**Midterm Project**

**Student ID：1011213**

**Name：謝凱靜**

　　通訊軟體越來越多，雖然多數人現在皆使用Line，但在Line之前，其實WhatsApp才是始祖，其初始版本為2010年1月，而Line則是2011年6月發表；2014年Facebook收購WhatsApp，2016年2月2日，Mark Zuckerberg於個人Facebook上表示WhatsApp每月活躍用戶突破10億，表示WhatsApp能夠有這麼高的使用人數、使用率一定有他的特點。

Objective：

* 透過抓取Twitter與WhatsApp相關的內容進行分析，得知大家對於WhatsApp的想法。
* 透過抓取Facebook中WhatsApp粉絲專業的文章內容進行分析。



　　WhatsApp是利用網路傳送簡訊的一種智慧型手機行動應用程式，能夠利用智慧型手機中的聯絡人資訊，尋找也有使用這個軟體的聯絡人。且原先是一款收費的通訊軟體，直到2016年1月18日，WhatsApp才在官方部落格宣布不再徵收年費。

# Data Source：Twitter

# Process：

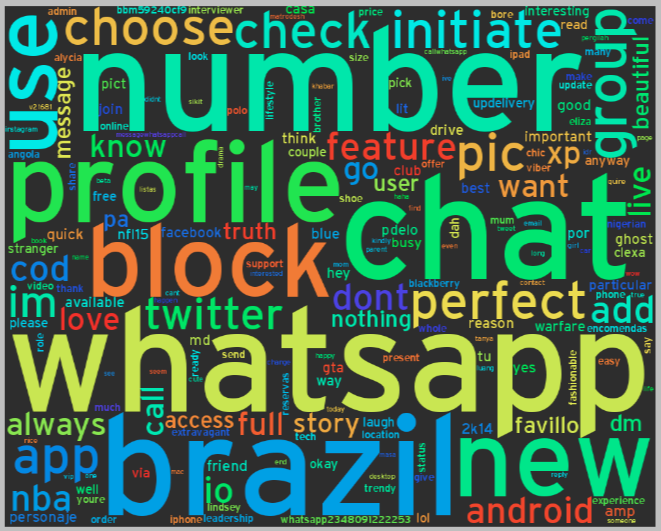


* **抓取資料：**使用s1011213\_tweepy.ipynb進行資料抓取，並且儲存於MongoDb中，資料包含其ID、內容。
* **清理資料：**使用s1011213\_cleaningCode.ipynb進行資料清理，由於每一筆所抓取的資料皆含有ID、表情符號、特殊符號……等，透過資料清理可以將影響分析的內容清除，以利之後分析。
* **分析資料：**使用s1011213\_classifier.ipynb進行分析資料，將之分成positive、neutral、negative三類，再透過s1011213\_mongo2wordcloud.ipynb將三類之字詞頻率繪出。

# 分析結果：

　　所分析的資料總數：1492筆資料，其中有448筆為positive、283筆為negative，761筆為neutral。

**Positive：**



**Neutral：**



**Negative：**

　　透過分類後將結果分成positive、negative、neutral；可以從最後的world cloud看出每個類別中出現次數最頻繁的是什麼字，由於這是透過關鍵字- WhatsApp進行抓取資料，當然每一筆資料都會包含WhatsApp字眼，所以它當然也出現在三個類別中，當然也是占有一定版面。

　　從此次抓取資料的結果發現，其抓取到的資料，有一定的重覆量，為了判別是此抓取方式出了異常，或者是因為抓取的對象是Twitter，有可能因為被推文，所以重覆的內容被抓取了很多次，在此期中報告中另外也抓取了Facebook粉絲專業的資料來相比較。

　　從Positive這張圖中，可以看到number、brazil、chat、profile皆占了很大的版面，利用這幾個關鍵字對分類後的pos\_tweets檔案進行搜尋，得到了以下兩個資訊，whatsapp number choose perfect profile pic initiate chat 此則訊息應該是廣告，而whatsapp brazil – whatsapp block brazil full story whatsapp block apps則是敘述著巴西的whatsapp支應用程式被阻止使用。

