Markdown_Week12_Analyzing_data

Aymeric Collart

1. Prepare the environment

1.1 Load the libraries

```
library(quanteda)
```

```
## Package version: 4.3.1
## Unicode version: 14.0
## ICU version: 71.1

## Parallel computing: disabled

## See https://quanteda.io for tutorials and examples.
```

```
library(quanteda.textstats)
library(jiebaR)
```

```
## Loading required package: jiebaRD
```

```
library(tidytext)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(jiebaR)
library(openxlsx)

#Sys.setlocale(category = "LC_ALL", locale = "cht")
```

1.2 Load the originally scraped data

```r
load(file = "ArticleETToday_CorpusCourse_CLEAN.Rdata")
```

2. Key Word In Context (KWIC)

2.1 Set the segmenter (for Chinese)

```r
seg_word <- worker(bylines = T,
                   symbol=T)

seg_POS <- worker(type = "tag",
                  symbol = F)
```

2.2 Prepare the dataset for the analyses

```r
Article_total2$docname <- paste0("text",
                                 1:nrow(Article_total2))

Article_tokens <- Article_total2$body %>%
  segment(jiebar = seg_word) %>%
  as.tokens
```

2.3 Perform the KWIC segmentation

2.3.1 Corpus with POS information on the following word

```r
kwic_data <- kwic(Article_tokens,
                  pattern = "有",
                  window = 1)

RightPost_Annot <- segment(kwic_data$post, seg_POS)

## Convert to dataframe
RightPost_Annot <- do.call(rbind,
                           lapply(RightPost_Annot,
                                  as.data.frame))

RightPost_Annot <- cbind(POS = rownames(RightPost_Annot),
                         RightPost_Annot)

rownames(RightPost_Annot) <- 1:nrow(RightPost_Annot)

names(RightPost_Annot)[2] <- "RightPost"
```

```
RightPost_Annot$POS <- gsub("[0-9]+", "", RightPost_Annot$POS)

RightPost_Annot <- RightPost_Annot[!duplicated(RightPost_Annot), ]

names(RightPost_Annot)[2] <- "post"

kwic_data <- right_join(kwic_data,
                        RightPost_Annot,
                        by = "post")
```

2.3.2 Corpus with longer sentences

```
kwic_data2 <- kwic(Article_tokens,
                   pattern = "有",
                   window = 15)
```

2.3.3 Combine the two datasets together

```
### Prepare the dataset with longer sentences
kwic_data2 <- as.data.frame(kwic_data2)

kwic_data2$Index <- paste0(kwic_data2$docname,
                           kwic_data2$from)

kwic_data2_selected <- kwic_data2 %>%
  select(docname, pre, post, Index)

### Prepare the dataset with the POS infomation
kwic_data <- as.data.frame(kwic_data)

names(kwic_data)[6] <- "post_1word"

kwic_data_selected <- kwic_data %>%
  select(docname, from, to, post_1word, keyword, POS)

kwic_data_selected$Index <- paste0(kwic_data_selected$docname,
                                   kwic_data_selected$from)

### Join the two datasets
kwic_data <- right_join(kwic_data_selected,
                   kwic_data2_selected,
                   by = "Index")
```

```
## Warning in right_join(kwic_data_selected, kwic_data2_selected, by = "Index"): Detected an unexpected
## i Row 49 of `x` matches multiple rows in `y`.
## i Row 51 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
### Change the location of the columns for convenience
kwic_data <- kwic_data %>%
  relocate(keyword, .after = pre)
kwic_data <- kwic_data %>%
  relocate(post_1word, .after = keyword)
```

2.3.4 (Optional) Select the sentences we are interested in

For this example, we are interested in the sentences where the character 有 'you' (to have) is followed immediately by a verb (to simplify in this exemple, when POS = 'v')

```
kwic_you_verb <- kwic_data[kwic_data$POS == "v", ]

table_YouVerb <- table(kwic_you_verb$post_1word)
table_YouVerb <- as.data.frame(table_YouVerb)

names(table_YouVerb)[1] <- "Verb"
table_YouVerb <- table_YouVerb %>%
  arrange(desc(Freq))
table_YouVerb_Top10 <- head(table_YouVerb, 10)
table_YouVerb_Top10
```

```
##    Verb Freq
## 1  可能  899
## 2  逃亡  372
## 3  相關  304
## 4  需要  254
## 5  看到  192
## 6  超過  186
## 7  🄵押  183
## 8  疑慮  161
## 9  發生  143
## 10 幫助  132
```

```
## As you can see, even if the POS tagging is useful, it's not completely reliable.
```

