**商業分析作業五**

**統計三 108304010 趙啟宏**

## 1. 請用上課的例子review資料集

Recommended IND: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.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。

##read in data##

data <- read.csv("Womens Clothing E-Commerce Reviews.csv",encoding = "UTF-8")

##reviewrs who are recommended##

c1 <- data[which(data$Recommended.IND==1),]

## Make a vector source and a corpus

library(tm)

x1=Corpus(VectorSource(c1$Review.Text))

##Clean text

x1 <- tm\_map(x1, content\_transformer(tolower))

x1 <- tm\_map(x1, removePunctuation) #remove punctuation

#Remove stopwords

x1 <- tm\_map(x1, removeWords, stopwords("english"))

x1 <- tm\_map(x1, stemDocument)

tdm1 <- TermDocumentMatrix(x1)

inspect(tdm1)

# Convert TDM to matrix

review\_m1 <- as.matrix(tdm1)

# Sum rows and frequency data frame

freq\_df1 <- rowSums(review\_m1)

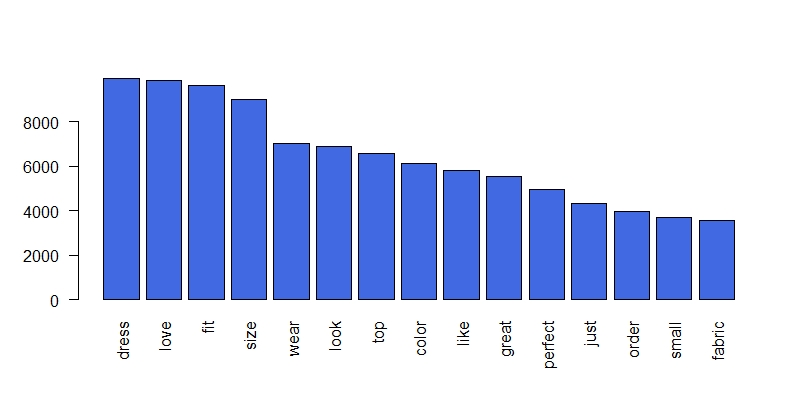
# Sort term\_frequency in descending order

freq\_df1 <- sort(freq\_df1, decreasing = T)

# View the top 15 most common words

freq\_df1[1:15]

barplot(freq\_df1[1:15], col = "royalblue", las = 2)



**詮釋:由上圖可知，那些會推薦產品的顧客們，他們的留言裡常常會出現( love, fit, great, perfect, top)等正面字眼，表示對於這項產品的高滿意度。**

summary1 <- data.frame(word = names(freq\_df1),

num = freq\_df1)

library(wordcloud2)

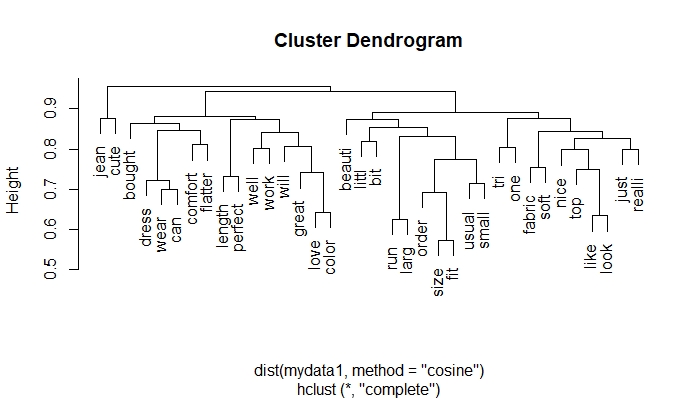
wordcloud2(summary1,size=0.5)

**詮釋:上圖為文字雲的結果，和直方圖呈現出來的結果大致相同。**

library("proxy")

hc <- hclust(d = dist(mydata1, method = "cosine"), method = "complete")

plot(hc)



**詮釋:上圖為字詞關係圖，除了先前看到的字詞外，還有(jeans, cute,bought, comfort)等字，而其中(jeans, cute)關聯性高，(bought, comfort)則不然。**

##reviewrs who are not recommended##

c2 <- data[which(data$Recommended.IND==0),]

## Make a vector source and a corpus

library(tm)

x2=Corpus(VectorSource(c2$Review.Text))