　　從Neutral這張圖中，可以看出time、see、face、reality、fellow為出現較頻繁之字眼，得到了whatsapp desktop launch document share feature heres use tech time latesttechnews這則類似推廣whatsapp的好的訊息、以及此則see whatsapp fellow nigerian time face reality訊息。

　　從Negative這張圖中，可以看出rivalry、aguero、fernandez、manchester、order字眼佔了比較大的版面，得到了federico fernandez reveals whatsapp rivalry sergio aguero ahead manchester此則訊息。

# Data Source：Facebook粉絲專業

# Process：

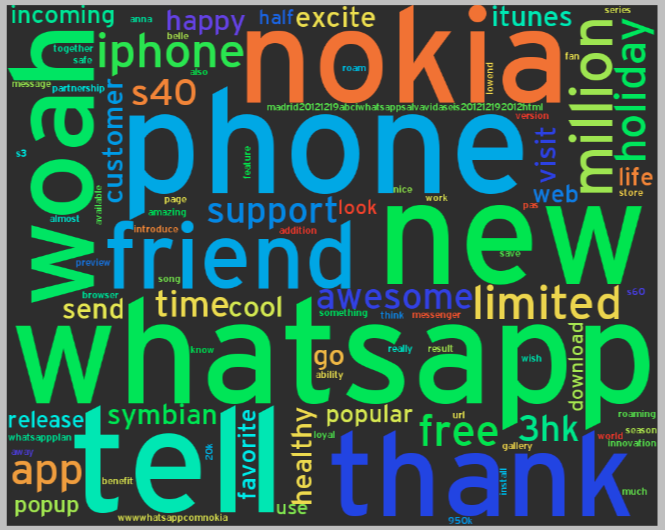


* **取得Facebook存取權杖**([Graph API Explorer](https://developers.facebook.com/tools/explorer/145634995501895))，但由於有時間的限制，可能過一陣子就必須再換一組Token。
* **抓取資料：**使用s1011213\_Whatsapp.ipynb至Facebook中WhatsApp的粉絲專業進行資料抓取，並且儲存於MongoDb中，資料包含其ID、內容。
* **清理資料：**使用s1011213\_cleaningCode.ipynb進行資料清理，由於每一筆所抓取的資料皆含有ID、表情符號、特殊符號……等，透過資料清理可以將影響分析的內容清除，以利之後分析。
* **分析資料：**使用s1011213\_classifier.ipynb進行分析資料，將之分成positive、neutral、negative三類，再透過s1011213\_mongo2wordcloud.ipynb將三類之字詞頻率繪出。

# 分析結果：

　　所分析的資料總數：24筆資料，其中有13筆為positive、4筆為negative，7筆為neutral。

**Positive：**



**Neutral：**



**Negative：**



　　透過分類後將結果分成positive、negative、neutral；可以從最後的world cloud看出每個類別中出現次數最頻繁的是什麼字，由於這是直接抓取Facebook中 WhatsApp的粉絲專業的文章，當然每一筆資料都會和WhatsApp相關，所以它當然也出現在三個類別中，當然也是占有一定版面。

　　從此次抓取資料的結果發現，其抓取到的資料，與在Twitter抓取的資料相比，沒有重覆的內容，也因為這個抓取的資料較少，可以直接透過其分好的結果，進行分析。

　　從Positive這張圖中，可以看到頻率出現高的字，直接進行分析此分類中的資料，可以看到，此13筆資料皆為正向的資料，皆是在稱讚WhatsApp的好，或是在宣布有多少人使用WhatsApp，此次分類的結果相當準確。

　　從Neutral這張圖中，可以看到頻率出現高的字，直接進行分析此分類中的資料，可以看到七句訊息中，皆屬於中性句子，算是分得相當準確的。

　　從Negative這張圖中，可以看到頻率出現高的字，直接進行分析此分類中的資料，可以看到其中有兩則訊息可能屬於中性的。

　　其結果與我當初所設想之目標有所落差，或許是因為所抓取的資料的關係，並沒有得到很多與WhatsApp相關的優點，或是需要改進的缺點，抓取到的資料較多為廣告等不太重要的訊息，但透過此次的期中報告，除了讓我們知道如何自行制定目標後進行資料抓取、分析以外，亦可以讓我們透過多方抓取，得到自己想要的資料。