Text Mining: Homework 1

學號: 410821204　　姓名: 杜昉紜　　系級: 資工三

作業要求:

1.以 PlainTextDocument 的形式讀入所有文檔；

2. 探索語料；

3. 準備語料庫，包括將文本轉換為小寫，去除數字和標點符號，去除停用詞，詞乾和識別同義詞；

4.創建文檔術語矩陣；

5. 通過將文檔術語矩陣轉換為矩陣並對列數求和來探索文檔術語矩陣；

6. 去除稀疏詞；

7. 識別頻繁項和關聯；

8. 繪製相關圖；

9. 繪製詞頻；

10. 畫詞云

11、對文本進行定量分析

R console 跑的内容

(1) Read in all the documents as PlainTextDocument

(2) Explore the corpus

```
> ## ----load_corpus-----------------------------------------------------------
> reut21578 <- system.file("texts", "crude", package = "tm")
> docs <- VCorpus(DirSource(reut21578), readerControl=list(reader = readReut21578XML$
> docs
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 20
> class(docs)
[1] "VCorpus" "Corpus"
> class(docs[[1]])
[1] "PlainTextDocument" "TextDocument"
> summary(docs)
    Length Class             Mode
127 2      PlainTextDocument list
144 2      PlainTextDocument list
191 2      PlainTextDocument list
194 2      PlainTextDocument list
211 2      PlainTextDocument list
236 2      PlainTextDocument list
237 2      PlainTextDocument list
242 2      PlainTextDocument list
246 2      PlainTextDocument list
248 2      PlainTextDocument list
273 2      PlainTextDocument list
349 2      PlainTextDocument list
352 2      PlainTextDocument list
353 2      PlainTextDocument list
368 2      PlainTextDocument list
489 2      PlainTextDocument list
502 2      PlainTextDocument list
543 2      PlainTextDocument list
704 2      PlainTextDocument list
708 2      PlainTextDocument list
```

```
> ## ----out.lines=26----------------------------------------------------------
> inspect(docs[16])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  16
Content:  chars: 876
```

文章內容 (可以看到有大寫/標點符號...)

```
> ## ------------------------------------------------------------------------
> viewDocs <- function(d, n) {d %>% extract2(n) %>% as.character() %>% writeLines()}
> viewDocs(docs, 16)
A study group said the United States
should increase its strategic petroleum reserve to one mln
barrels as one way to deal with the present and future impact
of low oil prices on the domestic oil industry.
    U.S. policy now is to raise the strategic reserve to 750
mln barrels, from its present 500 mln, to help protect the
economy from an overseas embargo or a sharp price rise.
    The Aspen Institute for Humanistic Studies, a private
group, also called for new research for oil exploration and
development techniques.
    It predicted prices would remain at about 15-18 dlrs a
barrel for several years and then rise to the mid 20s, with
imports at about 30 pct of U.S. consumption.
    It said instead that such moves as increasing oil reserves
and more exploration and development research would help to
guard against or mitigate the risks of increased imports.
 Reuter
>
```

(3) Prepare the corpus including converting the text to lower case, removing numbers and punctuation, removing stop words, stemming and identifying synonyms;

```
> ## ------------------------------------------------------------------------
> getTransformations()
[1] "removeNumbers"     "removePunctuation" "removeWords"       "stemDocument"
[5] "stripWhitespace"
>
> ## ----transform_slash-----------------------------------------------------
> toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
> docs <- tm_map(docs, toSpace, "/")
> docs <- tm_map(docs, toSpace, "@")
> docs <- tm_map(docs, toSpace, "\\|")
>
> ## ----eval=FALSE----------------------------------------------------------
> ## docs <- tm_map(docs, toSpace, "/|@|\\|")
>
> ## ----out.lines=26--------------------------------------------------------
> inspect(docs[16])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  16
Content:  chars: 876
```

lower case, removing numbers

```
> ## ------------------------------------------------------------------------
> docs <- tm_map(docs, content_transformer(tolower))
>
> ## ----out.lines=26--------------------------------------------------------
> inspect(docs[16])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  16
Content:  chars: 876

>
> ## ------------------------------------------------------------------------
> docs <- tm_map(docs, removeNumbers)
>
> ## ----out.lines=26--------------------------------------------------------
> viewDocs(docs, 16)
a study group said the united states
should increase its strategic petroleum reserve to one mln
barrels as one way to deal with the present and future impact
of low oil prices on the domestic oil industry.
    u.s. policy now is to raise the strategic reserve to
mln barrels, from its present  mln, to help protect the
economy from an overseas embargo or a sharp price rise.
    the aspen institute for humanistic studies, a private
group, also called for new research for oil exploration and
development techniques.
    it predicted prices would remain at about - dlrs a
barrel for several years and then rise to the mid s, with
imports at about  pct of u.s. consumption.
    it said instead that such moves as increasing oil reserves
and more exploration and development research would help to
guard against or mitigate the risks of increased imports.
 reuter
>
```