2.4 Save the data

2.4.1 Save as an Excel file

```
write.xlsx(kwic_you_verb, "ArticleETToday_KWIC_You.xlsx")
```

2.4.2 Save as an RData file

```
save(kwic_you_verb, file = "ArticleETToday_KWIC_You.Rdata")
```

3. Frequency tables

3.1 Create the overall frequency table

3.1.1 Creation of the first table

```
## We need to tranform the tokenized data into a 'dfm' dataset
Article_tokens_frequency <- dfm(Article_tokens)
Article_tokens_frequency <- textstat_frequency(Article_tokens_frequency)

table_AllWordsFreq_Top100 <- head(Article_tokens_frequency, 100)
table_AllWordsFreq_Top100
```

```
##     feature frequency rank docfreq group
## 1        ，   1644510    1  237409   all
## 2        的    370860    2  161174   all
## 3        。    337831    3  230147   all
## 4        、    232573    4  115421   all
## 5        「    153920    5   92773   all
## 6        」    153854    6   92734   all
## 7        在    138401    7   99935   all
## 8        是     98718    8   70585   all
## 9        也     85523    9   67916   all
## 10       有     77572   10   60208   all
## 11       日     73899   11   61032   all
## 12       後     66276   12   54287   all
## 13       與     59328   13   47584   all
## 14       🄵     56468   14   46797   all
## 15       他     53478   15   37078   all
## 16     表示     49438   16   47103   all
## 17       都     48572   17   38815   all
## 18       人     48010   18   37467   all
## 19       但     46043   19   41562   all
## 20       時     45949   20   38995   all
## 21       年     45537   21   32132   all
## 22       月     44994   22   33768   all
## 23       不     44304   23   36119   all
## 24        2     44074   24   34530   all
## 25       被     43016   25   35411   all
## 26       等     42082   26   35047   all
## 27       及     41314   27   32635   all
## 28       要     40613   28   31869   all
## 29       了     40276   29   32505   all
## 30              38709   30   11503   all
## 31       就     37864   31   31992   all
## 32       會     37195   32   30211   all
## 33       （     36863   33   30325   all
## 34       ）     36812   34   30289   all
## 35        1     36679   35   28854   all
## 36       陳     35683   36   23660   all
## 37        3     34953   37   28842   all
## 38       ？     34326   38   23458   all
## 39     警方    34271   39   26258   all
## 40       男    32284   40   20601   all
## 41       到    31541   41   27762   all
## 42       🄵    31060   42   27013   all
## 43     台灣    30730   43   20924   all
## 44       中    30631   44   27383   all
## 45       她    29945   45   18663   all
## 46       🄵    29901   46   27640   all
## 47   民進黨    29638   47   20901   all
## 48     🄵有    28581   48   23855   all
## 49     民🄵    28319   49   22482   all
## 50       對    28167   50   24378   all
## 51       讓    27701   51   24097   all
## 52       ;    27099   52   23338   all
## 53       歲    26212   53   18841   all
## 54     發生    25313   54   22034   all
## 55   國民黨    24924   55   17728   all
## 56       而    24849   56   23074   all
## 57       於    24782   57   21452   all
## 58       上    24503   58   22165   all
## 59       我    24467   59   14684   all
## 60     發現    23505   60   20672   all
## 61       以    23215   61   21098   all
## 62     指出    22959   62   22662   all
## 63        4    22918   63   19929   all
## 64       之    22418   64   14479   all
## 65     自己    22270   65   18894   all
## 66     立委    22238   66   16598   all
## 67     柯文    21775   67   13623   all
## 68        5    21646   68   19044   all
## 69       跟    21143   69   17572   all
## 70       和    20834   70   17259   all
## 71       將    20727   71   19307   all
## 72       已    20682   72   18945   all
## 73     男子    20610   73   15377   all
## 74     調查    20280   74   17599   all
## 75       🄵    19825   75   13703   all
## 76       10    18612   76   16702   all
## 77       黨    18545   77   12924   all
## 78        6    18502   78   16135   all
## 79     萬元    18322   79   13812   all
## 80     總統    18305   80   12866   all
## 81       才    18030   81   16679   all
## 82        8    17787   82   15900   all
## 83     相關    17758   83   15865   all
## 84       🄵    17579   84   16499   all
## 85       或    17538   85   13994   all
## 86       因    17246   86   16127   all
## 87       更    17227   87   15212   all
## 88       前    17182   88   15253   all
## 89     進行    17167   89   15401   all
## 90       向    16664   90   15359   all
## 91     可以    16515   91   14273   all
## 92     認🄵    15942   92   14762   all
## 93        仍     15882   93   14872   all
## 94      現場     15647   94   13401   all
## 95      就是     15632   95   13924   all
## 96      政府     15602   96   12846   all
## 97        台     15314   97   12904   all
## 98        哲     15242   98   10843   all
## 99         7     14980   99   13552   all
## 100       !      14928  100   12213   all
```

