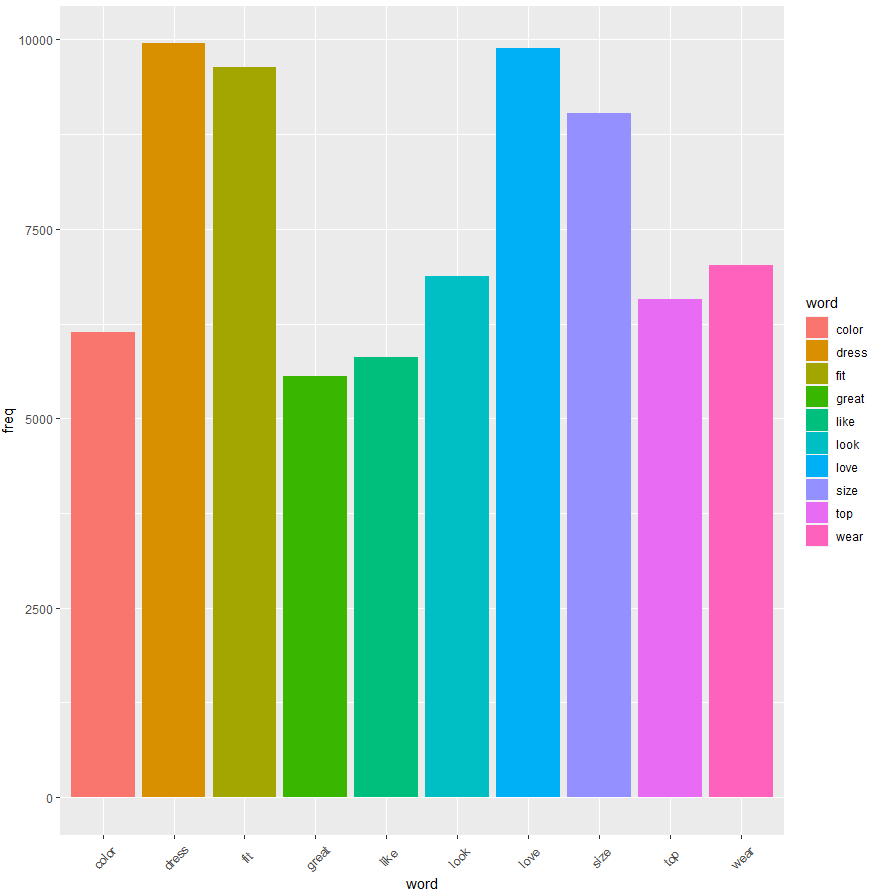
商業分析：SAS / R HW5

108208004 經濟三 白植允