removing stop words, stemming and identifying synonyms;

```
> ## ----remove_own_stopwords----------------------------------------------------
> docs <- tm_map(docs, removeWords, c("department", "email"))
>
> ## ----out.lines=26------------------------------------------------------------
> viewDocs(docs, 16)
a study group said the united states
should increase its strategic petroleum reserve to one mln
barrels as one way to deal with the present and future impact
of low oil prices on the domestic oil industry
    us policy now is to raise the strategic reserve to
mln barrels from its present  mln to help protect the
economy from an overseas embargo or a sharp price rise
    the aspen institute for humanistic studies a private
group also called for new research for oil exploration and
development techniques
    it predicted prices would remain at about  dlrs a
barrel for several years and then rise to the mid s with
imports at about  pct of us consumption
    it said instead that such moves as increasing oil reserves
and more exploration and development research would help to
guard against or mitigate the risks of increased imports
 reuter
```

```
> ## ----------------------------------------------------------------------------
> docs <- tm_map(docs, stripWhitespace)
>
> ## ----out.lines=26------------------------------------------------------------
> viewDocs(docs, 16)
a study group said the united states should increase its strategic petroleum reserve$
>
> ## ----specific_transforms-----------------------------------------------------
> toString <- content_transformer(function(x, from, to) gsub(from, to, x))
> docs <- tm_map(docs, toString, "harbin institute technology", "HIT")
> docs <- tm_map(docs, toString, "shenzhen institutes advanced technology", "SIAT")
> docs <- tm_map(docs, toString, "chinese academy sciences", "CAS")
>
> ## ----out.lines=26------------------------------------------------------------
> inspect(docs[16])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  16
Content:  chars: 826

>
> ## ----------------------------------------------------------------------------
> docs <- tm_map(docs, stemDocument)
>
> ## ----out.lines=26------------------------------------------------------------
> viewDocs(docs, 16)
a studi group said the unit state should increas it strateg petroleum reserv to one $
>
> ## ----create_document_term_matrix, out.lines=20-------------------------
> dtm <- DocumentTermMatrix(docs)
>
> dtm
<<DocumentTermMatrix (documents: 20, terms: 849)>>
Non-/sparse entries: 1878/15102
Sparsity           : 89%
Maximal term length: 16
Weighting          : term frequency (tf)
>
```

(4) Create a document term matrix

(5) Explore the Document Term Matrix by converting the document term matrix into a matrix and summing the column counts

(6) Remove Sparse Terms;

```
> ## ----dtm_matrix-------------------------------------------------------
> class(dtm)
[1] "DocumentTermMatrix"    "simple_triplet_matrix"
> dim(dtm)
[1]  20 849
>
> ## ----create_term_document_matrix, out.lines=20-------------------------
> tdm <- TermDocumentMatrix(docs)
> tdm
<<TermDocumentMatrix (terms: 849, documents: 20)>>
Non-/sparse entries: 1878/15102
Sparsity           : 89%
Maximal term length: 16
Weighting          : term frequency (tf)
>
> ## ---------------------------------------------------------------------
> freq <- colSums(as.matrix(dtm))
> length(freq)
[1] 849
>
> ## ----out.lines=10-----------------------------------------------------
> ord <- order(freq)
>
> # Least frequent terms.
> freq[head(ord)]
     abl   abroad   accept   across      add advantag
       1        1        1        1        1        1
>
> ## ---------------------------------------------------------------------
> # Most frequent terms.
> freq[tail(ord)]
  for price  said   and   oil   the
   52    63    73    77    85   232
>
> ## ---------------------------------------------------------------------
> # Frequency of frequencies.
> head(table(freq), 15)
freq
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
419 139  94  38  31  24  18  17  18   5   9   5   3   4   1
```

```
> tail(table(freq), 15)
freq
 18  21  23  24  26  28  30  31  47  52  63  73  77  85 232
  3   2   2   1   1   1   1   2   1   1   1   1   1   1   1
>
> ## ----dtm_to_m---------------------------------------------------------
> m <- as.matrix(dtm)
> dim(m)
[1]  20 849
```

```
> ## ----remove_sparse_terms----------------------------------------------------
> dim(dtm)
[1]  20 849
> dtms <- removeSparseTerms(dtm, 0.9)
> dim(dtms)
[1]  20 355
>
> ## ---------------------------------------------------------------------------
> inspect(dtms)
<<DocumentTermMatrix (documents: 20, terms: 355)>>
Non-/sparse entries: 1384/5716
Sparsity           : 81%
Maximal term length: 11
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  and for market mln oil opec price said that the
  144   9   5      5   4  12   15     6   11   10  19
  236   7   4      3   4   7    8     8   10    4  15
  237  11   4      0   1   3    1     1    1    1  30
  242   3   1      2   0   3    2     2    3    0   6
  246   9   6      0   0   5    2     2    5    2  18
  248   6   2     10   3   9    6    10    7    2  27
  273   5   4      1   9   5    5     5    8    0  21
  489   5   4      0   3   4    0     3    2    1   8
  502   6   5      0   3   5    0     3    2    1  13
  704   5   4      3   0   3    0     3    4    3  21
```

