

Twitter sentiment analysis using R with Sentiment140 dataset

姓名：顏懷瑾

學號：A4211050

學校：中國文化大學

學系：森保系五年級

研究主題

- 使用 R 語言以及其包含的機器學習工具包來進行機器學習中領域內自然語言處理的情緒分析。
- 透過這次的主題，可以學習到 R 語言的語法、機器學習庫的應用、自然語言處理的流程、情緒分析上的難題。

**Machine
Learning Using R**

什麼是情緒分析？

- 情緒分析在現在的應用上有很多種，文字的、聲音的、影像的。
- 我們使用的資料集是屬於文字的資料集，故本次的專題是以文字上的情緒分析為主。
- 以文字為主的情緒分析在目前又分為四個層次：
 1. 基於文件 (Documents-based) 的 ← 這次分析的層次
 2. 基於語句 (Sentences-based) 的
 3. 基於文字 (Words-based) 的
 4. 基於面向 (Aspect-based) 的

為何還要手動再來訓練一個情緒分析模型？

AUDIT | SENTIMENT ANALYSIS | SOCIAL LISTENING

03/31/2019

The 17 Best Sentiment Analysis Tools



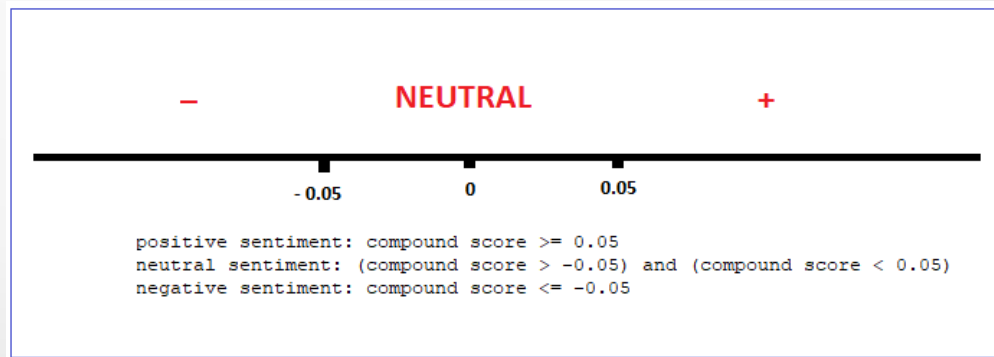
Kuba Rogalski

Community Manager at Brand24. When out of the office, I am probably chilling with my wife and kid, or out taking pictures with my camera. Football freak.



- 現成的工具也有，從分析正負向情緒的工具，到可以計算連續數值型的工具，前面提到的四種情緒分析階層工具一大堆，NLTK、VADER、SenticNet

Natural Language Tool Kit (NLTK) Basic Text Analytics



因為這些都是通用型的情緒分析工具

- 通用型的工具有什麼不好？ 隨插即用耶！ 功能還爆強！
- 自己訓練模型是腦袋有洞嗎？
- 舉個例子：英文單詞 Suck，有「吮吸、抽吸、糟糕」的意思
 - Suck it up. 忍著點。
 - My vacuum cleaner sucks. 我的吸塵器會吸喔。
 - You suck. 你很爛欸。
- 一個英文單字，可能帶有正向、中性、負向情緒。那麼通用型情緒分析工具是否能做到這樣的判斷？或許可以，或許不行。情緒分析對一個詞彙的解釋完全要看他的資料集。針對特定資料集去客製化模型最保險。

什麼是 Sentiment140?

- 該資料集的來源是 Stanford University 的三位學生為了發表論文而使用 Twitter API 收集整理得到的英文資料，爾後便命名該資料集為 Sentiment140。
- 該資料集也曾被用於 Kaggle 競賽中，許多專注於情緒分析的學生也會使用該資料集來做練習。
- Sentiment140 資料集的內容相當雜亂，因此用該資料及來做練習可以說相當的有挑戰性。

Sentiment140

General Information
Site Functionality
For Academics
API
Sentiment Analysis Sites
Contact Us

[Return to Sentiment140](#)

For Academics

Is the code open source?

