**任選2部電影，利用文字探勘方式探討評論，畫出兩部電影評論的文字雲，比較不同滿意指標或不同分數的評論文字情緒。**

# 載入要用到的套件

library(tm)

# install.packages("proxy")

library(proxy)

# install.packages("tidytext")

library(tidytext)

library(wordcloud2)

library(syuzhet)

# 載入檔案

setwd("C:\\Users\\ASUS\\Desktop\\五234 R\\HW5") #放你的路徑

data <- read.csv("IMDb\_Feature Film\_2022\_review\_data.csv")

data <- na.omit(data)

**選兩部分數、評論數較為相似的電影：Scream & The Adam Project**

# 第一部電影 Scream

data\_s <- subset(data,Title=="Scream")

# 創建 Corpus

x <- Corpus(VectorSource(data\_s$Review))

# 清理文本

x <- tm\_map(x, content\_transformer(tolower))

x <- tm\_map(x, removePunctuation)

# 針對 Scream 刪掉一些評論中常出現但無含意的字詞

myStopWords <- c(stopwords("english"),"just", "even", "will", "much",

"also", "still", "one", "can", "really") #remove words

x <- tm\_map(x, removeWords, myStopWords)

# 建立 Term-Document Matrix

x\_tdm <- TermDocumentMatrix(x)

# 轉換為矩陣

review\_mx <- as.matrix(x\_tdm)

# Sum rows and frequency data frame

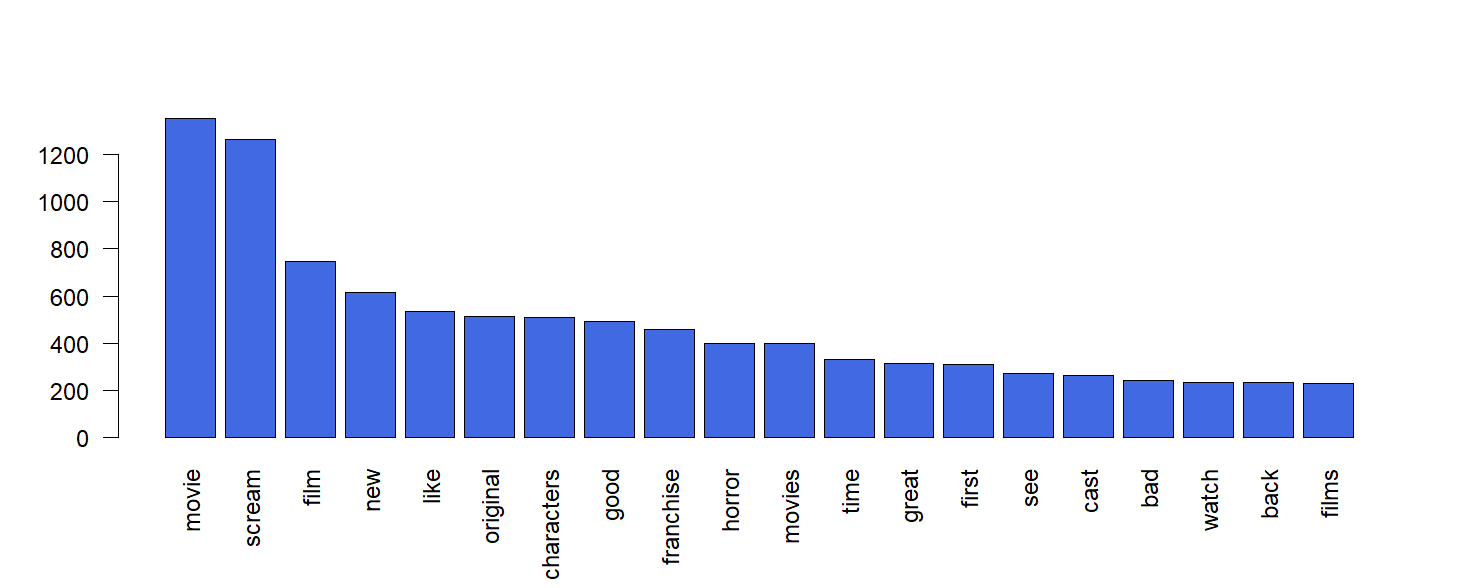
freq\_dfx <- rowSums(review\_mx)

# Sort term\_frequency in descending order

freq\_dfx <- sort(freq\_dfx, decreasing = T)

