Joey

2017.4

Introduction

這個作業主要是分析ptt mobile版的手機品牌和使用者的分析

Objective

主要有以下兩個目標

1. 用scrapy做一隻爬ptt mobile網頁的spider
2. 寫分析的程式分析其品牌keyword,熱門程度和個別使用者的愛好程度

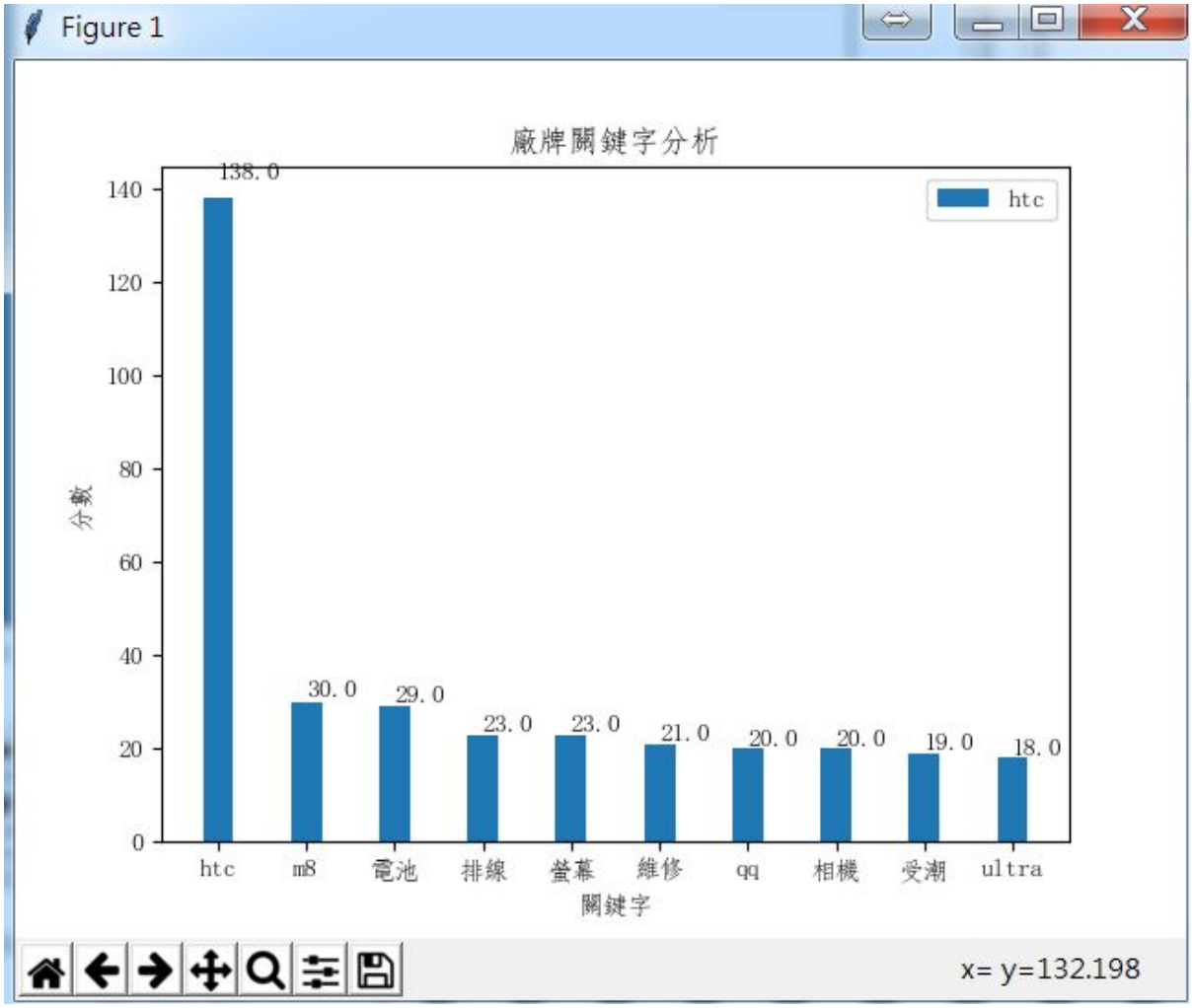
Actions

我們選用scrapy做我們爬網頁的主要工具,主程式為pttmobile.py.,最大的翻頁次數為60,用xpath定位網頁的element,取得文章標題和內容,作者,推文次數,和推文或噓文的人是那些,以下列出pttmobile的程式

```
import scrapy
from pttmobile.items import PttmobileItem


class PttMobileSpider(scrapy.Spider):
name = "pttmobile"
start_urls = [
'https://www.ptt.cc/bbs/MobileComm/index.html',
]
allowed_domains = ["ptt.cc"]
visit_pages=0
page_count=60
def parse_detail(self, response):
item = response.meta['item']
item['content']=response.xpath("//div[contains(@id,'main-content')]/text
()").extract()
positive_user=[]
negative_user=[]
for quote in response.css('div.push'):
if ('噓' in
quote.xpath("span[contains(@class,'push-tag')]/text()")[0].extract()):
username=quote.xpath("span[contains(@class,'push-userid')]/text()")[0].e
xtract()
if (username not in negative_user):
negative_user.append(username)
```

