HW5

2024-12-31

作業目標

請任選一位 Youtuber，針對他的影片簡介描述進行以下分析：

1. **文字清理**：
   - 移除不必要的符號、數字及停用詞。
   - 保留關鍵的中文字詞。
2. **文字探勘**：
   - 繪製文字雲。
   - 進行詞頻分析，了解該 Youtuber 使用字詞的習慣。

```
# Import the data
library(readr)
ytvideo <- read_csv("C:/Users/Ava/Desktop/R/HW5/ytvideo.csv")
```

```
## Rows: 314 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): yt, title
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(ytvideo)
```

```
# import the packages
library(dplyr)
```

```
##
## 載入套件：'dplyr'
```

```
## 下列物件被遮斷自 'package:stats':
##
##     filter, lag
```

```
## 下列物件被遮斷自 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: 套件 'ggplot2' 是用 R 版本 4.4.2 來建造的
```

```
library(tm)
```

```
## Warning: 套件 'tm' 是用 R 版本 4.4.2 來建造的
```

```
## 載入需要的套件：NLP
```

```
## Warning: 套件 'NLP' 是用 R 版本 4.4.2 來建造的
```

```
##
## 載入套件：'NLP'
```

```
## 下列物件被遮斷自 'package:ggplot2':
##
##     annotate
```

```
library(wordcloud)
```

```
## Warning: 套件 'wordcloud' 是用 R 版本 4.4.2 來建造的
```

```
## 載入需要的套件：RColorBrewer
```

```
library(RColorBrewer)
```

```
# classification
ytvideo_filtered <- ytvideo %>%
  filter(yt == "蔡阿嘎") %>%
  select(title)
```

這裡我做了兩個版本，一個有用到中文分詞套件，一個沒有。 以下為沒有用到中文分詞套件的版本。

```
# data cleaning
text_data <- paste(ytvideo_filtered$title, collapse = " ")

clean_text <- text_data %>%
  tolower() %>%
  gsub("[[:punct:]]", " ", .) %>%
  gsub("[^\\p{Han}]", " ", ., perl = TRUE) %>%
  gsub("[0-9]", " ", .) %>%
  gsub("[\\r\\n]", " ", .) %>%
  gsub("\\s+", " ", .) %>%
  trimws()

word_tokens <- unlist(strsplit(clean_text, "\\s+"))
```

```
# remove stop words
stop_words <- c("的", "是", "了", "在", "我", "也", "和", "有", "這", "他", "她", "就", "不")
filtered_words <- word_tokens[!word_tokens %in% stop_words]
filtered_words
```

```
