R語言與資料科學 小組期末專案

關於 TED Talk 的五件真相！

G15 蘆洲李陳

外文三 陳悅翔 B06102001

森林二 周昱雯 B07605012

會計四 盧信呈 B05702109

職治一 李昀茜 B08409009

目錄

1.簡介

1.1 研究動機

尋找研究素材的時候，我們希望能找到不需要花很多時間爬網站的資料，來進行資料處理與分析，後來在 Kaggle 上找到有人整理好的 TED Talks 資料集，資料不但乾淨清楚，符合我們的需求，而且 TED Talks 在近幾年來又蔚為風潮，感覺是個蠻新穎的題目。在好奇心之下，我們決定以此做為主題，並應用課堂所學的知識，開始進行研究。

1.2 研究目標

研究將從多種面向來探討 TED Talks，包含內文笑點、觀看次數、標題、主題分類、以及資料科學家的 TED Talks。期盼透過數據及文本分析，檢視我們的研究成果與大眾對於 TED Talks 的既定印象是否吻合，或是否有意外的發現。我們列出五個研究項目：

(1)觀看次數與「主題」和「標題」分別的關聯性為何？
(2)幽默的 TED Talk 的特性為何？
(3)歷年受歡迎的主題（Tag）流變
(4)相同主題影片不同觀看次數，有哪些性質的差異
(5)資料科學家在 TED Talk 中偏好使用的字詞

2.研究方法

2.1 資料取得

我們使用 Kaggle 這個網站上所提供的 TED Talks 資料。Kaggle 是全球知名的資料科學網站，提供眾多數據並舉辦許多數據分析競賽，影響深遠。而我們所使用的 TED Talks 資料，資料中包含到 2017 年 12 月底的兩千五百多則影片，內容包含每個影片的逐字稿、影片標題、評論數、分類標籤 tags 以及評分感想 rating。

2.2 原始碼運作說明

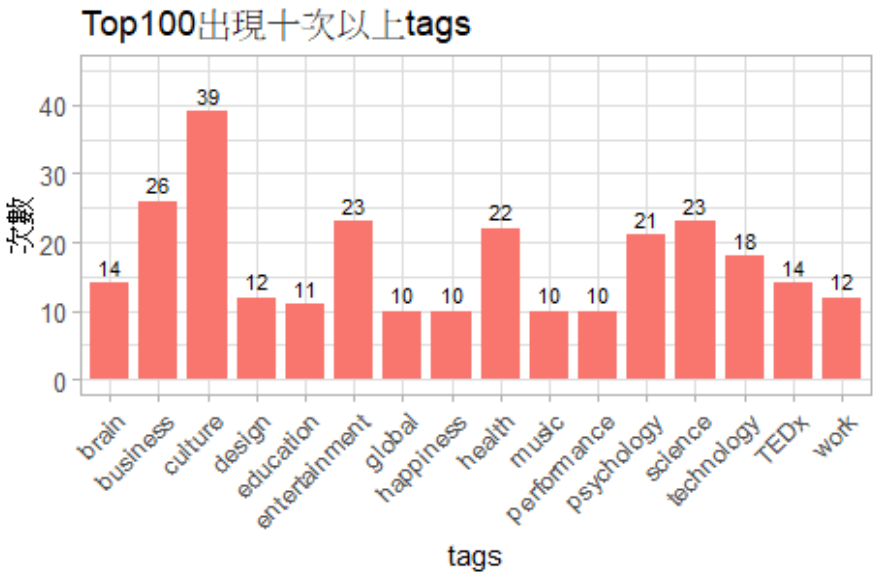
我們主要使用 R 語言來進行分析，主要使用的套件包含 tidyverse 以及 quanteda。詳細的說明請見 readme。

3. 結果

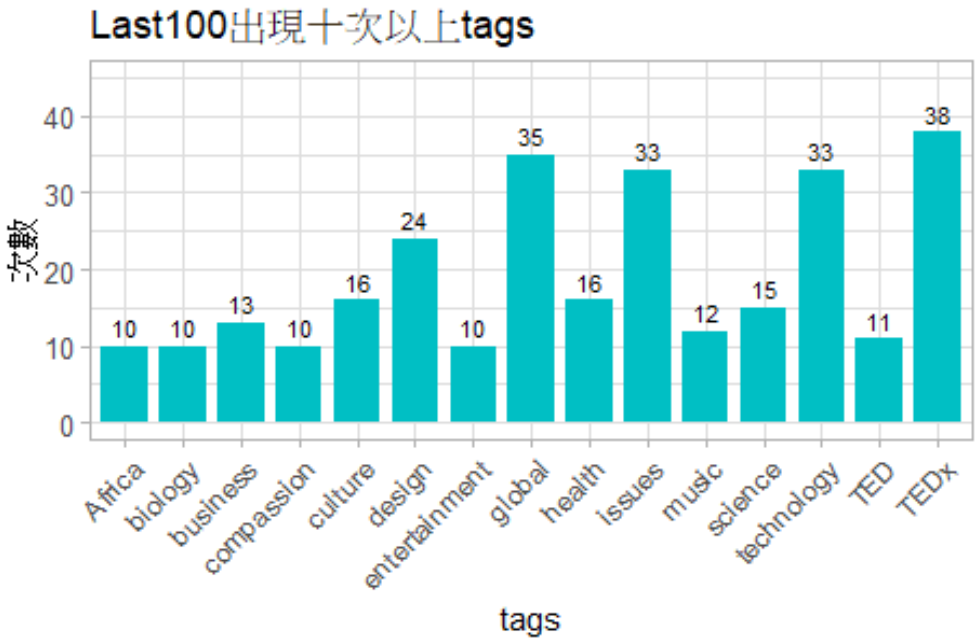
3.1 觀看次數分析

3.1.1 觀看次數與主題的分析

此部份我們旨在探討觀看次數與該影片 tags 的關係，以及出現相同 tag 但導致觀看次數不同的可能影響因素。我們將影片用觀看次數排名後，為求效果顯著，比較前後 100 名影片 tags 的使用差異，而非前 100 名與整體的比較。