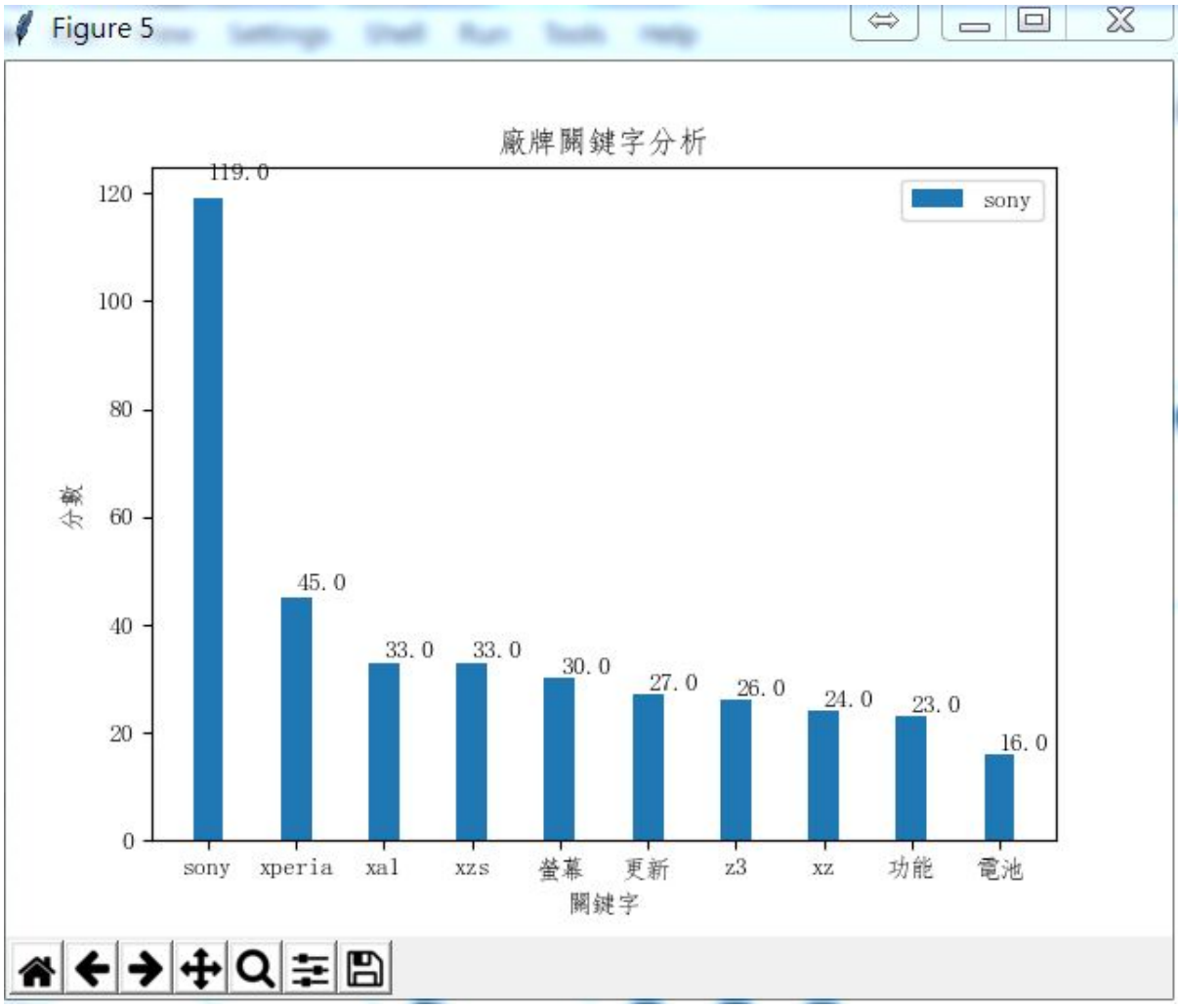
```
else:
username=quote.xpath("span[contains(@class,'push-userid')]/text()")[0].e
xtract()
if (username not in positive_user):
positive_user.append(username)
item['positive_user']=positive_user
item['negative_user']=negative_user
yield item
def parse(self, response):
if (self.visit_pages > self.page_count):
return
itemlist=[]
for quote in response.css('div.r-ent'):
title_list=list(quote.xpath("div[contains(@class,'title')]/text()"))
if (len(title_list)<2):
continue
item=PttmobileItem()
item['positive_count']=quote.xpath("div[contains(@class,'nrec')]/span[co
ntains(@class,'f2') or contains(@class,'f3')]/text()").extract_first()
if (item['positive_count']==None):
item['positive_count']='0'
item['negative_count']=quote.xpath("div[contains(@class,'nrec')]/span[co
ntains(@class,'f0')]/text()").extract_first()
if (item['negative_count']==None):
item['negative_count']='0'
item['title']=quote.xpath("div[contains(@class,'title')]/a/text()").extr
act()
item['content_url']=quote.xpath("div[contains(@class,'title')]/a/@href")
.extract_first()
item['date']=quote.xpath("div[contains(@class,'meta')]/div[contains(@cla
ss,'date')]/text()").extract_first()
item['author']=quote.xpath("div[contains(@class,'meta')]/div[contains(@c
lass,'author')]/text()").extract_first()
itemlist.append(item)
for myitem in itemlist:
next_page = response.urljoin(myitem['content_url'])
yield scrapy.Request(next_page,meta={'item': myitem}
,callback=self.parse_detail)

prev_page=response.urljoin(response.xpath("//div[contains(@class,'btn-gr
oup-paging')]/a/@href")[1].extract())
print("the prev page"+prev_page)
self.visit_pages=self.visit_pages+1
yield scrapy.Request(prev_page,callback=self.parse)
```

