關於TED Talk的5件真相！

研究目標

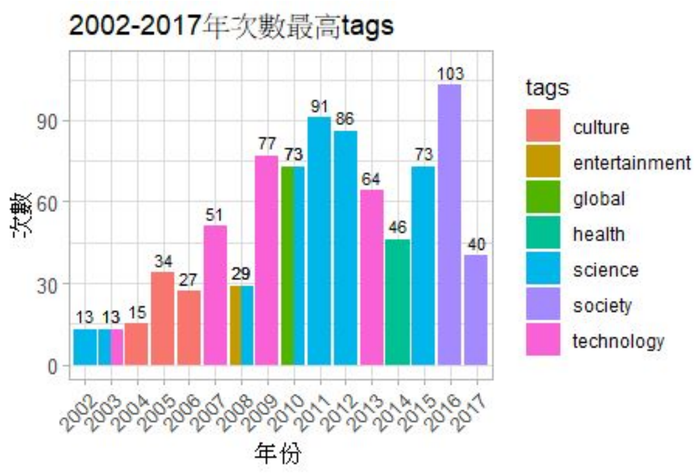
TEDTalk近年來成為指標性的平台，在年輕人中蔚為風潮。本研究致從三個主題來探討TEDTalk，分別為內文笑點分析、觀看次數相關分析、資料科學家的TEDTalk，期盼透過數據及文本分析，檢視大眾對TedTalk的既定印象。我們列出五個研究問題：

1) 觀看次數與「主題」和「標題」分別的關聯性為何？
2) 幽默的TEDTalk的特性為何？
3) 歷年受歡迎的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TEDTalk中偏好使用的字詞

研究方法

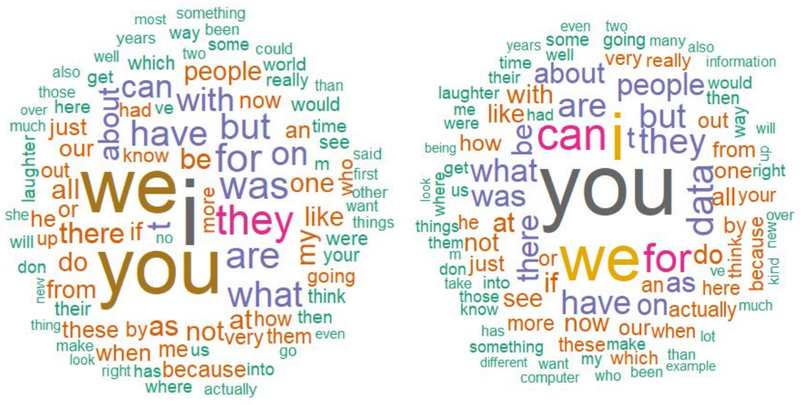
利用網路資料集，用R進行數據及文本分析。
本研究使用quanteda、dplyr、stringr、tidyr、tidytext等多種套件進行資料處理，並使用ggplot2和word cloud進行視覺化。