Sentiment140 isn't open source, but there are resources with open source code with a similar implementation:

- [Text Classification for Sentiment Analysis](#) by Jacob Perkins
- [TwitlGraph](#) by Ran Tavory
- [Twitter sentiment analysis using Python and NLTK](#) by Laurent Luce
- [Twitter Sentiment Corpus](#) by Niek Sanders

What algorithm are you using?

We are using a Maximum Entropy classifier. [Read about our machine learning approach.](#)

Where is the training data?

You can download our training data:

- [Stanford link](#)
- [Google Drive link](#)

Twitter Sentiment Classification using Distant Supervision

Alec Go
Stanford University
Stanford, CA 94305
alecmgo@stanford.edu

Richa Bhayani
Stanford University
Stanford, CA 94305
rbhayani@stanford.edu

Lei Huang
Stanford University
Stanford, CA 94305
leirocky@stanford.edu

開始分析：步驟 1 理解資料

• Sentiment140 資料集一共有六個欄位：

1. sentiment: 情緒的正負向 (0 代表負向情緒、1 代表正向情緒)

2. id: 發文者在 Twitter 上的 ID

3. date: 發文的日期

4. query: 整欄資料值全都是 NO_QUERY

5. user: 發文者在 Twitter 上的暱稱

6. text: 發文者的所撰寫的內容

• 資料筆數: 160 萬筆

```
1599975 "4", "2193578345", "Tue Jun 16 08:38:55 PDT 2009", "NO_QUERY", "Kristah_Diggs", "@yrcIndstnlvrahaha nooo you were just away from ev
1599976 "4", "2193578345", "Tue Jun 16 08:38:55 PDT 2009", "NO_QUERY", "CoachChic", "@BizCoachDeb Hey, I'm baack! And, thanks so much for a
1599977 "4", "2193578348", "Tue Jun 16 08:38:55 PDT 2009", "NO_QUERY", "serianna", "@mattycus Yeah, my conscience would be clear in that cas
1599978 "4", "2193578386", "Tue Jun 16 08:38:55 PDT 2009", "NO_QUERY", "TeamUKskyvixen", "@MayorDoriskWolfe Thats my girl - dishing out the 8
1599979 "4", "2193578386", "Tue Jun 16 08:38:55 PDT 2009", "NO_QUERY", "TeamUKskyvixen", "@MayorDoriskWolfe Thats my girl - dishing out the 8
1599980 "4", "2193578576", "Tue Jun 16 08:38:57 PDT 2009", "NO_QUERY", "angel_sammy04", "In the garden "
1599981 "4", "2193578576", "Tue Jun 16 08:38:57 PDT 2009", "NO_QUERY", "puchal_ek", "@myheartandmind jo jen by nemuselo zrovna té holce ael
1599982 "4", "2193578576", "Tue Jun 16 08:38:57 PDT 2009", "NO_QUERY", "youtubelatest", "Another Commenting Contest! [;: Yay!!! http://tiny
1599983 "4", "2193578739", "Tue Jun 16 08:38:57 PDT 2009", "NO_QUERY", "Mandi_Davenport", "@thrillmesoon i figured out how to see my tweets
1599984 "4", "2193578739", "Tue Jun 16 08:38:57 PDT 2009", "NO_QUERY", "xoAurixo", "@oxhot theri tomorrow, drinking coffee, talking about ou
1599985 "4", "2193578847", "Tue Jun 16 08:38:57 PDT 2009", "NO_QUERY", "RobFoxKerr", "You heard it here first -- We're having a girl. Hope i
1599986 "4", "2193578982", "Tue Jun 16 08:38:58 PDT 2009", "NO_QUERY", "LISKFEIST", "if ur the lead singer in a band, beware falling prey to
1599987 "4", "2193579012", "Tue Jun 16 08:38:58 PDT 2009", "NO_QUERY", "mami111", "mayqueen too much ads on my blog. "
1599988 "4", "2193579012", "Tue Jun 16 08:38:58 PDT 2009", "NO_QUERY", "mami111", "mayqueen too much ads on my blog. "
1599989 "4", "2193579191", "Tue Jun 16 08:38:59 PDT 2009", "NO_QUERY", "tallman", "@Roy_Everitt ha- good job. that's right - we gotta throw
1599990 "4", "2193579191", "Tue Jun 16 08:38:59 PDT 2009", "NO_QUERY", "tallman", "@Roy_Everitt ha- good job. that's right - we gotta throw
1599991 "4", "2193579249", "Tue Jun 16 08:38:59 PDT 2009", "NO_QUERY", "razzberry5594", "WOOOOO! Xbox is back "
1599992 "4", "2193579284", "Tue Jun 16 08:38:59 PDT 2009", "NO_QUERY", "AgustinaP", "@rmedina @LaTati Mmmm That sounds absolutely perfect..
1599993 "4", "2193579284", "Tue Jun 16 08:38:59 PDT 2009", "NO_QUERY", "AgustinaP", "@rmedina @LaTati Mmmm That sounds absolutely perfect..
1599994 "4", "2193579477", "Tue Jun 16 08:39:00 PDT 2009", "NO_QUERY", "ChloeAmisha", "@SCOOPY_GRITBOYS "
1599995 "4", "2193579477", "Tue Jun 16 08:39:00 PDT 2009", "NO_QUERY", "ChloeAmisha", "@SCOOPY_GRITBOYS "
1599996 "4", "2193601028", "Tue Jun 16 08:40:49 PDT 2009", "NO_QUERY", "AmandaMarie1028", "Just woke up. Having no school is the best feelin
1599997 "4", "2193601969", "Tue Jun 16 08:40:49 PDT 2009", "NO_QUERY", "TheWDBboards", "TheWDB.com - Very cool to hear old Walt interviews!
1599998 "4", "2193601991", "Tue Jun 16 08:40:49 PDT 2009", "NO_QUERY", "bpbabe", "Are you ready for your MoJo Makeover? Ask me for details "
1599999 "4", "2193602064", "Tue Jun 16 08:40:49 PDT 2009", "NO_QUERY", "tinydiamondz", "Happy 38th Birthday to my boo of alll time!!! Tupac
1600000 "4", "2193602129", "Tue Jun 16 08:40:50 PDT 2009", "NO_QUERY", "RyanTrevMorris", "happy #charitytuesday @theNSPCC @SparksCharity @Sp
```

