570', 'T1571', 'T1572', 'T1573', 'T1574', 'T1575', 'T1576', 'T1577', 'T1578', 'T1579', 'T1580', 'T1581', 'T1582', 'T1583', 'T1584', 'T1585', 'T1586', 'T1587', 'T1588', 'T
```

✕ ⓘ 001_analysis.txt

T1010, T1027, T1056, T1082, T1129, T1564

C0026, C0052, C0007, C0040, C0042, C0043, C0041, C0018

✕ ⓘ 002_analysis.txt

T1027, T1056, T1082, T1129

C0019, C0026, C0046, C0052, C0007, C0040, C0042, C0043, C0017, C0038, C0054, C0041, C0018

✕ ⓘ 003_analysis.txt

T1083, T1129

C0046, C0048, C0047, C0052, C0040, C0017, C0018

✕ ⓘ 004_analysis.txt

T1027, T1033, T1057, T1059, T1070, T1082, T1083, T1112, T1129, T1222, T1490, T1518, T1547,

C0027, C0031, C0029, C0028, C0045, C0047, C0049, C0063, C0051, C0050, C0052, C0007, C0042,

```
output > {} all2.json > {} train > [ ] data
```

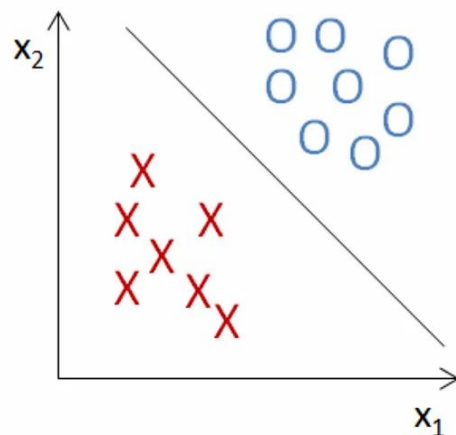
```

1  {
2      "train": {
3          "data": [
4
5              "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0027, C0021, C0028, C0060, C0026, C0063,
6              "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0021, C0060, C0063, C0051, C0038, C0018,
7              "T1027, T1083, T1129, C0001, C0027, C0031, C0021, C0028, C0032, C0060, C0026, C0046, C0049, C0051, C0052",
