# PSet 4

Jingpeng Hong

3/5/2022

### Question 1

#### a.

Load the data as a corpus.

```
texts = file.path("SimpleText_auto")
docs_raw = VCorpus(DirSource(texts))
```

#### b.

Clean the data.

```
docs = docs_raw %>%
  tm_map(content_transformer(tolower)) %>% # transform all characters to lowercase
  tm_map(removeWords, stopwords("english")) %>% # remove stop words
  tm_map(removeWords, c('table', 'figure', 'results', 'use', 'can', 'also')) %>%
  tm_map(removePunctuation) %>% # remove punctuation
  tm_map(removeNumbers) %>%
  tm_map(stripWhitespace) %>% # remove excess whitespace
  tm_map(stemDocument) # get to words' roots
```

Justify our answers.

```
docs_raw[[1]]$content[4]
```

```
## [1] "We prove an integrality property of the Chern character with values in Chow groups."
```

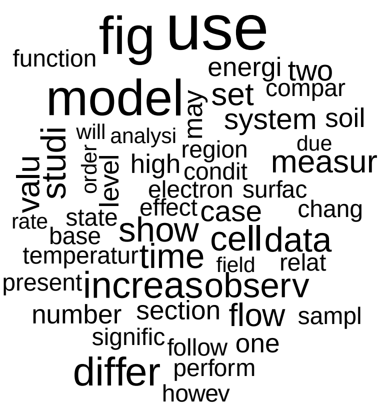
```
docs[[1]]$content[4]
```

```
## [1] "prove integr properti chern charact valu chow group"
```

#### c.

Present the 50 most frequently used words in the corpus in an Word Cloud.

```
wordcloud(docs, max.words = 50, scale=c(2, .2))
```



### d.

Fit a topic model on the corpus setting k equal to 2, 3, 5, 8, and 10.

```
set.seed(123)
dtm = DocumentTermMatrix(docs)

# define the function of topic models with k.
topic = function(k){
  lda = LDA(dtm, k = k, method = "Gibbs", control = list(burnin = 100, iter = 1000))
  topics = tidy(lda, matrix = "beta")
  topwords = topics %>%
    group_by(topic) %>%
    top_n(10, beta) %>%
    ungroup() %>%
    arrange(topic, -beta) # print the words with the highest beta from each topic
  for (i in 1:k) {
    print(topwords %>% filter(topic==i))
  }
}
```

```
topic(2)
```

```
## # A tibble: 10 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     1 model  0.00826
##  2     1 use    0.00734
##  3     1 fig    0.00732
##  4     1 flow   0.00496
##  5     1 time   0.00477
##  6     1 differ 0.00475
##  7     1 observ 0.00465
##  8     1 valu   0.00434
##  9     1 show   0.00408
## 10     1 energi 0.00399
## # A tibble: 10 x 3
##    topic term       beta
##    <int> <chr>     <dbl>
##  1     2 cell    0.00752
##  2     2 use     0.00631
##  3     2 soil    0.00587
##  4     2 studi   0.00407
##  5     2 express 0.00359
##  6     2 set     0.00339
##  7     2 plant   0.00334
##  8     2 increas 0.00330
##  9     2 gene    0.00327
## 10     2 follow  0.00323
```

The first topic is hard to be defined as a si

```
topic(3)
```

```
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 fig        0.00868
##  2     1 model      0.00798
##  3     1 use        0.00715
##  4     1 observ     0.00550
##  5     1 energi     0.00524
##  6     1 increas    0.00482
##  7     1 valu       0.00480
##  8     1 electron   0.00465
##  9     1 flow       0.00462
## 10     1 temperatur 0.00460
## # A tibble: 10 x 3
##    topic term       beta
##    <int> <chr>     <dbl>
##  1     2 cell    0.0115
##  2     2 soil    0.00903
##  3     2 use     0.00737
##  4     2 studi   0.00520
##  5     2 plant   0.00513
##  6     2 express 0.00509
##  7     2 gene    0.00502
##  8     2 differ  0.00401
##  9     2 level   0.00392
## 10     2 data    0.00376
## # A tibble: 10 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     3 set       0.00836
##  2     3 state     0.00631
##  3     3 node      0.00604
##  4     3 use       0.00599
##  5     3 cluster   0.00598
##  6     3 algorithm 0.00587
##  7     3 function  0.00536
##  8     3 time      0.00514
##  9     3 number    0.00508
## 10     3 let       0.00487
```