接著我們用jieba分詞定位文章的關鍵字並分類,主要的程式在handledata.py,因為程式過長我這邊不列出,以下的圖片是每個品牌前10名的關鍵字,我們先看看htc



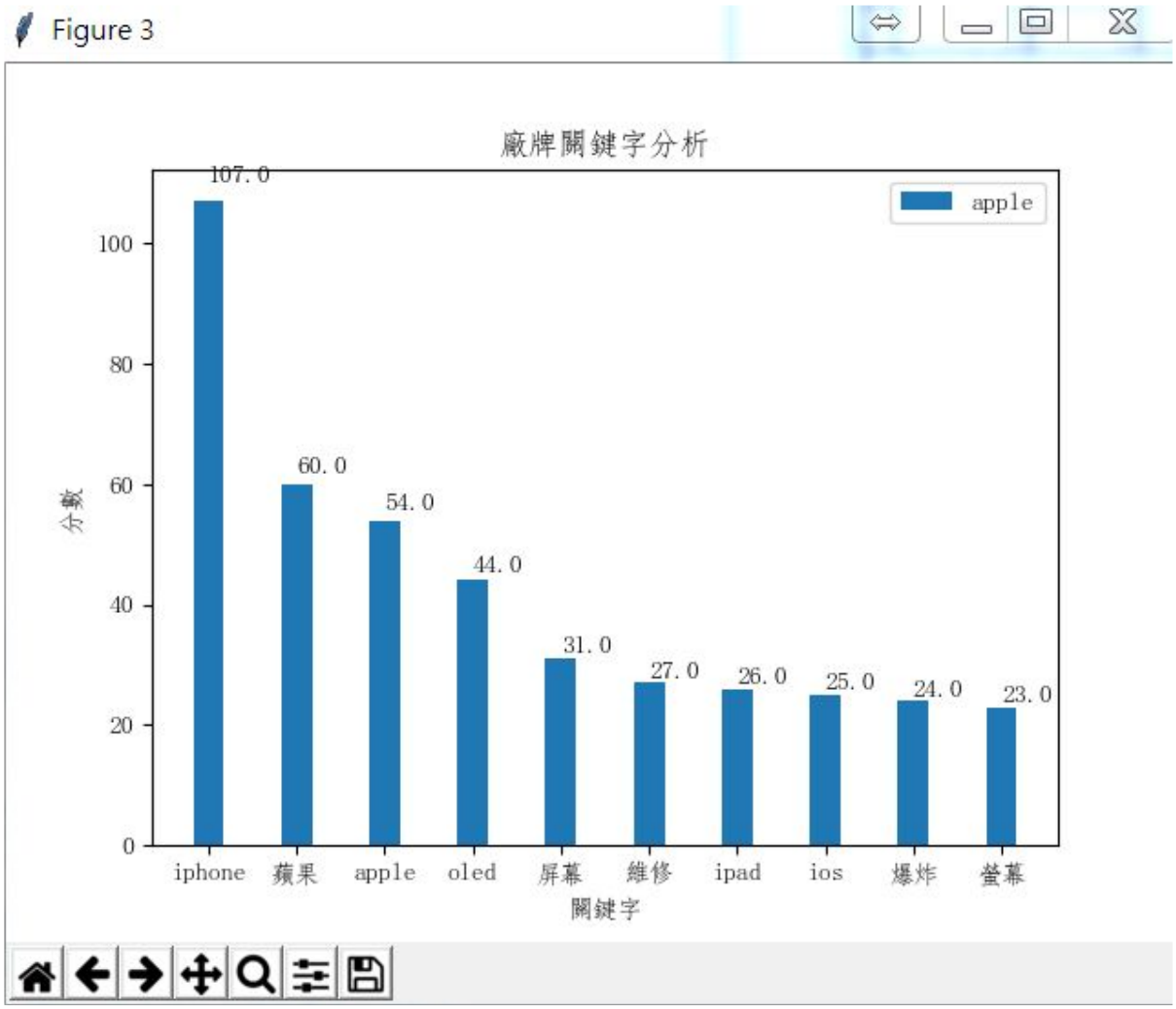
在htc的關鍵字排名裡面, htc這個詞無可厚非的排到第一名,M8這個兩三年的老機子居然排在第貳名可見經典不容乎視.有趣的是htc新機ultra的討論熱度比m8低,可見銷售狀況可能有點問題

