Twitter 資料分析

台灣科技大學社群媒體分析實務第一次作業

電機工程系｜M10407323｜王靖煊

第一大題

Docker 安裝

使用 macOS 10.12 Sierra 系統

- 下載 Docker 安裝檔
- 下載 Docker 配置檔
- 將 docker-compose.yml 檔案中的 logstash 區段用 # 註解起來。
- 使用終端機進入配置檔資料夾，執行 docker-compose up。
- 打開瀏覽器，輸入 127.0.0.1:8888，即可打開 ipython notebook。

- 在 shell 執行 docker ps，找到 jupyter 這個 Container 的代碼 (a87be8def877)。
- 在 shell 執行 docker exec –i –t a87be8def877 /bin/bash。
- 這樣就進入該 Container 了。

- 當打開 python notebooks 時，他的家目錄即為該 container 的 /opt/notebooks。
- 在 shell 下 cd /opt/notebooks，就會進入該目錄，可以 git 文件到這兒，並用 ipython notebook 操作。
- 在該 Container 中（ipython notebook 的 Terminal 中） /opt/notebooks 下 git clone <git 網址>

Elasticsearch 安裝

使用 macOS 10.12 Sierra 系統

- 下載 elasticsearch-2.4.1
- 修改設定檔，將 /config/elasticsearch.yml 檔案中的 cluster.name: MarkES 取消註解
- 執行 elasticsearch（無顯示副檔名的那個 elasticsearch.bat）
- 瀏覽器輸入 localhost:9200 測試是否成功。

然後就失敗了，去網路找到以下解法

- 從 https://github.com/mobz/elasticsearch-head 下載 ZIP 包。
- 在 elasticsearch 目錄下創建目錄 /plugins/head/
- 將剛剛解壓縮的 elasticsearch-head-master 目錄下所有內容複製到創建的目錄下即可。
- 瀏覽器輸入 localhost:9200/_plugin/head 測試是否成功。
- 最後在 Terminal 中執行 pip install pyes==0.99.4。

Logstash 安裝

使用 macOS 10.12 Sierra 系統

- 下載 logstash-2.3.4
- 解壓縮，在 bin 裡新建一個 twitter.config，輸入程式：

```
input{
    twitter{
        consumer_key => "d2SPjNwUuwuOXaB6Qt0ensRoP"
        consumer_secret => "ZXRWyFcPBDDLM9m8QUAWN6OjDK0iLSqmsUuwqpKcF4pUTjzNzi"
        oauth_token => "783260446975332352-3OGj2Oyn4hjTKGUcr09h9Ak0yyqCnhJ"
        oauth_token_secret => "TizqspclRoGpS4iuXhfayhbtXMJmV8ulPSaSEXld4nEIo"
        keywords => ["PPAP”]     #(要搜的關鍵字在這裡)
        languages => ["en"]
    }
}
output{
    elasticsearch{
        index => "twitter"
    }
}
```

- 開終端機，cd 到 bin 目錄下，執行 ./logstash agent –f twitter.config。
- http://localhost:9200/_plugin/head/ 就會看到抓取的資料。

第二大題

Q2(a), Q2(b)

所有 twitter 數量

In [1]:
```
import pyes
es_address='127.0.0.1:9200'

conn = pyes.es.ES(es_address)

q = pyes.query.MatchAllQuery()

result = conn.search(query=q , indices='twitter2' , doc_types='user')
len(result)
```

Out[1]: 113

所有 tweets 數量

In [2]:
```
import pyes
es_address='127.0.0.1:9200'
conn = pyes.es.ES(es_address)

q = pyes.query.MatchAllQuery()

result = conn.search(query=q , indices='twitter2' , doc_types='tweet')
len(result)
```