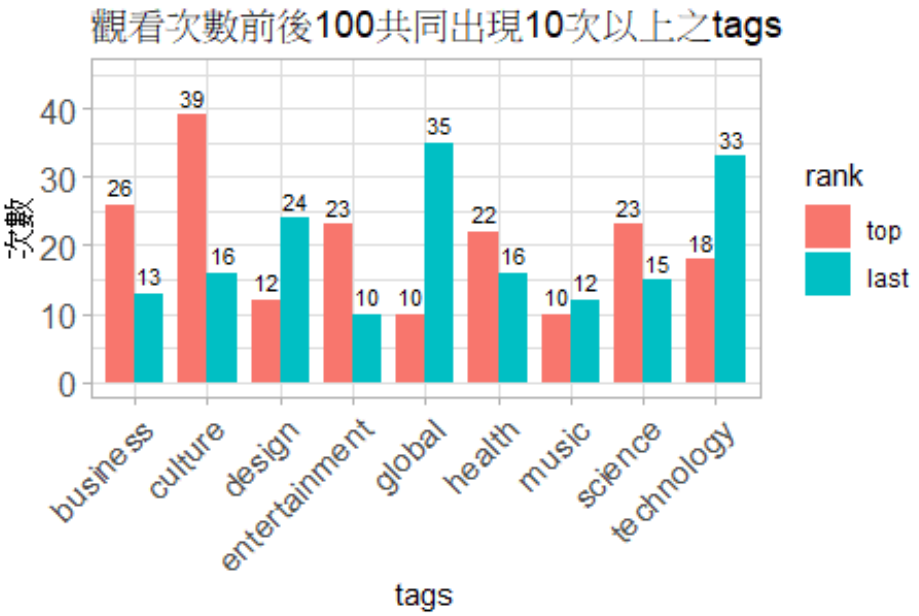


圖一、觀看次數前一百影片中累積出現十次以上之 tags



圖二、觀看次數後一百影片中累積出現十次以上之 tags

除了 Africa 我們推測是因為關於特定地域的相關影片的受眾比較少，所以連帶觀看次數較少，但其實前後 100 名的影片重複出現的 tags 比例不低，因此我們想繼續研究出現同樣 tag 的影片，觀看次數排名懸殊的影片是否有性質上的差異。

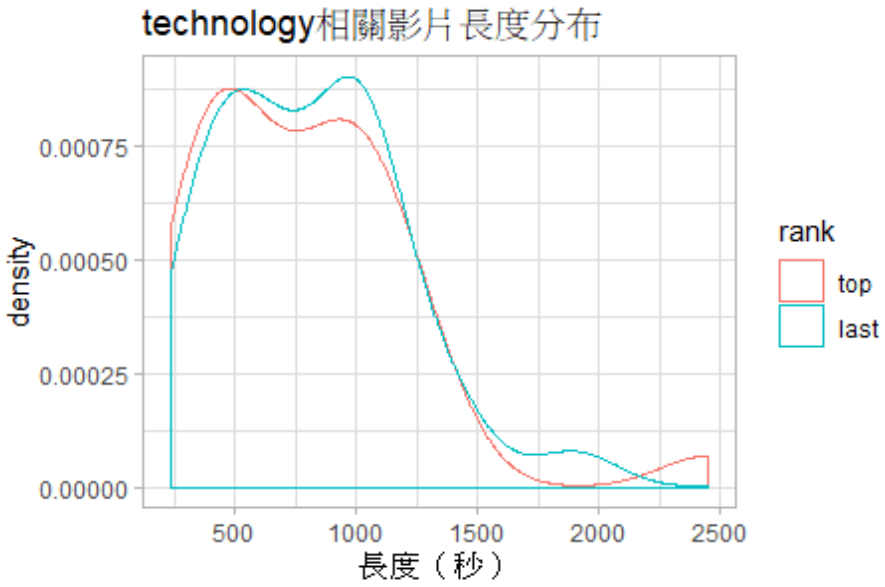


圖三、觀看次數後一百影片中共同出現十次以上之 tags

我們先挑出在前後一百名都出現十次以上的 tags，去除 TEDx 此類活動類型的 tag，再選擇其中幾個針對包含該 tag 的影片性質做分析。

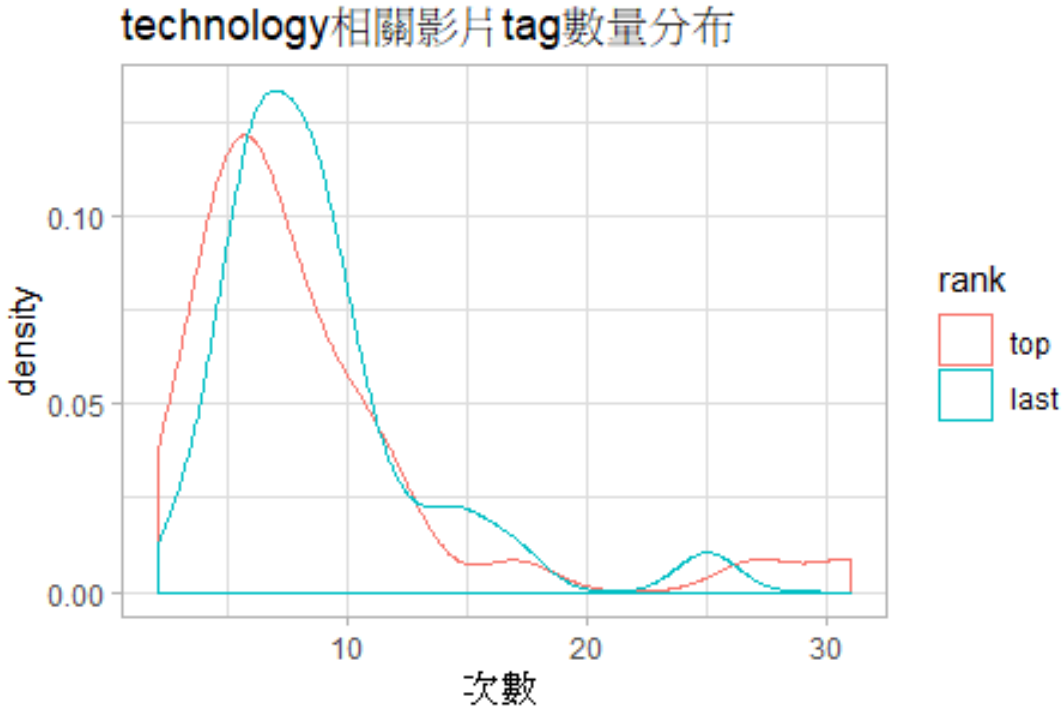
具有相同 tag 但觀看次數不同之影片性質分析 — 以 technology 為例