##Clean text

x2 <- tm\_map(x2, content\_transformer(tolower))

x2 <- tm\_map(x2, removePunctuation) #remove punctuation

#Remove stopwords

x2 <- tm\_map(x2, removeWords, stopwords("english"))

x2 <- tm\_map(x2, stemDocument)

tdm3 <- TermDocumentMatrix(x2)

inspect(tdm3)

# Convert TDM to matrix

review\_m2 <- as.matrix(tdm3)

# Sum rows and frequency data frame

freq\_df2 <- rowSums(review\_m2)

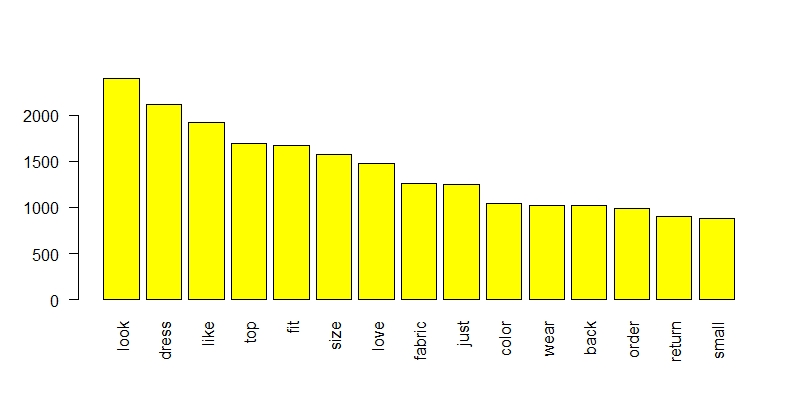
# Sort term\_frequency in descending order

freq\_df2 <- sort(freq\_df2, decreasing = T)

# View the top 15 most common words

freq\_df2[1:15]

barplot(freq\_df2[1:15], col = "yellow", las = 2)



**詮釋:由上圖可知，那些不會推薦產品的顧客們，他們的留言裡常常會出現( size, fabric, color, return)等名詞，可能表示在這些方面他們並不是滿意。**

library(wordcloud2)

wordcloud2(summary2,size=0.5)



**詮釋:上圖是文字雲的結果，和直方圖大致相同。**

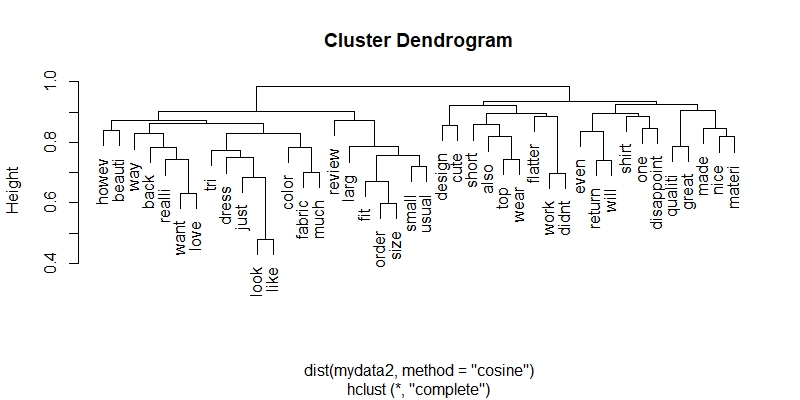
tdm4 <- removeSparseTerms(tdm3, sparse = 0.9)

mydata2 <- as.data.frame(as.matrix(tdm4))

library("proxy")

hc2 <- hclust(d = dist(mydata2, method = "cosine"), method = "complete")

plot(hc2)



**詮釋:上圖為字詞關係圖，除了先前看到的字詞外，還有(howev, short,qualiti等字，和會推薦產品的顧客們比起來，有更多的品質探討及語氣轉折。**

##Find the co-words##

good <- c(names(freq\_df1[1:10]))

bad <- c(names(freq\_df2[1:10]))

intersect(good,bad)

mystopwords <- c(intersect(good,bad))

**詮釋:在觀察兩個客群的常用字詞後，我們把最常出現的前十個字詞作比對，發現竟然有高達八個是相同的，這樣表示，這八個字詞雖然出現頻率高，對於分析顧客推薦產品與否卻沒有太大的幫助(因為不論怎樣都會出現)，所以我決定把這八個字刪除，再進行一次分析。**

