Perlombongan Data Teks

Data berbentuk tak berstruktur

1. Corpus Text
2. Pembersihan Data

```
text = readLines("G:/My Drive/Master-Data-Science/Semester_1/Data_Mining/Data/text.txt")
```

```
## Warning in readLines("G:/My
## Drive/Master-Data-Science/Semester_1/Data_Mining/Data/text.txt"): incomplete
## final line found on 'G:/My
## Drive/Master-Data-Science/Semester_1/Data_Mining/Data/text.txt'
```

```
class(text)
```

```
## [1] "character"
```

Jelmakan data kepada corpus data

```
docs = Corpus(VectorSource(text))
inspect(docs)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 46
##
##  [1]
##  [2] And so even though we face the difficulties of today and tomorrow, I still have a dream. It is a
##  [3]
##  [4] I have a dream that one day this nation will rise up and live out the true meaning of its creed
##  [5]
##  [6] We hold these truths to be self-evident, that all men are created equal.
##  [7]
##  [8] I have a dream that one day on the red hills of Georgia, the sons of former slaves and the sons
##  [9]
## [10] I have a dream that one day even the state of Mississippi, a state sweltering with the heat of i
## [11]
## [12] I have a dream that my four little children will one day live in a nation where they will not be
## [13]
## [14] I have a dream today!
## [15]
## [16] I have a dream that one day, down in Alabama, with its vicious racists, with its governor having
## [17]
## [18] I have a dream today!
## [19]
## [20] I have a dream that one day every valley shall be exalted, and every hill and mountain shall be
## [21]
## [22] This is our hope, and this is the faith that I go back to the South with.
## [23]
## [24] With this faith, we will be able to hew out of the mountain of despair a stone of hope. With thi
## [25]
## [26] And this will be the day, this will be the day when all of God s children will be able to sing w
## [27]
## [28] My country  tis of thee, sweet land of liberty, of thee I sing.
## [29] Land where my fathers died, land of the Pilgrim s pride,
## [30] From every mountainside, let freedom ring!
## [31] And if America is to be a great nation, this must become true.
## [32] And so let freedom ring from the prodigious hilltops of New Hampshire.
## [33] Let freedom ring from the mighty mountains of New York.
## [34] Let freedom ring from the heightening Alleghenies of Pennsylvania.
## [35] Let freedom ring from the snow-capped Rockies of Colorado.
## [36] Let freedom ring from the curvaceous slopes of California.
## [37]
## [38] But not only that:
## [39] Let freedom ring from Stone Mountain of Georgia.
## [40] Let freedom ring from Lookout Mountain of Tennessee.
## [41] Let freedom ring from every hill and molehill of Mississippi.
## [42] From every mountainside, let freedom ring.
## [43] And when this happens, when we allow freedom ring, when we let it ring from every village and ev
## [44] Free at last! Free at last!
## [45]
## [46] Thank God Almighty, we are free at last!
```

```
class(docs)
```

```
## [1] "SimpleCorpus" "Corpus"
```

Pembersihan teks

1. Keluarkan aksara khas daripada teks, iaitu simbol-simbol; /, @, | akan digantikan dengan ruang kosong.

```
toSpace = content_transformer(function(x, pattern)
  gsub(pattern, "",x))
```

gantikan semua simbol yang nak dikeluarkan daripada teks dengan ruang kosong

```
docs2 = tm_map(docs, toSpace, "!")
```

```
## Warning in tm_map.SimpleCorpus(docs, toSpace, "!"): transformation drops
## documents
```

```r
docs3 = tm_map(docs2, toSpace, ":")
```

```
## Warning in tm_map.SimpleCorpus(docs2, toSpace, ":"): transformation drops
## documents
```

```r
docs4 = tm_map(docs3, toSpace, ",")
```

```
## Warning in tm_map.SimpleCorpus(docs3, toSpace, ","): transformation drops
## documents
```

2. Tukar teks huruf besar kepada huruf kecil.

```r
docs5 = tm_map(docs4, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(docs4, content_transformer(tolower)):
## transformation drops documents
```

3. Keluarkan nombor-nombor.

```r
docs6 = tm_map(docs5, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(docs5, removeNumbers): transformation drops
## documents
```

4. Keluarkan kata henti (stopwords). Contoh kata henti dalam bahasa Inggeris “the, is, at, on”. Tiada senarai semesta (universal) kata henti yang digunakan dalam NLP.

```r
docs7 = tm_map(docs6, removeWords, stopwords("english"))
```

```
## Warning in tm_map.SimpleCorpus(docs6, removeWords, stopwords("english")):
## transformation drops documents
```

5. Keluaran tanda baca (punctuation).

```r
docs8 = tm_map(docs7, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(docs7, removePunctuation): transformation drops
## documents
```