此區我們選出 tag 包含 technology 的影片的觀看次數前後各 30 名進行比較，包括影片長度、tags 數量，以及詞頻分析。



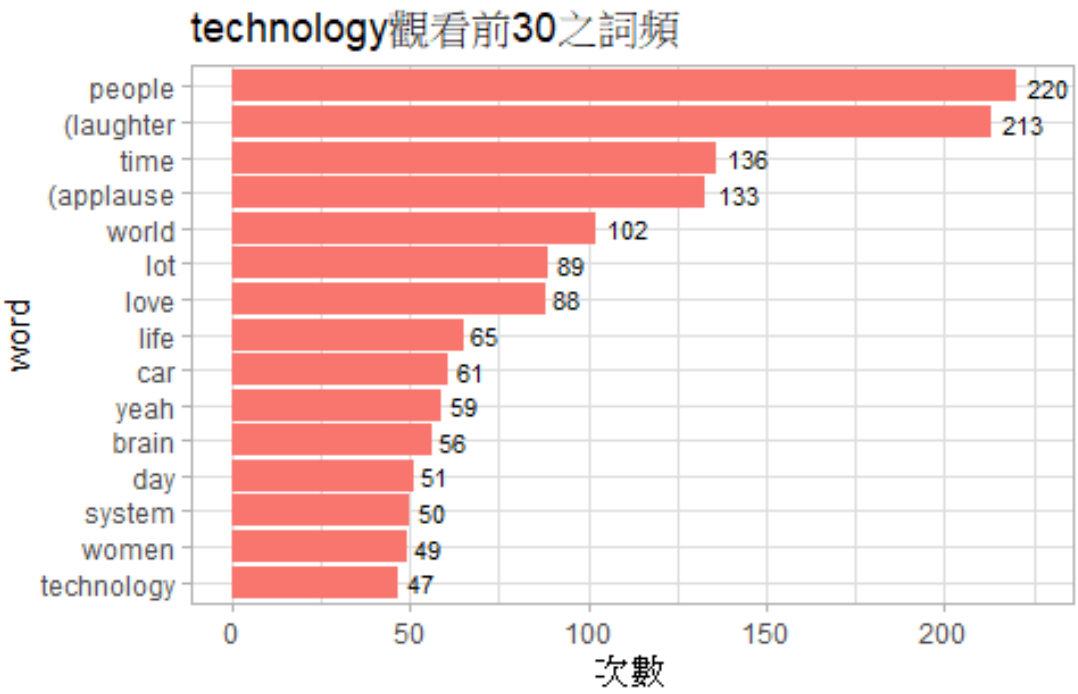
圖四、technology 影片觀看次數前後 30 名之影片長度分布圖

以影片長度而言，我們原本預期排名較高的影片長度會較長，也就是易於普羅大眾觀看吸收，數據（圖四）結果顯示 technology 相關影片中觀看次數前後 30 名平均時長僅差約 15 秒（前 30 名平均 794.13 秒；後 30 名平均 808.83 秒），兩者在分布的差異亦不大，推測影片長度並非該類別影片影響觀看次數的顯著相關因子。



圖五、technology 影片觀看次數前後 30 名之 tags 數量分布圖

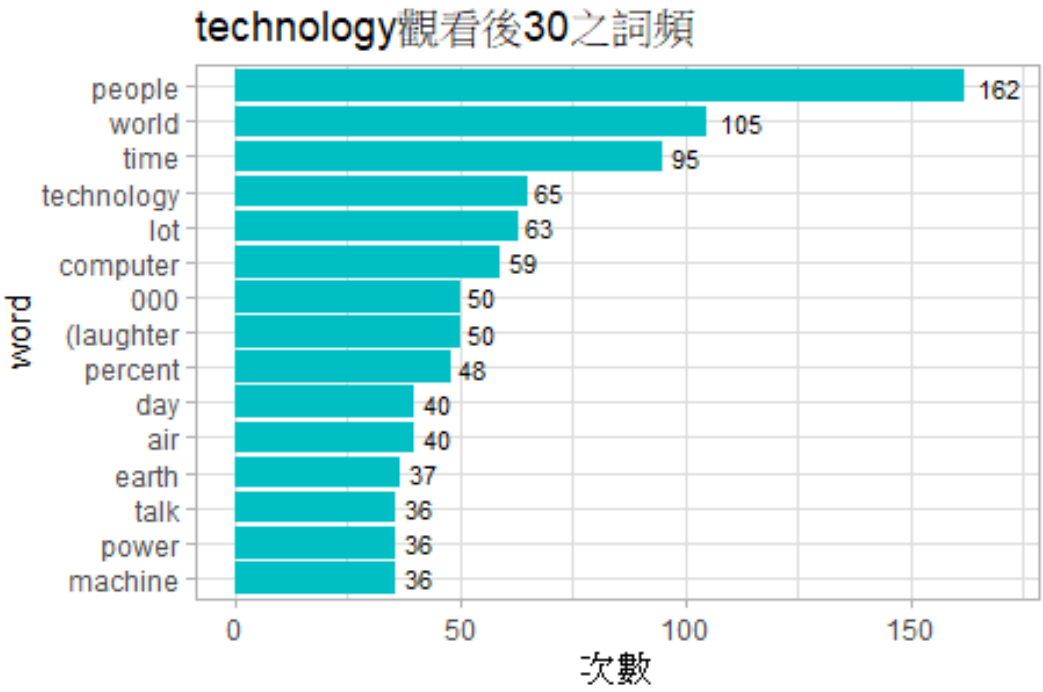
以影片的 tags 數量而言，我們原本預期排名較高的 tags 數量會較多，使該影片更容易在搜尋時被搜尋到，進而增加觀看次數，但我們的數據（圖五）顯示觀看次數前後 30 名的分布差異不大，平均也幾乎相同（前 30 名平均 8.53 個；後 30 名平均 8.76 個），因此推測 tags 的數量並非該類別影片影響觀看次數的顯著相關因子。



圖六、technology 影片觀看次數前 30 名詞頻

斷詞說明：

1. (laughter、(applause：觀眾的笑聲和鼓掌在逐字稿用(laughter)和(applause)方式呈現，為區分講者內容和觀眾反應，不用左括號分割，但因(laughter)和(applause)與下一個單字間無空格，因此依然用右括號斷詞。



圖七、technology 影片觀看次數後 30 名詞頻

斷詞說明：

1. 000：數字，因用逗號斷詞，四位數以上數的字會被分開，例如 5,000 會斷成 5 和 000。

我們利用詞頻的次數前 15 名做觀看次數前後 30 名的比較，前 30 名影片觀眾笑和鼓掌的次數明顯多於後 30 名的影片，且內容可能較容易與大眾達到共鳴（love, life, women），也有較多較輕鬆的口語用詞（yeah）；後 30 名出現較多大數據（000）以及較多與科技、數據相關字詞（computer, machine, percent）。推測觀看次數較高的影片內容理解的門檻較低、讓聽眾容易帶入生活情境，或是能夠以較輕鬆的與氣和方式進行演講。