8              "T1012, T1027, T1082, T1083, T1129, T1222, T1543, T1569, C0027, C0021, C0028, C0032, C0060, C0026, C0045,
9              "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0021, C0060, C0063, C0051, C0038, C0018,
10             "T1083, T1497, C0002, C0052, C0017",
11             "T1016, T1027, T1082, T1083, T1129, T1543, T1547, T1569, C0002, C0001, C0021, C0060, C0063, C0051, C0038,
12             "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0021, C0060, C0063, C0051, C0038, C0018,
13             "T1012, T1027, T1059, T1082, T1083, T1113, T1222, T1497, T1614, C0021, C0026, C0046, C0047, C0049, C0051,
14             "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0027, C0021, C0028, C0060, C0026, C0063,
15             "T1012, T1027, T1033, T1055, T1057, T1082, T1083, T1087, T1129, T1140, T1222, T1497, T1620, C0027, C0029,
16             "T1012, T1027, T1082, T1083, T1129, T1222, T1543, T1569, C0027, C0021, C0028, C0032, C0060, C0026, C0045,
17             "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0027, C0021, C0028, C0060, C0026, C0063,
18             "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0021, C0060, C0063, C0051, C0038, C0018,
19             "T1016, T1027, T1082, T1083, T1129, T1543, T1569, C0002, C0001, C0021, C0060, C0063, C0051, C0038, C0018,
20             "T1012, T1027, T1033, T1057, T1070, T1083, T1087, T1115, T1140, T1222, T1490, T1497, T1518, T1547, C0027

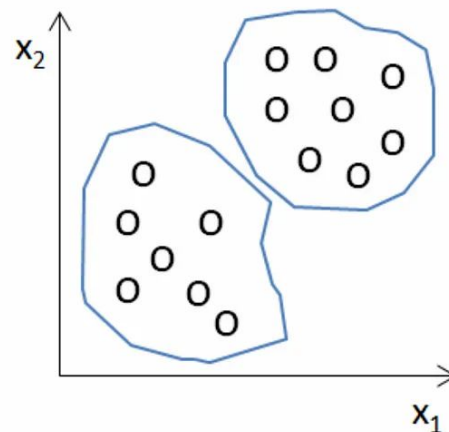
```

K-means介紹

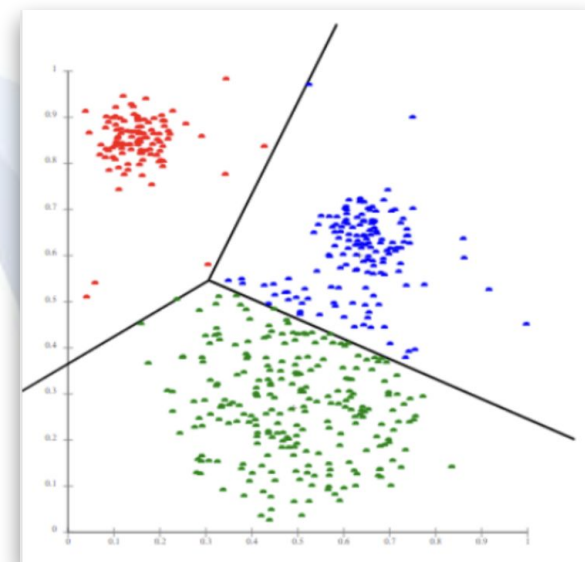
- 分群
- 非監督式學習



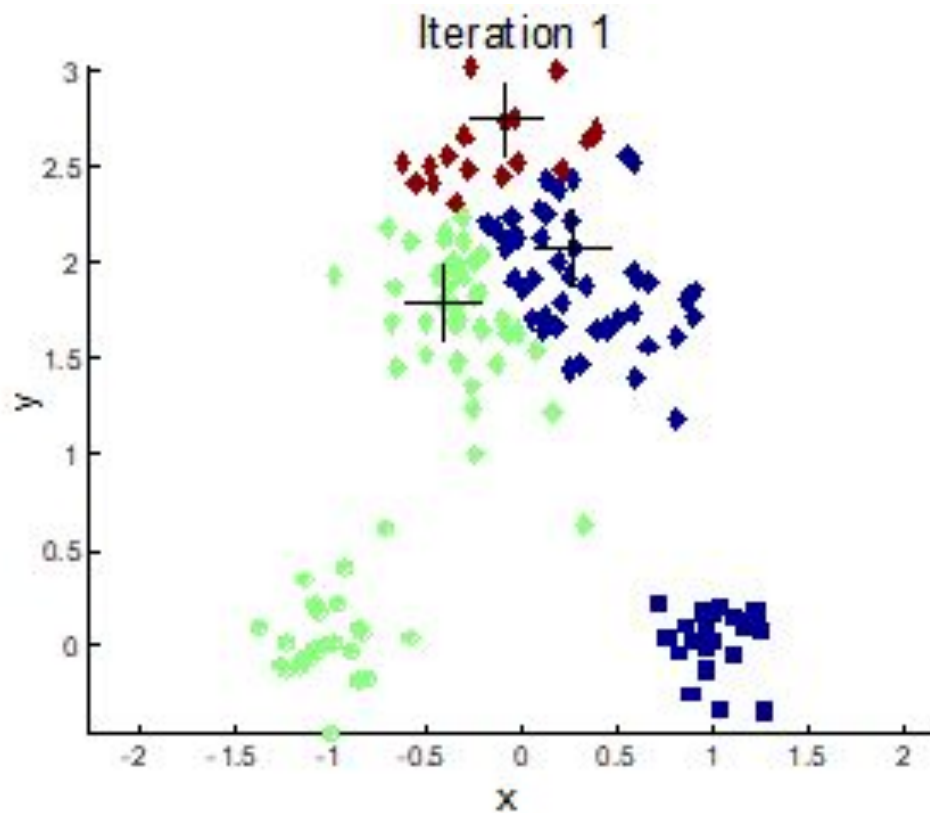
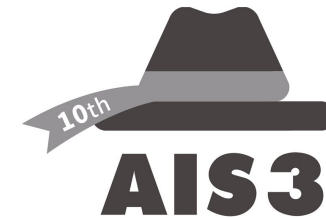
監督式學習



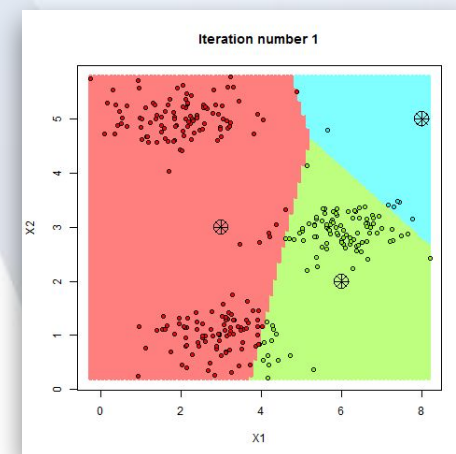
非監督式學習



K-means介紹



設 $k=3$



CountVectorizer

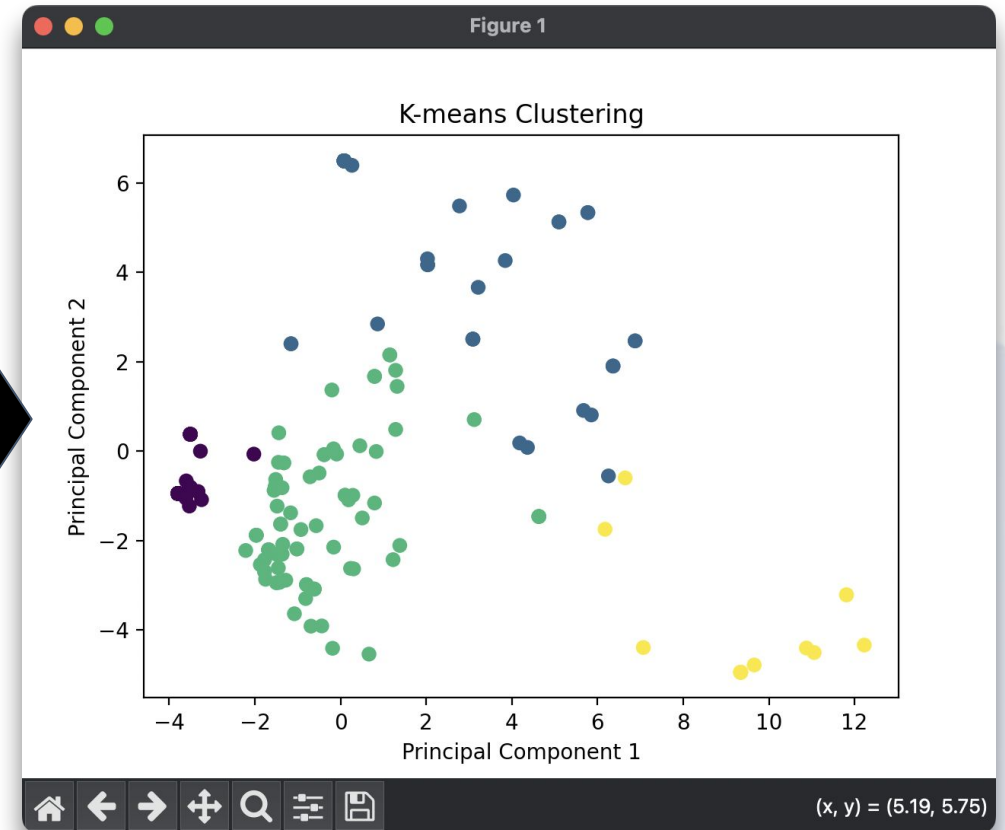


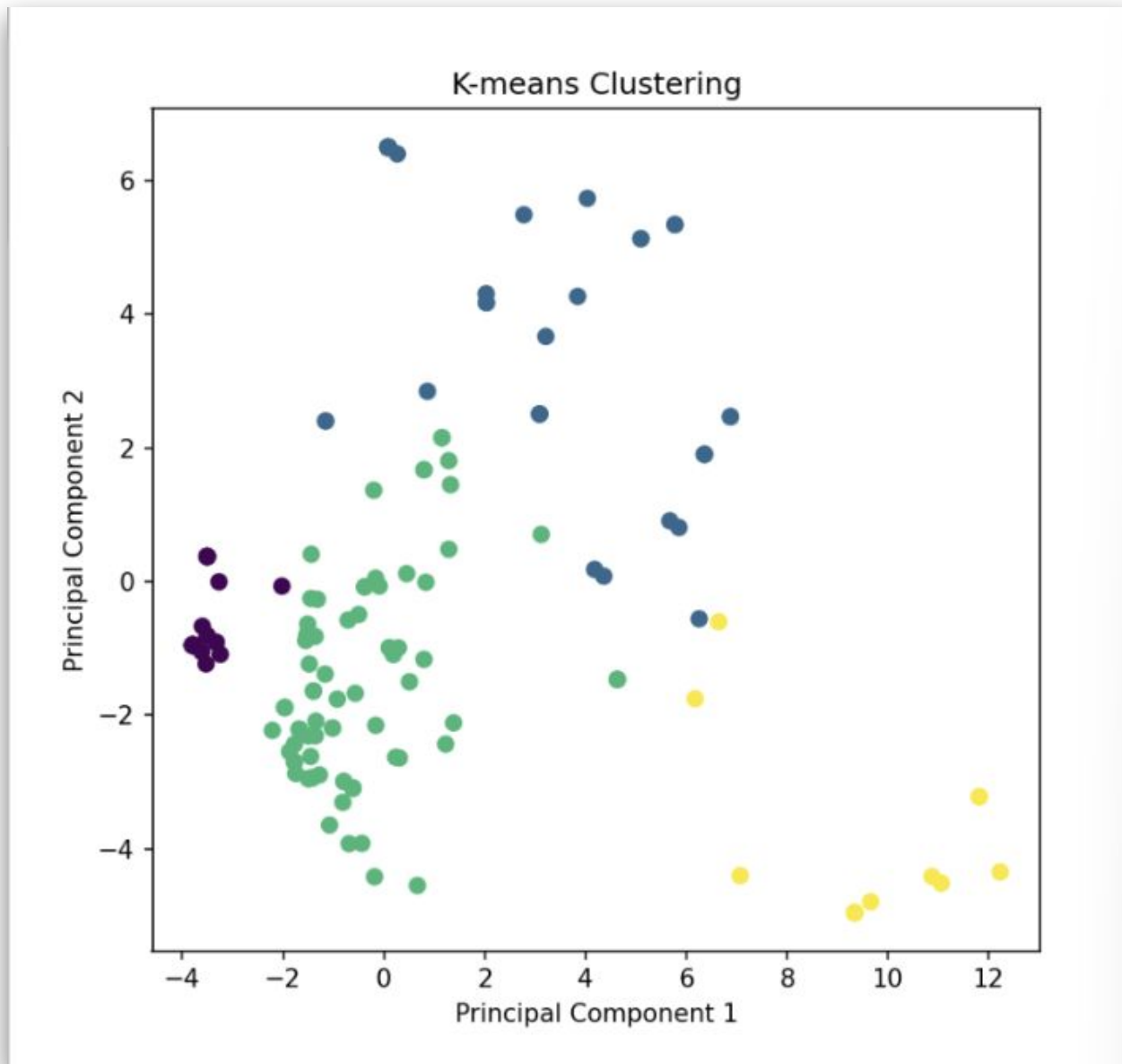
- 提取文本特徵
- 文本轉換為詞袋模型 (Bag-of-Words Model)
- 將文本分割為單詞(或詞組)後，統計每個單詞在文本中出現的次數，並用這些次數構建一個向量。

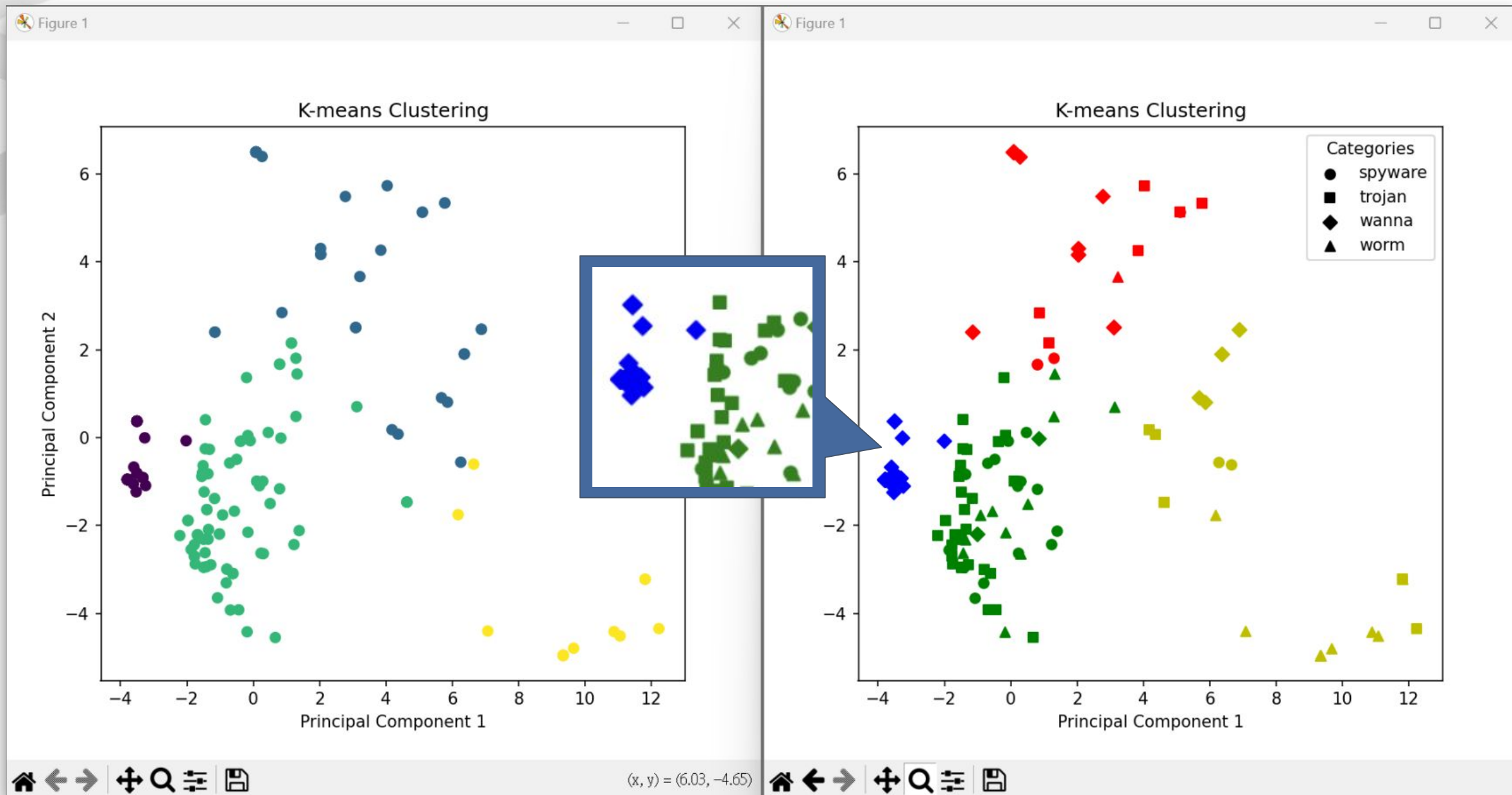
```

output > {} all2.json > {} train > [ ] data
1  {
2    "train": {
3      "data": [
4
5        "T1016, T1027, T1082, T1083, T1129, T1543,
6        "T1016, T1027, T1082, T1083, T1129, T1543,
7        "T1027, T1083, T1129, C0001, C0027, C0031,
8        "T1012, T1027, T1082, T1083, T1129, T1222,
9        "T1016, T1027, T1082, T1083, T1129, T1543,
