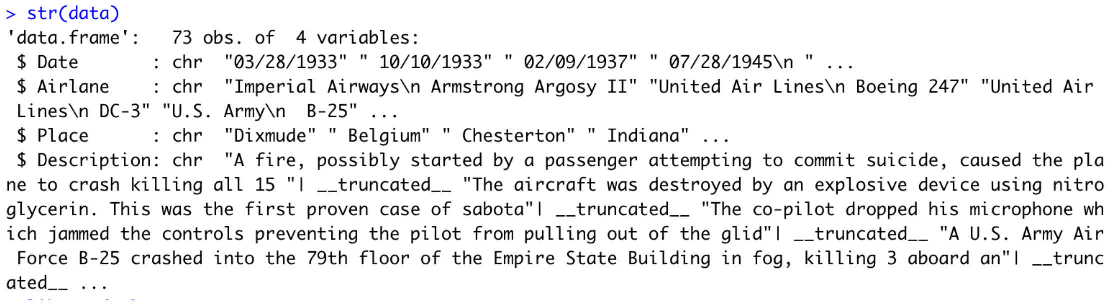
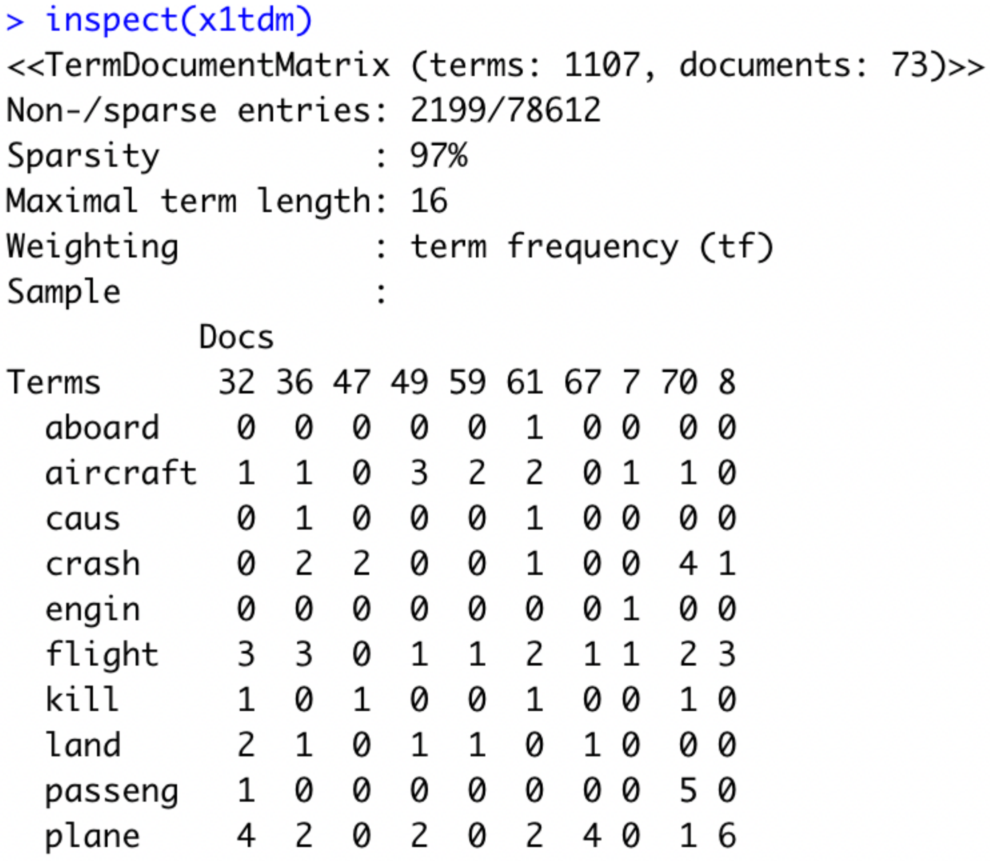
```
library(tidyverse)
#先讀入檔案並查看資料型態，因為我只需要 Description，儘管 Date 資料型態是錯誤的 chr，我只要確認 Description 是 chr 即可
data <- read.csv("airline.csv")
str(data)
```