最後根據數據分析結果，在 tag 中有 technology 的影片中，我們推得比起影片的長度和 tags 的數量，講者在演講中的用詞與觀看次數的關係相對明顯。

3.1.2 觀看次數與標題的分析

首先我們想了解標題的長度在受歡迎與不受歡迎的群組中是否存有差異。

	全體	Famous	Last
平均長度	6.63	6.56	5.19
中位數	6	6	5

由上可知，在不受歡迎的影片中，影片標題的長度明顯短於受歡迎的影片，甚至是也低於全體平均。

而我們發現，**標題中問句的使用情形在各群組中是有差異的**。

	全體	Famous	Last
WHY	5%	9%	1%
WHAT	4.7%	3%	3%
HOW	12%	12%	5%
總和	21.7%	24%	9%

在不受歡迎的群組中，使用問句的情形明顯低於全體以及受歡迎的影片；受歡迎的影片中問句使用情形比全體稍微高。而全體來說，使用問句的情形大概約兩成左右。

另外我們發現，使用 You/Your 以及大寫字母的比例，在個群組中有所不同。

You/Your 使用佔全部單詞的比例

全體	Famous	Last
1.3%	2.8%	0.9%

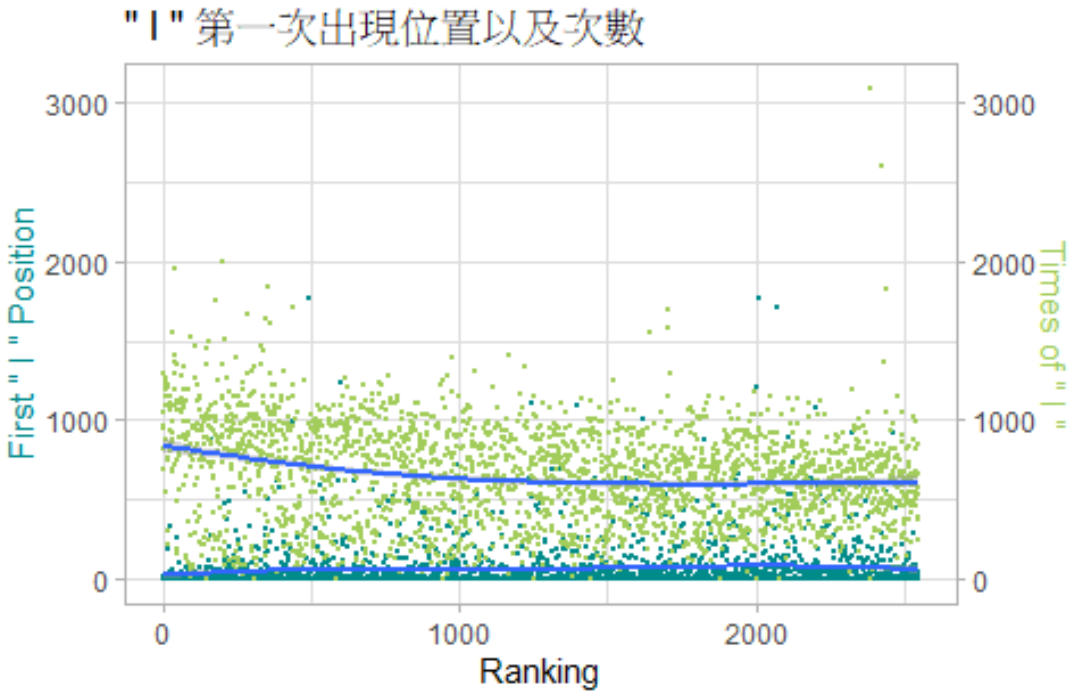
大寫字母使用佔全部單詞的比例

全體	Famous	Last
18.3%	17.7%	39.3%

在受歡迎的影片中，我們看到使用「你、你們」的比例較高，我們推測或許人們對於「你」這樣的敘述，比較具有吸引力。另外，大寫字母的使用，不受歡迎的影片中使用的比例比起其他群組來說，高出許多。我們認為這跟專有名詞的使用有關：不受歡迎的影片中有比較多的地區名稱、特定人物或事件，或是學術上專有名詞。這可能導致人們興趣的缺乏以及了解的困難（沒有太多背景知識難以理解），而使其不受到歡迎。

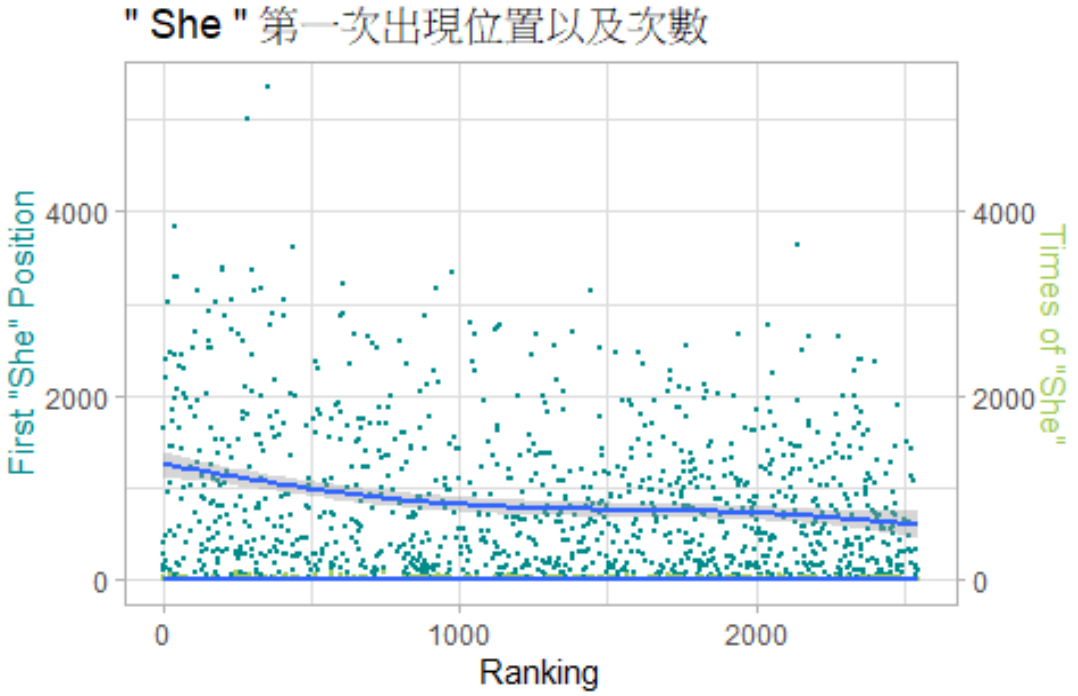
3.1.3 觀看次數與人稱的分析

此區我們想要探討人稱的使用以及觀看次數的關係。利用計算 I、She／she、He／he、You／you、They／they、We／we 和 It／it 第一次出現的位置以及出現的次數，試圖找出與影片觀看次數是否有相關。



圖八、I 在整體 TED talk 中第一次出現的位置和次數

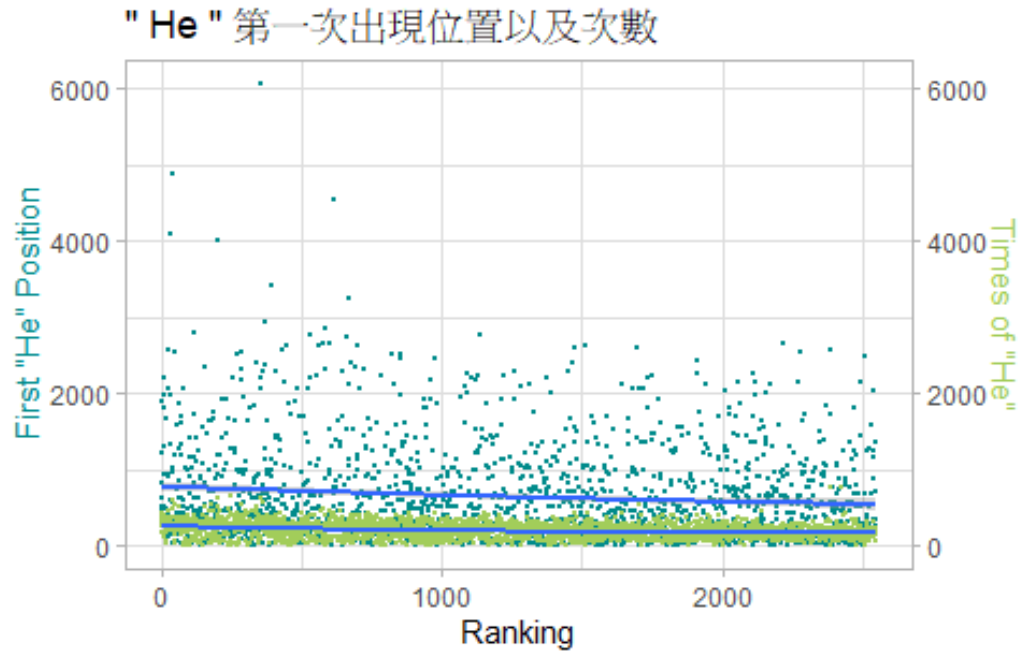
觀看次數高的影片（約前 500 名），I 出現的次數比起其餘影片，有些許較多的趨勢，但相關的關係並不明顯。