開始分析：步驟 2 使用 R 語言進行資料前處理

- 機器學習類別：監督式分類學習
- 使用的資料欄位：
 - sentiment: 情緒的正負向
 - text: 發文者的所撰寫的內容

```
Text_preprocess <- function(article) {  
  article <- iconv(article, "UTF-8", "ASCII") #要先移除所有emoji  
  article <- gsub(" ?(f|ht)(tp)(s?):(//)(.*)" ".|/](.*)", "", article) #移除URL  
  article <- gsub("\\$", "", article) # 移除錢號  
  article <- gsub("\\n", "", article) # 移除換行符號  
  article <- removeWords(article, stopwords("english"))  
  article <- gsub("[:punct:][:blank:][:digit:]]+", " ", article)  
  article <- trimws(article,  
    which = c("both", "left", "right")) #去除行初行尾空格  
  article <- tolower(article) #該函數如果遇到emoji會出現utf8towcs錯誤  
  return(article)  
}
```

1. 移除表情字符 (emoji)
2. 移除網址 (URL)
3. 移除錢號 (\$)
4. 移除換行符號 (\n)
5. 移除英文停用詞 (stop words)
6. 移除標點符號 (punctuation)
7. 移除多餘的空格 (Extra space)
8. 移除數字 (digit)