3.1.2 Clean it up a little bit

```
## Example with punctuation marks
table_FreqWord <- Article_tokens_frequency[-grep("，", Article_tokens_frequency$feature),]
table_FreqWord <- table_FreqWord[-grep("。", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("、", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("「", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("」", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("（", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("）", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("？", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("；", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("！", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("《", table_FreqWord$feature),]
table_FreqWord <- table_FreqWord[-grep("》", table_FreqWord$feature),]

## Example with numbers
table_FreqWord <- table_FreqWord[-grep("[[:digit:]]", table_FreqWord$feature),]
```

3.1.3 Final table, addition of the percentage

```
table_FreqWord_Top100 <- head(table_FreqWord, 100)
table_FreqWord_Top100
```

```
##      feature frequency rank docfreq group
## 2         的    370860    2  161174   all
## 7         在    138401    7   99935   all
## 8         是     98718    8   70585   all
## 9         也     85523    9   67916   all
## 10        有     77572   10   60208   all
## 11        日     73899   11   61032   all
## 12        後     66276   12   54287   all
## 13        與     59328   13   47584   all
## 14        🄵     56468   14   46797   all
## 15        他     53478   15   37078   all
## 16      表示     49438   16   47103   all
## 17        都     48572   17   38815   all
## 18        人     48010   18   37467   all
## 19        但     46043   19   41562   all
## 20        時     45949   20   38995   all
## 21        年     45537   21   32132   all
## 22       月   44994  22  33768  all
## 23       不   44304  23  36119  all
## 25       被   43016  25  35411  all
## 26       等   42082  26  35047  all
## 27       及   41314  27  32635  all
## 28       要   40613  28  31869  all
## 29       了   40276  29  32505  all
## 30            38709  30  11503  all
## 31       就   37864  31  31992  all
## 32       會   37195  32  30211  all
## 36       陳   35683  36  23660  all
## 39     警方   34271  39  26258  all
## 40       男   32284  40  20601  all
## 41       到   31541  41  27762  all
## 42       🄵   31060  42  27013  all
## 43     台灣   30730  43  20924  all
## 44       中   30631  44  27383  all
## 45       她   29945  45  18663  all
## 46       🄵   29901  46  27640  all
## 47   民進黨   29638  47  20901  all
## 48     🄵有   28581  48  23855  all
## 49     民🄵   28319  49  22482  all
## 50       對   28167  50  24378  all
## 51       讓   27701  51  24097  all
## 53       歲   26212  53  18841  all
## 54     發生   25313  54  22034  all
## 55   國民黨   24924  55  17728  all
## 56       而   24849  56  23074  all
## 57       於   24782  57  21452  all
## 58       上   24503  58  22165  all
## 59       我   24467  59  14684  all
## 60     發現   23505  60  20672  all
## 61       以   23215  61  21098  all
## 62     指出   22959  62  22662  all
## 64       之   22418  64  14479  all
## 65     自己   22270  65  18894  all
## 66     立委   22238  66  16598  all
## 67     柯文   21775  67  13623  all
## 69       跟   21143  69  17572  all
## 70       和   20834  70  17259  all
## 71       將   20727  71  19307  all
## 72       已   20682  72  18945  all
## 73     男子   20610  73  15377  all
## 74     調查   20280  74  17599  all
## 75       🄵   19825  75  13703  all
## 77       黨   18545  77  12924  all
## 79     萬元   18322  79  13812  all
## 80     總統   18305  80  12866  all
## 81       才   18030  81  16679  all
## 83     相關   17758  83  15865  all
## 84       🄵   17579  84  16499  all
## 85       或   17538  85  13994  all
## 86       因   17246  86  16127  all
## 87       更   17227  87  15212  all
## 88        前    17182   88   15253   all
## 89      進行    17167   89   15401   all
## 90        向    16664   90   15359   all
## 91      可以    16515   91   14273   all
## 92      認🄵    15942   92   14762   all
## 93        仍    15882   93   14872   all
## 94      現場    15647   94   13401   all
## 95      就是    15632   95   13924   all
## 96      政府    15602   96   12846   all
## 97        台    15314   97   12904   all
## 98        哲    15242   98   10843   all
## 101       再    14704  101   13627   all
## 102     目前    14474  102   13512   all
## 103   立法院    14440  103   11229   all
## 106       從    14154  106   13201   all
## 107       名    14116  107   11868   all
## 108       且    14051  108   13272   all
## 109       又    13972  109   12665   all
## 110       由    13840  110   12568   all
## 111     人員    13525  111   11281   all
## 112     因🄵    13423  112   12568   all
## 113     已經    13395  113   12304   all
## 115     希望    13141  115   11708   all
## 116       依    13123  116   12448   all
## 117     駕駛    13123  116    9729   all
## 118       這    13041  118   11926   all
## 119       遭    12767  119   11703   all
## 120       很    12570  120   10877   all
## 121     因此    12465  121   11939   all
## 122       大    12294  122   10741   all
```

```
table_FreqWord_Top100$percentage <-
↪  round(table_FreqWord_Top100$frequency/sum(table_FreqWord$frequency)*100, 4)
table_FreqWord_Top100
```

```