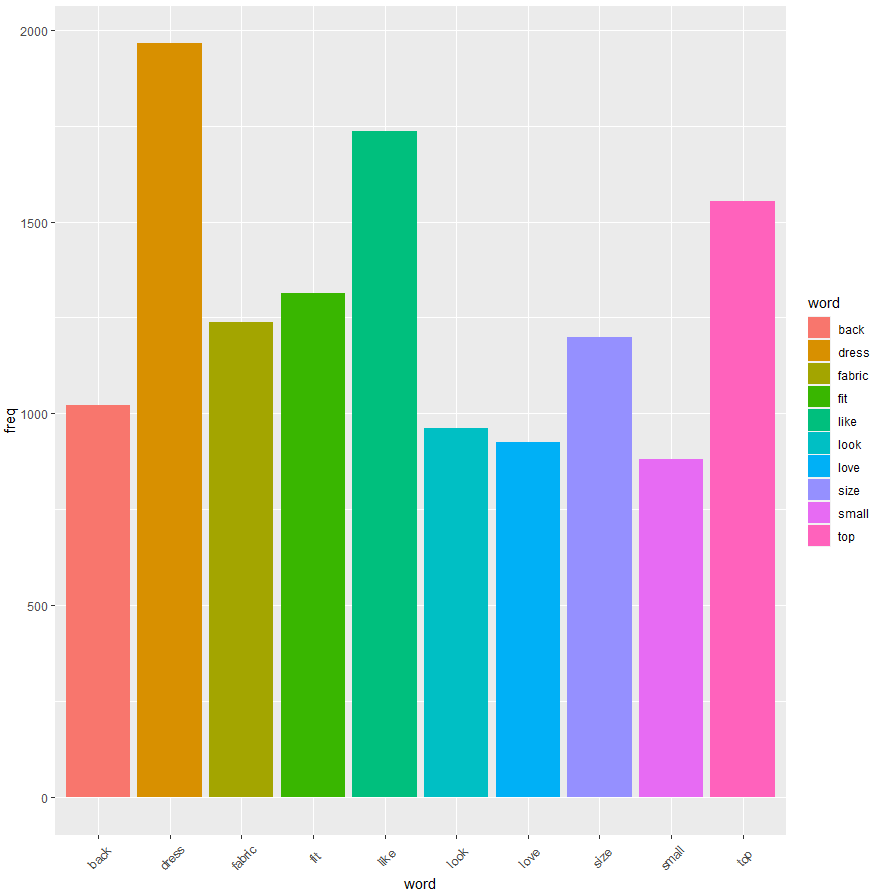
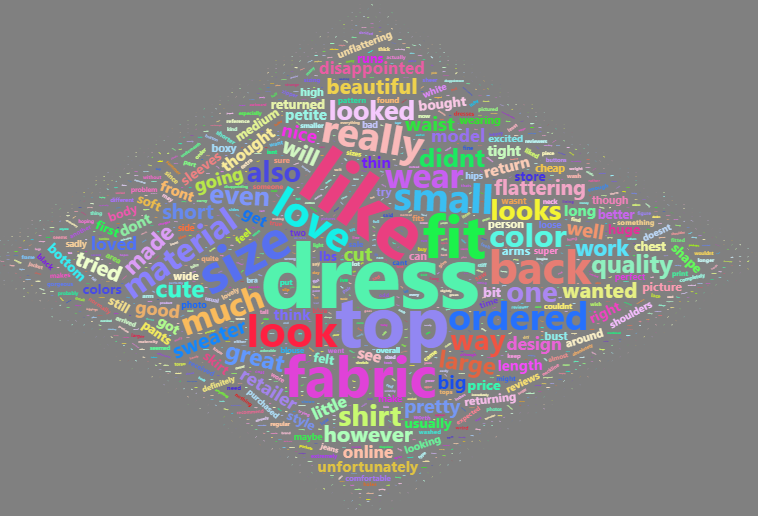
1. 請用上課的例子**review**資料集。