##   [1] "瘋狂"                          "小時挑戰賽"
##   [3] "台南人都吃這個"                "直接問台南人"
##   [5] "出賣"                          "間在地美食"
##   [7] "蔡阿嘎"                        "馬叔叔"
##   [9] "食尚玩嘎"                      "澳門賭場上線啦"
##  [11] "蔡阿嘎賭輸賭光流落街頭"        "食尚玩嘎"
##  [13] "到日本環球影城"                "買破萬元限定商品送大家"
##  [15] "瘋狂"                          "小時挑戰賽"
##  [17] "新竹美食沙漠"                  "直接問路人什麼好吃"
##  [19] "就連去吃"                      "間"
##  [21] "蔡阿嘎"                        "馬叔叔"
##  [23] "嘎慶君遊台灣"                  "元吃一餐"
##  [25] "元在北港玩一整天"              "瘋狂"
##  [27] "小時挑戰賽"                    "一日五塔"
##  [29] "環島台灣五極點"                "蔡阿嘎"
##  [31] "馬叔叔"                        "水陸鞋"
##  [33] "瘋狂"                          "小時挑戰賽"
##  [35] "一天來回東京"                  "完成"
##  [37] "個願望"                        "蔡阿嘎"
##  [39] "馬叔叔"                        "瘋狂"
##  [41] "小時挑戰賽"                    "征服"
##  [43] "家"                            "高雄大王"
##  [45] "蔡阿嘎"                        "馬叔叔"
##  [47] "三菱汽車"                      "食尚玩嘎"
##  [49] "不排隊不花錢玩東京迪士尼"      "蔡阿嘎"
##  [51] "招懶人玩法"                    "食尚玩嘎"
##  [53] "帛琉"                          "蔡阿嘎出發鞏固邦交國"
##  [55] "瘋狂"                          "小時挑戰賽"
##  [57] "吃"                            "間"
##  [59] "嘉義雞肉飯"                    "蔡阿嘎"
##  [61] "馬叔叔"                        "嘎慶君遊台灣"
##  [63] "花蓮好山好水"                  "蔡阿嘎私藏的深度玩法"
##  [65] "食尚玩嘎"                      "沖繩說走就走"
##  [67] "跟著蔡阿嘎從北玩到南"          "嘎慶君遊台灣"
##  [69] "慢城嘉義市"                    "蔡阿嘎帶你品嚐平價美食天堂"
##  [71] "日本自由行不用怕"              "大實用旅遊漢字"
##  [73] "蔡阿嘎來教你"                  "食尚玩嘎"
##  [75] "韓國首爾"                      "女人天堂"
##  [77] "男人地獄"                      "蔡阿嘎厭世代表作"
##  [79] "日本超級市場"                  "大必買好物"
##  [81] "蔡阿嘎真心推薦"                "食尚玩嘎"
##  [83] "桃園是美食沙漠"                "觀光墳場"
##  [85] "蔡阿嘎帶您破解"                "食尚玩嘎"
##  [87] "來香港一定要體驗的"            "件事"
##  [89] "蔡阿嘎真心不騙"                "日本超商"
##  [91] "大必買零食飲料"                "蔡阿嘎真心推薦"
##  [93] "食尚玩嘎"                      "澳洲黃金海岸"
##  [95] "凱恩斯"                        "跟蔡阿嘎飛到南半球曬太陽"
##  [97] "食尚玩嘎"                      "京都大阪篇"
##  [99] "蔡阿嘎教你日本旅遊的"          "大禁忌"
## [101] "食尚玩嘎"                      "一定要帶女朋友去約會的"
## [103] "東京"                          "大動漫主題樂園"
## [105] "跟著蔡阿嘎"                    "魔獸世界"
## [107] "玩正港台灣內地"                "南投"
## [109] "食尚玩嘎"                      "玩日本福岡只要"
```

```
## [111] "元"                                  "到處都是熊本熊"
## [113] "食尚玩嘎"                            "台東好山好水好好玩"
## [115] "蔡阿嘎帶你悄悄避開觀光客"            "食尚玩嘎"
## [117] "基隆不去流俗廟口"                    "蔡阿嘎帶你走訪"
## [119] "個巷弄美食景點"                      "食尚玩嘎"
## [121] "日本東京熱七天"                      "蔡阿嘎出國處男秀"
## [123] "食尚玩嘎"                            "蘭嶼"
## [125] "放鬆靈魂的遺世"                      "蔡阿嘎"
## [127] "馬叔叔"                              "食尚玩嘎"
## [129] "山海河港大高雄"                      "打狗先生"
## [131] "蔡阿嘎"                              "馬叔叔"
## [133] "首支單曲"                            "食尚玩嘎"
## [135] "彰化大佛故鄉有三寶"                  "跟著蔡阿嘎吃到飽嘟嘟"
## [137] "食尚玩嘎"                            "蔡阿嘎熱血七日機車環島台灣"
## [139] "去一輩子只會去一次的地方"            "食尚玩嘎"
## [141] "悠哉宜蘭城"                          "跟蔡阿嘎到蘭陽平原放空去"
## [143] "食尚玩嘎"                            "蛻變大台中"
## [145] "蔡阿嘎帶你探索都市與文化大熔爐"      "食尚玩嘎"
## [147] "屏東阿猴城"                          "猴囝仔蔡阿嘎來避寒囉"
## [149] "食尚玩嘎"                            "發現戀戀金門"
## [151] "跟蔡阿嘎大啖好菜好酒好風光"          "食尚玩嘎"
## [153] "新竹風城當追風少年"                  "不吃米粉貢丸"
## [155] "蔡阿嘎帶你吃巷內好味"                "食尚玩嘎"
## [157] "澎湖不防曬"                          "曬成黑面蔡阿嘎"
## [159] "食尚玩嘎"                            "板橋殘酷擂台"
## [161] "蔡阿嘎中肯挑戰全台灣人口第一區"      "食尚玩嘎"
## [163] "雲林來瘋"                            "蔡阿嘎報你美食瘋雲林"
## [165] "食尚玩嘎"                            "台南古都美食文化天堂"
## [167] "跟蔡阿嘎肚裡一起撐船"                "食尚玩嘎"