##     feature frequency rank docfreq group percentage
## 2        的    370860    2  161174   all     2.7226
## 7        在    138401    7   99935   all     1.0160
## 8        是     98718    8   70585   all     0.7247
## 9        也     85523    9   67916   all     0.6279
## 10       有     77572   10   60208   all     0.5695
## 11       日     73899   11   61032   all     0.5425
## 12       後     66276   12   54287   all     0.4866
## 13       與     59328   13   47584   all     0.4355
## 14       🄵     56468   14   46797   all     0.4145
## 15       他     53478   15   37078   all     0.3926
## 16     表示     49438   16   47103   all     0.3629
## 17       都     48572   17   38815   all     0.3566
## 18       人     48010   18   37467   all     0.3525
## 19       但     46043   19   41562   all     0.3380
## 20       時     45949   20   38995   all     0.3373
## 21       年     45537   21   32132   all     0.3343
## 22       月     44994   22   33768   all     0.3303
## 23        不    44304   23   36119   all     0.3252
## 25        被    43016   25   35411   all     0.3158
## 26        等    42082   26   35047   all     0.3089
## 27        及    41314   27   32635   all     0.3033
## 28        要    40613   28   31869   all     0.2982
## 29        了    40276   29   32505   all     0.2957
## 30              38709   30   11503   all     0.2842
## 31        就    37864   31   31992   all     0.2780
## 32        會    37195   32   30211   all     0.2731
## 36        陳    35683   36   23660   all     0.2620
## 39      警方    34271   39   26258   all     0.2516
## 40        男    32284   40   20601   all     0.2370
## 41        到    31541   41   27762   all     0.2316
## 42        🄵    31060   42   27013   all     0.2280
## 43      台灣    30730   43   20924   all     0.2256
## 44        中    30631   44   27383   all     0.2249
## 45        她    29945   45   18663   all     0.2198
## 46        🄵    29901   46   27640   all     0.2195
## 47    民進黨    29638   47   20901   all     0.2176
## 48      🄵有    28581   48   23855   all     0.2098
## 49      民🄵    28319   49   22482   all     0.2079
## 50        對    28167   50   24378   all     0.2068
## 51        讓    27701   51   24097   all     0.2034
## 53        歲    26212   53   18841   all     0.1924
## 54      發生    25313   54   22034   all     0.1858
## 55    國民黨    24924   55   17728   all     0.1830
## 56        而    24849   56   23074   all     0.1824
## 57        於    24782   57   21452   all     0.1819
## 58        上    24503   58   22165   all     0.1799
## 59        我    24467   59   14684   all     0.1796
## 60      發現    23505   60   20672   all     0.1726
## 61        以    23215   61   21098   all     0.1704
## 62      指出    22959   62   22662   all     0.1685
## 64        之    22418   64   14479   all     0.1646
## 65      自己    22270   65   18894   all     0.1635
## 66      立委    22238   66   16598   all     0.1633
## 67      柯文    21775   67   13623   all     0.1599
## 69        跟    21143   69   17572   all     0.1552
## 70        和    20834   70   17259   all     0.1529
## 71        將    20727   71   19307   all     0.1522
## 72        已    20682   72   18945   all     0.1518
## 73      男子    20610   73   15377   all     0.1513
## 74      調查    20280   74   17599   all     0.1489
## 75        🄵    19825   75   13703   all     0.1455
## 77        黨    18545   77   12924   all     0.1361
## 79      萬元    18322   79   13812   all     0.1345
## 80      總統    18305   80   12866   all     0.1344
## 81        才    18030   81   16679   all     0.1324
## 83      相關    17758   83   15865   all     0.1304
## 84        🄵    17579   84   16499   all     0.1291
## 85        或    17538   85   13994   all     0.1288
## 86        因    17246   86   16127   all     0.1266
## 87        更    17227   87   15212   all     0.1265
## 88        前    17182   88   15253   all     0.1261
## 89     進行   17167  89  15401   all     0.1260
## 90       向   16664  90  15359   all     0.1223
## 91     可以   16515  91  14273   all     0.1212
## 92     認🄵   15942  92  14762   all     0.1170
## 93       仍   15882  93  14872   all     0.1166
## 94     現場   15647  94  13401   all     0.1149
## 95     就是   15632  95  13924   all     0.1148
## 96     政府   15602  96  12846   all     0.1145
## 97       台   15314  97  12904   all     0.1124
## 98       哲   15242  98  10843   all     0.1119
## 101      再   14704 101  13627   all     0.1079
## 102    目前   14474 102  13512   all     0.1063
## 103  立法院   14440 103  11229   all     0.1060
## 106      從   14154 106  13201   all     0.1039
## 107      名   14116 107  11868   all     0.1036
## 108      且   14051 108  13272   all     0.1032
## 109      又   13972 109  12665   all     0.1026
## 110      由   13840 110  12568   all     0.1016
## 111    人員   13525 111  11281   all     0.0993
## 112    因🄵   13423 112  12568   all     0.0985
## 113    已經   13395 113  12304   all     0.0983
## 115    希望   13141 115  11708   all     0.0965
## 116      依   13123 116  12448   all     0.0963
## 117    駕駛   13123 116   9729   all     0.0963
## 118      這   13041 118  11926   all     0.0957
## 119      遭   12767 119  11703   all     0.0937
## 120      很   12570 120  10877   all     0.0923
## 121    因此   12465 121  11939   all     0.0915
## 122      大   12294 122  10741   all     0.0903
```