6. Buang semua ruang tambahan yang tidak perlu dalam teks.

```r
docs9 = tm_map(docs8, stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(docs8, stripWhitespace): transformation drops
## documents
```

```r
inspect(docs9)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 46
##
##  [1]
##  [2]  even though face difficulties today tomorrow still dream dream deeply rooted american dream
##  [3]
##  [4]  dream one day nation will rise live true meaning creed
##  [5]
##  [6]  hold truths selfevident men created equal
##  [7]
##  [8]  dream one day red hills georgia sons former slaves sons former slave owners will able sit toge
##  [9]
## [10]  dream one day even state mississippi state sweltering heat injustice sweltering heat oppressio
## [11]
## [12]  dream four little children will one day live nation will judged color skin content character
## [13]
## [14]  dream today
## [15]
## [16]  dream one day alabama vicious racists governor lips dripping words interposition nullification
## [17]
## [18]  dream today
## [19]
## [20]  dream one day every valley shall exalted every hill mountain shall made low rough places will
## [21]
## [22]  hope faith go back south
## [23]
## [24]  faith will able hew mountain despair stone hope faith will able transform jangling discords na
## [25]
## [26]  will day will day god s children will able sing new meaning
## [27]
## [28]  country tis thee sweet land liberty thee sing
## [29] land fathers died land pilgrim s pride
## [30]  every mountainside let freedom ring
## [31]  america great nation must become true
## [32]  let freedom ring prodigious hilltops new hampshire
## [33] let freedom ring mighty mountains new york
## [34] let freedom ring heightening alleghenies pennsylvania
## [35] let freedom ring snowcapped rockies colorado
## [36] let freedom ring curvaceous slopes california
## [37]
## [38]
## [39] let freedom ring stone mountain georgia
## [40] let freedom ring lookout mountain tennessee
## [41] let freedom ring every hill molehill mississippi
## [42]  every mountainside let freedom ring
## [43]  happens allow freedom ring let ring every village every hamlet every state every city will able
## [44] free last free last
## [45]
## [46] thank god almighty free last
```

Tokenisasi

Mewakili perkataan kepada format angka yang kemudiannya boleh digunakan dalam perlombongan teks.

To obtain bag of words

Boleh guna tokenisasi atau korpus data

Pembendungan Teks (Text Stemming)

Turunkan data kepada bentuk akar (root form)

```
docs10 = tm_map(docs9, stemDocument)
```

```
## Warning in tm_map.SimpleCorpus(docs9, stemDocument): transformation drops
## documents
```

Matriks Sebutan-Dokumen

```
dtm = TermDocumentMatrix(docs10)
m = as.matrix(dtm)
dim(m)
```

```
## [1] 162  46
```

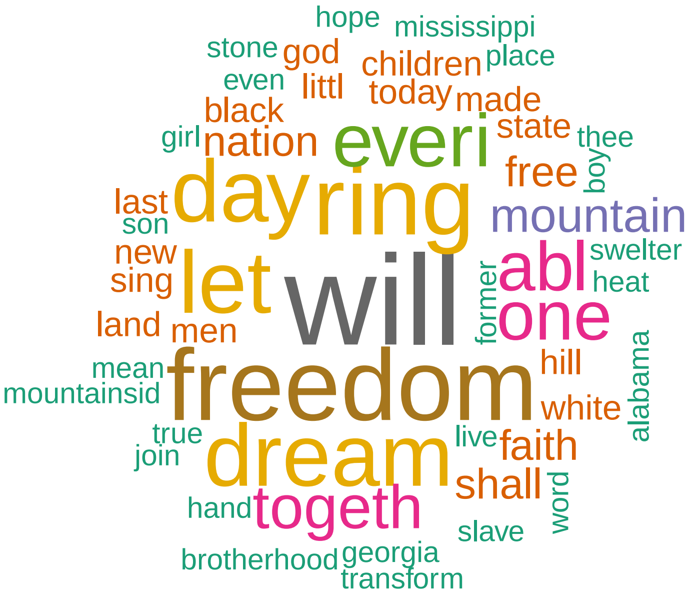
10 ayat paling kerap disebut dalam teks

```
v = sort(rowSums(m), decreasing=T)
d = data.frame(word=names(v), freq=(v))
head(d,10)
```

```
##            word freq
## will       will   17
## freedom freedom   13
## ring       ring   12
## dream     dream   11
## day         day   11
## let         let   11
## everi     everi    9
## one         one    8
## abl         abl    8
## togeth   togeth    7
```

Awan Perkataan

```
set.seed(12)
wordcloud(words=d$word, freq=d$freq, min.freq=2,
          max.words = 150, random.order=F, colors = brewer.pal(8, "Dark2"))
```



Perkaitan Perkataan (Word Association)

Contoh: Perkataan yang sering diserbut bersama 'freedom'?

```
findAssocs(dtm, terms='freedom', corlimit=0.3)$freedom
```

```
##         let        ring    mountain mississippi   transform       stone
##        0.89        0.86        0.40        0.34        0.34        0.34
## mountainsid       state       everi
##        0.34        0.32        0.32
```