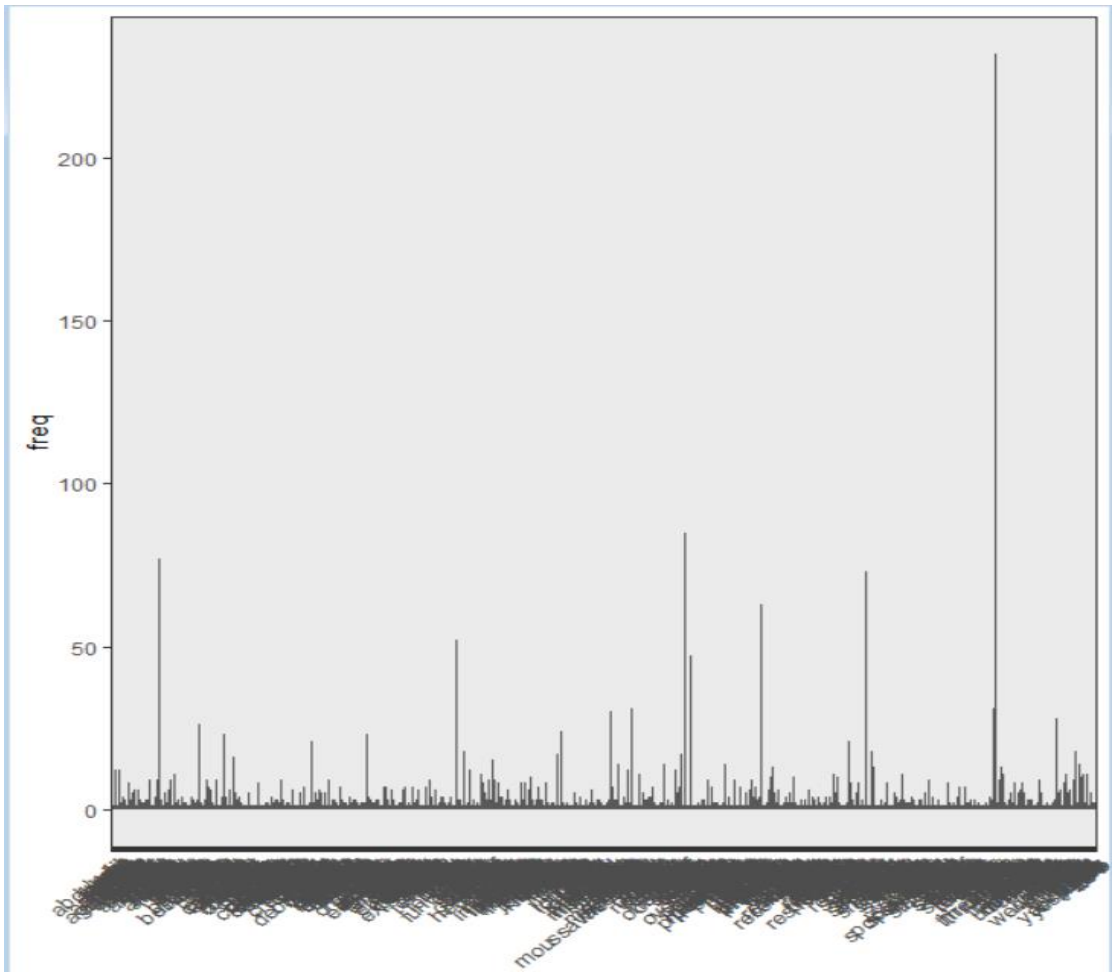
(7) Identify Frequent Items and Associations;

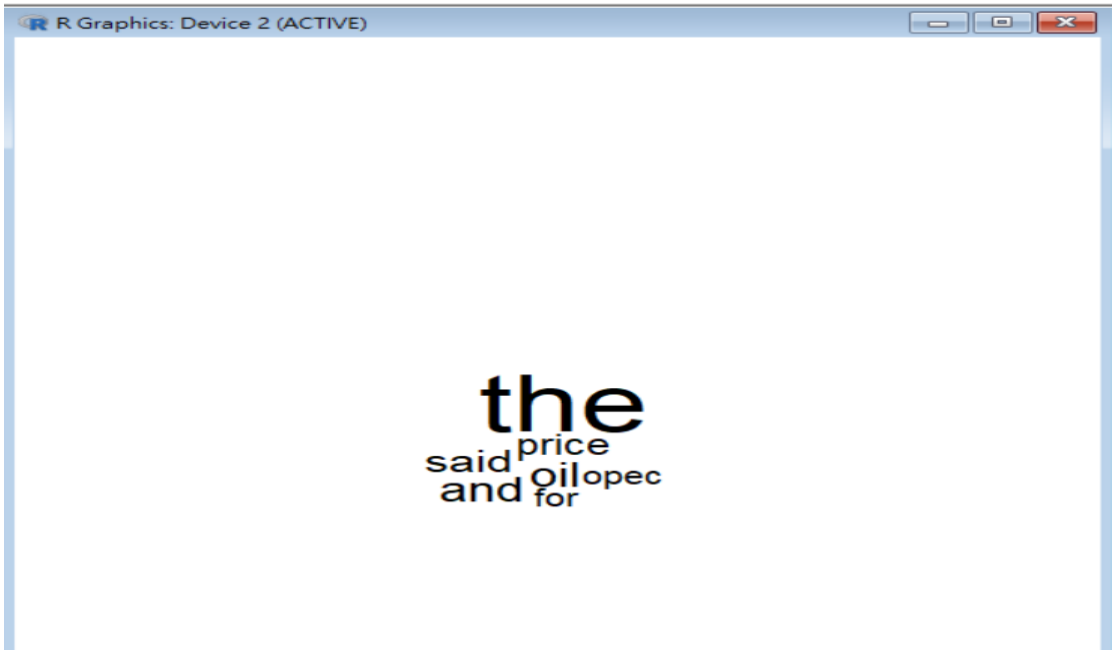
```
> freq <- colSums(as.matrix(dtms))
> freq
       abil       about        abov      accord       activ       adher       after
          6          12           8          12           2           3           8
    against       agenc        agre   agreement  agricultur         ali    alkhalifa
          3           5           3           6           3           6           3
        all       along      alqaba    alsabah        also       among     analyst
          3           2           3           3           9           4           9
        and         ani      announc       appar      appear      approv       april
         77           3           3           3           5           2           6
       arab      arabia     arabian   architect         are      around         ask
          5           9           3           2          11           3           4
      aspen      assign        back     bahrain      barrel        base         bbl
          2           2           3           2          26           2           3
     becaus        been       befor       below   benchmark     billion       boost
          9           8           7           6           2           9           3
       both         bpd       bring         but       buyer        call         can
          4          23           4          16           5           3           4
       cash        ceil       chang    circumst        cite       clear       close
          2           5           8           2           2           2           4
      coast        come      commit      commod     compani      compar      condit
          2           3           3           2           9           3           2
   consumpt     continu    contract        corp     countri       crude     current
          2           4           6           5           7          21           5
     custom         cut       daili         day        deal      decemb      declin
          3           6           5           5           2           9           3
    decreas       deliv      demand        deni      deputi     develop         did
          3           2           7           3           2           4           3
 differenti  difficulti   distribut         dlr        dlrs      dollar      domest
          3           3           2           2          23           4           3
       down        drop         due        each       earli      econom     economi
          2           2           3           3           2           7           7
    ecuador      effect     embargo       emerg        emir      energi       estim
          3           6           2           6           5           7           7
       even      exceed      excess      expect      explor      export        face
          2           2           3           7           4           9           4
       fall         feb    februari        fell       figur       first      fiscal
          6           2           4           4           2           2           4
        fix         for        forc     foreign        four        free        from
          3          52           3           3           3           4          18
      fulli       futur     general       given      govern       grade       group
          2          12           3           2          11           5           8
```