# View the top 20 most common words

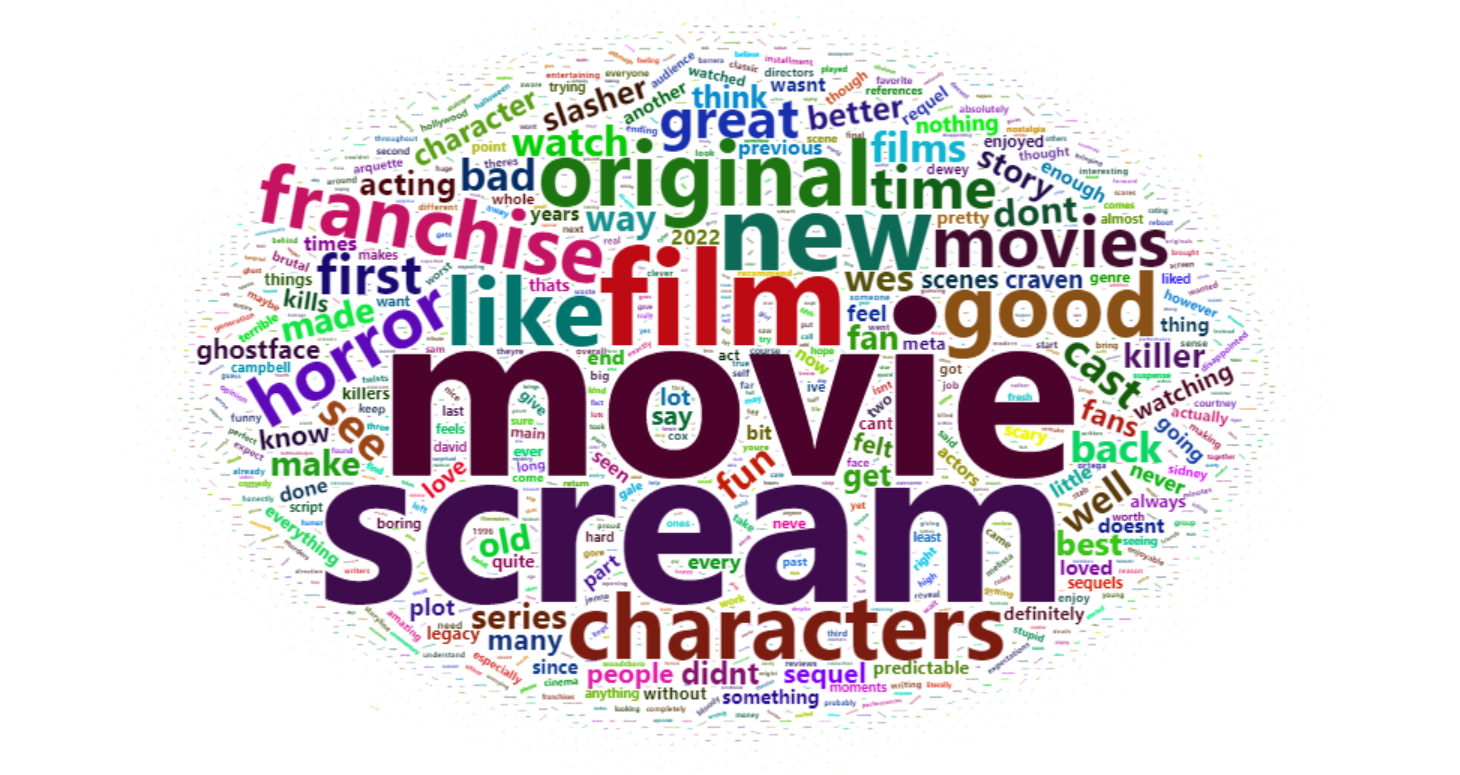
barplot(freq\_dfx[1:20], col = "royalblue", las = 2)



freq\_dfx <- data.frame(word = names(freq\_dfx), num = freq\_dfx)

# 生成文字雲

wordcloud2(freq\_dfx, size = 0.9)