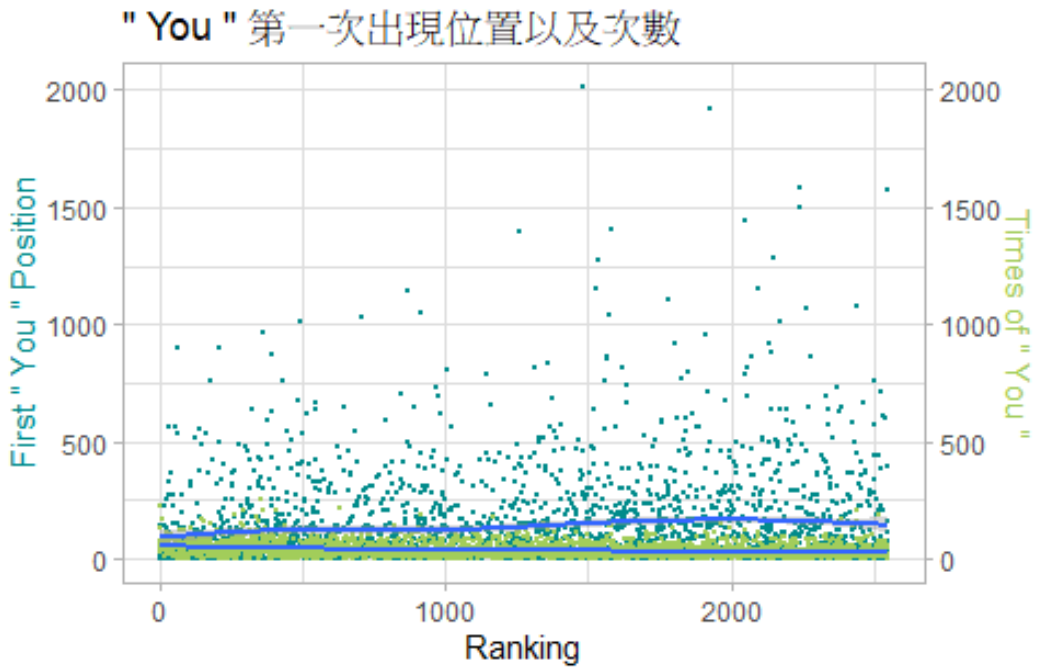


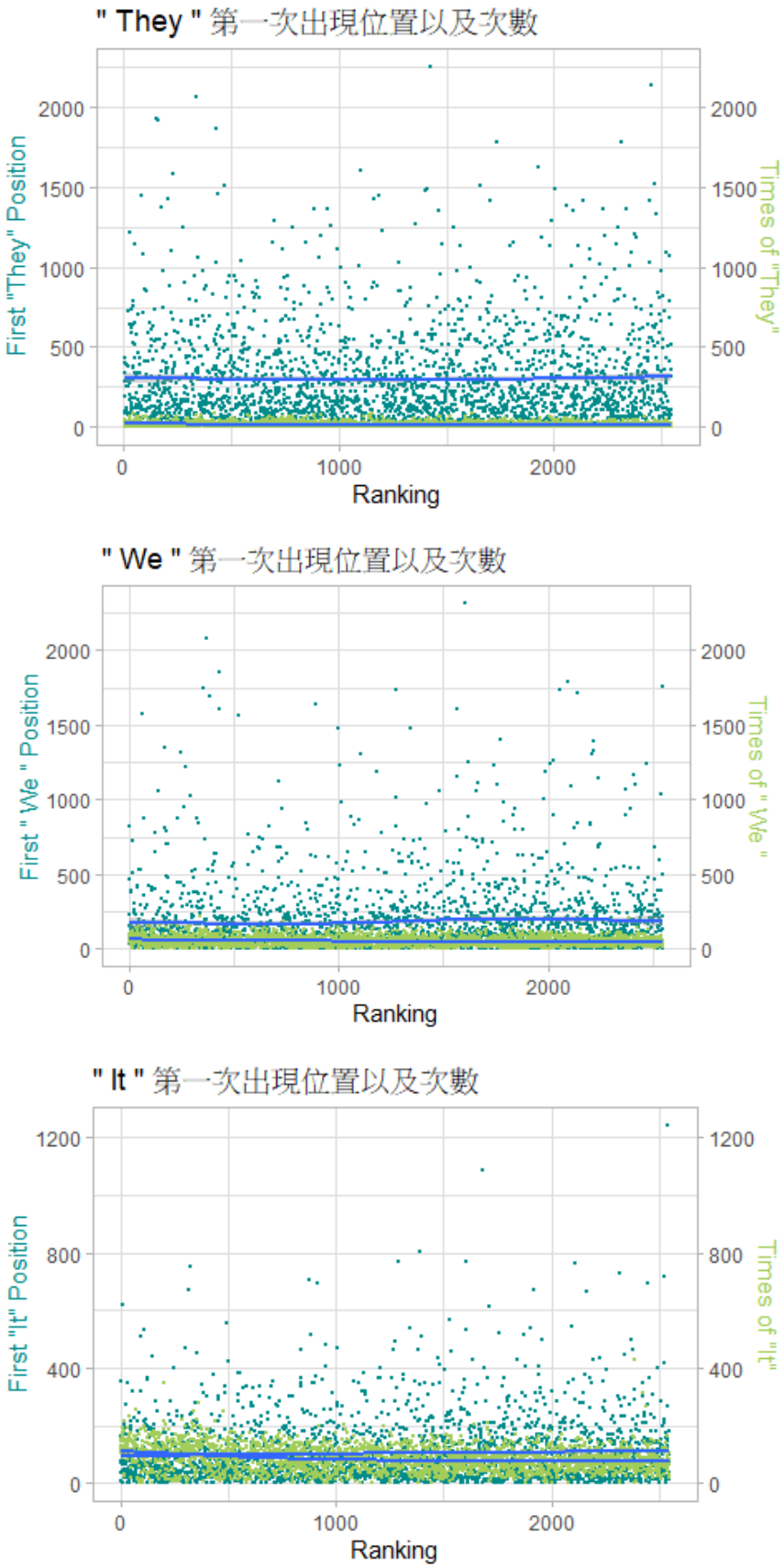
圖九、She／she 在整體 TED talk 中第一次出現的位置和次數

觀看次數高的影片（約前 1,000 名）中，She／she 出現的位置比起其餘影片出現的位置較晚，但一樣相關性不高。

其餘的人稱 He／he、You／you、They／they、We／we 和 It／it，與影片的排名幾乎沒有相關，一併於下方呈現。





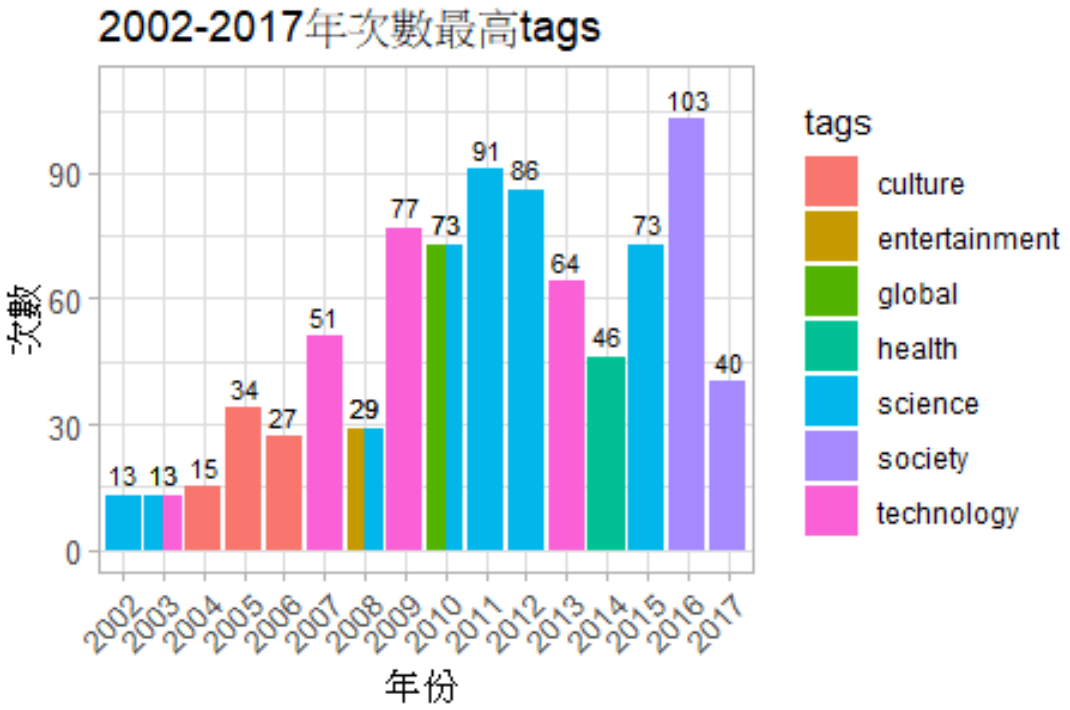


圖十、其餘人稱在整體 TED talk 中第一次出現的位置和次數

3.2tag 相關分析

除了 tag 與觀看次數的關係，我們也想了解 TED talk 歷年的最高使用次數 tag 是否有所變動，以及是否跟當代所關切的事件、氛圍相關。

在 2001 年以前的影片資料相當稀少，參考價值不高，因此我們僅針對 2002 年起至 2017 年 9 月的影片做計算。