3.2 Select only the 100 most frequent nouns

There are several ways to do it. We could use the raw data one more time. But since this is quite a large dataset, it will take a lot of time to process. An alternative way is to compute the 500 most frequent words, and hopefully there will be 100 nouns. [After trying 300 words, it was not enough. Tests with more words done after]

3.2.1 Set the segmenter

```
seg_POS_ByLines <- worker(type = "tag",
                          bylines = FALSE,
                          symbol = F)
```

3.2.2 Proceed to the segmentation and annotate

```
table_FreqWord_Top500 <- head(table_FreqWord, 500)

Top500_WordFreqPOS <- segment(table_FreqWord_Top500$feature,
                              seg_POS_ByLines)
```

```r
## Convert to dataframe
Top500_WordFreqPOS_Annotated <- do.call(rbind,
                                        lapply(Top500_WordFreqPOS,
                                               as.data.frame))

Top500_WordFreqPOS_Annotated <- cbind(POS = rownames(Top500_WordFreqPOS_Annotated),
                                      Top500_WordFreqPOS_Annotated)

rownames(Top500_WordFreqPOS_Annotated) <- 1:nrow(Top500_WordFreqPOS_Annotated)

names(Top500_WordFreqPOS_Annotated)[2] <- "Word"

Top500_WordFreqPOS_Annotated$POS <- gsub("[0-9]+", "", Top500_WordFreqPOS_Annotated$POS)
```

