劇本裡的千層套路

從情感分析談電影

第 19 組
曾理陽

2019-01-12

目錄

如何將此 Rmd 輸出成 PDF?
依序執行下方指令:

```
rmarkdown::render("index.Rmd")
pagedown::chrome_print("index.html")
```

1 簡介

看電影的時候，有沒有一種感覺好像這個橋段跟某部電影好像？ 男女主角是不是要親吻了？ 部隊這個時候是不是會被偷襲？ 怪物才不會那麼容易就死掉……你看牠果然還活著！

你們好奇同類型劇本是不是存在某種淺規則嗎？一起透過情感分析來一探究竟吧。

2 方法

1. 使用rvest進行網路爬蟲並存檔
2. 將文本斷成一句一句，存成csv檔
3. 利用斷好句的csv檔進行情感分析：

a. 將句斷成字
b. 使用tidytext內建停用詞表過濾不必要的字
c. 新增'paragraph'作為分析內容的分段，同類型電影的分段數量將統一
d. 產出情感分析後的結果，存成csv檔

1. 將同類型電影的情感分析結果集合成圖：

a. 讀取所有電影的情感分析結果
b. 新增'frame'作為GIF動態效果中切換圖表的依據
c. 使用gganimate做成GIF動態圖

1. 綜合分析情感波動趨勢的規律

2.1 資料取得

劇本網站：The Internet Movie Script Database (IMSDb) ## 原始碼運作說明

[main.Rmd] 1. 清理文本會使用到的正規表達式： 經過觀察，大多數的劇本會用數字標示來分篇章或是使用"INT., EXT., INSERT."標示場景轉移，我將利用這些標記作為斷點來分割文本。

主要使用函數： strsplit()

兩種正規式： a. 以數字標示作為分割 “[0-9]\.|[0-9][0-9]\.|[0-9][0-9][0-9]\.” 或 “[0-9] |[0-9][0-9] |[0-9][0-9][0-9]” b. 以場景轉移標示作為分割 “INT.|EXT.|INSERT.”

1. 統一化每部劇本的段落長度： 進行情感字典比對後，以line來計算每種情緒出現次數，並且利用line分段。我想統一每部劇本的段落能落在30，因此用「無條件捨去法除line的總數，再加上1」的方式，統一所有劇本的段落長度

原始碼示意： tidy_script %>% anti_join(custom_stop_words) %>% inner_join(get_sentiments(“nrc”)) %>% count(line, sentiment) %>% mutate(paragraph = (line %/% 5.9) + 1)

[Analysis.Rmd] 1. 整理每部電影的data.frame內容： 內容必須包含： a. 情感分析結果 b. frame

※ frame = rep(“Beauty_and_the_Beast”,length(Beauty_and_the_Beast$line))

1. 數值校正： 分析發現，有些電影的情感強度太強，會讓ggplot直方圖的y軸級距過大，導致其他情感強度較弱的電影在GIF圖片中的變化趨勢難以觀察。

為了使彼此y軸級距相近，需要進行校正→將情感的出現次數進行乘除運算

1. GIF動畫圖產出： 先將不同電影類型的所有data.frame分別集合成總表，產出ggplot的直方圖，並用transition_states函數製作GIF。

在transition_states函數中， a. 圖層的轉換： 依照frame類似標籤的功能切換圖層 b. transition_length：轉換所花的時長 c. state_length： 圖層保持的時長 d. ease_aes： 轉換的效果，預設為'linear'

請參照： transition_states(frame, transition_length = 200, state_length = 500) + ease_aes("linear")

3 結果

1. 好萊塢劇本呈現三幕劇形式： 根據情感分析的變化趨勢，可以發現其高低起伏的頻率明顯都有三次循環。

2. 驚悚片： 根據情感分析的變化趨勢，驚悚片在時間軸到了10%、30%、70%和90%，最容易進入高張的劇情。

3. 戰爭片： 根據情感分析的變化趨勢，可以發現戰爭片通常在一開始便是高張力的情感呈現，尤其以anger和sadness為甚，表示戰爭片不同於其他類型電影，少以鋪陳開始其劇情，而是進入刺激性情節吸引觀眾注意。

4. 愛情片： 根據情感分析的變化趨勢，愛情片中愛情的過程中不可能完全沒有阻礙，本專案觀察anger、sadness、joy和trust的組合，去判斷它們表現出的一致性(anger和sadness的同時上升與joy和trust的同時下降，代表此時正是衝突的劇情)，最終，發現大部分劇本會歷經3~4次衝突。

4 討論與貢獻

1. 除了語言，還有什麼能表達情感？ 情感可能從肢體動作、眼神所傳達出來，傷心難過也許靠啜泣傳達、生氣透過怒目瞪視、快樂會手舞足蹈，勾肩搭背通常表示信任，這些都不屬於語言的部分，因此，我們尚有被忽略的因素沒有兼顧到。然而，若要兼顧這些點，就必須加入電影內容的視覺分析，這是我目前還不具備的能力。

2. 藉由情感分析找出電影類別潛藏的規律，若運用機器學習的方法建構出分析模型，未來能夠預測並分類電影的種類，不必透過觀看整部影片或猜測片名的寓意。

3. 如果擔心直接看劇情內容會「暴雷」，但又很在乎劇情走向，可以透過情感分析，間接猜測出劇情的過程與結果。

5 附錄

5.1 組員分工

曾理陽： 資料蒐集、預處理、斷詞、情感分析、作圖、製成分析報告。