The first topic

```
topic(5)
```

```
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 cell       0.0178
##  2     1 use        0.00837
##  3     1 express    0.00787
##  4     1 gene       0.00769
##  5     1 particl    0.00598
##  6     1 cultur     0.00525
##  7     1 differenti 0.00518
##  8     1 studi      0.00460
##  9     1 human      0.00420
## 10     1 polym      0.00406
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     2 cluster  0.0152
##  2     2 set      0.0136
##  3     2 let      0.0118
##  4     2 function 0.00890
##  5     2 order    0.00763
##  6     2 theorem  0.00746
##  7     2 case     0.00718
##  8     2 follow   0.00709
##  9     2 oper     0.00705
## 10     2 lemma    0.00642
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     3 soil     0.0128
##  2     3 increas  0.00661
##  3     3 fig      0.00658
##  4     3 use      0.00593
##  5     3 plant    0.00544
##  6     3 site     0.00535
##  7     3 sampl    0.00489
##  8     3 sediment 0.00485
##  9     3 concentr 0.00466
## 10     3 chang    0.00456
## # A tibble: 10 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     4 state     0.00841
##  2     4 node      0.00786
##  3     4 algorithm 0.00764
##  4     4 use       0.00727
##  5     4 time      0.00701
##  6     4 model     0.00646
##  7     4 network   0.00643
##  8     4 differ    0.00576
##  9     4 perform   0.00529
## 10     4 number    0.00516
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     5 model      0.0101
##  2     5 fig        0.00878
##  3     5 energi     0.00733
##  4     5 observ     0.00709
##  5     5 use        0.00694
##  6     5 flow       0.00660
##  7     5 electron   0.00653
##  8     5 temperatur 0.00568
##  9     5 measur     0.00526
## 10     5 wind       0.00521
```

```
topic(8)
```

```
## # A tibble: 10 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     1 flow   0.0135
##  2     1 turbin 0.00997
##  3     1 fig    0.00889
##  4     1 wind   0.00769
##  5     1 use    0.00706
##  6     1 speed  0.00687
##  7     1 differ 0.00656
##  8     1 power  0.00613
##  9     1 model  0.00605
## 10     1 veloc  0.00605
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     2 model      0.0124
##  2     2 electron   0.0116
##  3     2 observ     0.0112
##  4     2 energi     0.0103
##  5     2 fig        0.00736
##  6     2 temperatur 0.00729
##  7     2 atmospher  0.00693
##  8     2 region     0.00685
##  9     2 flux       0.00657
## 10     2 field      0.00595
## # A tibble: 10 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     3 cluster   0.0158
##  2     3 program   0.0105
##  3     3 use       0.00905
##  4     3 soil      0.00889
##  5     3 size      0.00681
##  6     3 model     0.00669
##  7     3 depend    0.00656
##  8     3 algorithm 0.00615
##  9     3 memori    0.00573
## 10     3 number    0.00571
## # A tibble: 10 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     4 state  0.0142
##  2     4 modul  0.0115
##  3     4 node   0.0109
##  4     4 time   0.00863
##  5     4 use    0.00673
##  6     4 beach  0.00629
##  7     4 model  0.00609
##  8     4 featur 0.00606
##  9     4 system 0.00542
## 10     4 port   0.00500
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     5 let      0.0150
##  2     5 set      0.0123
##  3     5 theorem  0.00932
##  4     5 order    0.00923
##  5     5 oper     0.00893
##  6     5 follow   0.00838
##  7     5 lemma    0.00821
##  8     5 element  0.00747
##  9     5 proof    0.00747
## 10     5 function 0.00741
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     6 soil     0.0104
##  2     6 fig      0.00752
##  3     6 increas  0.00715
##  4     6 sediment 0.00683
##  5     6 site     0.00679
##  6     6 sampl    0.00599
##  7     6 chang    0.00571
##  8     6 use      0.00558
##  9     6 concentr 0.00524
## 10     6 carbon   0.00515
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     7 particl    0.00959