10       "T1083, T1497, C0002, C0052, C0017",
11       "T1016, T1027, T1082, T1083, T1129, T1543,
12       "T1016, T1027, T1082, T1083, T1129, T1543,
13       "T1012, T1027, T1059, T1082, T1083, T1113,
14       "T1016, T1027, T1082, T1083, T1129, T1543,
15       "T1012, T1027, T1033, T1055, T1057, T1082,
16       "T1012, T1027, T1082, T1083, T1129, T1222,
17       "T1016, T1027, T1082, T1083, T1129, T1543,
18       "T1016, T1027, T1082, T1083, T1129, T1543,
19       "T1016, T1027, T1082, T1083, T1129, T1543,
20       "T1012, T1027, T1033, T1057, T1070, T1083,

```







SentenceTransformer



- 用於句子和段落 Embedding
- 基於BERT及其變體(如 RoBERTa、DistilBERT 等)的預訓練模型
- 生成之 Embedding能夠捕捉到句子的語義和語境

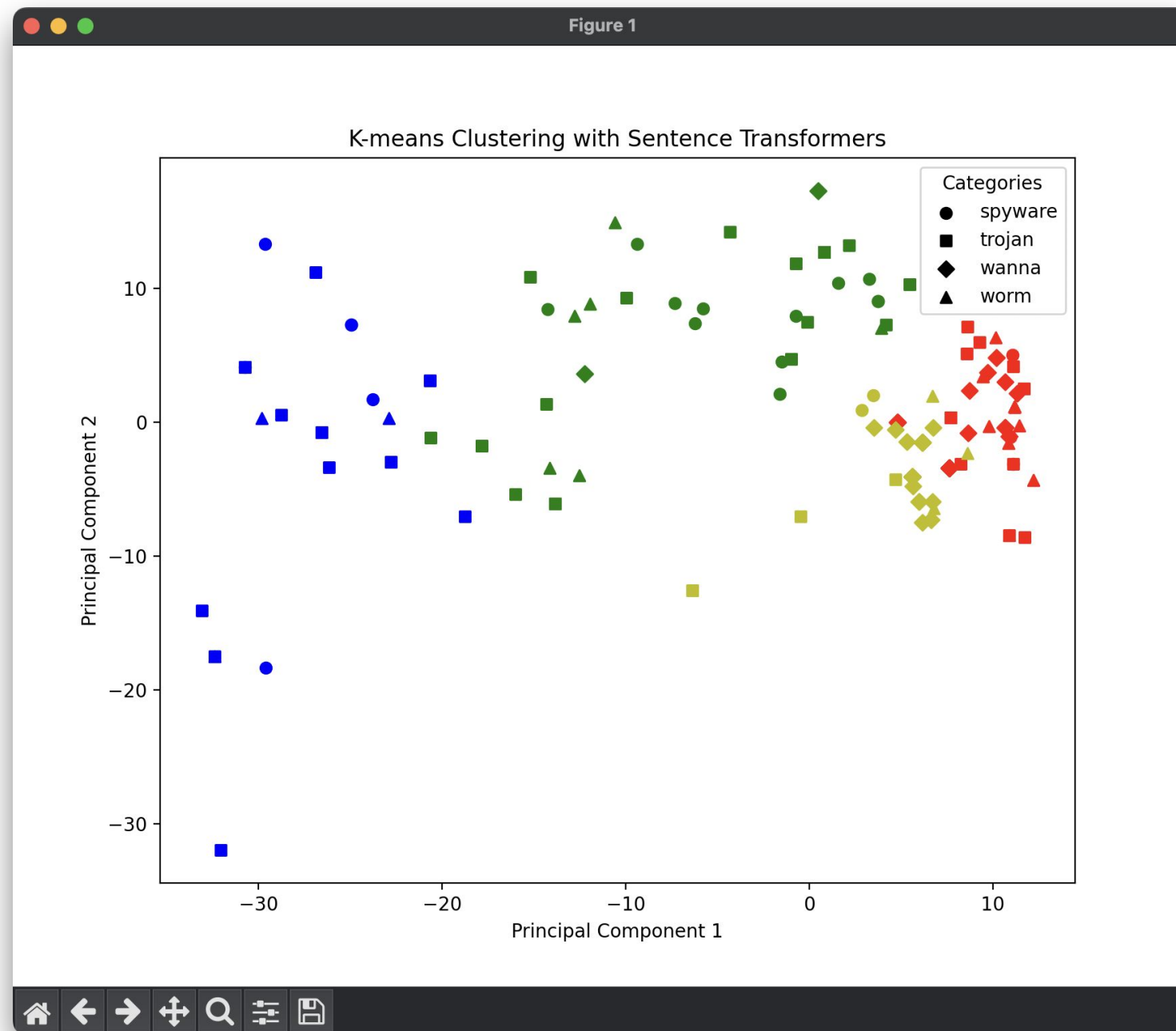
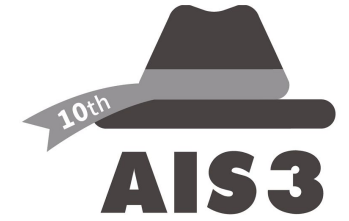
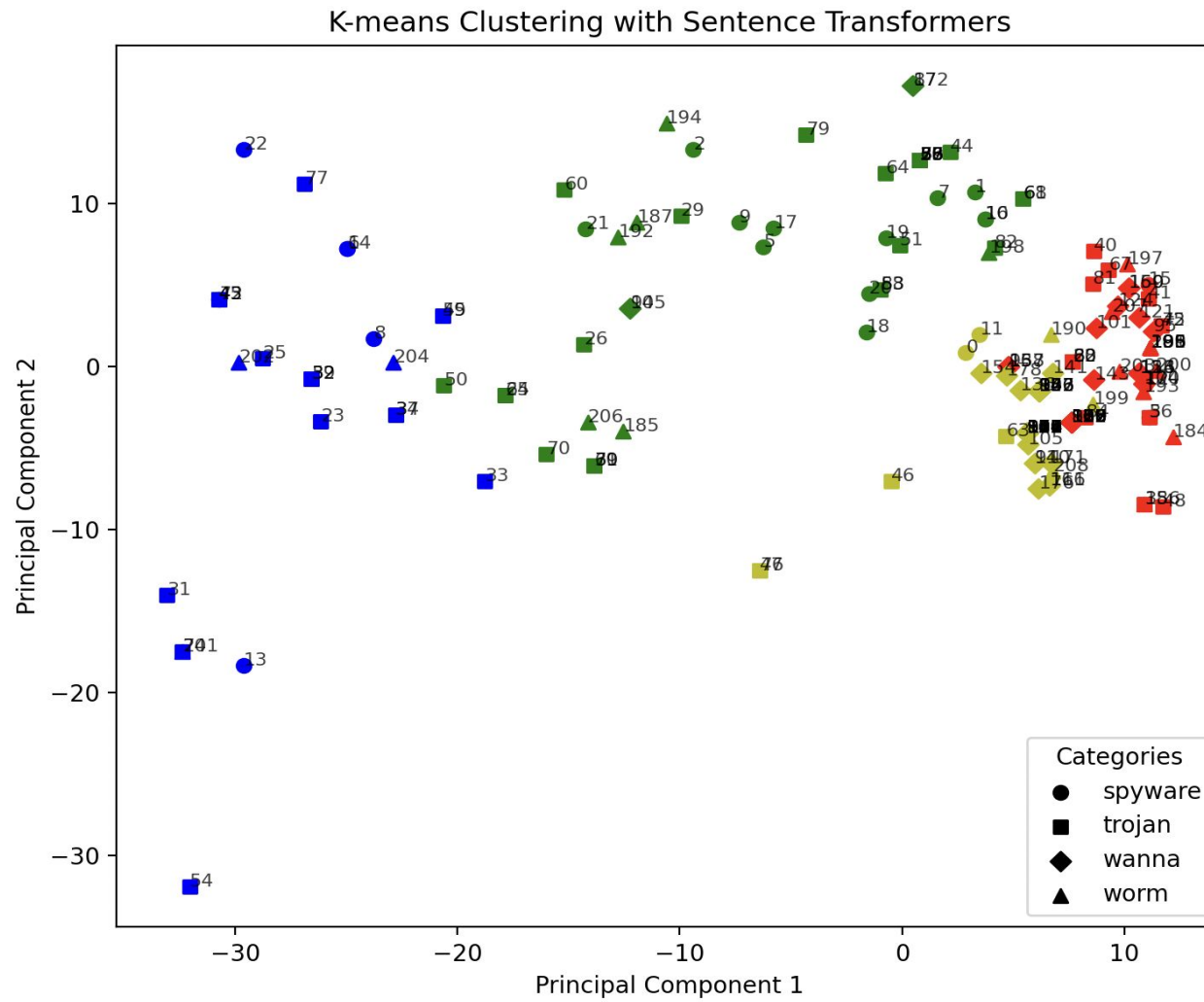


Figure 1



...QQ



方法三

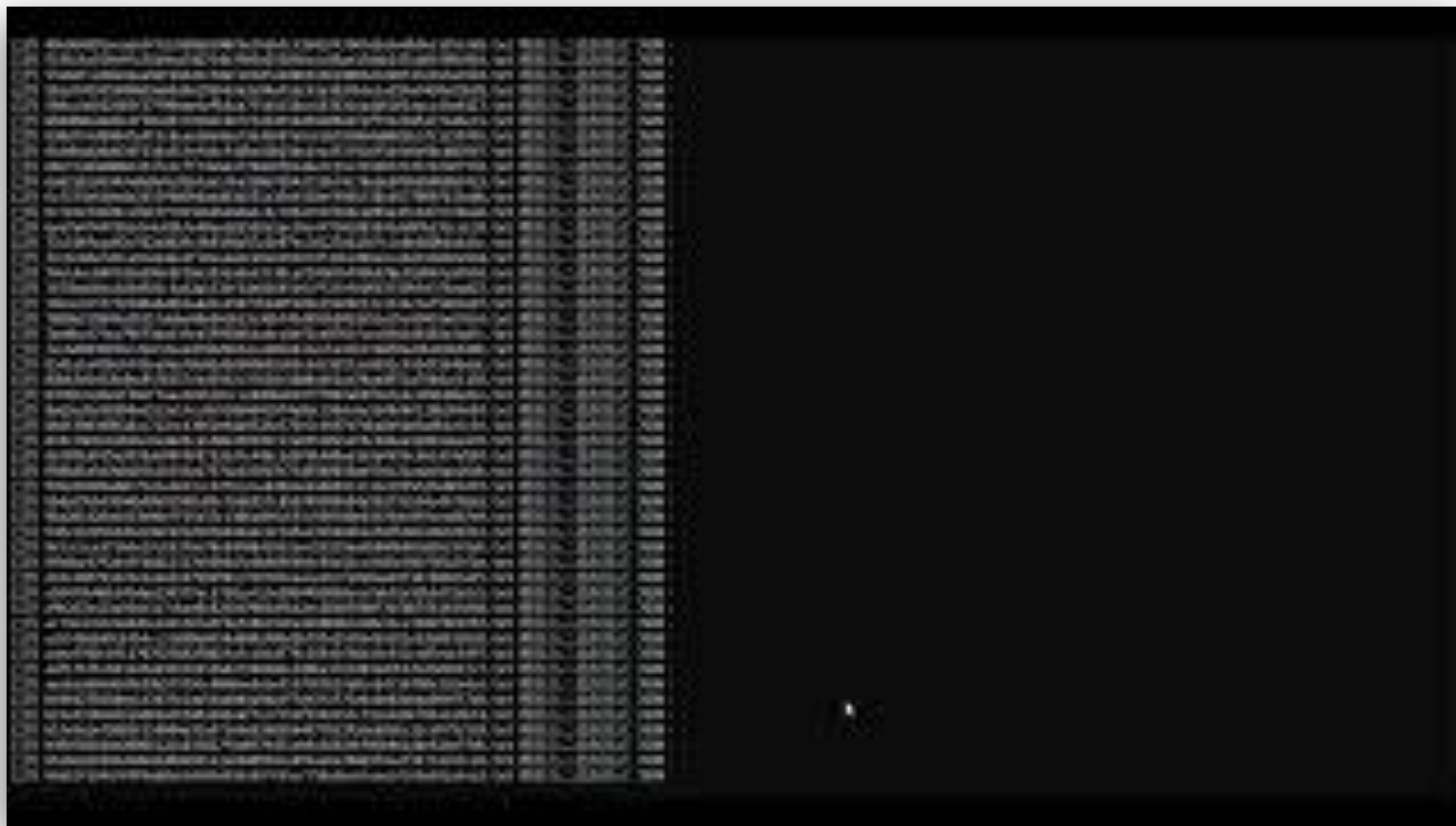
最終採納的方式



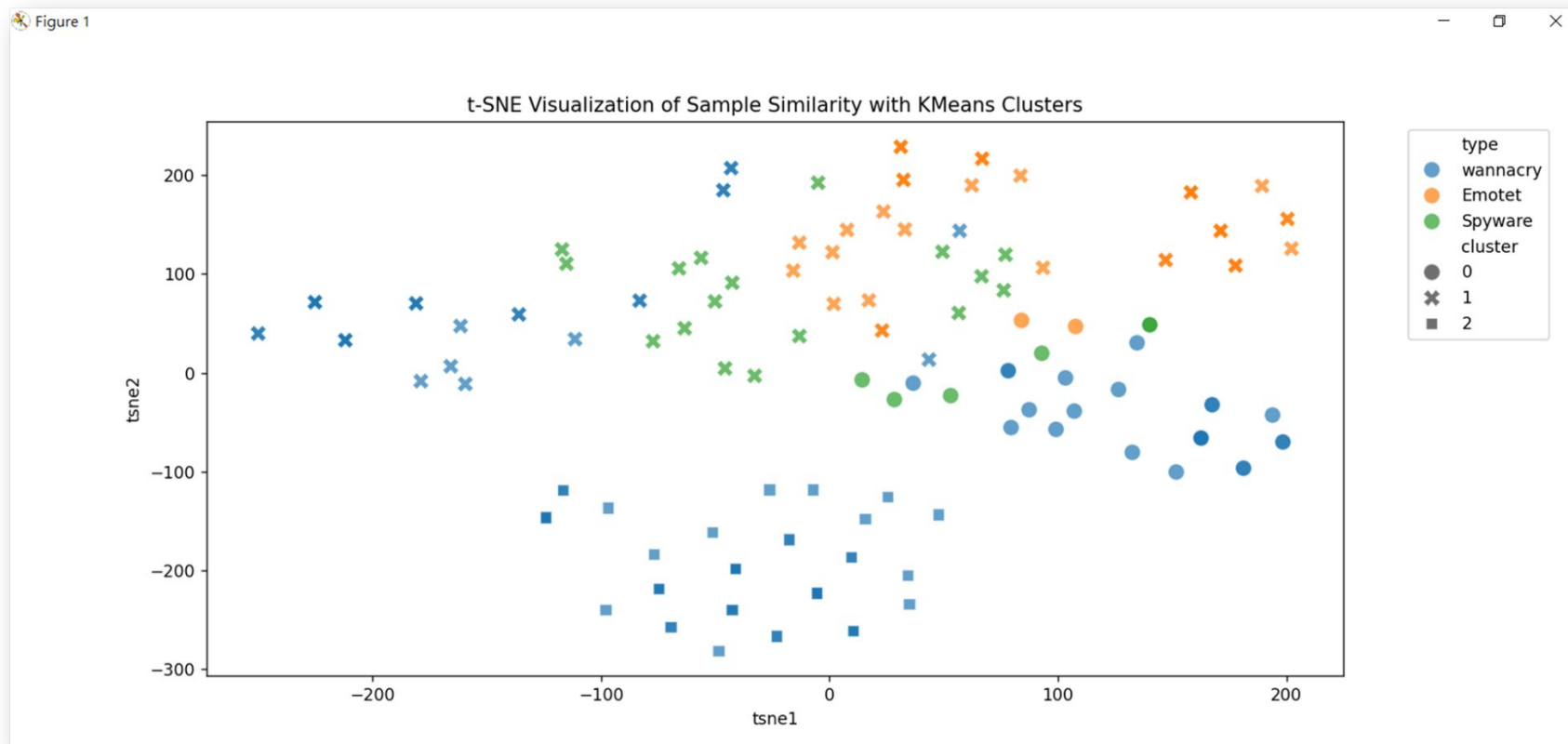
1. 改變資料前處理方式。
2. 使用UMAP將相似度矩陣降維。
3. 繪製UMAP聚類圖，並標註特定樣本。

```
{  
  "attack_ids": [  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    0,  
    1,  
    0
```

提取和分析

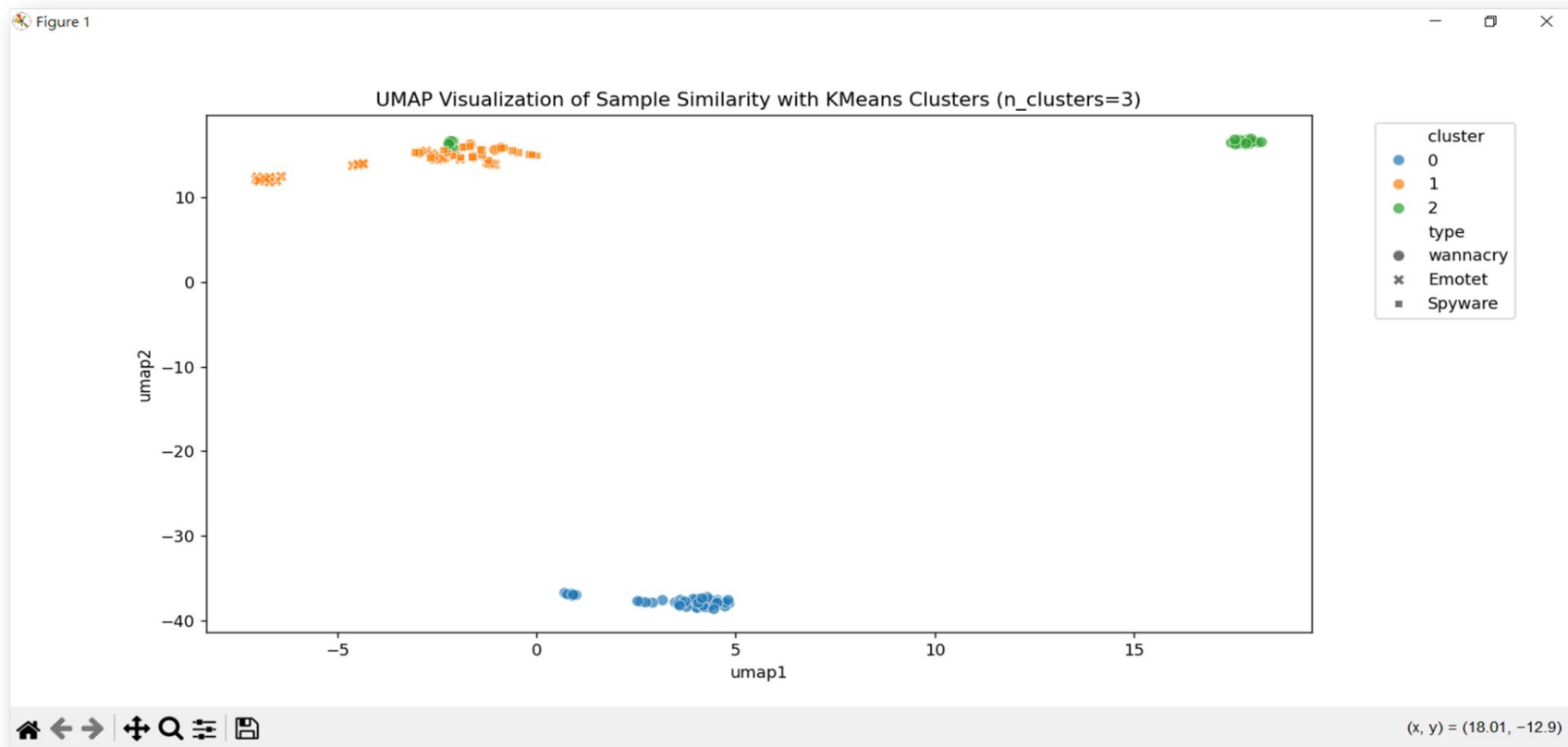