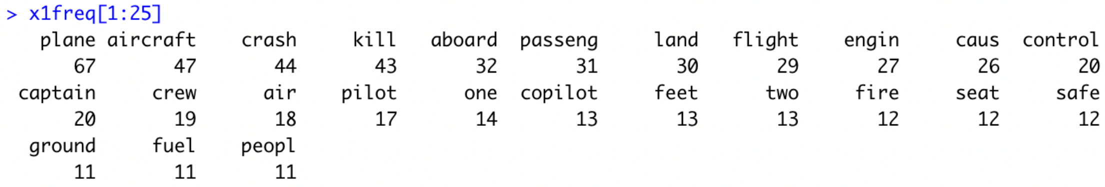
```
> str(data)
'data.frame':	73 obs. of  4 variables:
 $ Date       : chr  "03/28/1933" " 10/10/1933" " 02/09/1937" " 07/28/1945\n " ...
 $ Airlane    : chr  "Imperial Airways\n Armstrong Argosy II" "United Air Lines\n Boeing 247" "United Air
 Lines\n DC-3" "U.S. Army\n  B-25" ...
 $ Place      : chr  "Dixmude" " Belgium" " Chesterton" " Indiana" ...
 $ Description: chr  "A fire, possibly started by a passenger attempting to commit suicide, caused the pla
ne to crash killing all 15 "| __truncated__ "The aircraft was destroyed by an explosive device using nitro
glycerin. This was the first proven case of sabota"| __truncated__ "The co-pilot dropped his microphone wh
ich jammed the controls preventing the pilot from pulling out of the glid"| __truncated__ "A U.S. Army Air
 Force B-25 crashed into the 79th floor of the Empire State Building in fog, killing 3 aboard an"| __trunc
ated__ ...
```

```
#引入 text mining 套件
#進行資料處理：英文轉小寫、移除標點符號、移除和空難原因無關的連接詞
#轉成 matrix 並用 inspect()觀察
library(tm)
x1 <- Corpus(VectorSource(data$Description))
x1 <- tm_map(x1,tolower)
x1 <- tm_map(x1,content_transformer(tolower))
x1 <- tm_map(x1,removePunctuation)
x1StopWords <- c(stopwords(),"the","and","this","that","was","but","for")
x1 <- tm_map(x1,removeWords,x1StopWords)
x1 <- tm_map(x1,stemDocument)
x1tdm <- TermDocumentMatrix(x1)
inspect(x1tdm)
```



```
> inspect(x1tdm)
<<TermDocumentMatrix (terms: 1107, documents: 73)>>
Non-/sparse entries: 2199/78612
Sparsity           : 97%
Maximal term length: 16
Weighting          : term frequency (tf)
Sample             :
          Docs
Terms      32 36 47 49 59 61 67 7 70 8
  aboard    0  0  0  0  0  1  0 0  0 0
  aircraft  1  1  0  3  2  2  0 1  1 0
  caus      0  1  0  0  0  1  0 0  0 0
  crash     0  2  2  0  0  1  0 0  4 1
  engin     0  0  0  0  0  0  0 1  0 0
  flight    3  3  0  1  1  2  1 1  2 3
  kill      1  0  1  0  0  1  0 0  1 0
  land      2  1  0  1  1  0  1 0  0 0
  passeng   1  0  0  0  0  0  0 0  5 0
  plane     4  2  0  2  0  2  4 0  1 6
```

```
x1review <- as.matrix(x1tdm)
x1freq <- rowSums(x1review)
x1freq <- sort(x1freq, decreasing=T)
x1freq[1:25]
```