##  2     7 use        0.00935
##  3     7 temperatur 0.00609
##  4     7 measur     0.00574
##  5     7 materi     0.00567
##  6     7 surfac     0.00547
##  7     7 fig        0.00529
##  8     7 increas    0.00529
##  9     7 film       0.00527
## 10     7 polym      0.00516
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     8 cell       0.0234
##  2     8 express    0.00795
##  3     8 cultur     0.00722
##  4     8 use        0.00708
##  5     8 differenti 0.00700
##  6     8 studi      0.00592
##  7     8 human      0.00580
##  8     8 gene       0.00566
##  9     8 mutat      0.00507
## 10     8 neuron     0.00495
```

```
topic(10)
```

```
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     1 cell       0.0293
##  2     1 cultur     0.00882
##  3     1 express    0.00882
##  4     1 differenti 0.00856
##  5     1 use        0.00704
##  6     1 human      0.00690
##  7     1 mutat      0.00626
##  8     1 neuron     0.00614
##  9     1 gene       0.00607
## 10     1 hipsc      0.00549
## # A tibble: 10 x 3
##    topic term       beta
##    <int> <chr>     <dbl>
##  1     2 soil    0.0140
##  2     2 plant   0.0110
##  3     2 gene    0.00786
##  4     2 tree    0.00701
##  5     2 use     0.00605
##  6     2 speci   0.00596
##  7     2 level   0.00580
##  8     2 increas 0.00563
##  9     2 differ  0.00530
## 10     2 effect  0.00495
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     3 let      0.0150
##  2     3 set      0.0120
##  3     3 theorem  0.00956
##  4     3 order    0.00883
##  5     3 oper     0.00877
##  6     3 lemma    0.00810
##  7     3 follow   0.00807
##  8     3 element  0.00767
##  9     3 function 0.00739
## 10     3 proof    0.00736
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     4 soil     0.0180
##  2     4 sediment 0.0113
##  3     4 data     0.0103
##  4     4 ratio    0.00828
##  5     4 object   0.00679
##  6     4 plant    0.00635
##  7     4 use      0.00627
##  8     4 root     0.00622
##  9     4 sampl    0.00557
## 10     4 month    0.00552
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     5 particl    0.0121
##  2     5 use        0.0109
##  3     5 fig        0.00723
##  4     5 materi     0.00692
##  5     5 film       0.00680
##  6     5 polym      0.00666
##  7     5 measur     0.00661
##  8     5 surfac     0.00659
##  9     5 temperatur 0.00645
## 10     5 sampl      0.00626
## # A tibble: 10 x 3
##    topic term          beta
##    <int> <chr>        <dbl>
##  1     6 electron   0.0139
##  2     6 observ     0.0126
##  3     6 energi     0.0121
##  4     6 model      0.0115
##  5     6 temperatur 0.00720
##  6     6 region     0.00682
##  7     6 fig        0.00680
##  8     6 field      0.00666
##  9     6 flux       0.00643
## 10     6 ion        0.00637
## # A tibble: 10 x 3
##    topic term       beta
##    <int> <chr>     <dbl>
##  1     7 cluster 0.0186
##  2     7 use     0.0108
##  3     7 studi   0.0107
##  4     7 program 0.00986
##  5     7 set     0.00730
##  6     7 depend  0.00701
##  7     7 data    0.00675
##  8     7 network 0.00670
##  9     7 featur  0.00662
## 10     7 slice   0.00638
## # A tibble: 10 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     8 fig      0.00955
##  2     8 increas  0.00737
##  3     8 carbon   0.00647
##  4     8 deposit  0.00588
##  5     8 element  0.00537
##  6     8 site     0.00531
##  7     8 concentr 0.00524
##  8     8 model    0.00485
##  9     8 chang    0.00478
## 10     8 may      0.00439
## # A tibble: 10 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     9 flow   0.0134
##  2     9 turbin 0.0124
##  3     9 wind   0.00994
##  4     9 fig    0.00986
##  5     9 model  0.00978
##  6     9 veloc  0.00861
##  7     9 use    0.00859
##  8     9 speed  0.00828
##  9     9 power  0.00761
## 10     9 simul  0.00594
## # A tibble: 10 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1    10 state     0.0121
##  2    10 node      0.0111
##  3    10 algorithm 0.0106
##  4    10 time      0.00898
##  5    10 model     0.00624
##  6    10 execut    0.00620
##  7    10 modul     0.00612
##  8    10 memori    0.00578
##  9    10 number    0.00578
## 10    10 perform   0.00578
```