K-means繪圖



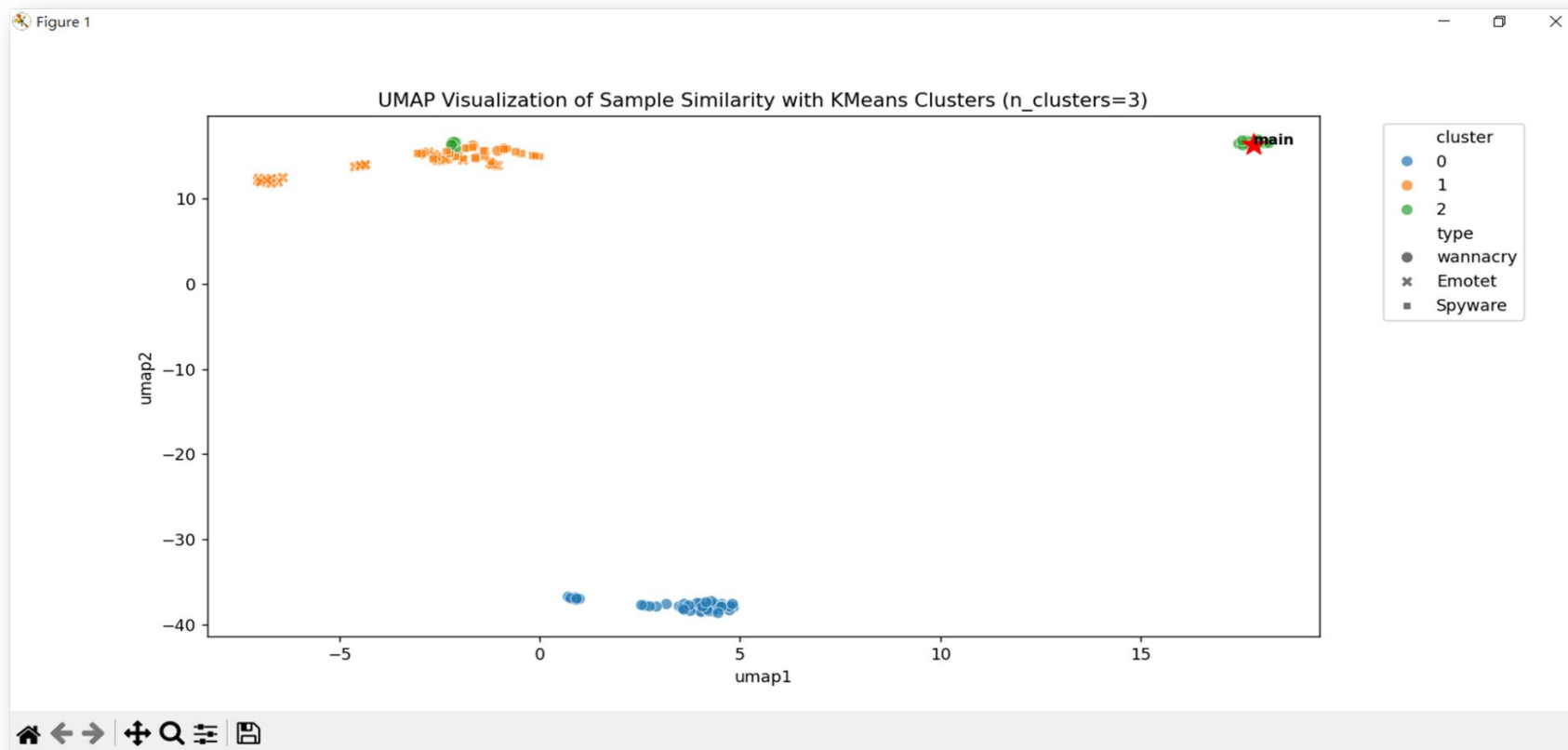
使用t-SNE降維畫出的圖

K-means繪圖



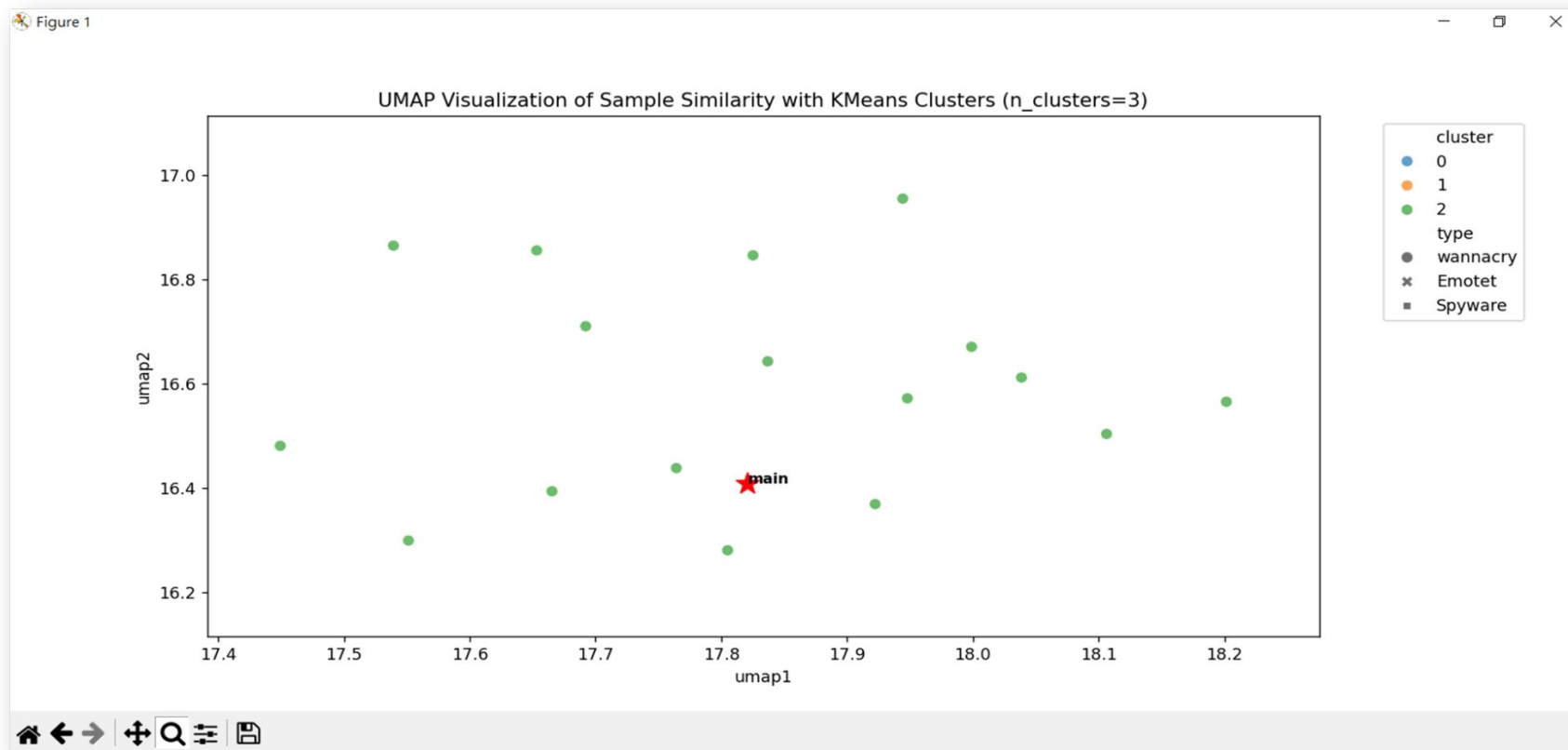
使用UMAP降維畫出的
圖

標記



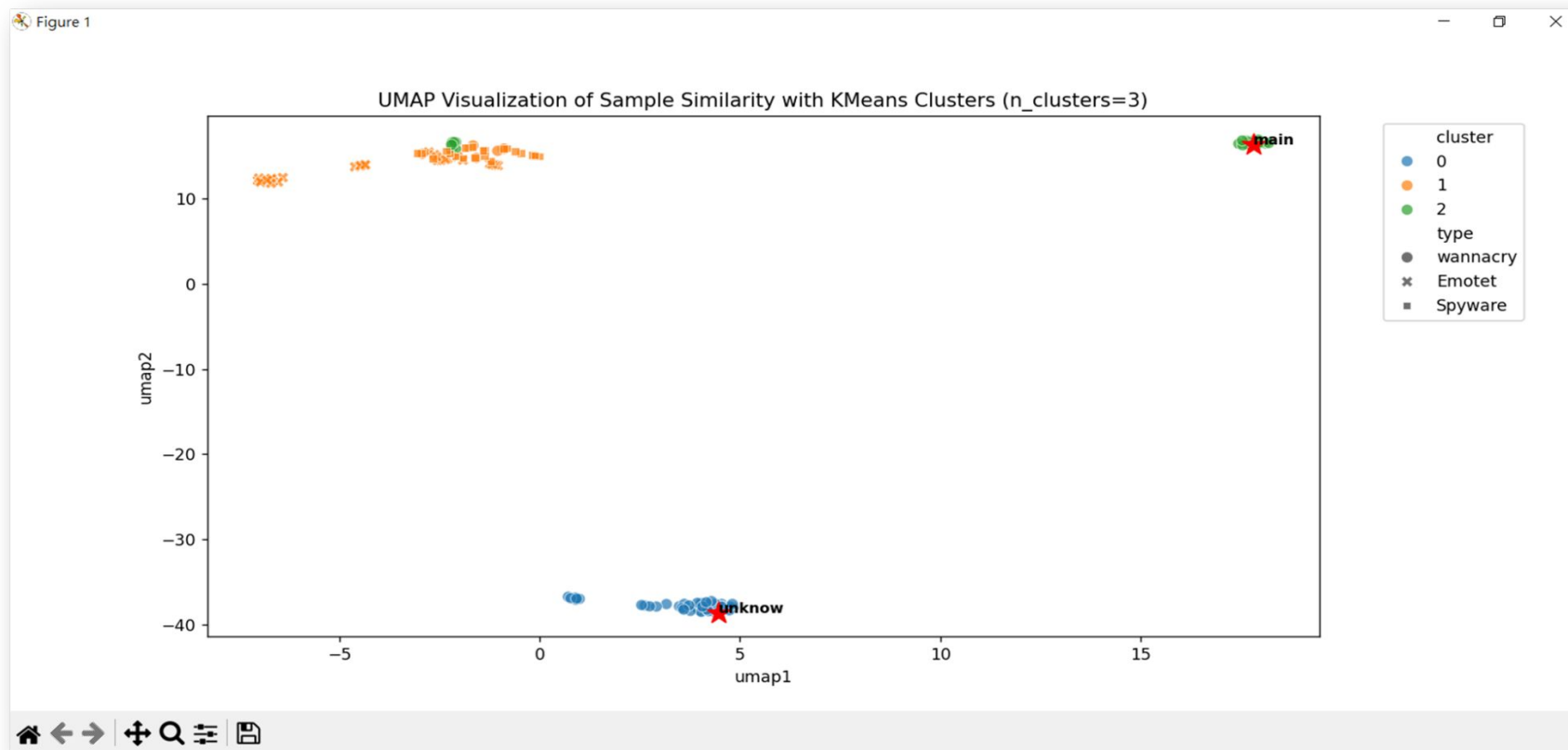
標記Wannacry原珠位置

標記



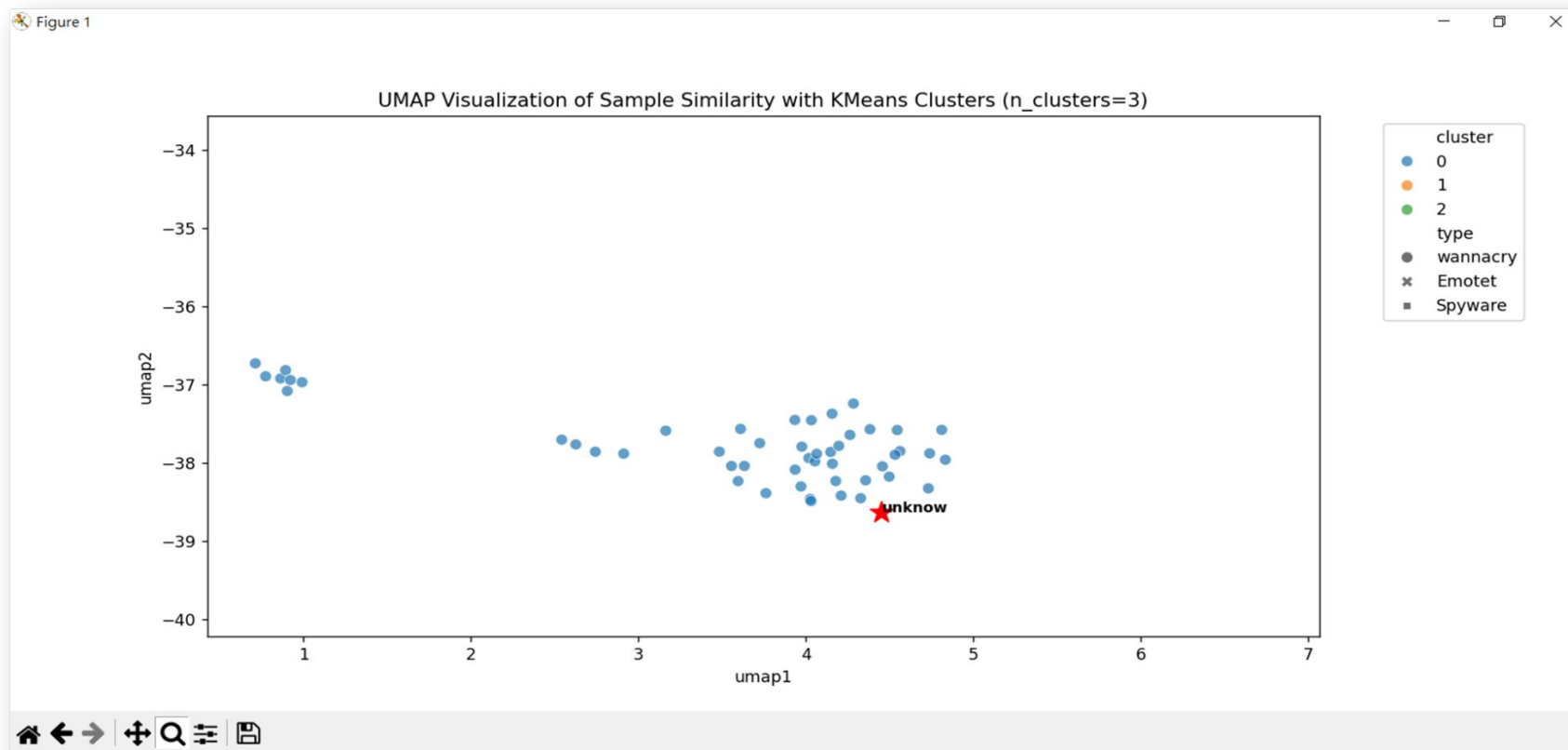
標記Wannacry原珠位置

成果展示



測試放入一筆未知種類的檔案

成果展示



測試放入一筆未知種類的檔案

準嗎？



File analysis interface for file `b15c431868dfe6d02a565442e46aaf7ecf75df3341e55cf1ce42dc761aa19bf2`. The interface shows a community score of 70/74, indicating it is flagged as malicious by 70/74 security vendors. The file is identified as `peexe` and has a size of 3.35 MB. The last analysis date is "a moment ago".

Popular threat label: `ransomware.wannacry/wannacryptor`. Threat categories: `ransomware`, `trojan`. Family labels: `wannacry`, `wannacryptor`, `wanna`.

Security vendors' analysis:

Vendor	Detection	Vendor	Detection
Acronis (Static ML)	Suspicious	AhnLab-V3	Trojan/Win32.WannaCryptor.R200571
Alibaba	Ransom:Win32/WannaCrypt.3a0	AliCloud	RansomWare:Win/Wannacryptor
ALYac	Trojan.Ransom.WannaCryptor	Antiy-AVL	Trojan(Ransom)/Win32.WannaCryptor
Arcabit	Trojan.Ransom.WannaCryptor.A	Avast	Win32:WanaCry-A [Trj]