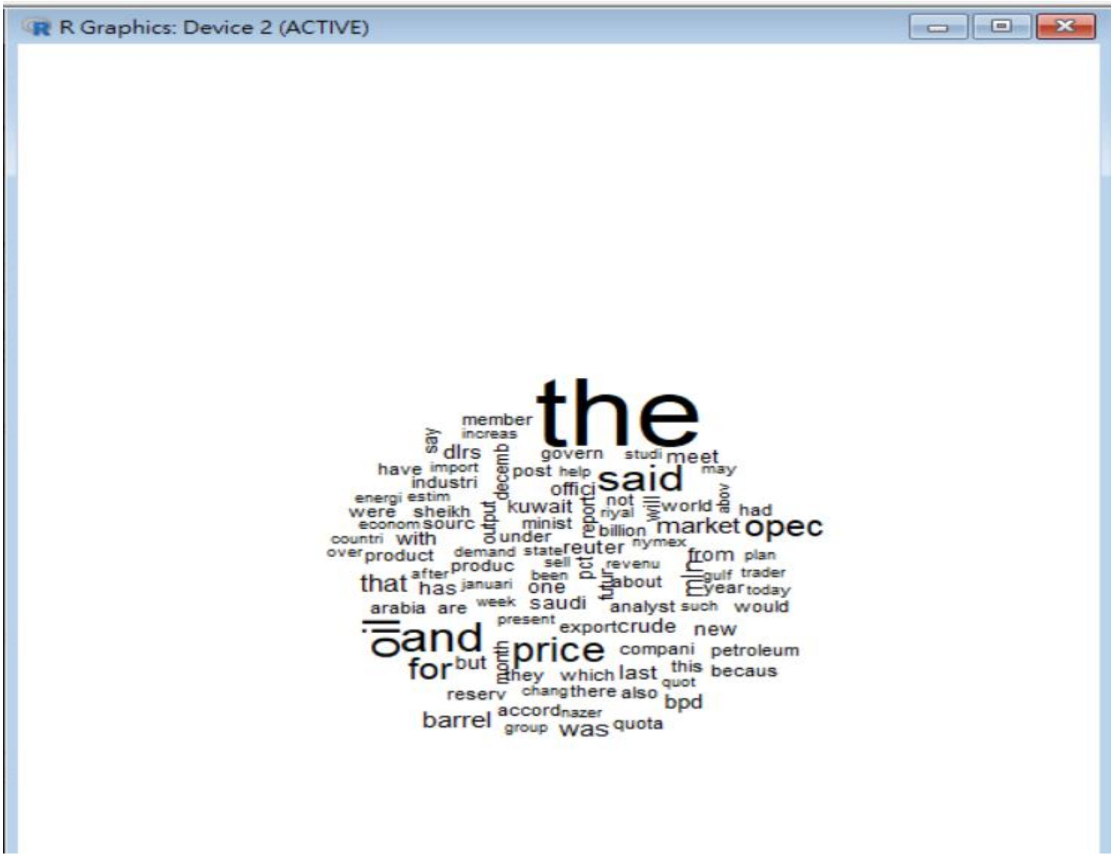
```
> table(freq)
freq
  2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  21  23  24  26
 89  82  35  26  20  17  17  18   5   9   5   3   4   1   1   3   3   2   2   1   1
 28  30  31  47  52  63  73  77  85 232
  1   1   2   1   1   1   1   1   1   1
>
> ## ----freq_terms_1000------------------------------------------------------
> findFreqTerms(dtm, lowfreq=1000)
character(0)
>
> ## ----freq_terms_100-------------------------------------------------------
> findFreqTerms(dtm, lowfreq=100)
[1] "the"
>
> ## ----assoc----------------------------------------------------------------
> findAssocs(dtm, "data", corlimit=0.6)
$data
numeric(0)
```