因為該資料集的特性，處理順序必須按照步驟

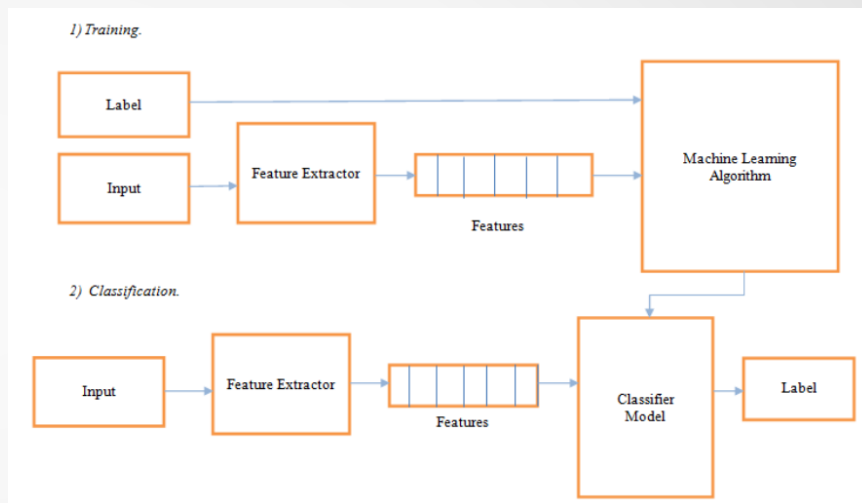
開始分析：步驟 3 檢查資料集完整性

- 文字前處理後檢視資料發現有一些 NA 值在裡面，這種帶有標籤 (label) 卻沒有 tweet 的必須要去除
- 使用 stats 函式庫的 `complete.cases()` 來作清除

736553	0	2301859136	Tue Jun 23 16:00:32 PDT 2009	NO_QUERY	Rimfyre	NA
736554	0	2301859337	Tue Jun 23 16:00:33 PDT 2009	NO_QUERY	Ryanjhughes	buckingham palace wi fi
736555	0	2301859717	Tue Jun 23 16:00:35 PDT 2009	NO_QUERY	favadi	NA
736556	0	2301860269	Tue Jun 23 16:00:38 PDT 2009	NO_QUERY	fairy_windy	NA
736557	0	2301860738	Tue Jun 23 16:00:39 PDT 2009	NO_QUERY	funmsdrebirth	verastic one hour late gosh cant believe

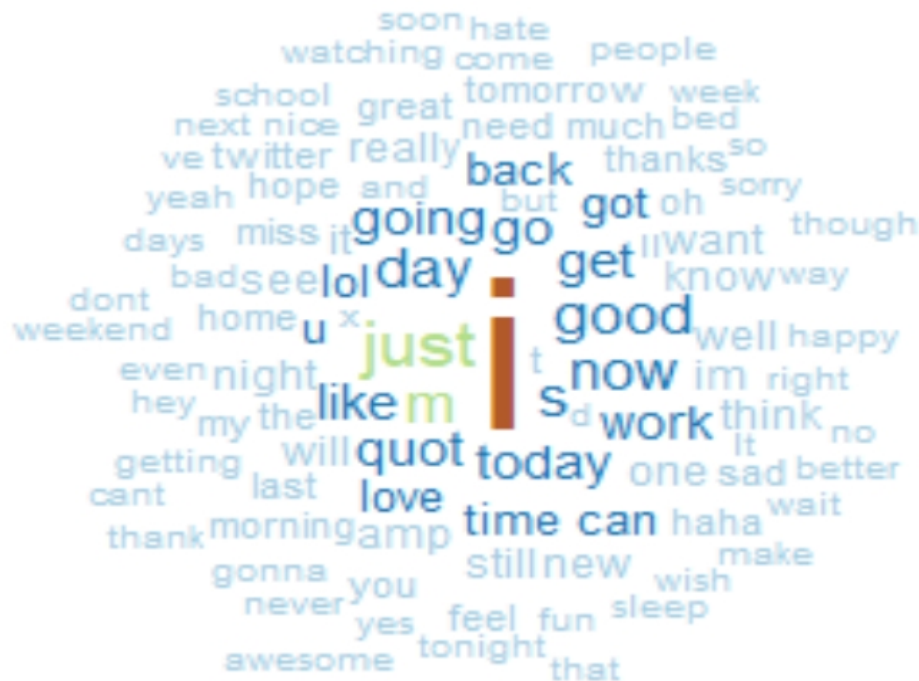
開始分析：步驟 4 機器學習中文字分析的典型流程（一）

- 分割訓練用資料集與測試用資料集， 80% 用於訓練， 20% 用於測試。
- 定義文檔術語矩陣 (document-term matrix) 與分詞 (tokenization) 用的配置。
- 建立詞彙表與文檔術語矩陣 (document-term matrix)



開始分析：文字雲

- 未經過 TF-IDF 處理之前的文字雲，可以看出很多的語句，都會以 i(我)開頭，或許和人類心理常以自我為中心來表達意見有關連性。



開始分析：步驟 5 機器學習中文字分析的典型流程（二）

- 利用 TF-IDF 抽取真正有意義的詞
- 明明已經做過了停用詞 (stop words) 去除
- 為何還要用 TF-IDF 再篩選一次？
- 舉例：如果今天一位正常的人爆粗口，那麼他說的這個詞彙就有很高的 TF-IDF 權重；相反的如果是館長的發言，那麼同樣的詞彙的 TF-IDF 權重就會很低。