歷年流行tag流變



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

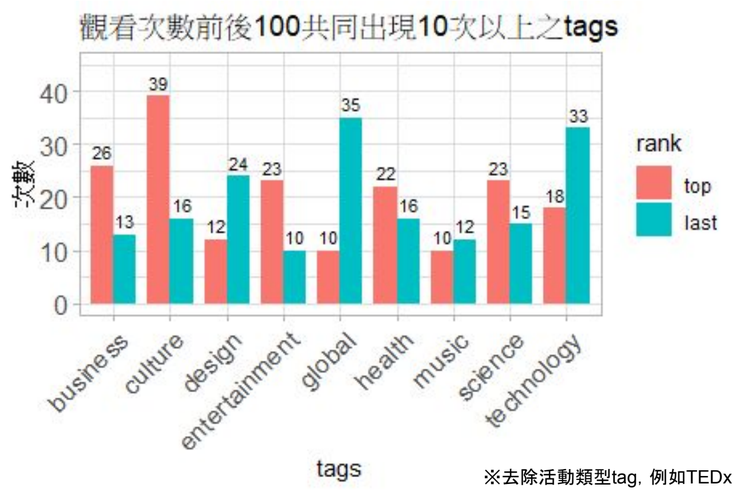
資料科學家文字雲分析



左圖：所有Ted Talk文字雲　右圖：資料科學家Ted Talk文字雲

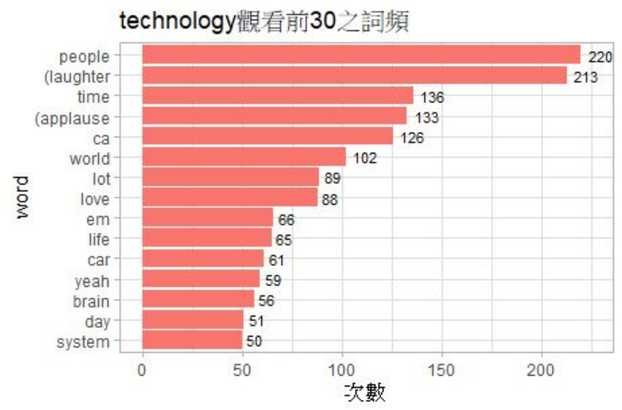
1. 資料科學家在「You」的使用次數，明顯高於「I」、「We」。
2. 資料科學家在「Can」的使用次數較高。
3. 出現於資料科學領域的常用名詞，只有「Data」使用次數較多。

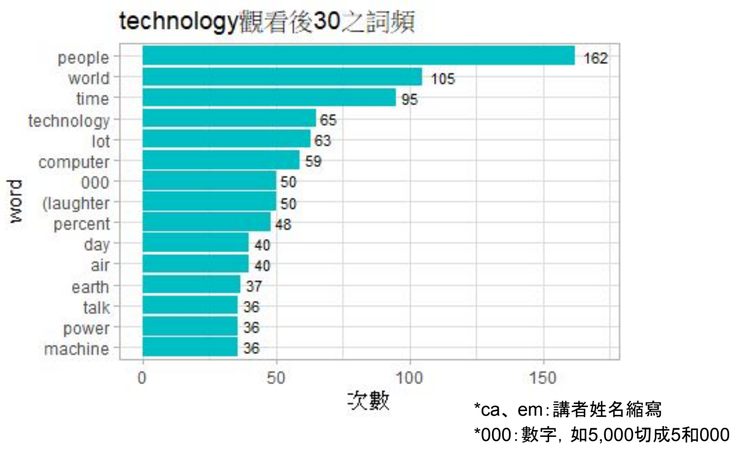
觀看次數相關分析



※去除活動類型tag，例如TEDx

共同tag觀看次數不同之詞頻差異





*ca、em：講者姓名縮寫
*000：數字，如5,000切成5和000

前30名的影片觀眾笑和鼓掌的次數明顯多於後30名的影片；且後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）。

標題分析

	全體	Top 100	Last 100
標題長度平均	6.63	6.56	5.19
中位數	6	6	5

標題的長度在Last100中較短，中位數也較小

開頭	全體	Top 100	Last 100
WHY	5%	9%	1%
WHAT	4.7%	3%	3%
HOW	12%	12%	5%
總和	21.7%	24%	9%