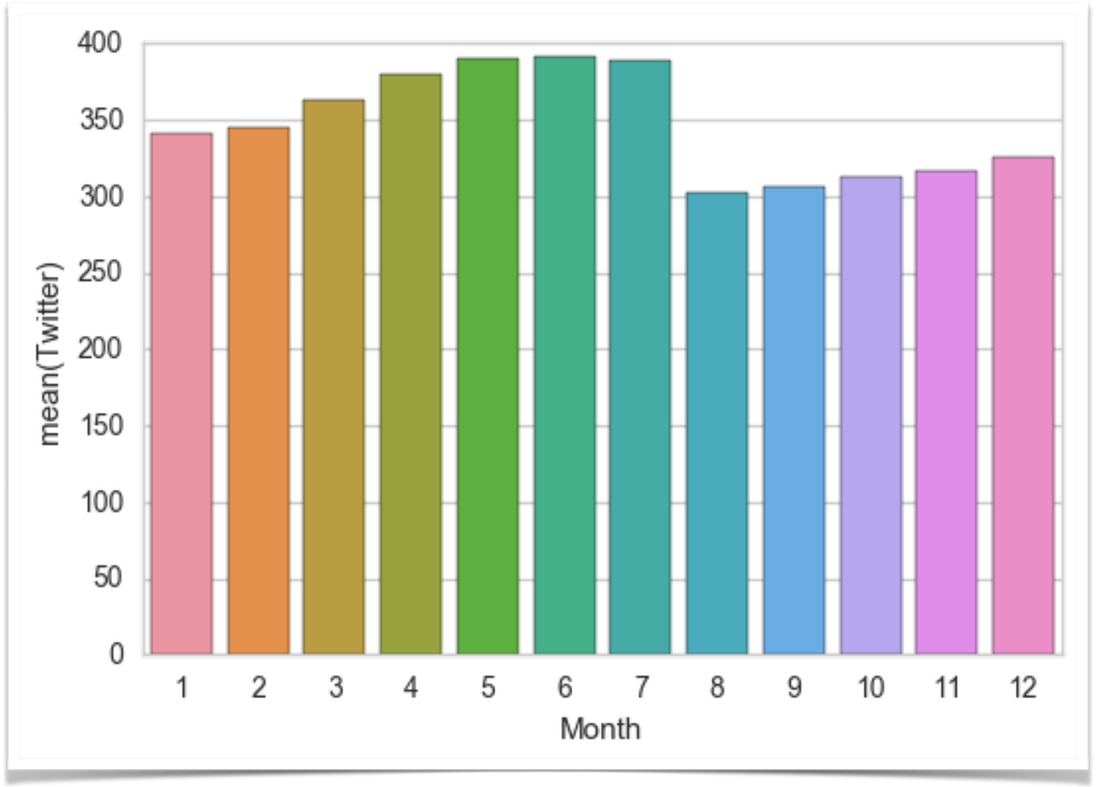
Out[2]: 392792

Q2(a), Q2(b)