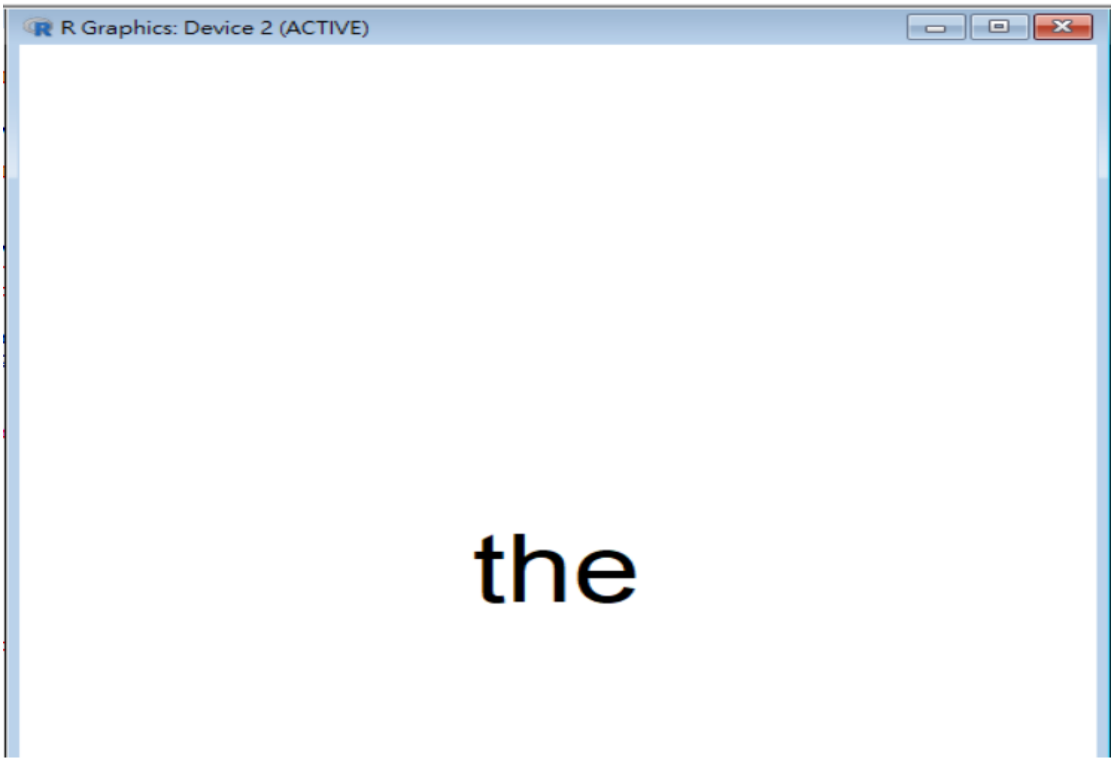
```
> ## ----word_count-----------------------------------------------------------
> freq <- sort(colSums(as.matrix(dtm)), decreasing=TRUE)
> head(freq, 14)
   the    oil    and   said  price    for   opec    mln   that market    was barrel
   232     85     77     73     63     52     47     31     31     30     28     26
  last    bpd
    24     23
> wf   <- data.frame(word=names(freq), freq=freq)
> head(wf)
      word freq
the    the  232
oil    oil   85
and    and   77
said  said   73
price price   63
for    for   52
```

(8) Draw Correlations Plots;
(9) Plot Word Frequencies;
(10) Draw Word Clouds
(11) Perform quantitative analysis of text;