AVG	Win32:WanaCry-A [Trj]	Avira (no cloud)	TR/Ransom.Gen
Baidu	Win32:Trojan.WannaCry.c	BitDefender	Trojan.Ransom.WannaCryptor.A
BitDefenderTheta	AI-Packer.D67BB80D1F	Bkav Pro	W32.WanaCryptBTTc.Worm
ClamAV	Win.Ransomware.Wannacryptor-994018...	CrowdStrike Falcon	Win/malicious_confidence_100% (W)

未知檔案

File analysis interface for file `ed01ebfbc9eb5bbea545af4d01bf5f1071661840480439c6e5babe8e080e41aa`. The interface shows a community score of 70/74, indicating it is flagged as malicious by 70/74 security vendors. The file is identified as `diskpart.exe` and has a size of 3.35 MB. The last analysis date is "1 hour ago".

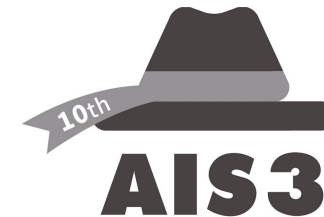
Popular threat label: `ransomware.wannacry/wannacryptor`. Threat categories: `ransomware`, `trojan`. Family labels: `wannacry`, `wannacryptor`, `wanna`.

Security vendors' analysis:

Vendor	Detection	Vendor	Detection
AhnLab-V3	Trojan/Win32.WannaCryptor.R200571	Alibaba	Ransom:Win32/WannaCry.ali1020010
AliCloud	RansomWare	ALYac	Trojan.Ransom.WannaCryptor
Antiy-AVL	Trojan(Ransom)/Win32.Scatter	Arcabit	Trojan.Ransom.WannaCryptor.A
Avast	Win32:WanaCry-A [Trj]	AVG	Win32:WanaCry-A [Trj]
Avira (no cloud)	TR/Ransom.JB	Baidu	Win32:Trojan.WannaCry.c