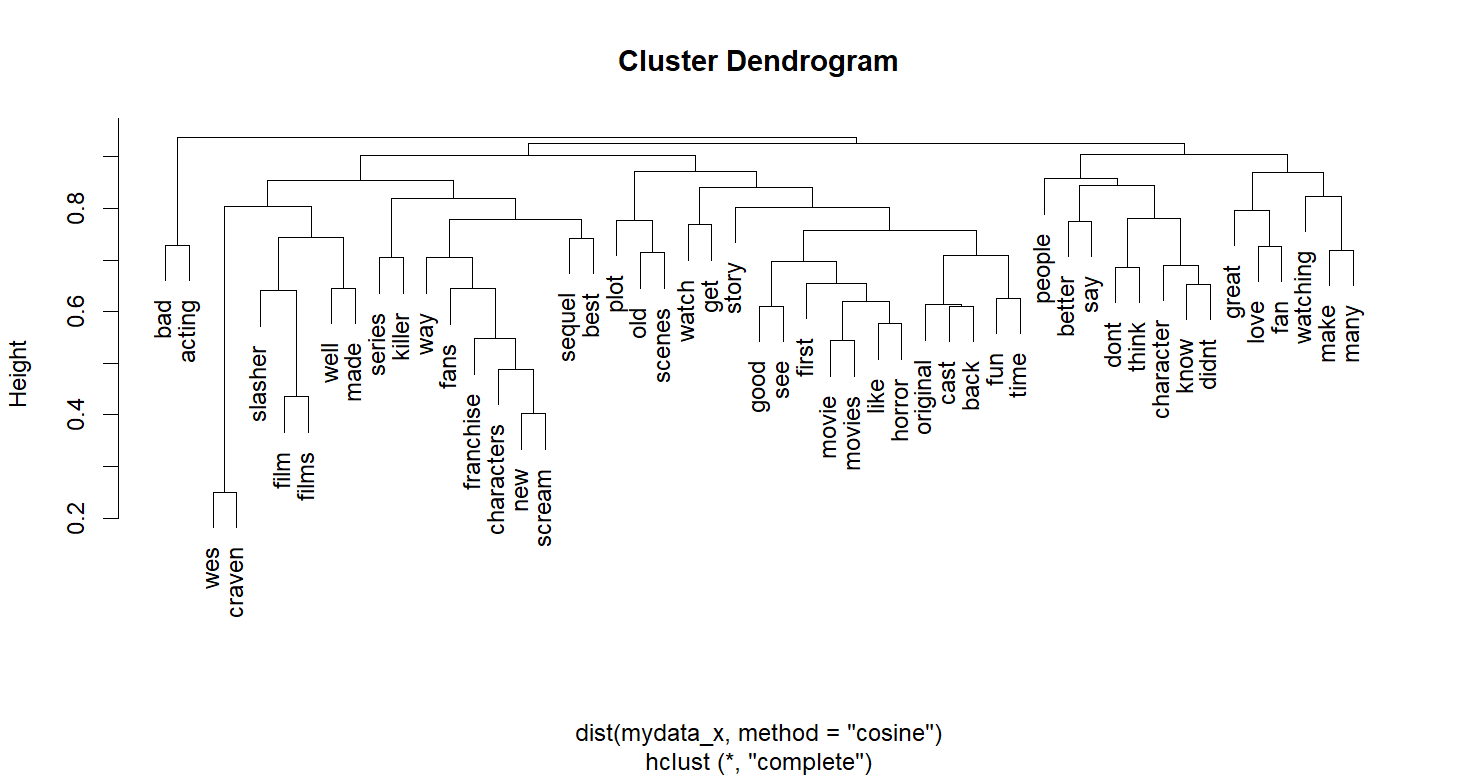
# 計算相似度

x\_tdm2 <- removeSparseTerms(x\_tdm, sparse = 0.9)

mydata\_x <- as.data.frame(as.matrix(x\_tdm2))

hc\_x <- hclust(d=dist(mydata\_x, method="cosine"), method="complete")

plot(hc\_x)



#sentiments

get\_sentiments("bing")

bing\_word\_countsx <- freq\_dfx %>%

inner\_join(get\_sentiments("bing"))

bing\_word\_countsx

table(bing\_word\_countsx$sentiment)

bing\_word\_countsx %>%

filter(sentiment == "positive") %>%

select(word,num)%>%

wordcloud2()

# 第二部電影 The Adam Project

data\_a <- subset(data,Title=="The Adam Project")

y <- Corpus(VectorSource(data\_a$Review))

y <- tm\_map(y, content\_transformer(tolower))

y <- tm\_map(y, removePunctuation)

myStopWords <- c(stopwords("english"), "just", "can", "really", "will",

"still","one","every","even","much" )

y <- tm\_map(y, removeWords, myStopWords)