## [169] "苗栗客家庄大喀草莓"                  "蔡阿嘎"
## [171] "恁仔細承蒙您"                        "恁久好否"
## [173] "食尚玩嘎"                            "馬祖勇士們教蔡阿嘎的愛國熱血"
## [175] "食尚玩嘎"                            "蔡阿嘎要顛覆你對嘉義的想像"
## [177] "食尚玩嘎"                            "蔡阿嘎教你花"
## [179] "元去高雄義大世界當大爺"              "食尚玩嘎"
## [181] "跟蔡阿嘎到北投一日在地輕旅行"        "食尚玩嘎"
## [183] "蔡阿嘎推薦你花蓮不流俗在地玩法"
```

從這個分詞沒有比較仔細的資料中可以看出，蔡阿嘎的影片可能以旅遊型的為主。

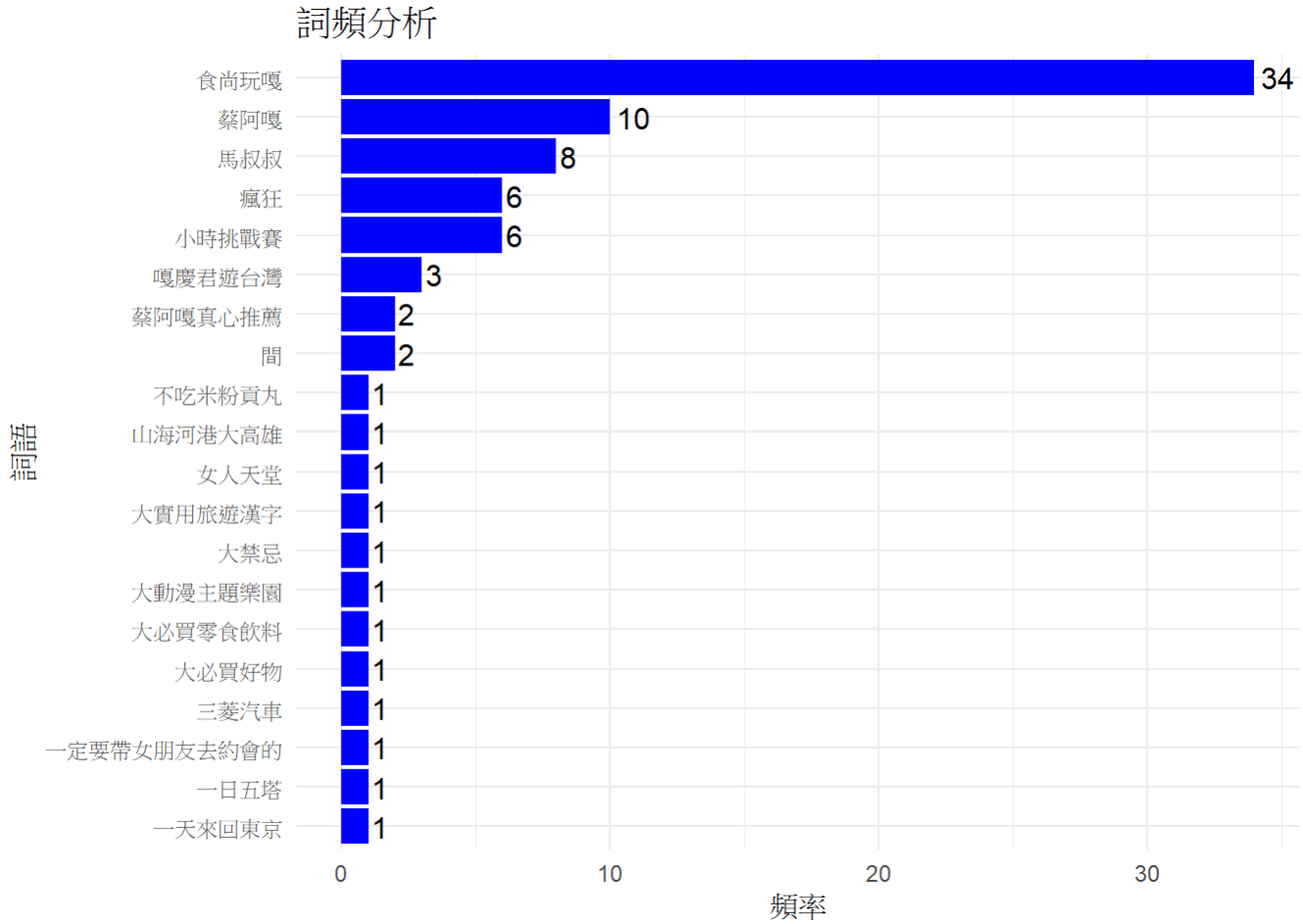
```
# wordcloud
word_freq <- table(filtered_words)
wordcloud(names(word_freq),
          freq = word_freq,
          min.freq = 2,
          random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))
```

因為分詞沒有整理得太乾淨，所以這裡的文字雲資料顯示的並不多，不過從這個可以看的出來: 1.主題很明顯是以"食尚玩嘎"為標題 2.蔡阿嘎本人的名字也常出現在影片標題中。 3.形式首要是以小時挑戰賽進行，真心推薦為次要。 4.瘋狂出現了好幾次，可能這個影片不像正常旅遊行程。 5.馬叔叔出現了好幾次，經查詢後得知為另一名youtuber，他們可能常常出合作影片。

```
# visualization
word_freq_df <- as.data.frame(word_freq, stringsAsFactors = FALSE)
colnames(word_freq_df) <- c("word", "freq")

top_words <- word_freq_df %>%
  arrange(desc(freq)) %>%
  head(20)

ggplot(top_words, aes(x = reorder(word, freq), y = freq)) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = freq), hjust = -0.2, size = 4) +
  coord_flip() +
  labs(title = "詞頻分析", x = "詞語", y = "頻率") +
  theme_minimal()
```



以上圖表為字詞出現頻率的長條圖，和文字雲相比，長條圖可以更明顯從圖表中看出實際出現幾次，但文字雲很美觀，可以用在設計和行銷上。

接下來是有用到中文分詞套件的版本

```
# data cleaning
text_data <- paste(ytvideo_filtered$title, collapse = " ")

clean_text <- text_data %>%
  tolower() %>%
  gsub("[[:punct:]]", " ", .) %>%
  gsub("[^\\p{Han}]", " ", ., perl = TRUE) %>%
  gsub("[0-9]", " ", .) %>%
  gsub("[\\r\\n]", " ", .) %>%
  gsub("\\s+", " ", .) %>%
  trimws()

library(jiebaR)
```

```
## Warning: 套件 'jiebaR' 是用 R 版本 4.4.2 來建造的
```

```
## 載入需要的套件：jiebaRD
```

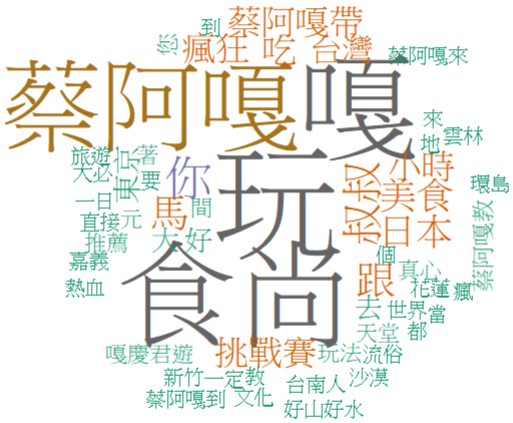
```
cutter <- worker()
word_tokens <- cutter[clean_text]
```

```