##Dropping these co-words##

x1 <- tm\_map(x1, removeWords,mystopwords)

tdm1 <- TermDocumentMatrix(x1)

inspect(tdm1)

##Doing everything again##

# Convert TDM to matrix

review\_m1 <- as.matrix(tdm1)

# Sum rows and frequency data frame

freq\_df1 <- rowSums(review\_m1)

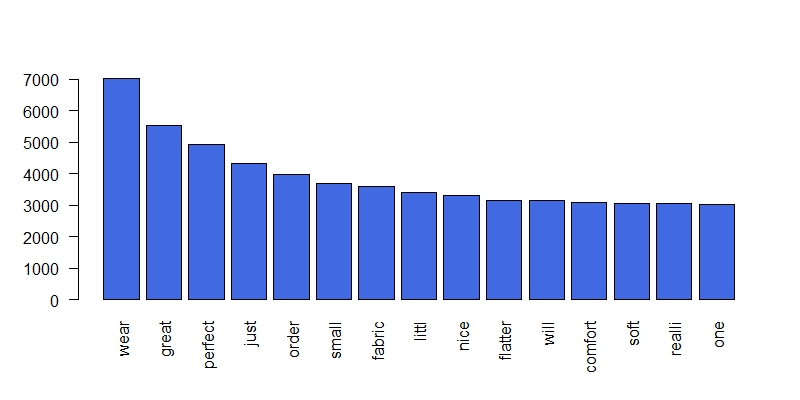
# Sort term\_frequency in descending order

freq\_df1 <- sort(freq\_df1, decreasing = T)

# View the top 15 most common words

freq\_df1[1:15]

barplot(freq\_df1[1:15], col = "royalblue", las = 2)



summary1 <- data.frame(word = names(freq\_df1),

num = freq\_df1)

wordcloud2(summary1,size=0.5)



**詮釋:可以發現，在去除重複字眼後，對於那些會推薦商品的顧客，(wear, perfect, great, comfort, beauti)等詞出現的頻率最高，而這些詞多半都是正向的形容詞。**

##Dropping these co-words##

x2 <- tm\_map(x2, removeWords,mystopwords)

##Doing everthing again##

tdm3 <- TermDocumentMatrix(x2)

inspect(tdm3)

# Convert TDM to matrix

review\_m2 <- as.matrix(tdm3)

# Sum rows and frequency data frame

freq\_df2 <- rowSums(review\_m2)

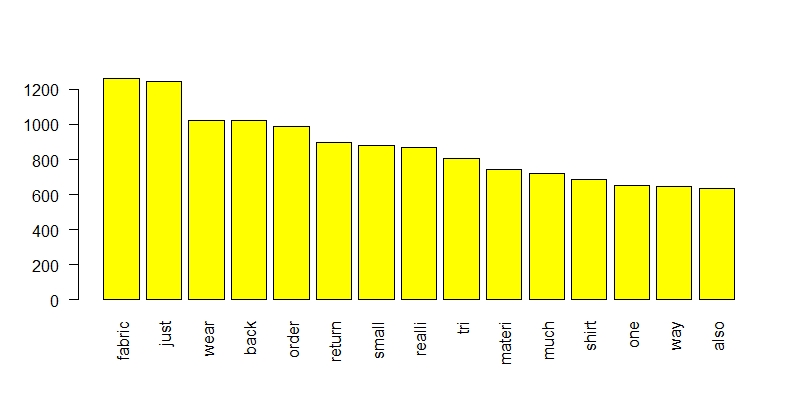
# Sort term\_frequency in descending order

freq\_df2 <- sort(freq\_df2, decreasing = T)

# View the top 15 most common words

freq\_df2[1:15]

barplot(freq\_df2[1:15], col = "yellow", las = 2)



summary2 <- data.frame(word = names(freq\_df2),

num = freq\_df2)

wordcloud2(summary2,size=0.5)



**詮釋:在去除重複字眼後，對於那些不會推薦商品的顧客，(fabric, materi, return, small, just)等詞出現的頻率最高，這些詞彙能夠讓我們清楚知道這些顧客是不滿意商品的哪個部分而選擇不推薦。**