3.2.3 Extract the nouns (POS = n, to make it simple) and annotate

```r
TopFreqNoun <- Top500_WordFreqPOS_Annotated[Top500_WordFreqPOS_Annotated$POS == "n", ]
TopFreqNoun$Index <- "TopNouns"

names(table_FreqWord)[1] <- "Word"
TopFreqNoun <- right_join(TopFreqNoun,
                          table_FreqWord,
                          by = "Word")

TopFreqNoun$Percentage <- round(TopFreqNoun$frequency/sum(TopFreqNoun$frequency)*100, 4)

TopFreqNoun <- TopFreqNoun[+grep("TopNouns", TopFreqNoun$Index),]

table_FreqNoun_Top100 <- head(TopFreqNoun, 100)
table_FreqNoun_Top100 <- table_FreqNoun_Top100 %>%
  arrange(desc(frequency))
table_FreqNoun_Top100
```

```
##      POS   Word    Index frequency rank docfreq group Percentage
## 1      n     人 TopNouns     48010   18   37467   all     0.3525
## 2      n   警方 TopNouns     34271   39   26258   all     0.2516
## 3      n     男 TopNouns     32284   40   20601   all     0.2370
## 4      n   男子 TopNouns     20610   73   15377   all     0.1513
## 5      n     黨 TopNouns     18545   77   12924   all     0.1361
## 6      n   總統 TopNouns     18305   80   12866   all     0.1344
## 7      n   現場 TopNouns     15647   94   13401   all     0.1149
## 8      n   政府 TopNouns     15602   96   12846   all     0.1145
## 9      n     哲 TopNouns     15242   98   10843   all     0.1119
## 10     n   人員 TopNouns     13525  111   11281   all     0.0993
## 11     n   分局 TopNouns     12237  125   10362   all     0.0898
## 12     n   公司 TopNouns     11992  131    8170   all     0.0880
## 13     n   國會 TopNouns     11801  134    8217   all     0.0866
## 14     n   大家 TopNouns     11582  139    9814   all     0.0850
## 15     n   國家 TopNouns     11264  143    9123   all     0.0827
## 16     n   市長 TopNouns     11252  145    8707   all     0.0826
## 17    n   問題 TopNouns    11013  152   9618   all    0.0808
## 18    n   法官 TopNouns    10815  155   9100   all    0.0794
## 19    n   政治 TopNouns    10357  167   8522   all    0.0760
## 20    n   媒體 TopNouns    10217  169   8956   all    0.0750
## 21    n     🄵 TopNouns    10181  173   9454   all    0.0747
## 22    n   原因 TopNouns     9894  183   9026   all    0.0726
## 23    n   社會 TopNouns     9695  190   8305   all    0.0712
## 24    n     案 TopNouns     9390  196   7994   all    0.0689
## 25    n   時間 TopNouns     9327  201   8440   all    0.0685
## 26    n     處 TopNouns     9207  205   8013   all    0.0676
## 27    n   主席 TopNouns     9099  208   7533   all    0.0668
## 28    n   檢方 TopNouns     9041  210   7703   all    0.0664
## 29    n   部分 TopNouns     8834  214   7798   all    0.0649
## 30    n   員警 TopNouns     8766  216   6217   all    0.0644
## 31    n   無法 TopNouns     8763  217   8164   all    0.0643
## 32    n   結果 TopNouns     8383  230   7803   all    0.0615
## 33    n   黨團 TopNouns     8342  231   6193   all    0.0612
## 34    n   地方 TopNouns     8325  232   6750   all    0.0611
## 35    n   女子 TopNouns     8263  235   6157   all    0.0607
## 36    n   法院 TopNouns     8155  239   7162   all    0.0599
## 37    n   規定 TopNouns     7999  243   7044   all    0.0587
## 38    n   院長 TopNouns     7962  244   5828   all    0.0585
## 39    n 委員會 TopNouns     7875  246   5909   all    0.0578
## 40    n   過程 TopNouns     7861  247   7463   all    0.0577
## 41    n   民主 TopNouns     7694  251   5686   all    0.0565
## 42    n   醫院 TopNouns     7664  254   6453   all    0.0563
## 43    n 檢察官 TopNouns     7551  257   6228   all    0.0554
## 44    n   全案 TopNouns     7532  260   7325   all    0.0553
## 45    n   案件 TopNouns     7460  261   6592   all    0.0548
## 46    n   事件 TopNouns     7376  264   6612   all    0.0541
## 47    n   機車 TopNouns     7326  267   5435   all    0.0538
## 48    n   人民 TopNouns     