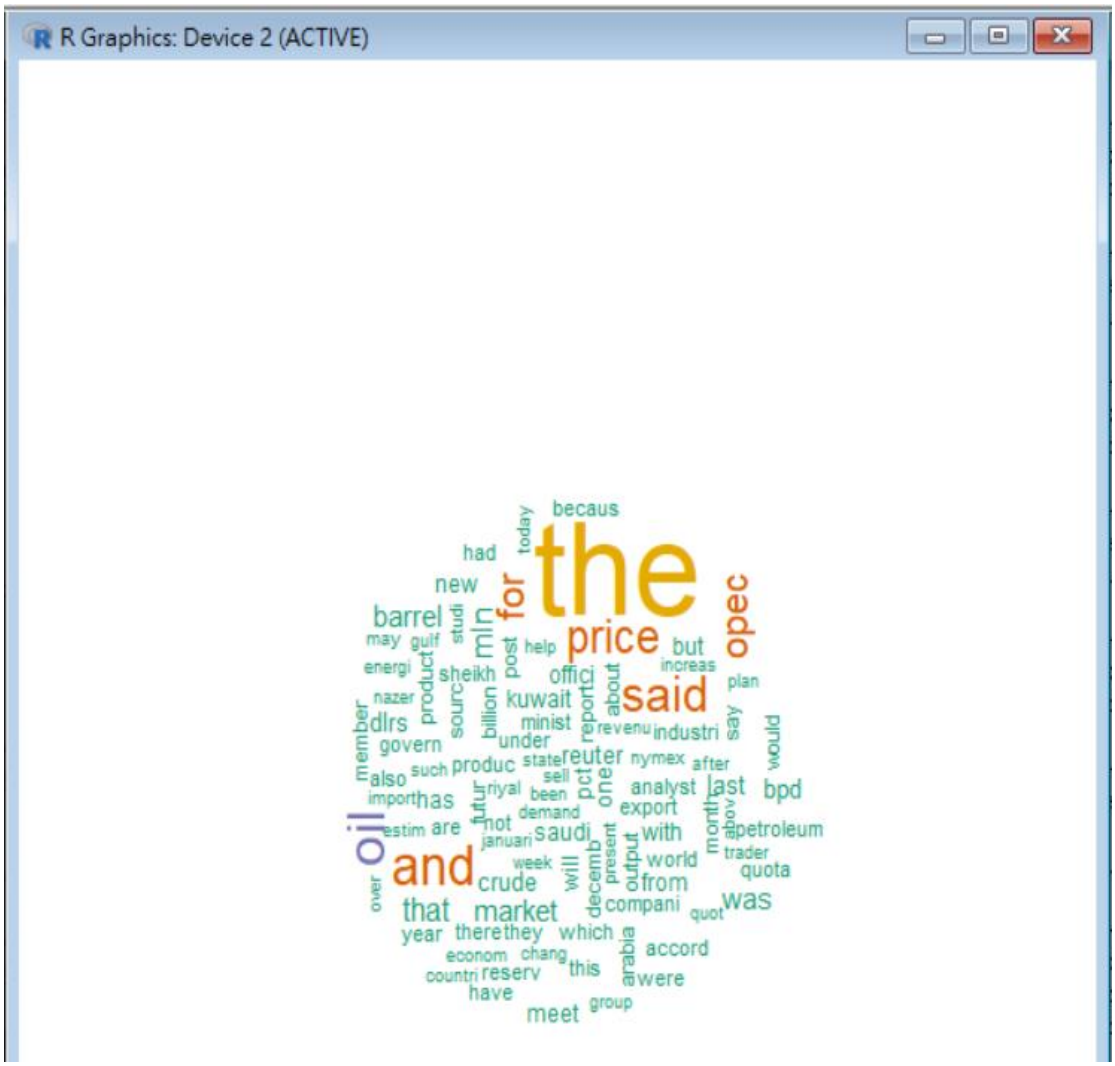



R Graphics: Device 2 (ACTIVE)
the
said
opec
oil
and
price
for
that
market
kuwait
mln
from
member
increas
say
dlrs
govern
studi
meet
have
import
post
help
may
industri
offici
energi
estim
not
will
world
abov
had
were
sheikh
decemb
output
riyal
report
econom
sourc
minist
billion
countri
with
under
nymex
over
product
demand
state
reuter
plan
sell
revenu
gulf
trader
after
produc
been
pct
futur
year
today
has
januari
one
about
arabia
are
week
saudi
analyst
such
would
present
export
crude
new
companie
petroleum
but
month
they
which
last
this
becaus
reserv
chang
there
also
quot
barrel
accord
nazer
bpd
group
was
quota


R Graphics: Device 2 (ACTIVE)
the


R Graphics: Device 2 (ACTIVE)
the
said
oil
and
price
for
opec
market
that
has
dlrs
member
increas
govern
meet
import
industri
post
help
offici
kuwait
reporti
billion
world
had
sheikh
sourc
minist
under
reuter
nymex
from
product
demand
state
sell
revenu
produc
been
about
year
today
januari
one
saudi
analyst
such
would
arabia
are
week
present
export
crude
new
compani
petroleum
month
they
which
last
this
becaus
reserv
there
also
bpd
barrel
accord
group
was
quota