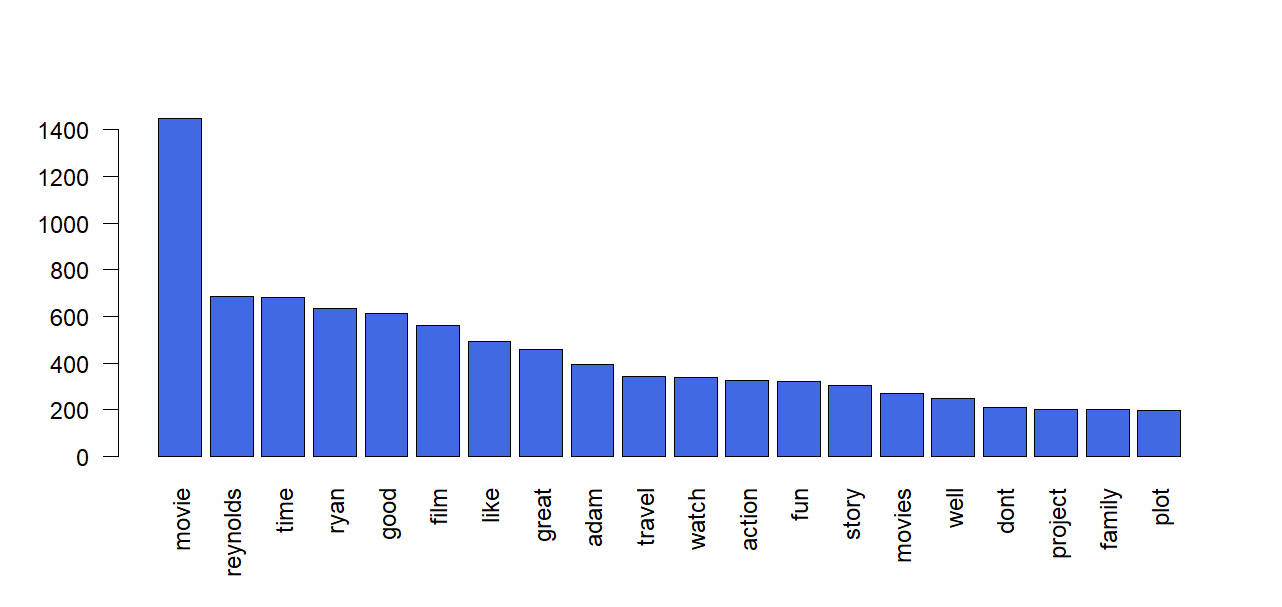
y\_tdm <- TermDocumentMatrix(y)

review\_my <- as.matrix(y\_tdm)

freq\_dfy <- rowSums(review\_my)

freq\_dfy <- sort(freq\_dfy, decreasing = T)

barplot(freq\_dfy[1:20], col = "royalblue", las = 2)



freq\_dfy <- data.frame(word = names(freq\_dfy), num = freq\_dfy)

wordcloud2(freq\_dfy, size = 0.9)