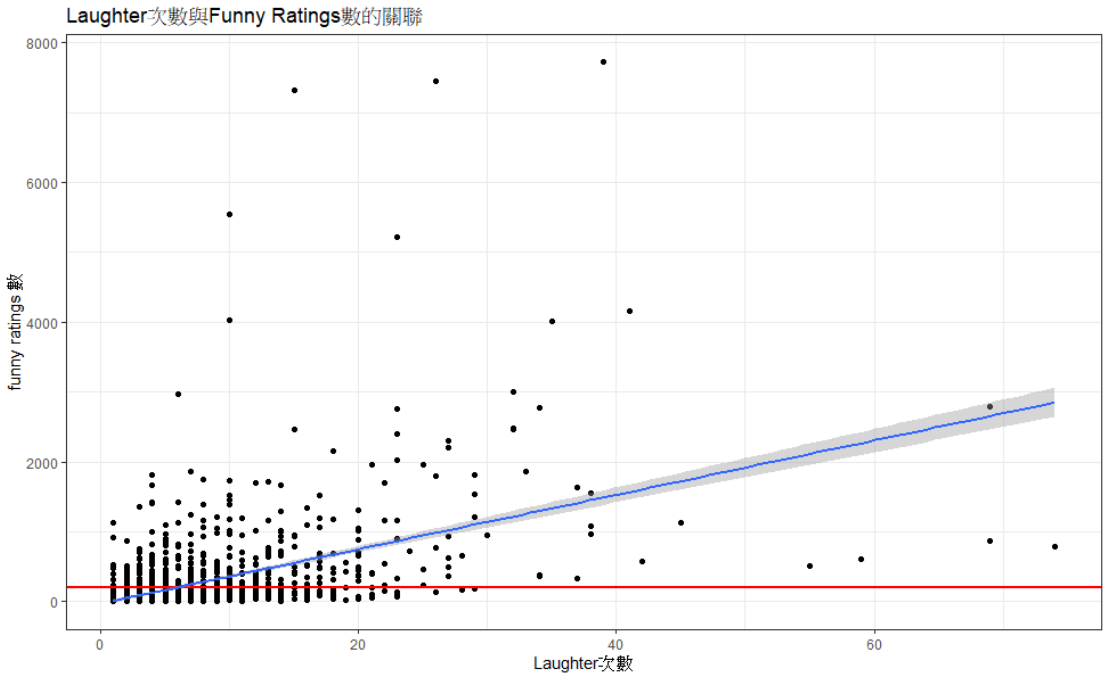


圖十一、2002 至 2017 年最高使用次數 tags

在 2016 前，第一名的多半是科技和科學，到 2016-2017，社會的 tag 升到第一，且 2016 年的次數高於前幾年第一名 tag 的數量。

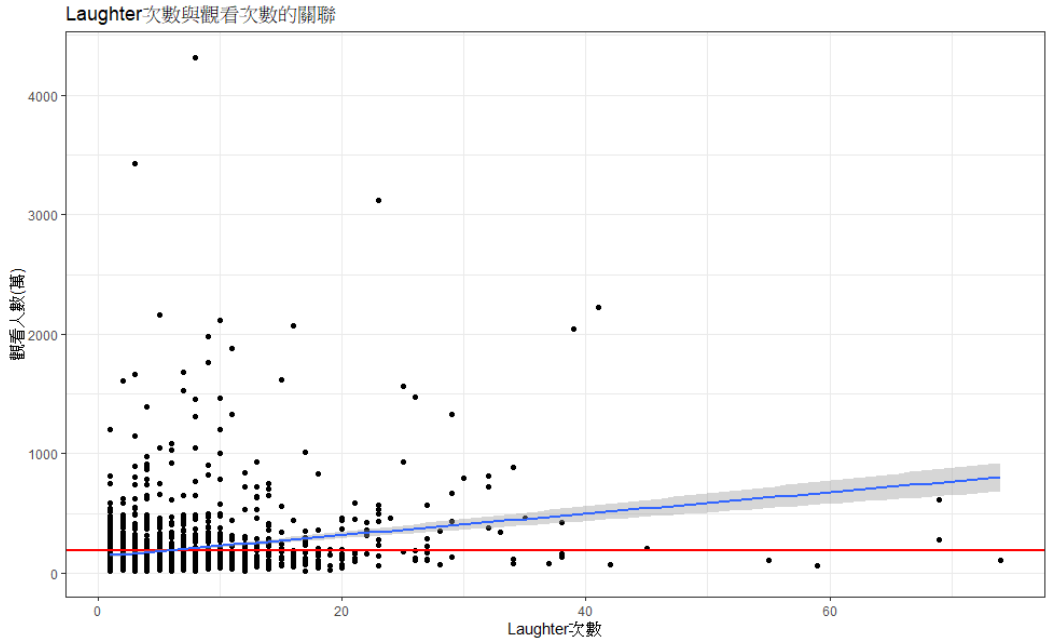
3.3 幽默 TED Talk 分析

在 TED Talk 中只要觀眾有發出笑聲，逐字稿就會以（Laughter）顯示，我們將它視為影片的「幽默」指標來進行分析，以下為研究結果：



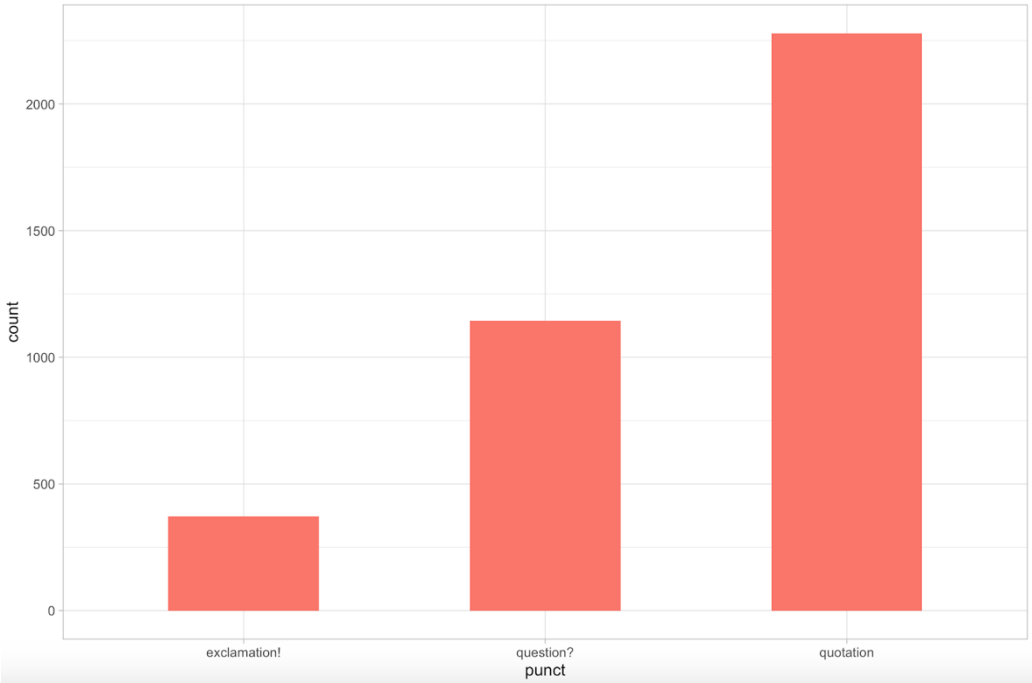
圖十二、Laughter 次數與 Funny Ratings 數的關聯（紅線為平均，藍線為回歸線）

整體而言，Laughter 次數愈高的影片，Funny Ratings 次數愈高，符合我們原先的假設。而且 Laughter 次數高於 30 的影片，其 Funny Ratings 皆高於平均（192.4）。