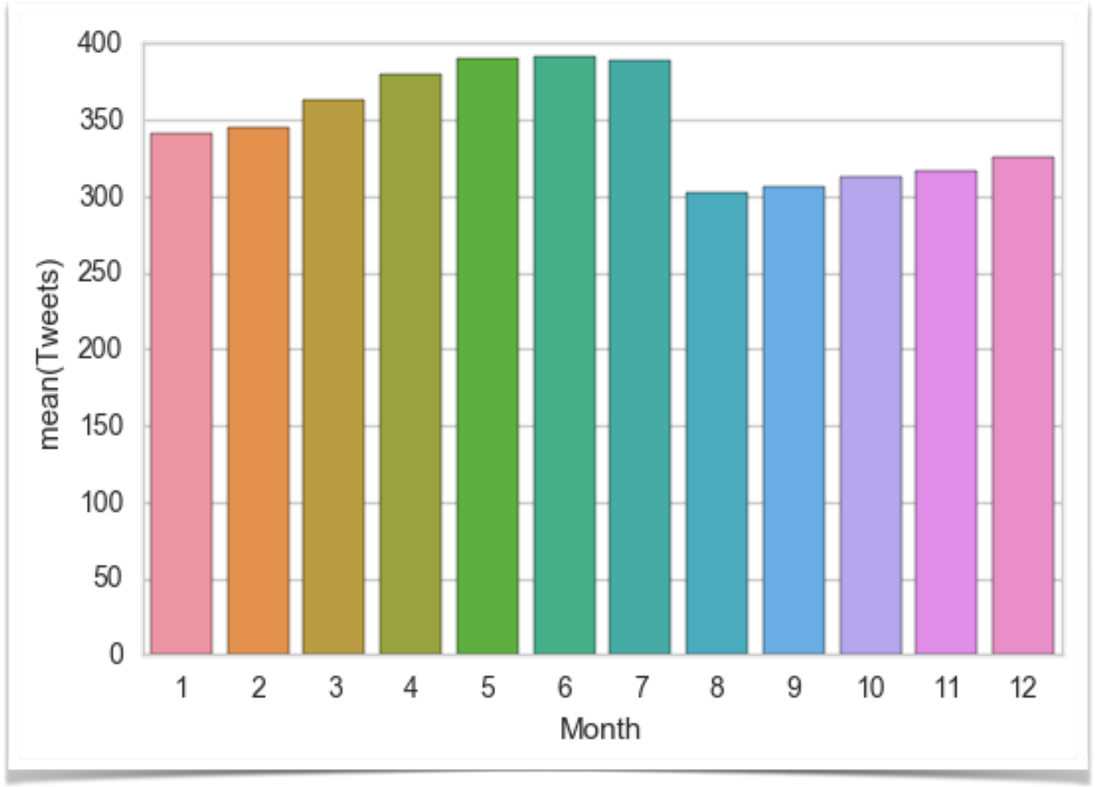
延伸

參與 CVE 關鍵字議題討論？	Tweets 數量：644 Twitter 數量：71
參與 Vulnerability 關鍵字議題討論？	Tweets 數量：2242 Twitter 數量：96
參與 Exploit 關鍵字議題討論？	Tweets 數量：1969 Twitter 數量：89

Q2(c), Q2(d)