請將資料分成會推薦及不會推薦來比較，分別做wordcloud及直方圖，分析這兩種顧客的留言差異。

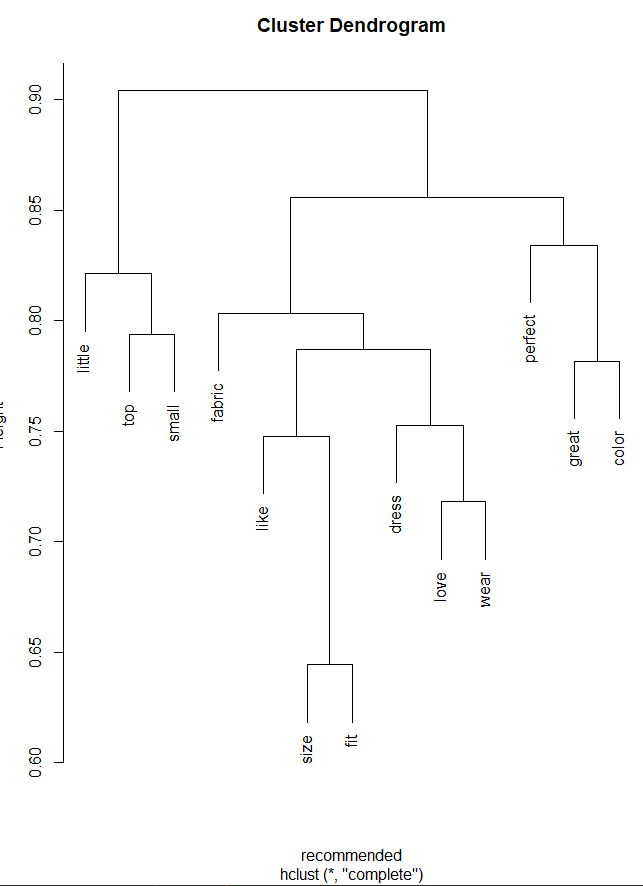
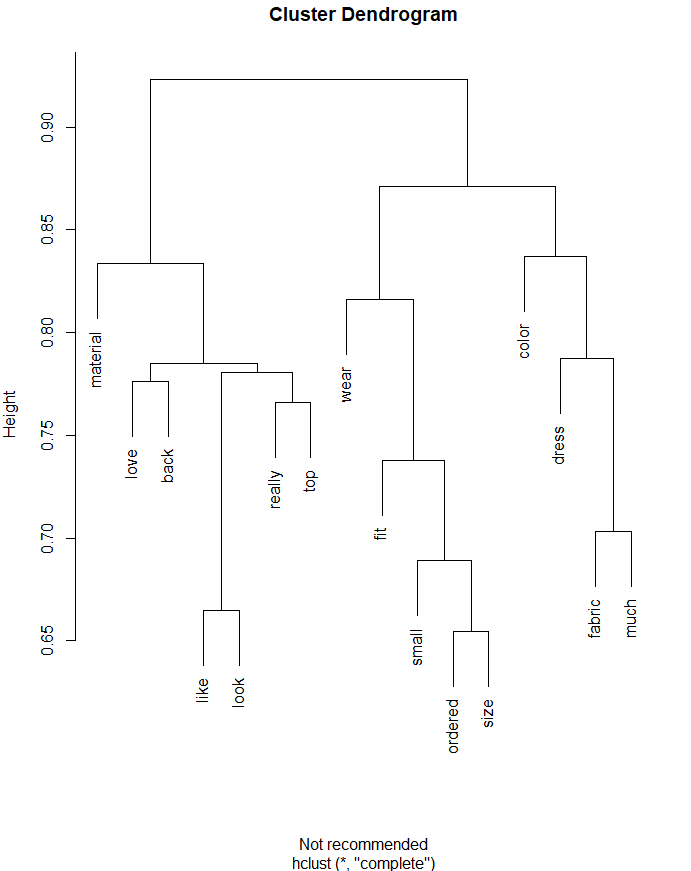
Recommend-



Not recommend –



=>最多出現的字都是dress，可見洋裝是客人最重視的商品，另外，從推薦者的worldcloud可推論顧客喜歡的點是size合適(size,fit)、顏色好看(look,color,great)，而不推薦的人不喜歡的原因推論可能是材質不好(fabric)，size不合(size,small)

1. 利用上課或TA課(或其他你會的)網路爬蟲方式，任選一筆資料整理，做出wordcloud。



=>中國跟美國對財經市場影響很大，疫情也有影響，華為最近可能有新動態，大選似乎影響不大

**附錄:R程式碼**

library(tm)

library(tmcn)

library(devtools)

library(jiebaR)

library(tidyverse)

library(wordcloud2)

library(proxy)

library(rvest)

library(stringr)

library(httr)

library(jsonlite)

library(devtools)

install.packages("RSelenium")

library(RSelenium)