接下來我們看看Sony



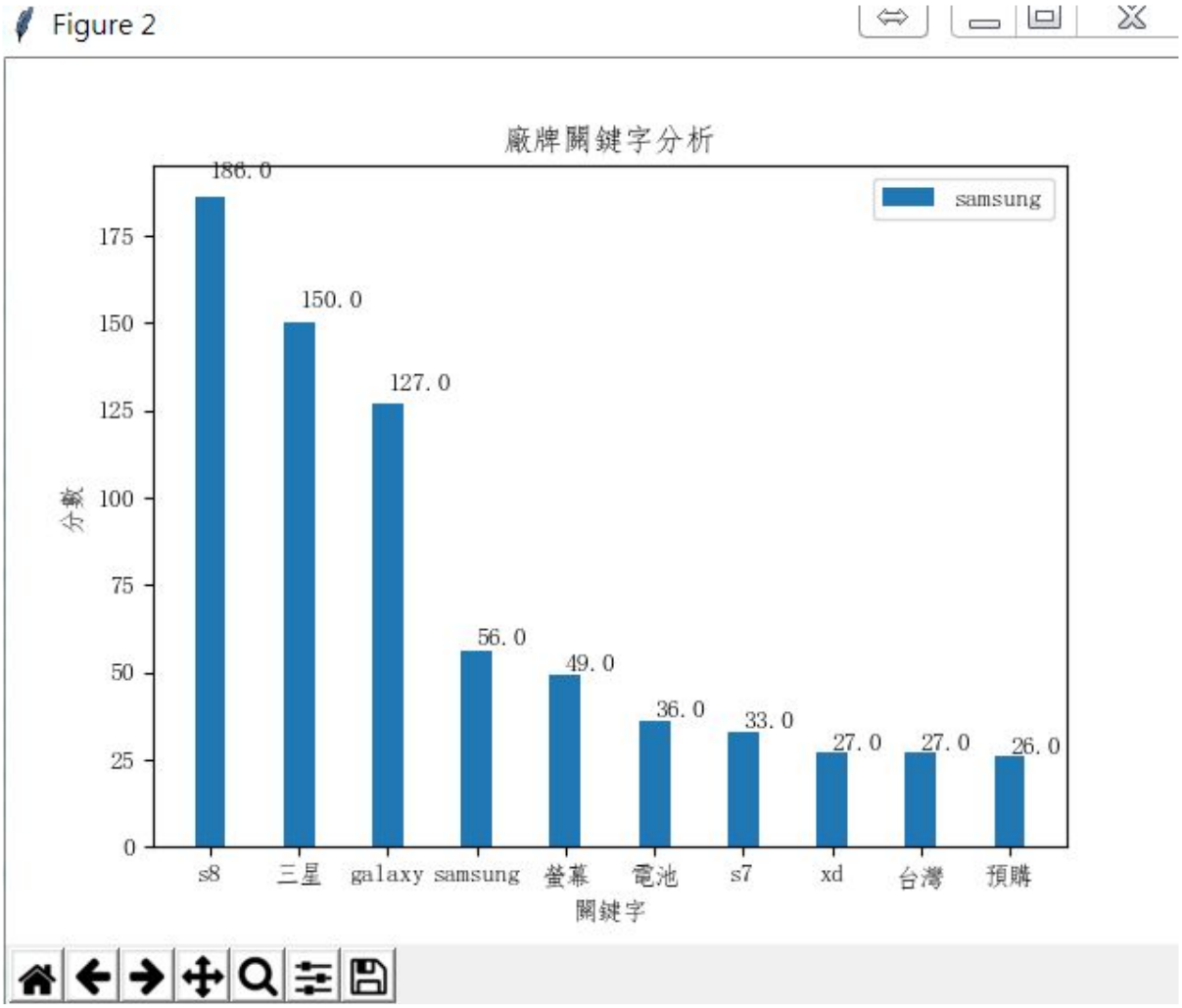
Xperia系列是今年最熱的topic,尤其是xzs, xz這幾個月剛出的新機,這邊可以看到手機的螢幕和電池不止在htc品牌被提及,在sony也是討論的重點

接下是apple



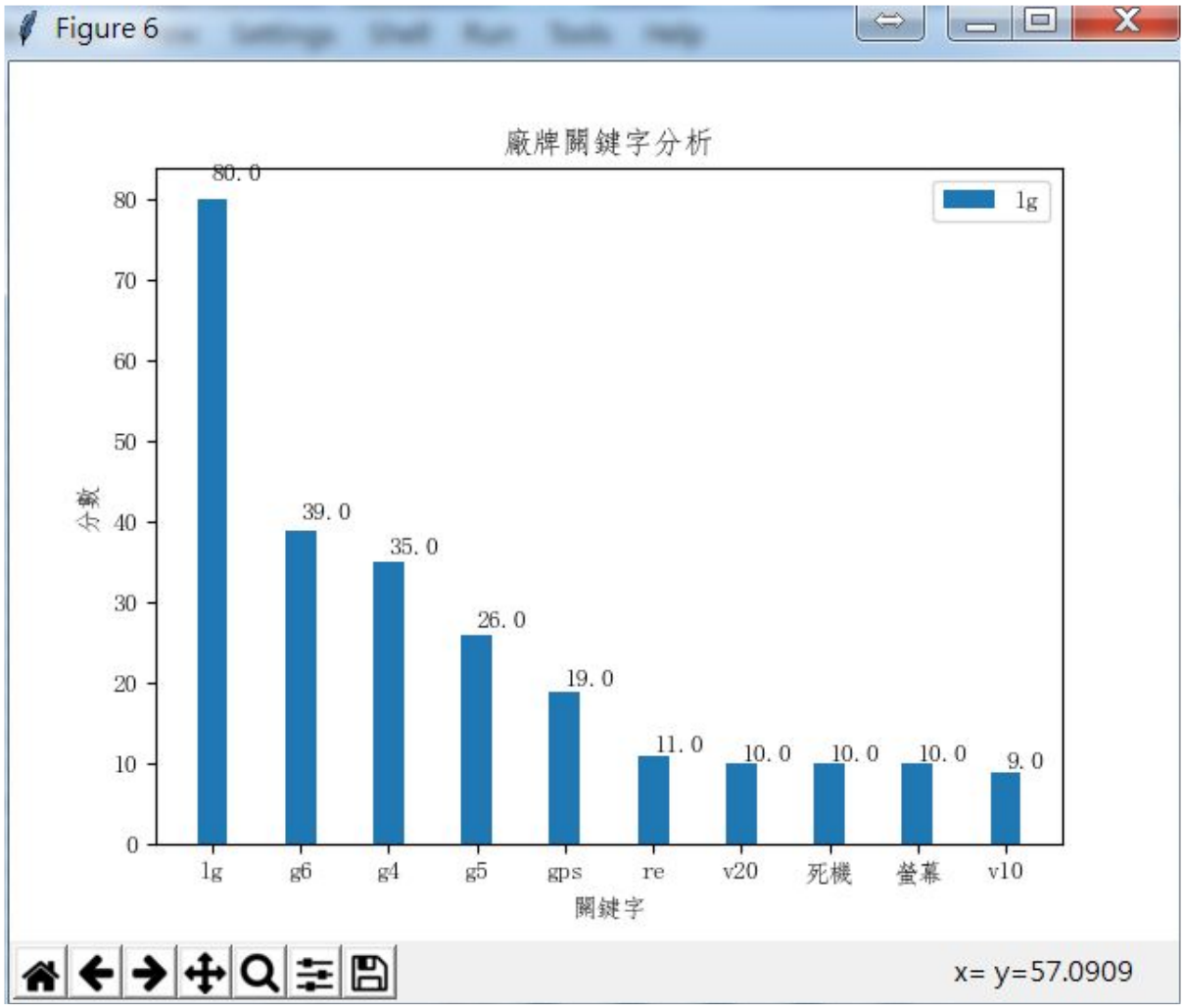
oled, ipad和奇怪的爆炸討論度很高,維修似乎也是個很熱的topic,可能要進一步檢視文章才能得知到底是手機有問題維修太多,還是維修內容的討論