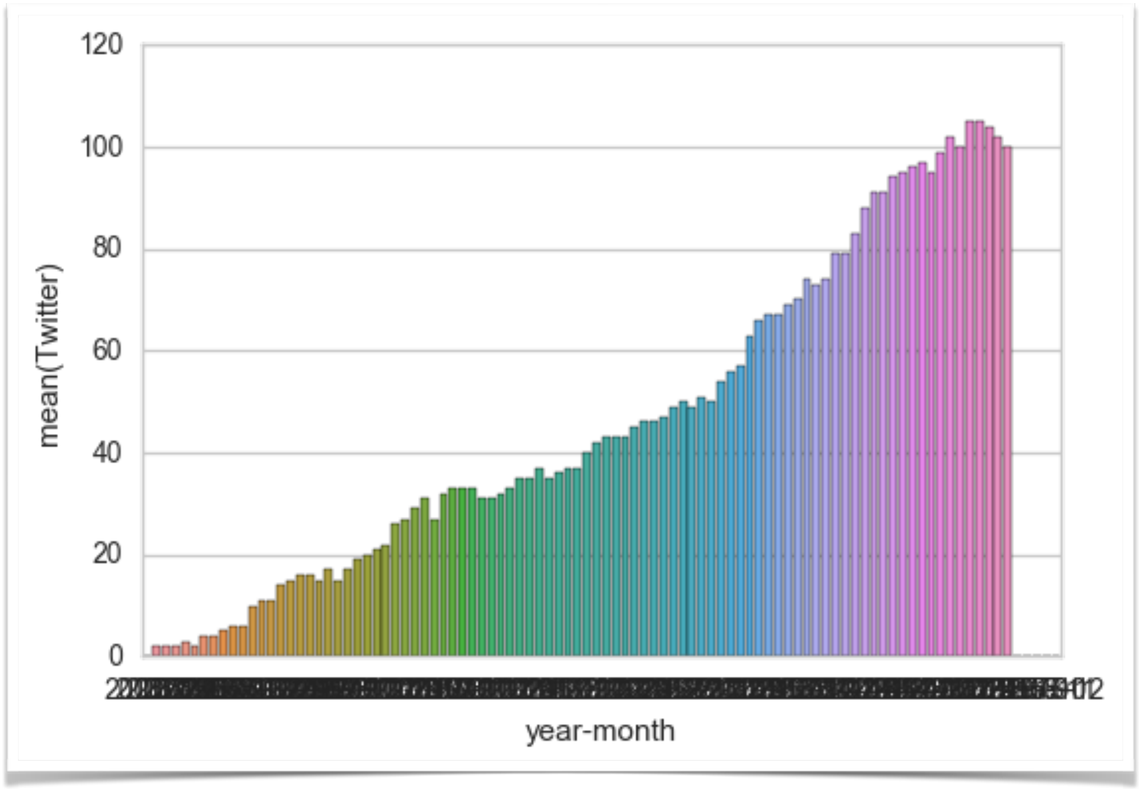
不論年份的
每月參與 Twitter 數量長條圖



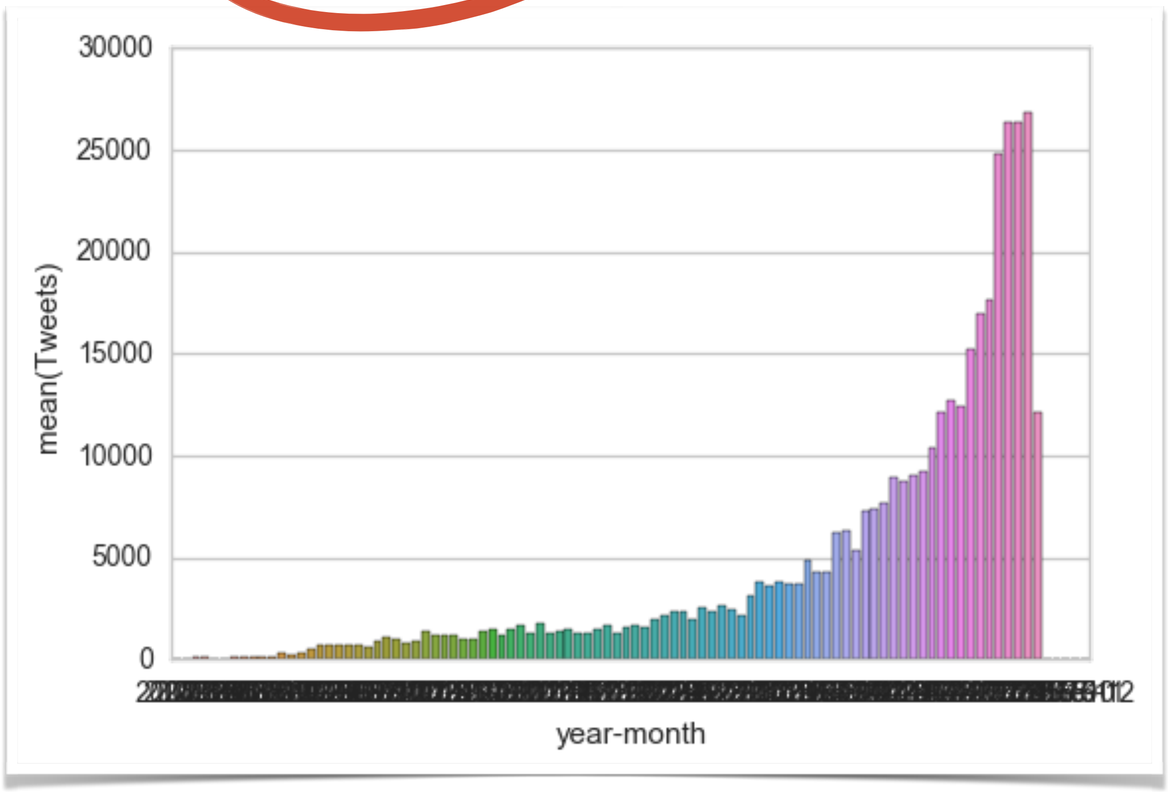
不論年份的
每月參與 Tweets 數量長條圖

Q2(c), Q2(d)