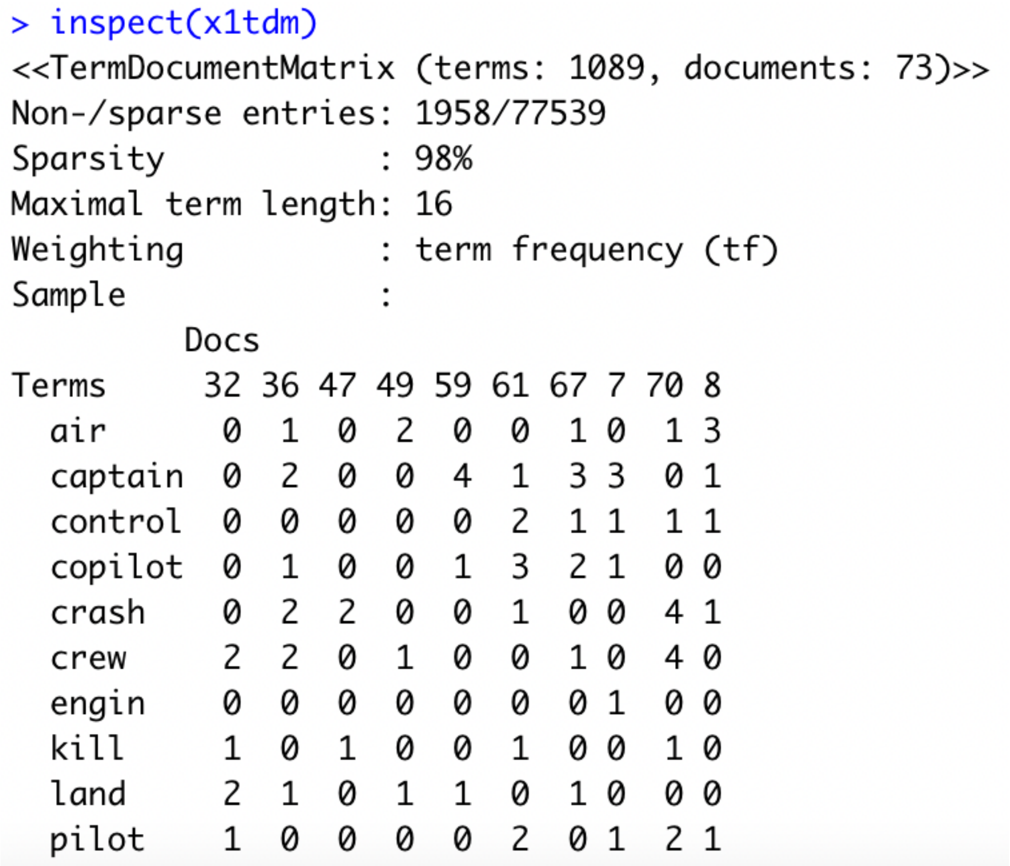


```
> x1freq[1:25]
   plane aircraft    crash     kill   aboard  passeng     land   flight    engin     caus  control 
      67       47       44       43       32       31       30       29       27       26       20 
 captain     crew      air    pilot      one  copilot     feet      two     fire     seat     safe 
      20       19       18       17       14       13       13       13       12       12       12 
  ground     fuel    peopl 
      11       11       11 
```

#由 inspect 結果看出現多次的字詞再用 freq 方便檢視，我認為還是有很多和空難原因不直接相關的字詞如 plane, aircraft 等，因此再做一次文字探勘，這次將之前認為不太相關的字詞刪去

```
x1StopWords <- c(stopwords(),"plane","aircraft","aboard","passeng","flight","caus","one","feet","two","seat","safe","peopl")
x1 <- tm_map(x1,removeWords,x1StopWords)
x1 <- tm_map(x1,stemDocument)

x1tdm <- TermDocumentMatrix(x1)
inspect(x1tdm)
```



```
> inspect(x1tdm)
<<TermDocumentMatrix (terms: 1089, documents: 73)>>
Non-/sparse entries: 1958/77539
Sparsity           : 98%
Maximal term length: 16
Weighting          : term frequency (tf)
Sample             :
         Docs
Terms     32 36 47 49 59 61 67 7 70 8
  air      0  1  0  2  0  0  1 0  1 3
  captain  0  2  0  0  4  1  3 3  0 1
  control  0  0  0  0  0  2  1 1  1 1
  copilot  0  1  0  0  1  3  2 1  0 0
  crash    0  2  2  0  0  1  0 0  4 1
  crew     2  2  0  1  0  0  1 0  4 0
  engin    0  0  0  0  0  0  0 1  0 0
  kill     1  0  1  0  0  1  0 0  1 0
  land     2  1  0  1  1  0  1 0  0 0
  pilot    1  0  0  0  0  2  0 1  2 1
```

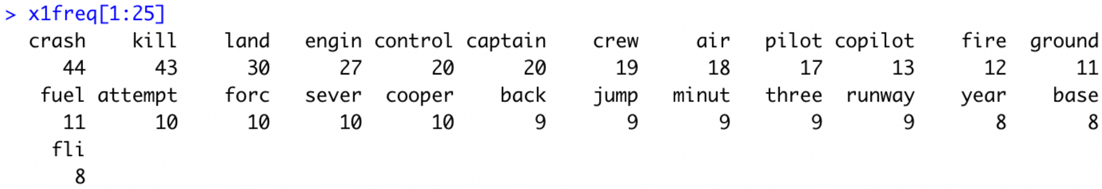
x1review <- as.matrix(x1tdm)

x1freq <- rowSums(x1review)

x1freq <- sort(x1freq, decreasing=T)