7206  272   5976   all    0.0529
## 49    n     區 TopNouns     7113  277   6579   all    0.0522
## 50    n   方式 TopNouns     6919  284   6453   all    0.0508
## 51    n   車輛 TopNouns     6838  290   5600   all    0.0502
## 52    n   雙方 TopNouns     6810  292   6118   all    0.0500
## 53    n   議員 TopNouns     6735  295   5408   all    0.0494
## 54    n     著 TopNouns     6652  298   6260   all    0.0488
## 55    n   被告 TopNouns     6502  303   3997   all    0.0477
## 56    n   市府 TopNouns     6493  304   4954   all    0.0477
## 57    n   狀🄵 TopNouns     6437  310   6044   all    0.0473
## 58    n 行政院 TopNouns     6426  312   5124   all    0.0472
## 59    n   毒品 TopNouns     6292  315   3659   all    0.0462
## 60    n   報案 TopNouns     6162  323   5758   all    0.0452
## 61    n   事故 TopNouns     6158  324   5149   all    0.0452
## 62    n   🄵容 TopNouns     6007  331   5475   all    0.0441
## 63    n   家屬 TopNouns     5996  333   4819   all    0.0440
## 64    n   條例 TopNouns     5985  335   5286   all    0.0439
## 65    n   影片 TopNouns     5982  336   4701   all    0.0439
## 66    n   報告 TopNouns     5832  344   4770   all    0.0428
## 67    n   國際 TopNouns     5798  349   4597   all    0.0426
## 68    n   單位 TopNouns     5681  358   5186   all    0.0417
## 69    n   集團 TopNouns     5639  363   4536   all    0.0414
## 70    n   對方 TopNouns     5500  369   4721   all    0.0404
## 71    n   學生 TopNouns      5427  373    3518   all     0.0398
## 72    n   通報 TopNouns      5425  374    4853   all     0.0398
## 73    n     路 TopNouns      5415  375    4800   all     0.0398
## 74    n   憲法 TopNouns      5346  380    3535   all     0.0392
## 75    n   法案 TopNouns      5275  385    4245   all     0.0387
## 76    n   手機 TopNouns      5274  386    4288   all     0.0387
## 77    n   委員 TopNouns      5254  389    4054   all     0.0386
## 78    n     珊 TopNouns      5207  392    2150   all     0.0382
## 79    n   事情 TopNouns      5188  395    4685   all     0.0381
## 80    n   律師 TopNouns      5176  397    3722   all     0.0380
## 81    n 監視器 TopNouns      5134  399    4708   all     0.0377
## 82    n 派出所 TopNouns      5116  402    4470   all     0.0376
## 83    n   程序 TopNouns      5066  404    4404   all     0.0372
## 84    n   司法 TopNouns      5046  406    4148   all     0.0370
## 85    n   政策 TopNouns      4943  412    4173   all     0.0363
## 86    n   畫面 TopNouns      4930  413    4469   all     0.0362
## 87    n     警 TopNouns      4926  416    4679   all     0.0362
## 88    n   關🄵 TopNouns      4924  417    4456   all     0.0361
## 89    n   車禍 TopNouns      4904  421    4310   all     0.0360
## 90    n   資料 TopNouns      4873  424    4129   all     0.0358
## 91    n     團 TopNouns      4859  429    4133   all     0.0357
## 92    n   中央 TopNouns      4857  431    3952   all     0.0357
## 93    n 被害人 TopNouns      4746  442    3235   all     0.0348
## 94    n   網友 TopNouns      4671  449    3889   all     0.0343
## 95    n   中心 TopNouns      4603  454    3899   all     0.0338
## 96    n   報警 TopNouns      4579  459    4454   all     0.0336
## 97    n   大陸 TopNouns      4559  461    3403   all     0.0335
## 98    n   罪嫌 TopNouns      4505  466    4307   all     0.0331
## 99    n   肇事 TopNouns      4501  467    3896   all     0.0330
## 100   n   依法 TopNouns      4488  469    4267   all     0.0329
```

```
## Again, you can see that some entries are a little bit weird. There are several ways to
↪ handle them, but we will not cover them in class
```

3.3 Save the data

3.3.1 Save as an Excel file

```
write.xlsx(table_FreqNoun_Top100, "ArticleETToday_Top100nouns.xlsx")
```

3.3.2 Save as an RData file

```
save(table_FreqNoun_Top100, file = "ArticleETToday_Top100nouns.Rdata")
```