# remove stop words
stop_words <- c("的", "是", "了", "在", "我", "也", "和", "有", "這", "他", "她", "就", "不")
filtered_words <- word_tokens[!word_tokens %in% stop_words]
filtered_words
```

```
##   [1] "瘋狂"       "小時"       "挑戰賽"     "台南人"     "都"         "吃"
##   [7] "這個"       "直接"       "問"         "台南人"     "出賣"       "間"
##  [13] "地"         "美食"       "蔡阿嘎"     "馬"         "叔叔"       "食尚"
##  [19] "玩"         "嘎"         "澳門"       "賭場"       "上線"       "啦"
##  [25] "蔡阿嘎"     "賭輸"       "賭光"       "流落"       "街頭"       "食尚"
##  [31] "玩"         "嘎"         "到"         "日本"       "環球"       "影城"
##  [37] "買破"       "萬元"       "限定"       "商品"       "送"         "大家"
##  [43] "瘋狂"       "小時"       "挑戰賽"     "新竹"       "美食"       "沙漠"
##  [49] "直接"       "問路"       "人"         "什麼"       "好吃"       "連去"
##  [55] "吃"         "間"         "蔡阿嘎"     "馬"         "叔叔"       "嘎慶君遊"
##  [61] "台灣"       "元"         "吃"         "一餐"       "元"         "北港"
##  [67] "玩"         "一整天"     "瘋狂"       "小時"       "挑戰賽"     "一日"
##  [73] "五塔"       "環島"       "台灣"       "五"         "極點"       "蔡阿嘎"
##  [79] "馬"         "叔叔"       "水陸"       "鞋"         "瘋狂"       "小時"
##  [85] "挑戰賽"     "一天"       "來回"       "東京"       "完成"       "個"
##  [91] "願望"       "蔡阿嘎"     "馬"         "叔叔"       "瘋狂"       "小時"
##  [97] "挑戰賽"     "征服"       "家"         "高雄"       "大王"       "蔡阿嘎"
## [103] "馬"         "叔叔"       "三菱"       "汽車"       "食尚"       "玩"
## [109] "嘎"         "排隊"       "花錢"       "玩"         "東京"       "迪士尼"
## [115] "蔡阿嘎"     "招"         "懶人"       "玩法"       "食尚"       "玩"
## [121] "嘎"         "帛"         "琉"         "蔡阿嘎"     "出發"       "鞏固"
## [127] "邦交國"     "瘋狂"       "小時"       "挑戰賽"     "吃"         "間"
## [133] "嘉義"       "雞肉飯"     "蔡阿嘎"     "馬"         "叔叔"       "嘎慶君遊"
## [139] "台灣"       "花蓮"       "好山好水"   "蔡阿嘎"     "私藏"       "深度"
## [145] "玩法"       "食尚"       "玩"         "嘎"         "沖繩"       "說走就走"
## [151] "跟"         "著"         "蔡阿嘎"     "從"         "北玩到"     "南"
## [157] "嘎慶君遊"   "台灣"       "慢城"       "嘉義市"     "蔡阿嘎帶"   "你"
## [163] "品嚐"       "平價"       "美食"       "天堂"       "日本"       "自由"
## [169] "行"         "不用"       "怕"         "大"         "實用"       "旅遊"
## [175] "漢字"       "蔡阿嘎來"   "教"         "你"         "食尚"       "玩"
## [181] "嘎"         "韓國"       "首爾"       "女人"       "天堂"       "男人"
## [187] "地獄"       "蔡阿嘎"     "厭世"       "代表作"     "日本"       "超級市場"
## [193] "大必"       "買好"       "物"         "蔡阿嘎"     "真心"       "推薦"
## [199] "食尚"       "玩"         "嘎"         "桃園"       "美食"       "沙漠"
## [205] "觀光"       "墳場"       "蔡阿嘎帶"   "您"         "破解"       "食尚"
## [211] "玩"         "嘎"         "來"         "香港"       "一定"       "要"
## [217] "體驗"       "件"         "事"         "蔡阿嘎"     "真心"       "不騙"
## [223] "日本"       "超商"       "大必"       "買"         "零食"       "飲料"
## [229] "蔡阿嘎"     "真心"       "推薦"       "食尚"       "玩"         "嘎"
## [235] "澳洲"       "黃金海岸"   "凱恩斯"     "跟"         "蔡阿嘎"     "飛到"
## [241] "南半球"     "曬太陽"     "食尚"       "玩"         "嘎"         "京都"
## [247] "大阪"       "篇"         "蔡阿嘎教"   "你"         "日本"       "旅遊"
## [253] "大"         "禁忌"       "食尚"       "玩"         "嘎"         "一定"
## [259] "要"         "帶"         "女朋友"     "去"         "約會"       "東京"
## [265] "大"         "動漫"       "主題樂園"   "跟"         "著"         "蔡阿嘎"
## [271] "魔獸"       "世界"       "玩正"       "港台"       "灣"         "內地"
## [277] "南投"       "食尚"       "玩"         "嘎"         "玩"         "日本"
## [283] "福岡"       "只要"       "元"         "到處"       "都"         "熊本"
## [289] "熊"         "食尚"       "玩"         "嘎"         "台東"       "好山好水"
## [295] "好"         "好玩"       "蔡阿嘎帶"   "你"         "悄悄"       "避開"
## [301] "觀光客"     "食尚"       "玩"         "嘎"         "基隆"       "不去"
## [307] "流俗"       "廟口"       "蔡阿嘎帶"   "你"         "走訪"       "個"
## [313] "巷弄"       "美食"       "景點"       "食尚"       "玩"         "嘎"
## [319] "日本"       "東京"       "熱"         "七天"       "蔡阿嘎"     "出國"
## [325] "處男"       "秀"         "食尚"       "玩"         "嘎"         "蘭嶼"
```