#可以觀察到更多關於空難原因的字詞，如 engin(可能是引擎問題)、fire、fuel 等

x1freq[1:25]



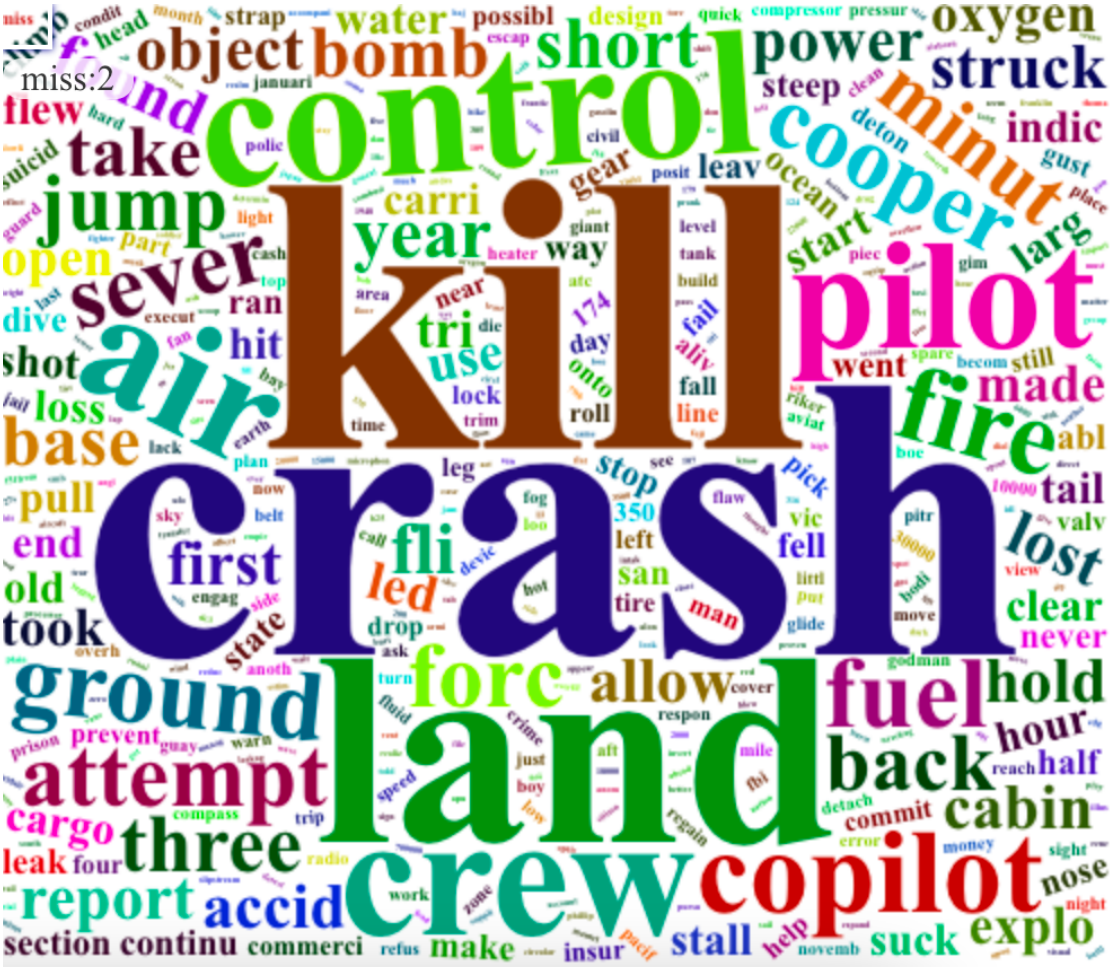
```
> x1freq[1:25]
   crash     kill     land    engin  control  captain     crew      air    pilot  copilot     fire   ground
      44       43       30       27       20       20       19       18       17       13       12       11
    fuel  attempt     forc    sever   cooper     back     jump    minut    three   runway     year     base
      11       10       10       10       10        9        9        9        9        9        8        8
     fli
       8
```

#產生文字雲以看到更多

library(wordcloud2)

x1freqframe <- data.frame(word=names(x1freq),num=x1freq)

wordcloud2(x1freqframe,size=1)



#可以觀察到更多空難原因相關的字眼如 control、bomb、struck、explo(可能和爆炸或爆裂物相關)等等