在經過 TF-IDF 處理後有意義的詞彙表

[1] "aclocal"	[26] "takjing"	[51] "lucrecerb"	[76] "meema"
[2] "ambermarion"	[27] "jennywynter"	[52] "nascaraddict"	[77] "breathmynt"
[3] "whitesoul"	[28] "worriededed"	[53] "flgel"	[78] "lovaahh"
[4] "tjv"	[29] "kellyjordan"	[54] "jstwtg"	[79] "valinyaozhen"
[5] "alexcg"	[30] "erronocampo"	[55] "petetaylor"	[80] "xixixi"
[6] "healthfair"	[31] "playniki"	[56] "notoriouskitsch"	[81] "mebee"
[7] "artdance"	[32] "neyomfriday"	[57] "miltr"	[82] "oamcortney"
[8] "montemplar"	[33] "tendont"	[58] "jasooo"	[83] "woooppee"
[9] "foxdream"	[34] "jennyvilleda"	[59] "lesesa"	[84] "otherniceman"
[10] "suzimcdowell"	[35] "saywhatx"	[60] "bigdaddycoolj"	[85] "ahoykatrina"
[11] "uncomark"	[36] "doughed"	[61] "shaneya"	[86] "mothsex"
[12] "wesdunn"	[37] "lucdew"	[62] "madisonapril"	[87] "damdams"
[13] "kachnajunior"	[38] "zaccy"	[63] "swaggaboom"	[88] "earrr"
[14] "pfeffior"	[39] "lcict"	[64] "ashbee"	[89] "joelllllllll"
[15] "douglassiter"	[40] "xurs"	[65] "jordandwagner"	[90] "promisedddd"
[16] "badaette"	[41] "truw"	[66] "stephenwalker"	[91] "sacbeejp"
[17] "chrisolsen"	[42] "leoguy"	[67] "itilac"	[92] "unairconditioned"
[18] "eliaskeppens"	[43] "hollace"	[68] "definedfinesse"	[93] "sayasaras"
[19] "brookwood"	[44] "darrengreene"	[69] "briannalina"	[94] "jaimito"
[20] "fakingly"	[45] "inspirative"	[70] "gigglesnort"	[95] "saekuto"
[21] "reviling"	[46] "varberg"	[71] "kaivari"	[96] "kelliemurfski"
[22] "depper"	[47] "chandlergrace"	[72] "seeeein"	[97] "catecorbitt"
[23] "rocketella"	[48] "noleafclover"	[73] "deafness"	[98] "kimeaglestone"
[24] "feew"	[49] "highclasswhore"	[74] "drnking"	[99] "melcam"
[25] "mabulay"	[50] "scottcarefoot"	[75] "iamfase"	[100] "mondaymadness"