然而經由 lm 函數所畫出來的回歸線來看，相關係數只有 0.2571，表示兩者的關聯沒有很高，我們推測是因為現場觀眾跟事後看影片的觀眾受到環境影響，在現場的觀眾較容易受到氣氛的渲染，反應比較熱烈，才會有這樣的結果差別。



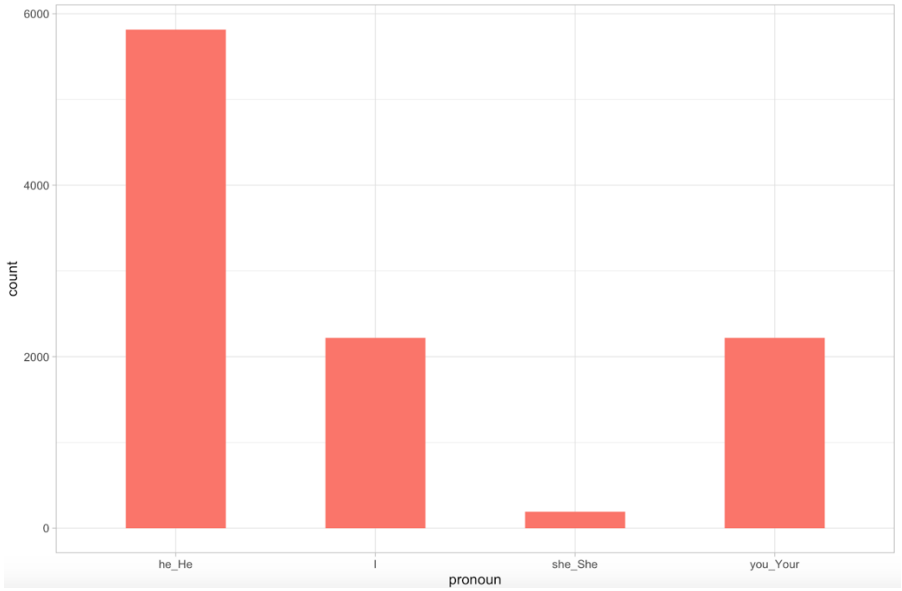
圖十三、Laughter 次數與觀看次數的關聯（紅線為平均，藍線為回歸線）

Laughter 次數跟觀看次數的相關係數只有 0.0507，可以說兩者幾乎沒有任何關聯，表示一個好笑的影片並不會因此被很多人推薦而獲得較高的觀看次數。但如果好笑的影片，標題不吸引人的話，對觀看次數也是會有影響的。



圖十四、Laughter 前的標點符號（！、？、“）

除了上述的分析之外，我們也很好奇是什麼樣的言論才會讓觀眾笑，所以我們擷取了 Laughter 出現前的字句來進行分析。我們發現，Laughter 前出現的標點符號中，以「引號」最多，代表講者可能是在講述某個故事或者是引用他人的話語。