#1

data = read.csv("reviews.csv")

recommend <- data %>%

filter(Recommended.IND == 1) %>%

select(Review.Text)

not\_recommend <- data %>%

filter(Recommended.IND == 0) %>%

select(Review.Text)

x = Corpus(VectorSource(recommend$Review.Text))

x = tm\_map(x,removeNumbers)

x = tm\_map(x,removePunctuation)

x = tm\_map(x,removeWords,c(stopwords("english"),"just"))

x = tm\_map(x,tolower)

x = tm\_map(x,removeWords,c(stopwords("english"),"just"))

x = tm\_map(x,stripWhitespace)

x\_tdm <- TermDocumentMatrix(x)

inspect(x\_tdm)

x\_matrix <- as.matrix(x\_tdm)

x\_v <- sort(rowSums(x\_matrix), decreasing = TRUE)

x\_d <- data.frame(word = names(x\_v), freq = x\_v)

wordcloud2(x\_d,size=0.5,color = "random-light", backgroundColor = "grey",shape = 'diamond')

ggplot(aes(x = word,y = freq,fill = word),data = x\_d[1:10,])+

geom\_bar(stat = "identity")+

theme(axis.text.x = element\_text(angle = 45, hjust = 0.5, vjust = 0.5))

y = Corpus(VectorSource(not\_recommend$Review.Text))

y = tm\_map(y,removeNumbers)

y = tm\_map(y,removePunctuation)

y = tm\_map(y,removeWords,c(stopwords("english"),"just"))

y = tm\_map(y,tolower)

y = tm\_map(y,removeWords,c(stopwords("english"),"just"))

y = tm\_map(y,stripWhitespace)

y\_tdm <- TermDocumentMatrix(y)

inspect(y\_tdm)

y\_matrix <- as.matrix(y\_tdm)

y\_v <- sort(rowSums(y\_matrix), decreasing = TRUE)

y\_d <- data.frame(word = names(y\_v), freq = y\_v)

wordcloud2(y\_d,size=0.5,color = "random-light", backgroundColor = "grey",shape = 'diamond')

ggplot(aes(x = word,fill = word),data = y\_d[1:10,])+

geom\_bar()+

theme(axis.text.x = element\_text(angle = 45, hjust = 0.5, vjust = 0.5))

x\_tdm2 <- removeSparseTerms(x\_tdm, sparse = 0.85)

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

xhc <- hclust(d = dist(x\_mydata, method = "cosine"), method = "complete")

plot(xhc,xlab = 'recommended')

y\_tdm2 <- removeSparseTerms(y\_tdm, sparse = 0.85)

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

yhc <- hclust(d = dist(y\_mydata, method = "cosine"), method = "complete")

plot(yhc,xlab = 'Not recommended')

#2

url <- "http://blog.moneydj.com/news/"

doc <- read\_html(url, encoding = "UTF-8")

article.all <- c()

df.all <- data.frame()

for(i in 1:4) {

url <- paste0("http://blog.moneydj.com/news/", "page/", i)

doc <- read\_html(url, encoding = "UTF-8")

header <- doc %>%

html\_nodes(".entry-title.mh-loop-title") %>%

html\_nodes("a") %>%

html\_text()

href <- doc %>%

html\_nodes(".entry-title.mh-loop-title") %>%

html\_nodes("a") %>%

html\_attr("href")

article.page <- c()

for(i in 1:length(href)) {

doc.a <- read\_html(href[i])

article <- doc.a %>%

html\_nodes("div.entry-content.mh-clearfix") %>%

html\_nodes("p") %>%

html\_text() %>%

str\_c(collapse = "")

article <- ifelse(str\_length(article) < 10 || rlang:::is\_empty(article), NA, article)

article.all <- append(article.all, article)

}

df <- data.frame(title = header, content = article.all) %>%

na.omit() %>%

mutate(title = as.character(title),

content = as.character(content))

}

cc <- worker()

word <-cc[df[,2]]

word\_df <- as.data.frame(table(word))

word\_df %>%

filter(!str\_detect(word, "[a-zA-Z0-9]+")) %>%

filter(nchar(as.character(word)) > 1) %>%

filter( Freq > 10) ->temp

wordcloud2(temp,size = 0.5,color = "random-dark", backgroundColor = "white",shape = 'diamond')