BitDefender	Trojan.Ransom.WannaCryptor.A	BitDefenderTheta	Gen:NN.Zexaf.36810.wt0@aGEmS3di
Bkav Pro	W32.WanaCryptBTTc.Worm	ClamAV	Win.Ransomware.Wannacryptor-994018...
CrowdStrike Falcon	Win/malicious_confidence_100% (W)	Cybereason	Malicious.5a5d21

原珠

結論



透過Capa以及Jaccard
自動分析報告、提取特徵並計算相似度，
有效協助研究人員快速識別和分類
未知病毒近似於何種現有樣本，
以此來加快、簡化惡意軟體分析速度。

參考資料



- CAPA - [[連結](#)]
- MITRE ATT&CK® - [[連結](#)]
- MBC - [[連結](#)]
- MalwareBazza - [[連結](#)]
- Jaccard Index介紹 - [[連結](#)]

THANKS !

Capa、巴卡阿卡、Jaccard、咪卡巴卡、噶(๑>๓<๑)

Capa、巴卡阿卡、Jaccard、咪卡巴卡、姆(๑>~<)و

心得分享 - 錯誤問題計畫過程

希望我們的經驗能給你們一些幫助和提示。

以下我們會分享一些

1. 做專題過程中，嘗試後卻失敗，從失敗中的經驗，所學的知識！
2. 做專題過程中有趣的部分分享！

該先做的事

1. 了解組員

- 一開始大家要互相了解一下熟悉的領域和知識有哪些
- 可以從有人的專長開始延伸，這樣成果會比較有深度

2.DC好棒 ->

3.跟助教打好關係



遇過的問題

- 想不到題目 (最可怕的事)
- 想到題目結果都不會做
- 抓惡意程式遇到的障礙
- Capa報告資訊太複雜
- Jaccard index切分詞要如何選擇
- Capa區分不了變體
- AI問題 (萊姆)
- 格式轉換問題+合併程式碼 (萊姆)

想不到題目

我們組是B組中倒數第二晚想出題目的，第二天晚上兩位助教陪著我們到十二點一點。我們前前後後提了二十多種吧。

會遇到的問題(被打槍的種種原因):

- 有人做過一樣的功能 / 專案
- 用Rule-base就可以了
- 太難了 - 助教覺得我們做不出來 / 我們自己覺得做不出來
- 要做惡意軟體相關, 但沒有人會逆向
- 不想做AI

如果大家跟這領域不熟真的可以直接請助教幫忙生題目

抓惡意程式遇到的障礙

手動下在速度太慢，還要一筆一筆拿去分析，乾脆寫個程式自動下載後自動使用 Capa 分析。

但是!!