```
## [331] "放鬆"       "靈魂"       "遺世"       "蔡阿嘎"     "馬"         "叔叔"
## [337] "食尚"       "玩"         "嘎"         "山海"       "河港"       "大高雄"
## [343] "打狗"       "先生"       "蔡阿嘎"     "馬"         "叔叔"       "首支"
## [349] "單曲"       "食尚"       "玩"         "嘎"         "彰化"       "大佛"
## [355] "故鄉"       "三寶"       "跟"         "著"         "蔡阿嘎"     "吃"
## [361] "到"         "飽"         "嘟嘟"       "食尚"       "玩"         "嘎"
## [367] "蔡阿嘎"     "熱血"       "七日"       "機車"       "環島"       "台灣"
## [373] "去"         "一輩子"     "只會"       "去"         "一次"       "地方"
## [379] "食尚"       "玩"         "嘎"         "悠哉"       "宜蘭"       "城"
## [385] "跟"         "蔡阿嘎到"   "蘭陽"       "平原"       "放空"       "去"
## [391] "食尚"       "玩"         "嘎"         "蛻變"       "大台"       "中"
## [397] "蔡阿嘎帶"   "你"         "探索"       "都市"       "與"         "文化"
## [403] "大"         "熔爐"       "食尚"       "玩"         "嘎"         "屏東"
## [409] "阿猴城"     "猴"         "囝"         "仔"         "蔡阿嘎來"   "避寒"
## [415] "囉"         "食尚"       "玩"         "嘎"         "發現"       "戀戀"
## [421] "金門"       "跟"         "蔡阿嘎大"   "啖"         "好"         "菜"
## [427] "好"         "酒"         "好"         "風光"       "食尚"       "玩"
## [433] "嘎"         "新竹"       "風城"       "當"         "追風"       "少年"
## [439] "不吃"       "米粉"       "貢丸"       "蔡阿嘎帶"   "你"         "吃"
## [445] "巷"         "內"         "好味"       "食尚"       "玩"         "嘎"
## [451] "澎湖"       "防曬"       "曬成"       "黑面"       "蔡阿嘎"     "食尚"
## [457] "玩"         "嘎"         "板橋"       "殘酷"       "擂台"       "蔡阿嘎"
## [463] "中肯"       "挑戰"       "全台灣"     "人口"       "第一"       "區"
## [469] "食尚"       "玩"         "嘎"         "雲林"       "來"         "瘋"
## [475] "蔡阿嘎報"   "你"         "美食"       "瘋"         "雲林"       "食尚"
## [481] "玩"         "嘎"         "台南"       "古都"       "美食"       "文化"
## [487] "天堂"       "跟"         "蔡阿嘎"     "肚裡"       "一起"       "撐船"
## [493] "食尚"       "玩"         "嘎"         "苗栗"       "客家"       "庄大"
## [499] "喀"         "草莓"       "蔡阿嘎"     "恁"         "仔細"       "承蒙"
## [505] "您"         "恁久"       "好否"       "食尚"       "玩"         "嘎"
## [511] "馬祖"       "勇士"       "們"         "教"         "蔡阿嘎"     "愛國"
## [517] "熱血"       "食尚"       "玩"         "嘎"         "蔡阿嘎要"   "顛覆"
## [523] "你"         "對"         "嘉義"       "想像"       "食尚"       "玩"
## [529] "嘎"         "蔡阿嘎教"   "你"         "花"         "元去"       "高雄義大"
## [535] "世界"       "當"         "大爺"       "食尚"       "玩"         "嘎"
## [541] "跟"         "蔡阿嘎到"   "北投"       "一日"       "地輕"       "旅行"
## [547] "食尚"       "玩"         "嘎"         "蔡阿嘎"     "推薦"       "你"
## [553] "花蓮"       "流俗"       "地"         "玩法"
```

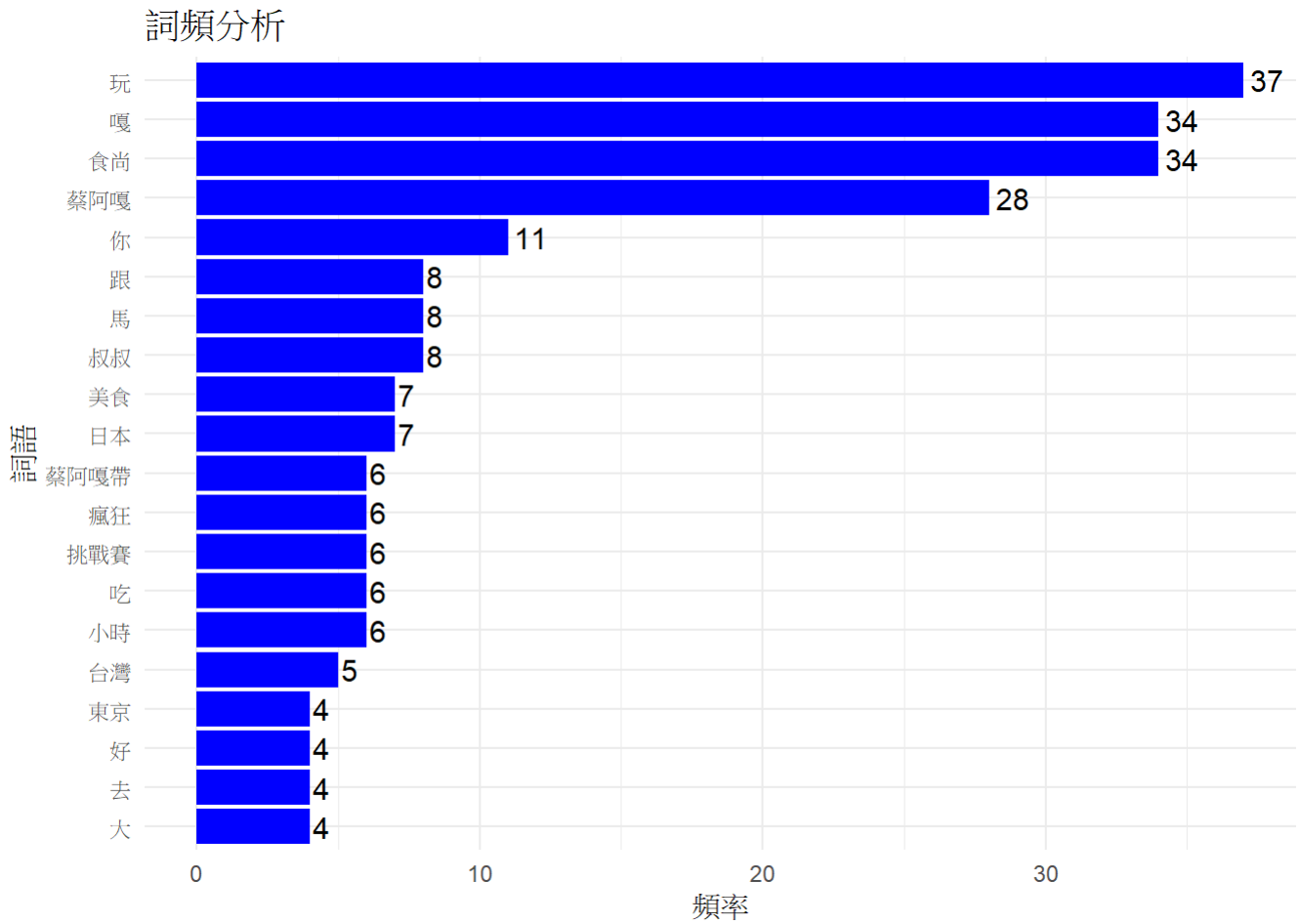
```
# wordcloud
word_freq <- table(filtered_words)
wordcloud(names(word_freq),
          freq = word_freq,
          min.freq = 2,
          random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))
```




```
# visualization
word_freq_df <- as.data.frame(word_freq, stringsAsFactors = FALSE)
colnames(word_freq_df) <- c("word", "freq")

top_words <- word_freq_df %>%
  arrange(desc(freq)) %>%
  head(20)

ggplot(top_words, aes(x = reorder(word, freq), y = freq)) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = freq), hjust = -0.2, size = 4) +
  coord_flip() +
  labs(title = "詞頻分析", x = "詞語", y = "頻率") +
  theme_minimal()
```



仔細切分過後並和前面結果相比，可以發現前幾高的還是食尚玩嘎和蔡阿嘎、馬叔叔等，和前面結果不同的是美食、日本、台灣、東京次數也不算低，由此可知影片選定的旅遊地點可能多在這幾個地方。此外，瘋狂和挑戰賽出現的次數也不少，同樣的也可以看出影片可能並非一般的旅遊vlog。