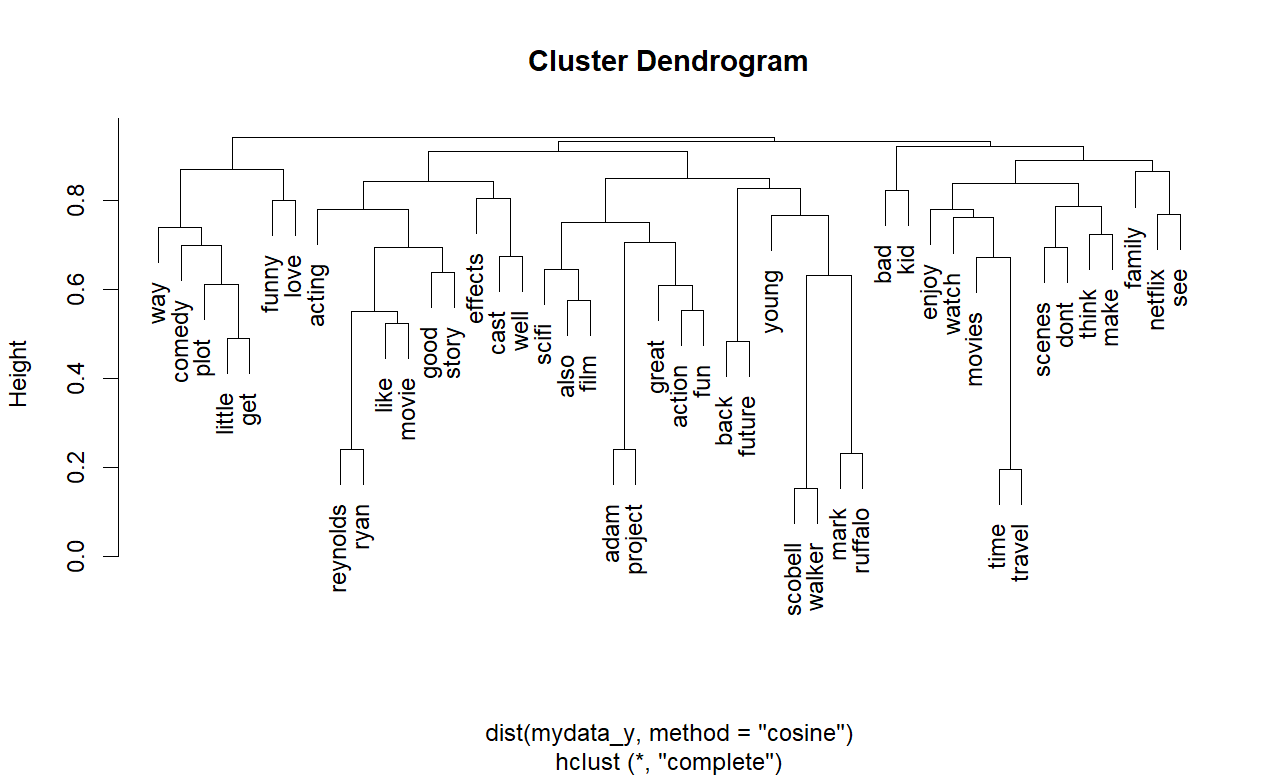


y\_tdm2 <- removeSparseTerms(y\_tdm, sparse = 0.9)

mydata\_y <- as.data.frame(as.matrix(y\_tdm2))

hc\_y <- hclust(d = dist(mydata\_y, method = "cosine"), method = "complete")

plot(hc\_y)



bing\_word\_countsy <- freq\_dfy %>%

inner\_join(get\_sentiments("bing"))

bing\_word\_countsy

table(bing\_word\_countsy$sentiment)

bing\_word\_countsy %>%

filter(sentiment == "positive") %>%

select(word,num)%>%

wordcloud2()

# 再進一步分析Review Sentiment和Satisfaction Index的關係

# 第一部電影

# 執行情感分析，將情感分數添加到 movie\_data

data\_s$Sentiment\_Score <- get\_sentiment(data\_s$Review, method = "afinn")

# 第二部電影

data\_a$Sentiment\_Score <- get\_sentiment(data\_a$Review, method = "afinn")