延伸

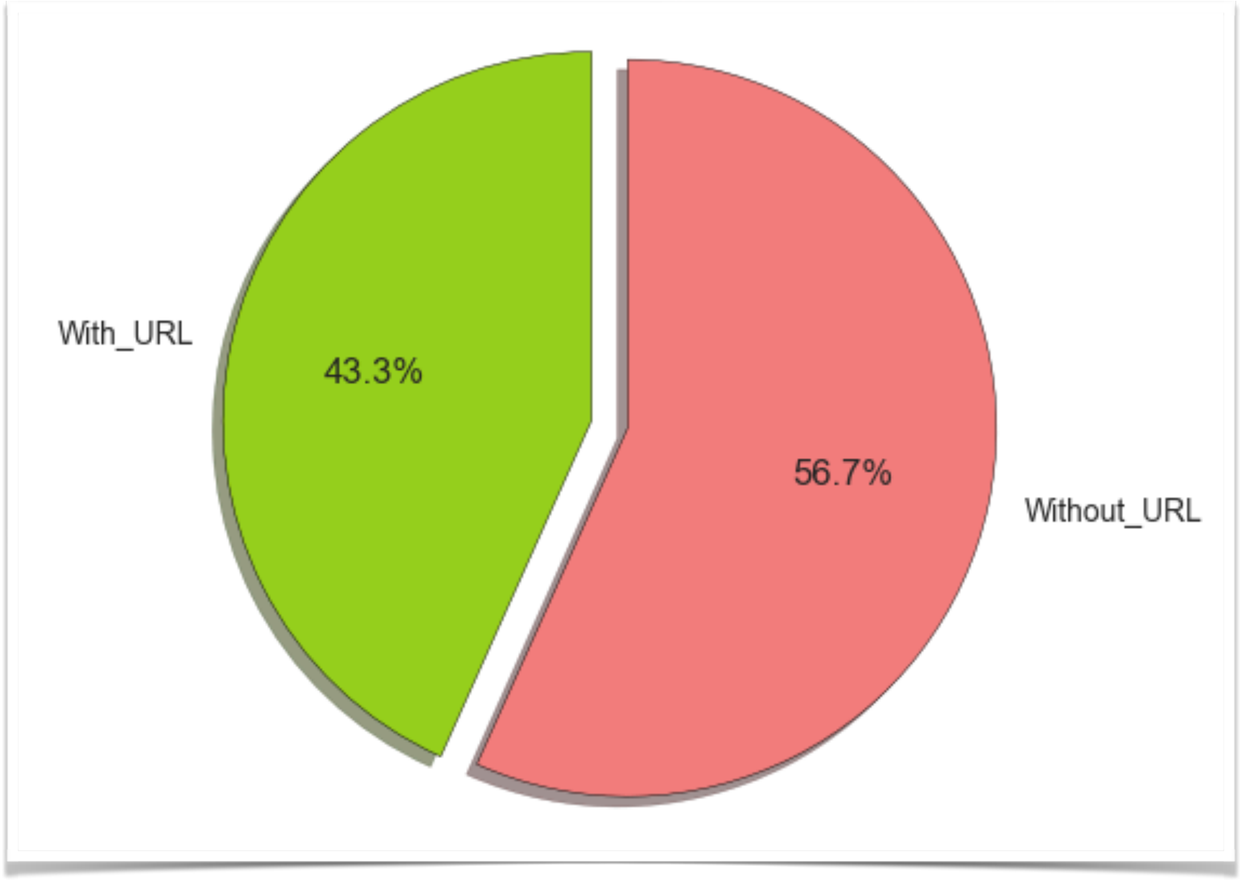


每年每月
參與的 Twitter 數量長條圖



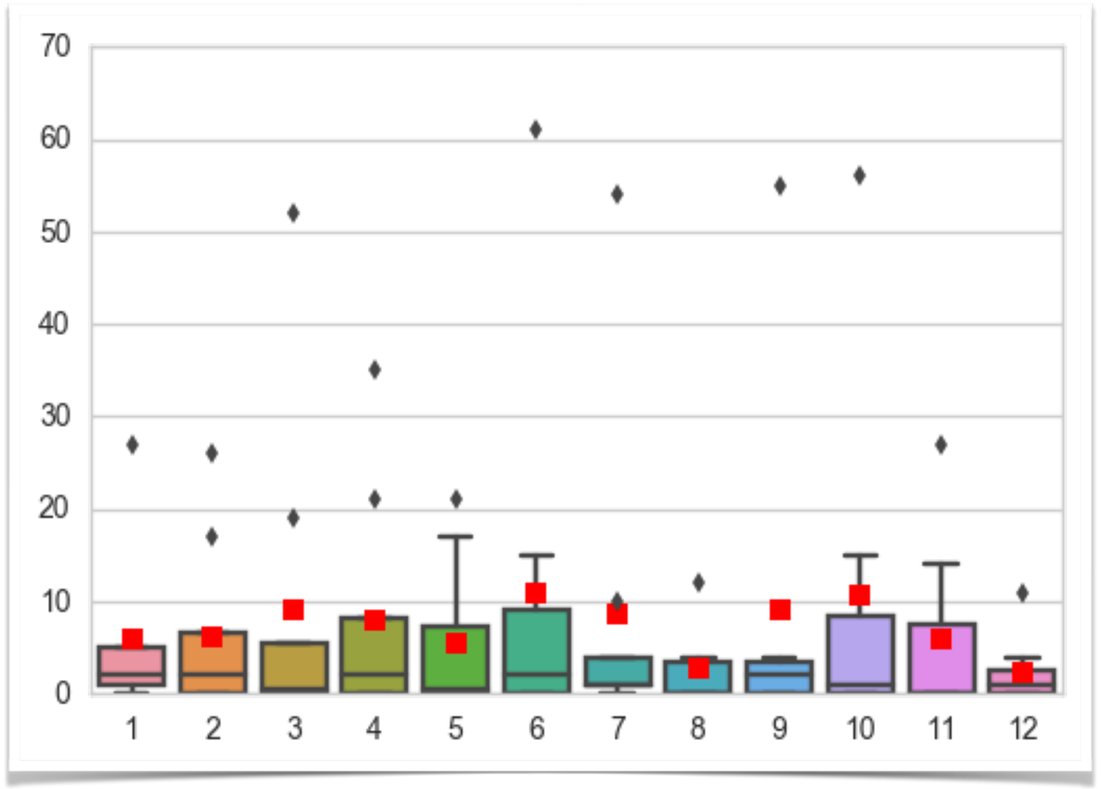