我發現程式也是一筆一筆下載再一筆一筆分析，這樣不知道要跑到民國幾年。

最後使用異步處理、多執行續，加快整個下載分析的過程。

Capa報告資訊太複雜

從Capa所提取出的報告中有很多資訊，一開始認真看了很久 MBC 跟 MITRE ATT&CK框架，想來想去感覺每個類別都有它的重要性 很難去做去捨

所以最後就直接全部採納了 

Jaccard index切分詞要如何選擇

一開始用文字且以每個空格作為切分詞來做比較，發現出來了效果不好，分類出來的差距不明顯，但當我們提起出標籤並且改用豆好做為切分詞後效果明顯的升了很多

AI - Dataset - 格式轉換問題+合併程式碼

在要將資料喂進去前要先將資料做好前處理，我們在進行資料轉換的過程中，最終是零零散散的寫了五支程式在跑（非常難過）。這樣會導致許多問題，像是每次都要改一堆引入文件目錄，還有每個環節都可能出錯卡住。總之後來花了一個多小時，把他們都整合再一起了。

↓ 生出capa報告.txt

用array[0or1]表示有無之所有Tag

把兩個array併成一個

a.轉成array[0orTag] .txt

b.轉成array[Tag] .txt

↓ 將所有.txt合併成一份json

__genCapa.py(?)

__trans.py(?)

__check_format.py

__mergeAll.py

__extract.py

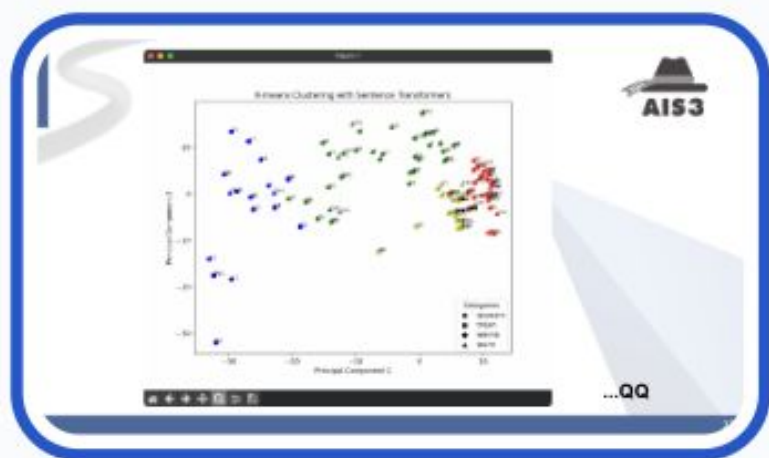
__toJSON.py

capa.rar ->一堆capa.txt

best.py ->一個json

AI - 跑出來不是我們想要的

37



就是一個難過的故事。

助教說我們要去調可能要搞一個禮拜，於是就是直接放棄方法二，直接轉成用方法三了。

K-means繪圖

一開始用文字且以每個空格作為切分詞來做比較，發現出來了效果不好，分類出來的差距不明顯，但當我們提起出標籤並且改用豆好做為切分詞後效果明顯的升了很多

有趣的部分 （這一part是不是有點諷刺）

- 前面再跑 capa腳本分析 148個變體時，直接跑了兩節課剛好跑完
- 跟隊友玩
- 一群人塞在一個地方一起爆肝到凌晨連續五天
- 玩了兩次大老二 XD
- 第四天通宵學了一下惡意軟體的編譯 結果最後沒用到 🤡 但其實也沒學會就是了 🤡🤡🤡

Temp

TODO(萊姆)

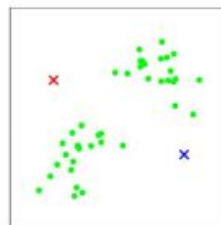
- 報一遍
- 流程圖

流程圖

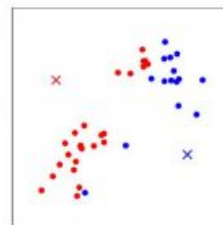
用sample了解我們做了什麼事、一張圖 一張文字
人不要一直換、格式要統一、



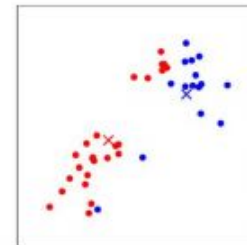
(a)



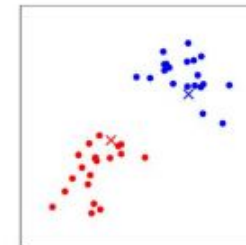
(b)



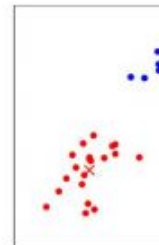
(c)



(d)



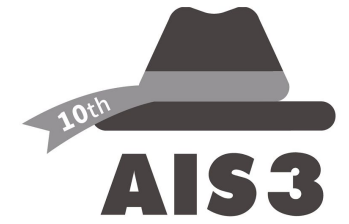
(e)



(f)



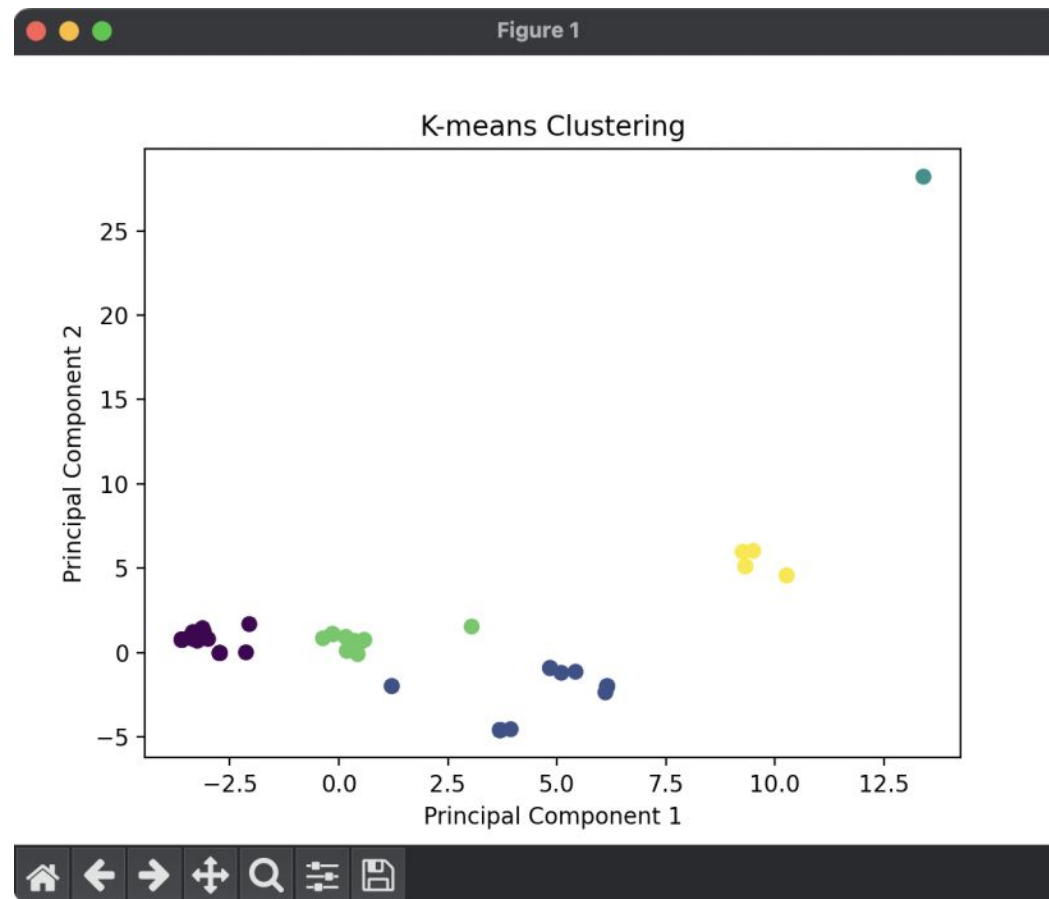
展望？



K-means 分群



1. 分WannaCry變體



[illegible][illegible]

目錄



動機
目標
整體架構

方法1

Jaccard index介紹
Capa介紹
成果1

方法2

K-means

方法3

成果展示

總結