然候是三星



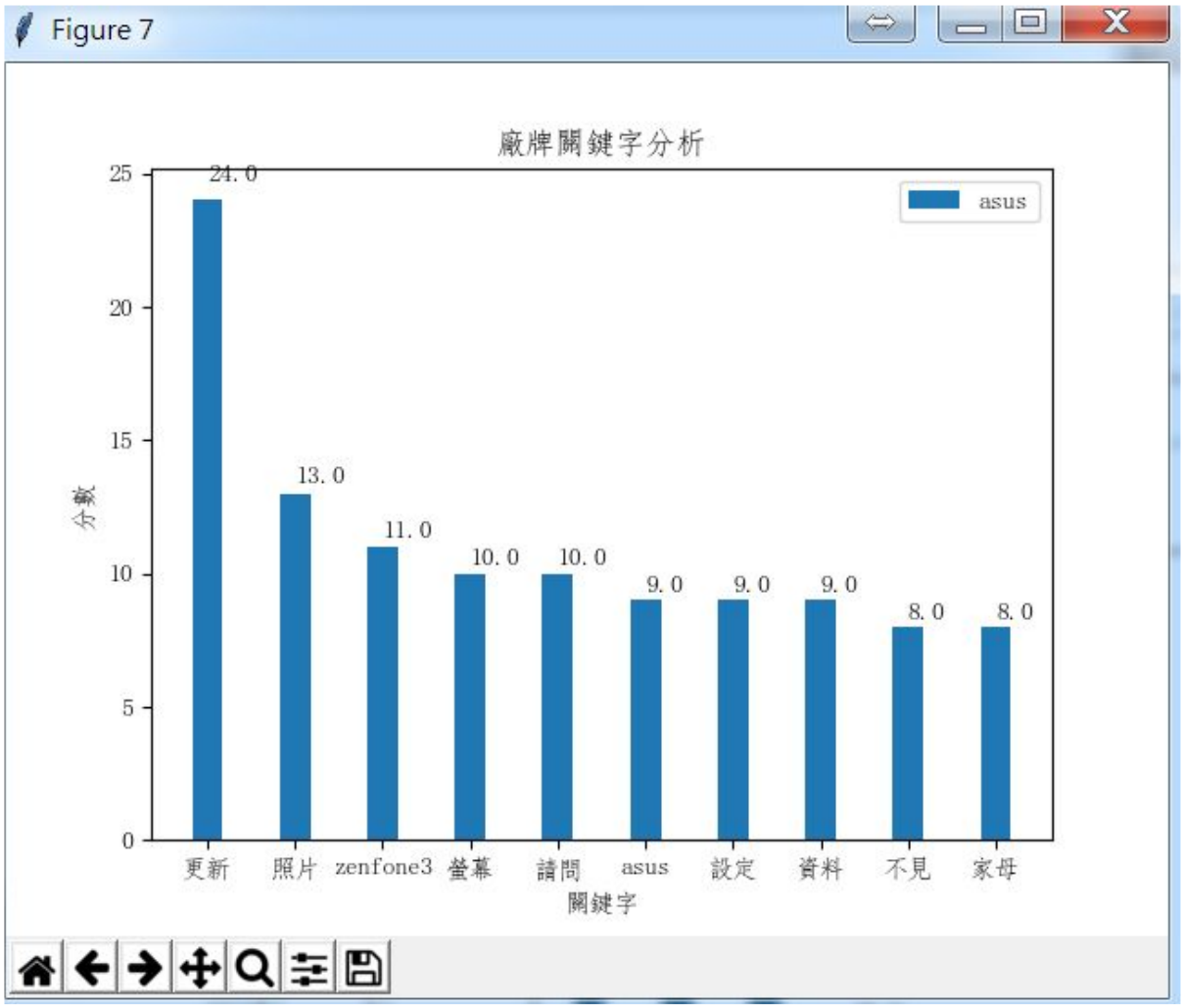
S8是機皇,討論的熱度比三星這個關鍵字還高

再下來是LG



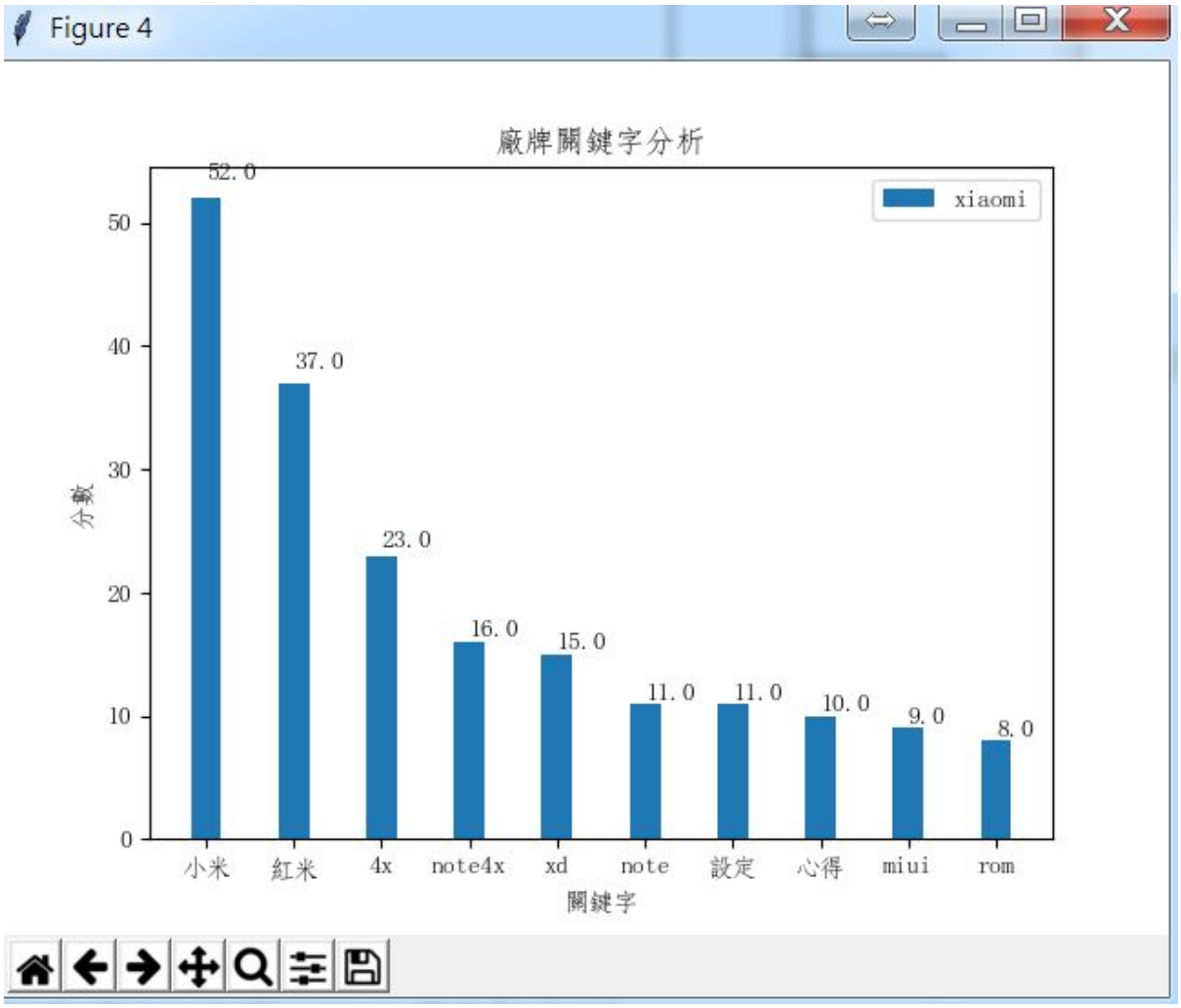