使用問句的情形　Top100中使用問句的情形明顯高Last100，而全體標題使用問句的情形來到21.7%。

標題You／Your 詞頻

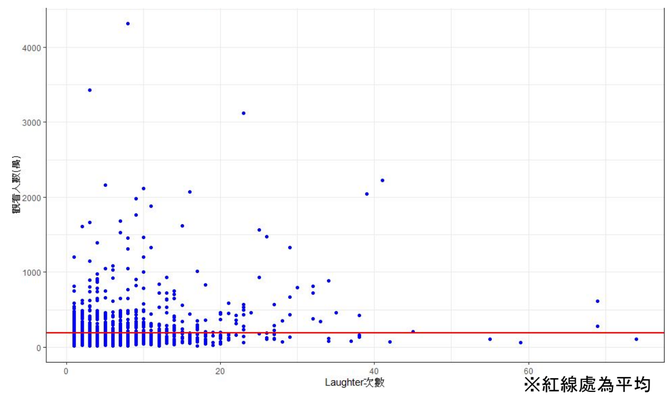
使用「你／你們」的情形 在Top100使用比例較高。

	全體	Top 100	Last 100
佔全體字數比例	1.3%	2.8%	0.6%

標題使用大寫字母的情形

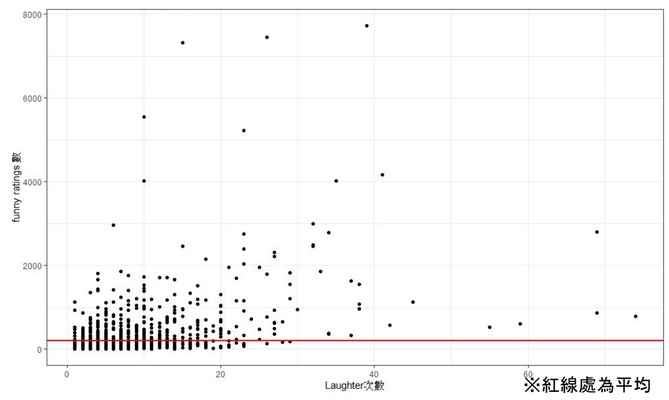
	全體	Top 100	Last 100
佔全體字數比例	18.8%	17.7%	39.3%

幽默 TED Talk 相關分析



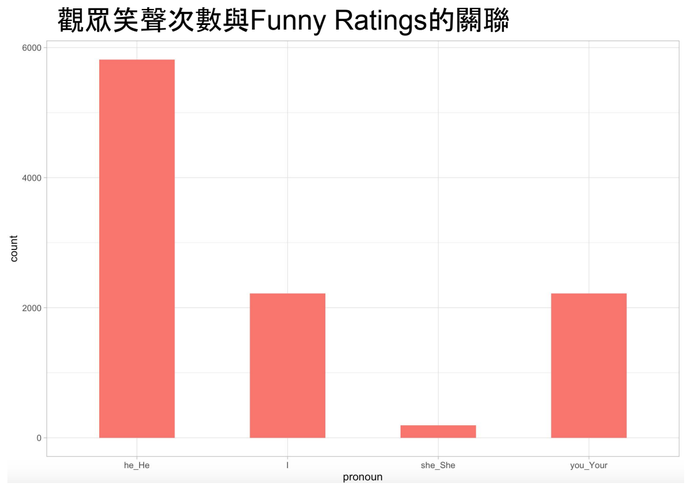
※紅線處為平均

觀眾笑聲次數與觀看次數的關聯



※紅線處為平均

觀眾笑聲次數與Funny Ratings的關聯



觀眾笑聲前一句(punchline)中的代名詞

結論

針對我們提出的問題，整理出的結論：

1)　標題使用問句的數量有一定比例，這當中How的使用最多。我們也發現熱門的影片標題用了比較多「你」，不熱門影片標題有比較多的大寫專有名詞。

2)　觀眾笑聲前一句話中出現He的頻率遠大於She，推測是因為敘事時通常都使用He作為代名詞。而觀眾笑聲次數跟Funny_Ratings大致呈正相關，符合我們原先的想法。

3)　資料科學家的常用字詞並不會比較艱澀，推測他們的演講內容多為資料科學的推廣。

資料來源

Kaggle dataset　TED Talks
TED Talk data collected from TED.com.
Collected until Sep. 21st, 2017

課名／聯絡資料

R語言與資料科學導論
Introduction to Data Science with R

授課老師 謝舒凱
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