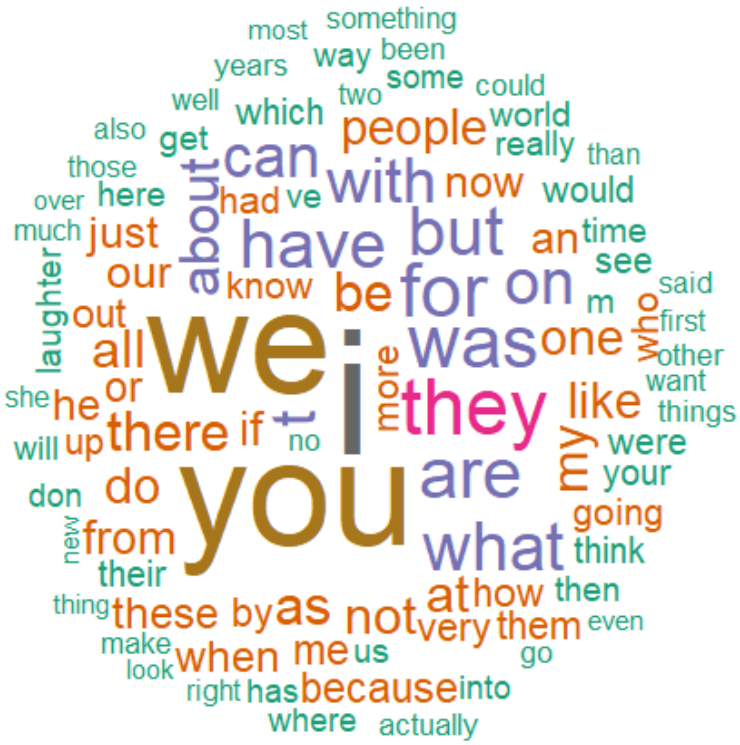
圖十五、Laughter 前的代名詞（He、I、She、You）

繼續上方標點符號分析的發現，我們又做了代名詞的分析。結果發現，Laughter 前面的字句中的男性代名詞(He)遠大於女性(She)。結合標點符號中最多的引號，我們認為這可能跟故事中常出現的角色為男性，以及男性代名詞較為常用有關。

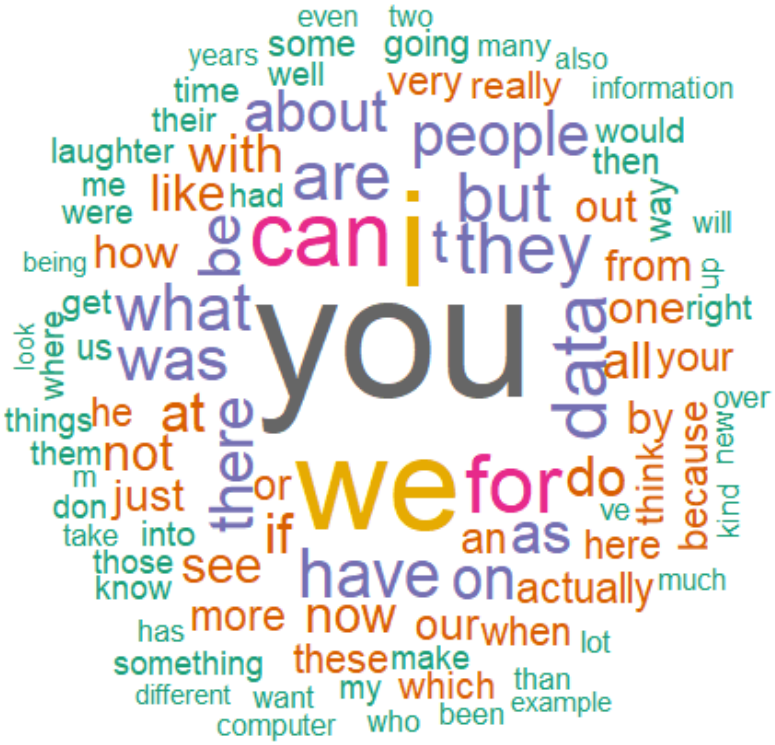
最後，我們也有發現在講者的「起手式」（演講開頭）中，只有 380 篇演講的觀眾在開頭有笑（這裡的開頭我們以前五句話作為標準）。而 2068 名講者的開頭並沒有人笑。

3.4 資料科學家分析

我們對於資料科學家好奇的地方在於，他們講的內容、字彙，是否會因為他們的職業關係而跟其他 TED Talks 有顯著的不同。以下為去掉共同出現次數很高的停用詞後，所做出來的文字雲：



圖十六、所有 Ted Talks 文字雲



圖十七、資料科學家文字雲

	所有 TED Talks： 每篇平均使用次數	資料科學家 TED Talks： 每篇平均使用次數
I	39.5	29.8
You	32.9	41.6
We	32.2	29.7

從以上的圖和表格歸納可以看到，資料科學家在「You」的使用次數，明顯高於「I」、「We」，然後在「Can」的使用次數也較高。而出現於資料科學領域的常用名詞，只有「Data」使用次數較多，其實資料科學家的 TED Talk 內容用字並不會特別艱澀，推測他們的演講內容只有講到關於 Data 的基本概念。

4. 討論與貢獻

海報展之後，統整當天收到的意見，以及事後組員對於內容的檢討，以下我們歸納出幾項之後我們的主題和內容可以改進或更深入探討的事項。

1. 觀看次數與人稱的分析：人稱應該考慮英文人稱的變化，例如受格和所有格，還有縮寫的影響。

2. Tag 相關分析：除了找出歷年流行 tag 的轉變，應加上對於當年或是前幾年討論較熱絡之事件比對是否有相關，同時，此次討論的時間維度太短，僅有 15 年的資料，或許無法做出太多有討論意義的分析。

3. Laughter 次數與觀看次數和 Funny Ratings 的關聯分析：海報上的圖表除了做出散布圖，應該加上回歸線以利判讀兩者之間關係。此闕漏已在本書面報告改善。

4. Punchline 分析：人稱缺少討論 it；除了前一句的人稱外，可再加深找出關鍵使人發笑的句子性質以及內容，做出更關鍵的分析。

5. 資料科學家分析：使用詞彙的探討可以去除人稱進行再次比較，或增加與其他不同職業的影片演講用詞的比較。

6. 分析的結果不顯著的可能原因為 TED Talk 本身有相關的規定，例如影片長度應少於 18 分鐘，導致影片長度的分析結果差異小。

7. 未來可就更多可能與觀看次數以及 TED talk 幽默成份相關的因子進行研究，例如納入拍攝年份、上傳時間長度、講者名氣、講者職業特質等等，讓本研究臻於完善。

這次研究，我們選出了和觀看次數有所關連的一些可能因子進行分析，也嘗試研究幽默在 TED talk 中的特性。我們認為這份研究可能的貢獻如下：

1. 提供高觀看次數之影片特性以利欲產出高觀看演講者參考。

2. 發現逐字稿中附註觀眾笑聲的部份，有助於分析 TED talk 的幽默內容特質，未來研究可就此進行更多關於幽默以及 punchline 的探勘。

5.附錄

5.1 組員分工

學號	系級	姓名	負責項目
B06102001	外文三	陳悅翔	Laughter 前一句話的標點符號、代名詞分析(3.3) 起手式中的 Laughter 分析(3.3)
B07605012	森林二	周昱雯	觀看次數與主題的分析（3.1.1） 觀看次數與人稱的分析（3.1.3） Tag 相關分析（3.2）
B05702109	會計四	盧信呈	各群組標題長度分析 各群組使用問句比例分析 各群組使用 You/Your 及大寫字母比例分析
B08409009	職治一	李昀茜	Laughter 次數與觀看次數和 Funny Ratings 的關聯分析（3.3） 資料科學家分析（3.4）

＊海報及書面報告由四位組員們共同製作