開始分析：步驟 6 訓練機器學習模型

- 使用的模型：對數機率回歸 (Logistic Regression)
- 交叉驗證分割數：5 折
- 模型訓練時間：46 分鐘

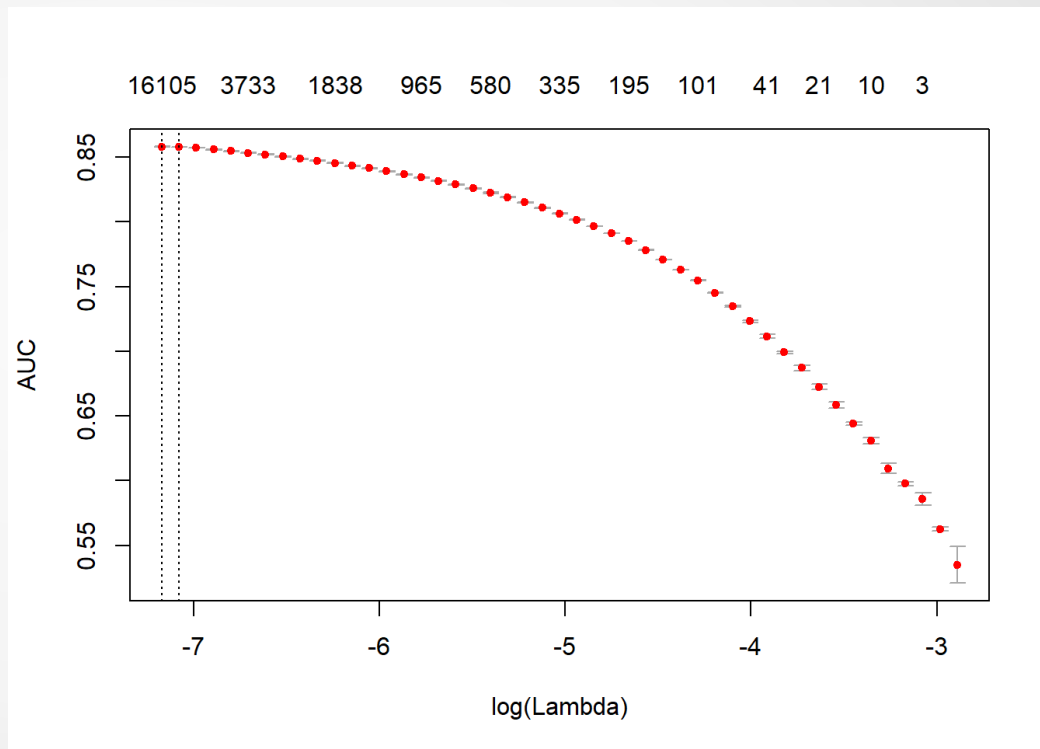
```
print(difftime(Sys.time(), t1, units = 'mins'))
```

```
## Time difference of 46.64771 mins
```

```
# train the model
t1 <- Sys.time()
glmnet_classifier <- cv.glmnet(x = dtm_train_tfidf, y = t
  family = 'binomial',
  # L1 penalty
  alpha = 1,
  # interested in the area under ROC curve
  type.measure = "auc",
  # 5-fold cross-validation
  nfolds = 5,
  # high value is less accurate, but has faster training
  thresh = 1e-3,
  # again lower number of iterations for faster training
  maxit = 1e3)
```

開始分析：步驟 7 使用曲線下面積（AUC）評估模型

- 何為 AUC?
- AUC 的值必在 0~1 之間
- AUC 值越高代表模型預測率越準確
- 模型內的評估：max AUC = 0.8579
- 20% 測試資料集的 AUC = 0.8566717



心得

- 快樂的時間總是過得特別快，課堂上非常感激蔡老師的指導，課後也很感謝助教們細心的評閱我的作業，也替我解惑。就這樣在一次次的課程中慢慢進步，因為我非本科系的學生，在學習 R 語言程式設計上顯得特別吃力，下課後幾乎用了所有的時間在碰撞中學習，卻從中漸漸的感覺到有趣，越練越開心，今天熬到了最後的專題報告，累積起來的知識可以說是收穫豐富，但學海無涯，還是要再次感謝蔡老師與助教們替我打開這道資料分析的大門。

HW1 建議 #1

 MiccWan opened this issue 9 days ago · 0 comments




MiccWan commented 9 days ago

https://github.com/alien410/Allen/blob/027a520b81f291408bf3ea1d2f9df45e039abc3d/Week1/108_%E5%85%A8%E5%9C%8B%E5%A4%8F%E5%AD%A3%E5%AD%B8%E9%99%A2_7%E6%9C%8811%E6%97%A5_Clas2.Rmd

建議

- 說明很詳盡，可以利用 Rmd 的特性在程式碼的外面寫說明，不用寫成註解
- 兩個資料集沒有關聯很可惜

HW3 建議 #2

 MiccWan opened this issue 3 days ago · 0 comments



MiccWan commented 3 days ago · edited

https://github.com/alien410/Allen/blob/ea09534edd262842eda74dc3f677951919e7e66a/Week2/108_%E5%85%A8%E5%9C%8B%E5%A4%8F%E5%AD%A3%E5%AD%B8%E9%99%A2_7%E6%9C%8818%E6%97%A5_Clas4.Rmd

建議

- 如果要檢查一個函數是不是已經存在，可以用 `find(...)`，如果存在的話會回傳該函數被宣告的位置，不存在的話會回傳 character(0)