LG的討論都集中在近期機種如G6,G4,G5

然後是華碩



ASUS的照相功能常被討論,其zenfone3也是討論熱點

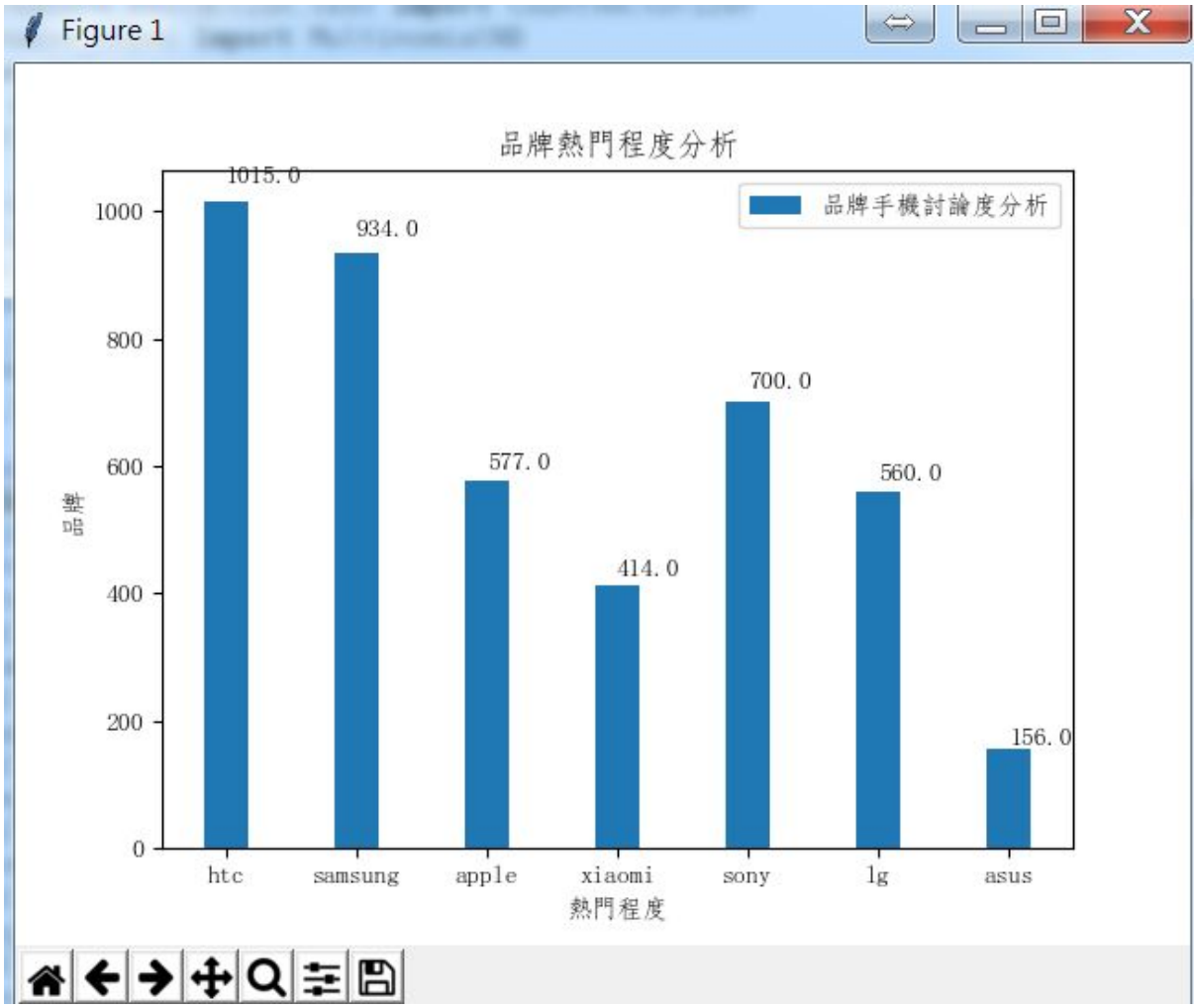
最後是小米