每年每月
參與的 Tweets 數量長條圖

Q2(e)

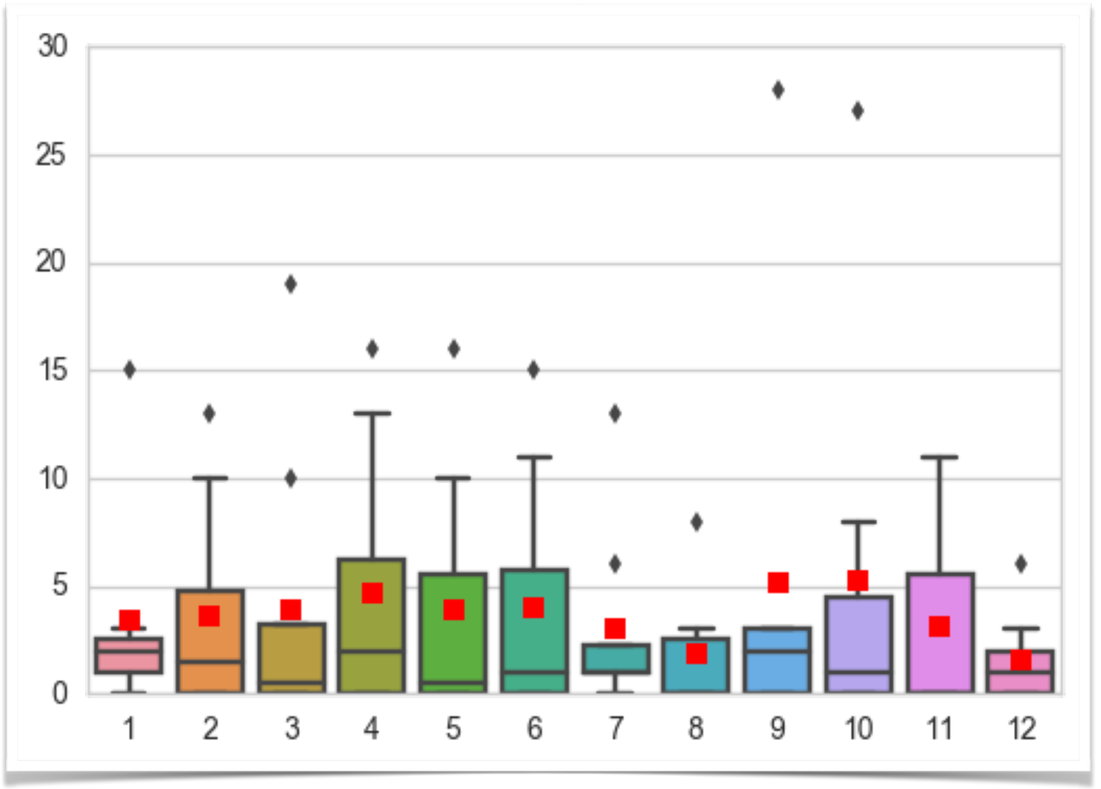


有 URL / 無 URL 的 Tweets 的比例圓餅圖

Q2(f)