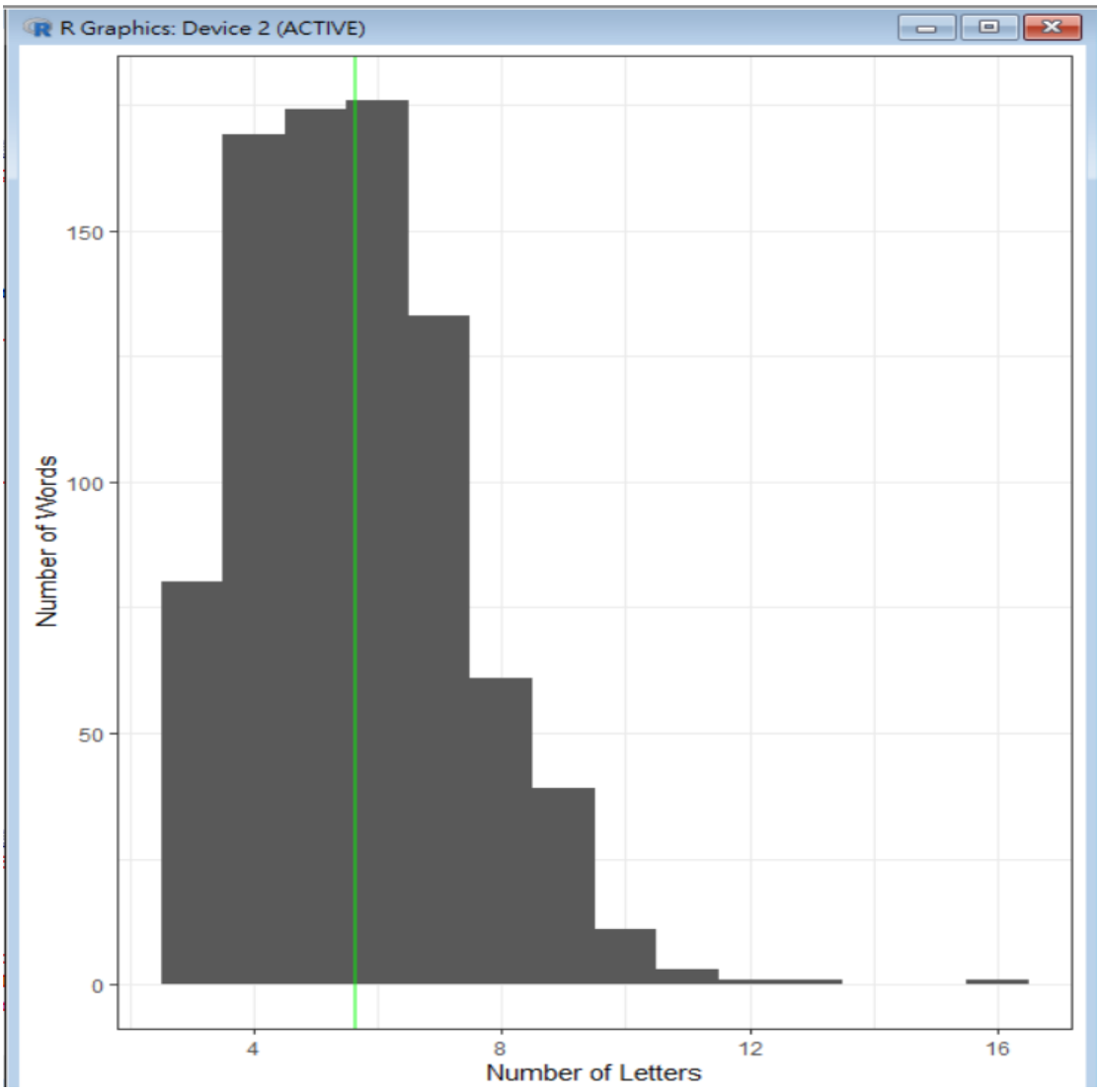

R Graphics: Device 2 (ACTIVE)
oil
opec
price
said
for
the
and
market
barrel
that
was
bpd
dlrs
last
reuter
crude