紅米很熱,幾乎每篇都有它的身影

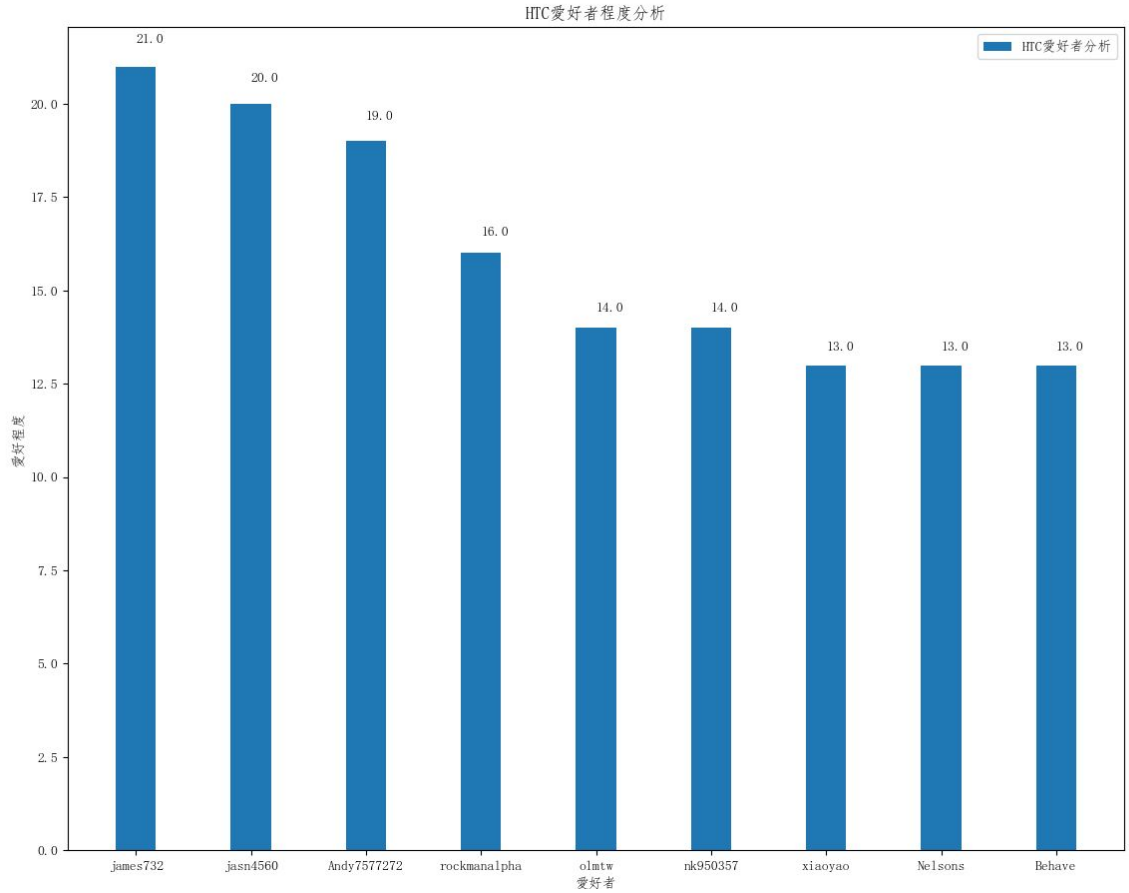
這邊可以做個小結論,不管那牌手機,螢幕永遠都是討論前十名,除了小米外,所以它會是品牌手機的最主要戰場,照相的功能反而還好