各 Tweets 所提到的 CVE 箱型圖



各 Twitter 所提到的 CVE 箱型圖

Q2(g)

```
Rank 1 - [u'Kevin Mitnick'] Has 156953 followers.
Rank 2 - [u'Eugene Kaspersky'] Has 137956 followers.
Rank 3 - [u'Mikko Hypponen'] Has 103281 followers.
Rank 4 - [u'briankrebs'] Has 102361 followers.
Rank 5 - [u'Threatpost'] Has 92687 followers.
Rank 6 - [u'Schneier Blog'] Has 69928 followers.
Rank 7 - [u'McAfee Labs'] Has 59201 followers.
Rank 8 - [u'Christopher Soghoian'] Has 49406 followers
Rank 9 - [u'Jeremiah Grossman'] Has 48695 followers.
Rank 10 - [u'Charlie Miller'] Has 43646 followers.
Rank 11 - [u'The Dark Tangent'] Has 43524 followers.
Rank 12 - [u'SophosLabs'] Has 42771 followers.
Rank 13 - [u'Dan Kaminsky'] Has 41892 followers.
Rank 14 - [u'Dejan Kosutic'] Has 38406 followers.
Rank 15 - [u'Graham Cluley'] Has 37767 followers.
Rank 16 - [u'Naked Security'] Has 37502 followers.
Rank 17 - [u'Dr. Eric Cole'] Has 35334 followers.
Rank 18 - [u'Jeff Barr'] Has 33148 followers.
Rank 19 - [u'Richard Bejtlich'] Has 33056 followers.
Rank 20 - [u'Help Net Security'] Has 30583 followers.
```

以 followers_count 數量作為
專家帳號的評斷標準

第三大題

Q3-1

◀ 選出前 20 名追隨人數最多者

```
Top 20 user's id:
14666934
297856522
23566038
22790881
18789893
18476766
19206209
14669471
14181505
65845659
14924745
16730420
8917142
64677310
11791512
198365324
38956896
48443
17767238
14293266
```

```
[('RT', 12612), ('I', 9747), ('The', 4601), ('Security', 3076), ('New', 1552), ('How', 1495), ('U
S', 1391), ("I'm", 1338), ('A', 1329), ('This', 1194), ('You', 1143), ('If', 1115), ('Google', 108
2), ('We', 1060), ('ISO', 1054), ('NSA', 1039), ('What', 983), ('It', 888), ('Facebook', 836), ("I
t's", 804), ('In', 798), ('Affairs', 732), ('Internet', 640), ('My', 635), ('Just', 607), ('Appl
e', 600), ('FBI', 597), ('Microsoft', 595), ('From', 585), ('Thanks', 575), ('Cyber', 559), ('Twit
ter', 553), ('Squid', 549), ('Not', 545), ('Is', 538), ('Data', 529), ('Your', 524), ('Android', 5
22), ('No', 517), ('Amazon', 496)]
```

▲ 搜尋這 20 位使用者的 40 個最常用字詞

Q3-2

	Security	ISIS	Terrorism	ISO	Class
0	13	0	0	0	Security
1	61	1	0	0	Security
2	12	3	0	0	Security
3	23	0	0	0	Security
4	54	0	1	0	Security
5	269	3	22	0	News_follower
6	244	0	0	0	Security
7	41	12	0	0	News_follower
8	55	1	0	0	Security
9	22	0	0	0	Security
10	32	0	0	0	Security
11	168	1	0	0	Security
12	14	3	1	6	Standard_Organization
13	1555	9	2	1044	Standard_Organization
14	72	10	0	0	News_follower
15	175	1	0	0	Security
16	92	0	0	0	Security
17	44	0	0	1	Security
18	78	8	0	2	Security
19	52	0	0	1	Security