# 繪製 Review 和 Satisfaction\_Index 的散點圖

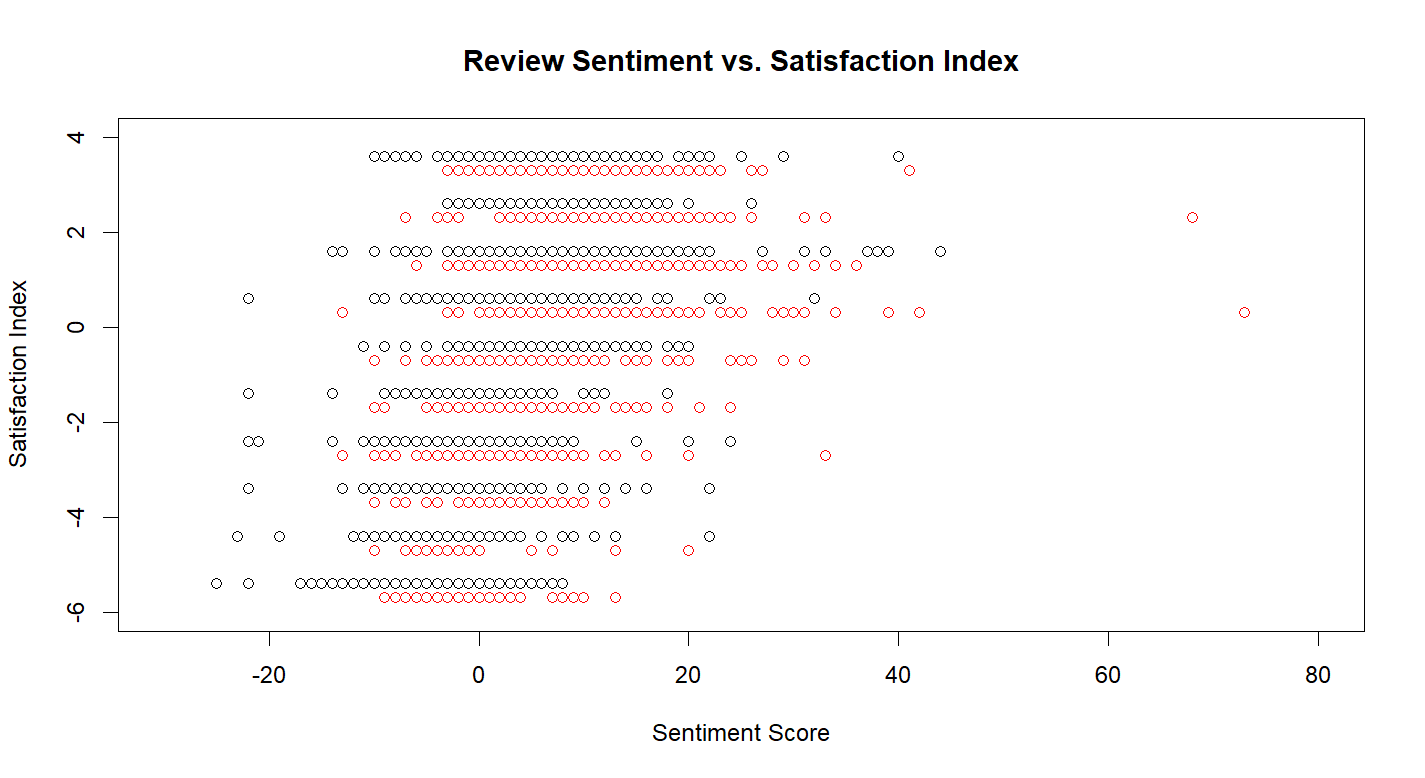
plot(data\_s$Sentiment\_Score, data\_s$Satisfaction\_Index,

xlab = "Sentiment Score", ylab = "Satisfaction Index",

main = "Review Sentiment vs. Satisfaction Index",

xlim = c(-30, 80), ylim = c(-6, 4))

points(data\_a$Sentiment\_Score, data\_a$Satisfaction\_Index, col = "red")



**針對兩部電影文字雲的部分，將”movie”和”film”等代表電影的字彙刪除（經常出現在評論中，但不具正負面意涵），並將電影名稱中包含的字彙刪除（Sceram、Adam、Project），再重新生成一次文字雲（只做文字雲的部分，不做後續分群及情緒分析），以便做進一步分析**

|  |  |
| --- | --- |
| **Scream** | **The Adam Project** |
|  |  |
| **是一個系列電影（franchise、series），但具有其原創性（new、original）且是一部有殺手和鬼臉出現的恐怖電影（killer、ghostface、horror）** | **應該有萊恩雷洛斯參演（ryan、reynolds），並且是一部關於穿越時空的科幻電影(travel、time、future、scifi)，能在網飛（netflix）上看到** |
| **兩部電影皆為正評居多（like、good、great、fun、enjoy、love、well）但仍帶有些負評（bad），與其電影分數的狀況相符（Scream 6.4、The Adam Project 6.7），此外還可看出對評論者來說電影卡司（character）很重要，在兩部影評中都大量被提及。** | |
|  |  |
| 1. 部份的人認為演技不好 2. 衛斯克萊文製作這部跟鐮刀有關的電影受好評 3. 被認為是Scream系列電影最好的續集 4. 是留言者看或喜歡的第一部恐怖電影 5. 前面系列的演員有回歸演出 | 1. 觀影者喜歡這部有趣的喜劇電影 2. 喜歡這部電影的人是因為萊恩雷洛斯的演出及好故事 3. Walker Scobell和Mark Ruffalo出演的角色與時間旅行（back、future）有關 4. 喜歡這部時空旅行的電影 5. 覺得不合理 6. 可以和家人在netflix上觀看 |
|  |  |
| 評分相當的兩部電影，文字情緒也相當，甚至常出現的字彙也差度多（good、like、great、well、love、fun、instersting、awesome） | |
|  | |
| 從散布圖可以看出Satisfaction越低的評論Sentiment越低，兩者間大致呈現正相關。  再分析兩部電影的結果，明顯第一部電影的Sentiment score偏左，散佈在-30~50之間。  而第二部電影的Sentiment score較偏向右邊，散佈範圍在-20~80之間，右偏極值較多，Satisfaction低的評論者也明顯較少。 | |