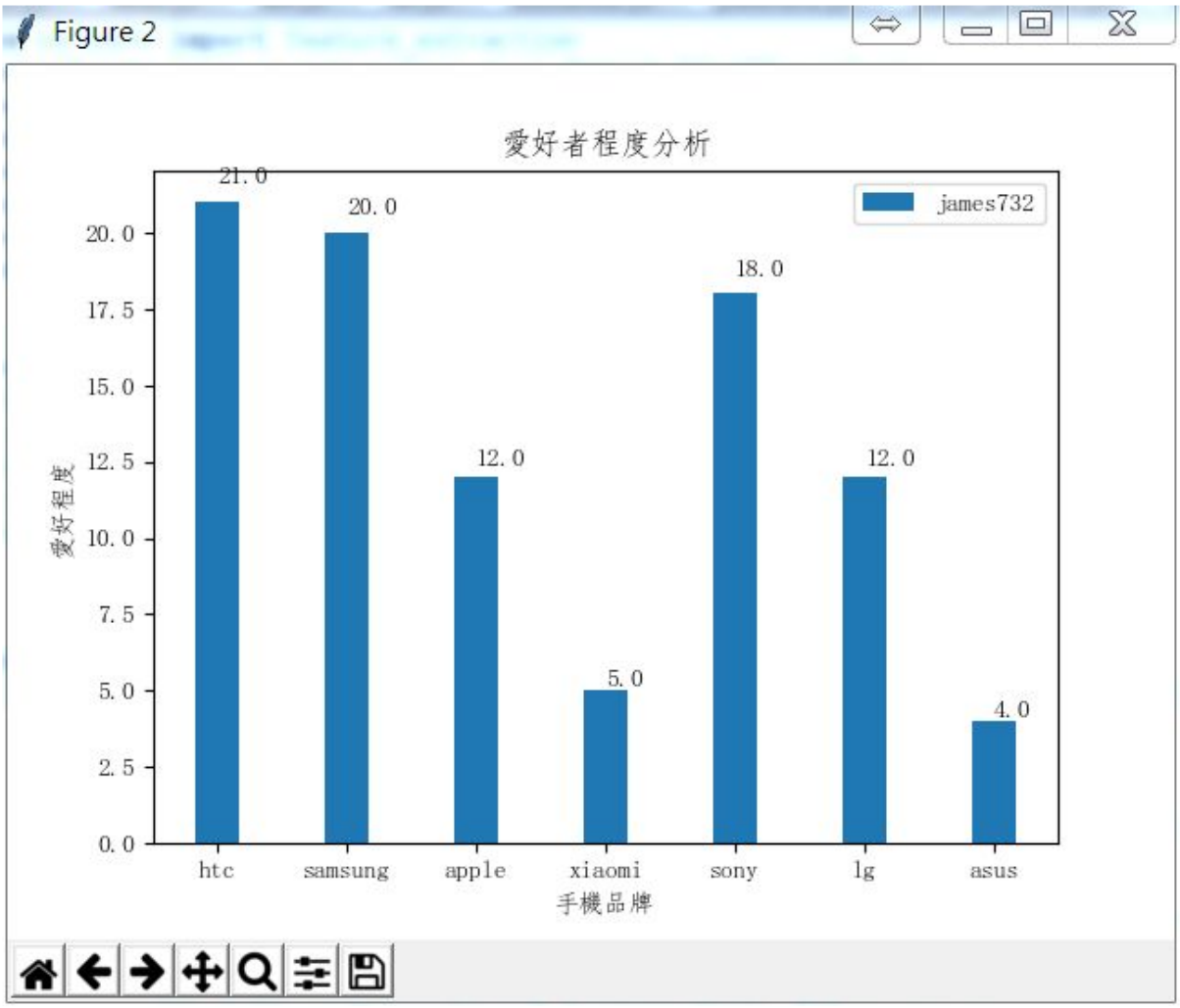
接下來我們看看品牌推文熱度



HTC居然拿第一, ASUS敬陪末座,有可能是分析的資料不夠多,所以不夠客觀,我本來以為apple應該比sony討論熱度高才是,不過也很有可能apple新機還沒出,sony已經從2Q開始推出一系列新手機

接下來我們看看htc手機的討論者群中的推文前十名和單一使用者的手機品牌熱度


HTC爱好者程度分析
HTC爱好者分析
21.0
20.0
19.0
16.0
14.0
14.0
13.0
13.0
13.0
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
爱好程度
james732
jasn4560
Andy7577272
rockmanalpha
olmtw
nk950357
xiaoyao
Nelsons
Behave
爱好者



這樣子我們大概可以知道對那些使用者要推播那些品牌廣告,這些資料可以當作分群的基礎,再進一步的延伸