◀ 採用 Security、ISIS、Terrorism、ISO 4 個字詞
將使用者分類為

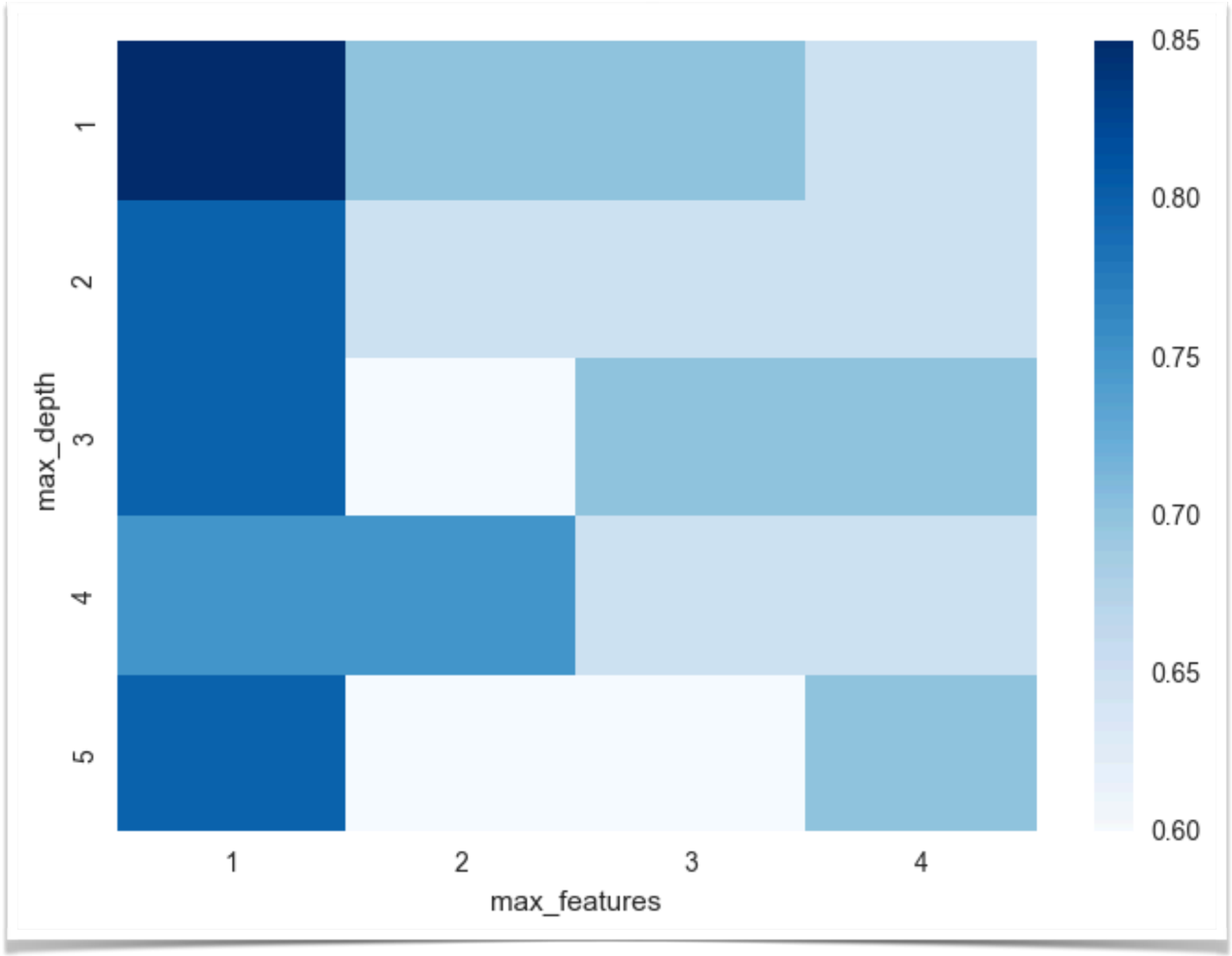
Security（資安專家）

News_follower（新聞關注者）

Standard_Organization（國際標準化組織）

3 個標籤 (Class)

Q3-3



◀ Heat Map

```
Best score: 0.85
Best parameters: {'max_features': 1, 'max_depth': 1}
```

◀ Decision Tree