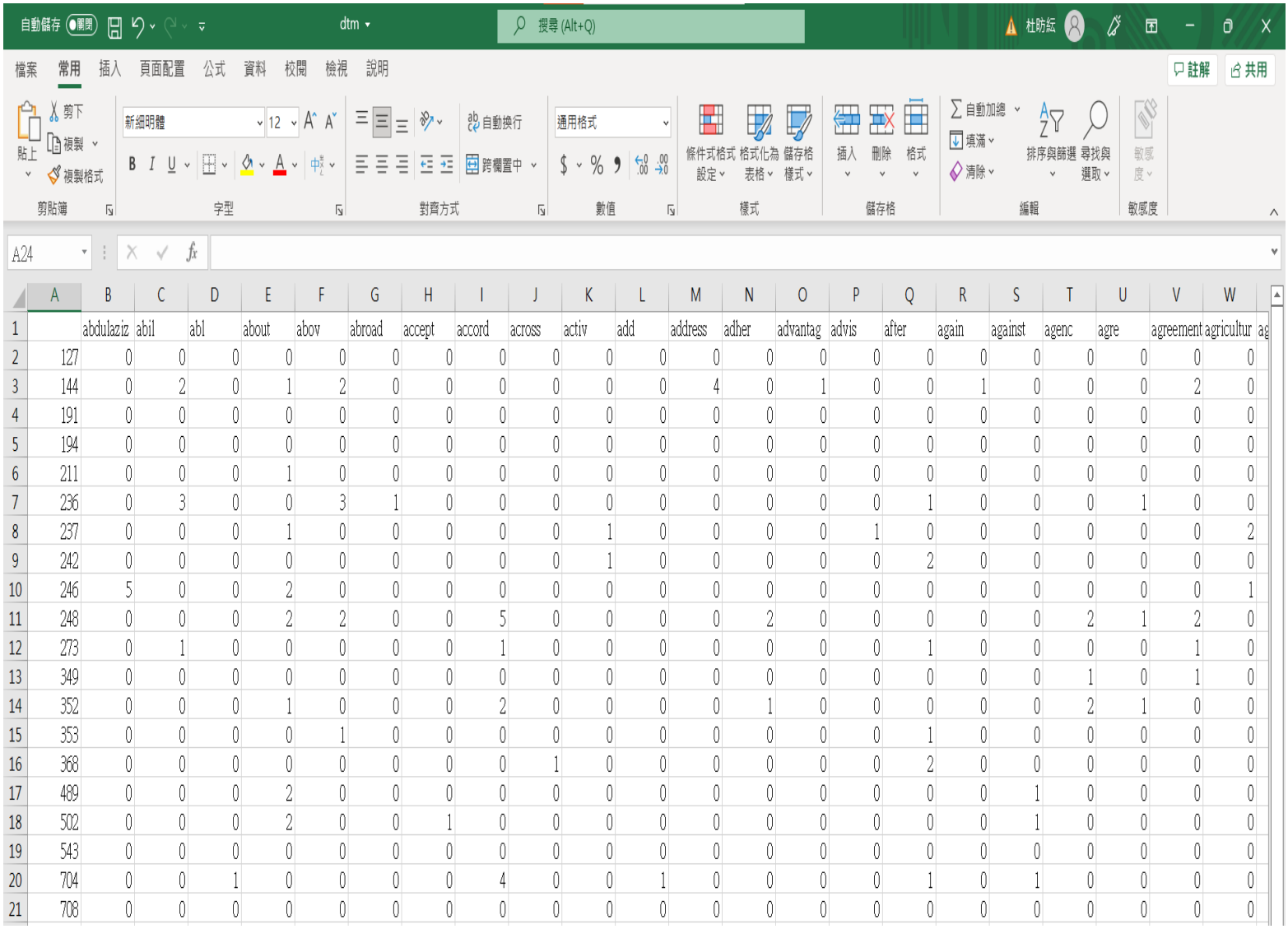


Create words list

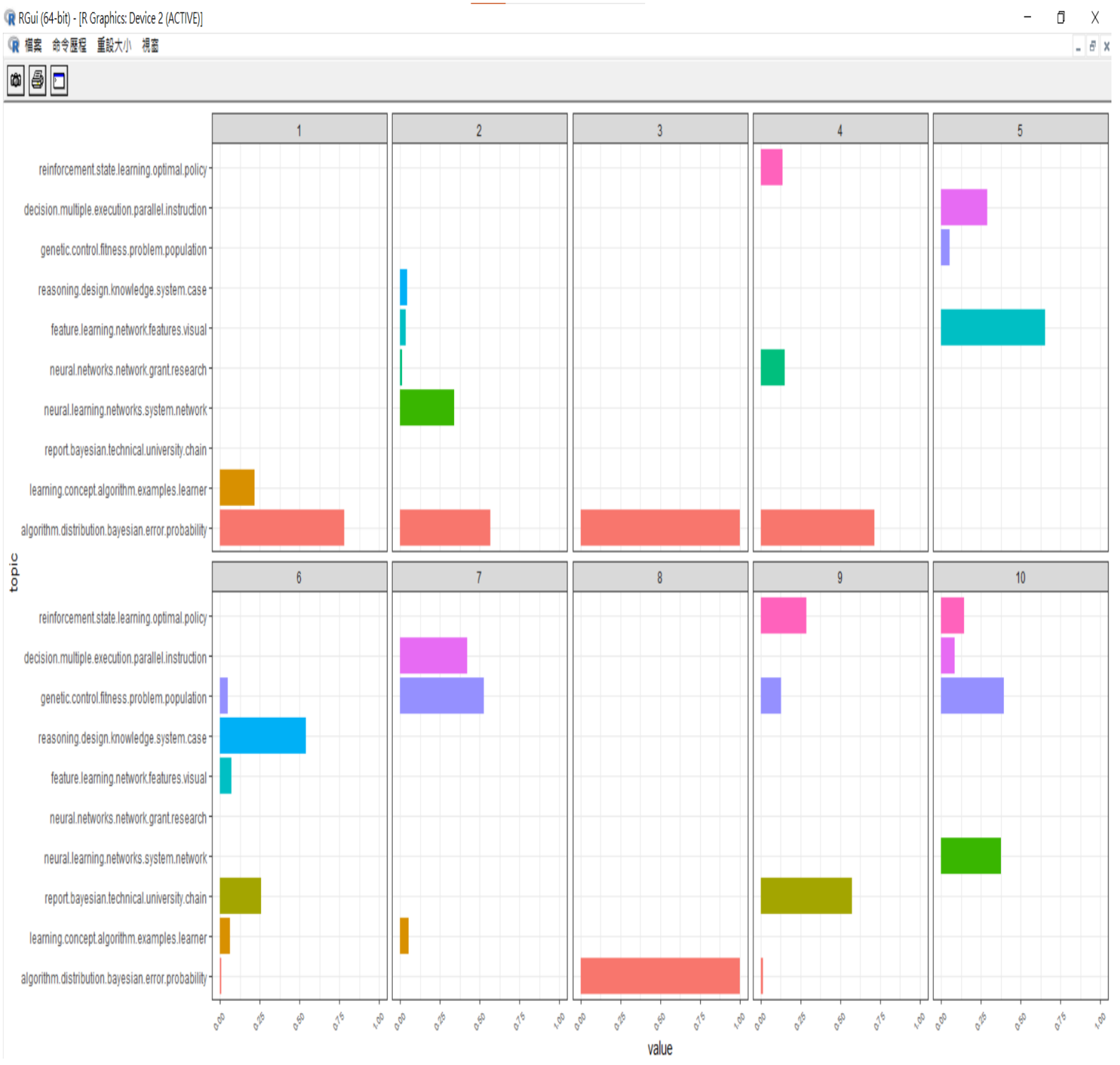
```
> ## ----qdap_create_word_list-------------------------------------------------
> words <- dtm                                                            %>%
+   as.matrix                                                             %>%
+   colnames                                                              %>%
+   (function(x) x[nchar(x) < 20])
>
> ## ----qdap_word_length, out.lines=11----------------------------------------
> length(words)
[1] 849
> head(words, 15)
 [1] "abdulaziz" "abil"      "abl"       "about"     "abov"      "abroad"
 [7] "accept"    "accord"    "across"    "activ"     "add"       "address"
[13] "adher"     "advantag"  "advis"
> summary(nchar(words))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   6.000   5.649   7.000  16.000
> table(nchar(words))

  3   4   5   6   7   8   9  10  11  12  13  16
 80 169 174 176 133  61  39  11   3   1   1   1
```


R Graphics: Device 2 (ACTIVE)
150
100
50
0
Number of Words
4
8
12
16
Number of Letters



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		abdulaziz	abil	abl	about	abov	abroad	accept	accord	across	activ	add	address	adher	advantag	advis	after	again	against	agenc	agre	agreement	agricultur	ag
2	127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	144	0	2	0	1	2	0	0	0	0	0	0	4	0	1	0	0	1	0	0	0	2	0
4	191	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	194	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	211	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	236	0	3	0	0	3	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
8	237	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	2
9	242	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0
10	246	5	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	248	0	0	0	2	2	0	0	5	0	0	0	0	2	0	0	0	0	0	2	1	2	0
12	273	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0
13	349	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
14	352	0	0	0	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	2	1	0	0
15	353	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
16	368	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0
17	489	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
18	502	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
19	543	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	704	0	0	1	0	0	0	0	4	0	0	1	0	0	0	0	1	0	1	0	0	0	0
21	708	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0


RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]
檔案 命令歷程 重設大小 視窗
1
2
3
4
5
6
7
8
9
10
reinforcement.state.learning.optimal.policy
decision.multiple.execution.parallel.instruction
genetic.control.fitness.problem.population
reasoning.design.knowledge.system.case
feature.learning.network.features.visual
neural.networks.network.grant.research
neural.learning.networks.system.network
report.bayesian.technical.university.chain
learning.concept.algorithm.examples.learner
algorithm.distribution.bayesian.error.probability
topic
0.00
0.25
0.50
0.75
1.00
value