Cari hubungan ayat yang berlaku sekurang kurangnya 10 kali

```
findAssocs(dtm, terms=findFreqTerms(dtm,lowfreq=10), corlimit=0.5)
```

```
## $dream
##   american     deepli difficulti       face       root      still     though
##       0.74       0.74       0.74       0.74       0.74       0.74       0.74
##   tomorrow       even      today
##       0.74       0.67       0.67
##
## $day
##        one        abl   children      littl      black      white       mean
##       0.78       0.60       0.56       0.56       0.56       0.56       0.52
##       hand       join       word    alabama        boy    brother       drip
##       0.52       0.52       0.52       0.51       0.51       0.51       0.51
##       girl   governor interposit        lip     nullif     racist      right
##       0.51       0.51       0.51       0.51       0.51       0.51       0.51
##     sister    vicious
##       0.51       0.51
##
## $will
##         abl      togeth      beauti     despair     discord         hew
##        0.80        0.68        0.62        0.62        0.62        0.62
##        jail       jangl        know        pray       stand     struggl
##        0.62        0.62        0.62        0.62        0.62        0.62
##    symphoni        work    children         one       faith brotherhood
##        0.62        0.62        0.60        0.58        0.57        0.52
##   transform
##        0.52
##
## $freedom
## numeric(0)
##
## $let
## numeric(0)
##
## $ring
##   everi   allow  cathol    citi  gentil  hamlet  happen     jew   negro     old
##    0.57    0.53    0.53    0.53    0.53    0.53    0.53    0.53    0.53    0.53
## protest   speed spiritu  villag
##    0.53    0.53    0.53    0.53
```

Analisis Sentimen

```
library(sentimentr)
```

```
##
## Attaching package: 'sentimentr'

## The following object is masked from 'package:syuzhet':
##
##     get_sentences
```

```
x = "Sentiment analysis is super fun"
```

```
sentiment(x)
```

```
## Key: <element_id, sentence_id>
##    element_id sentence_id word_count sentiment
##         <int>       <int>      <int>     <num>
## 1:          1           1          5 0.6708204
```

```
y = "Sentiment analysis is super boring. I do love working with R"
sentiment(y)
```

```
## Key: <element_id, sentence_id>
##    element_id sentence_id word_count  sentiment
##         <int>       <int>      <int>      <num>
## 1:          1           1          5 -0.1118034
## 2:          1           2          6  0.4082483
```

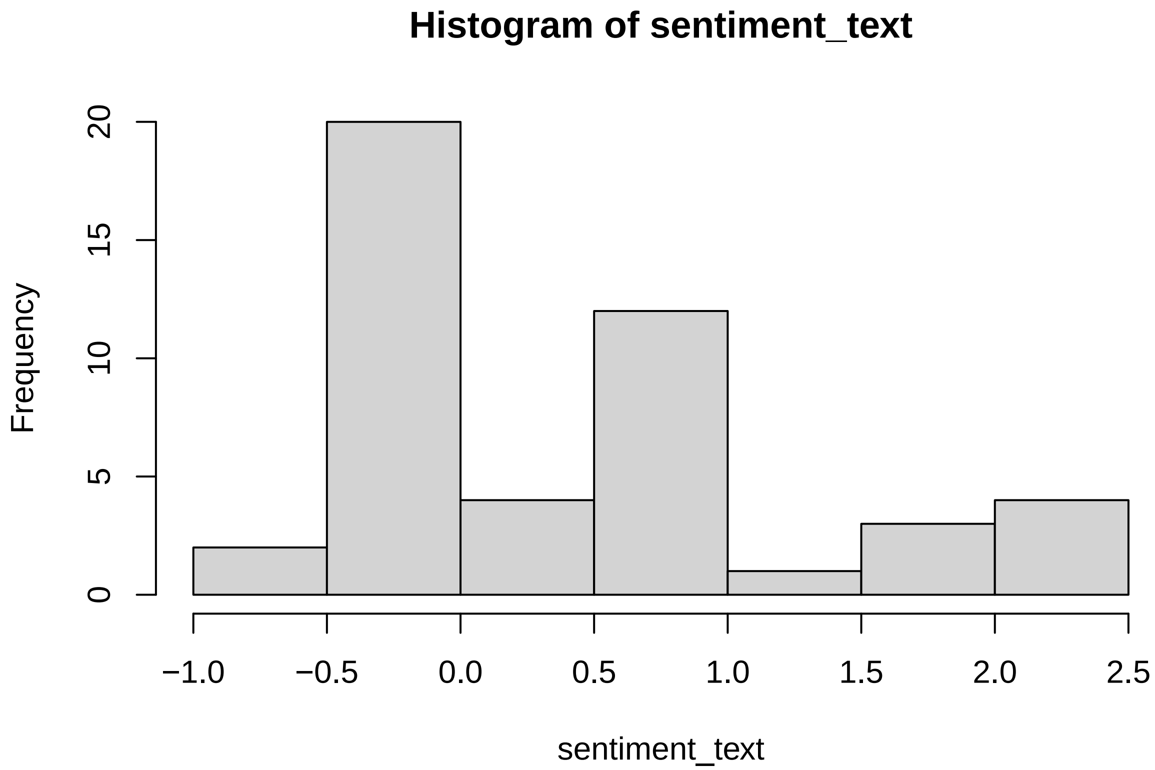
```
sentiment_text = get_sentiment(text, method='syuzhet')
sentiment_text
```

```
##  [1]  0.00 -0.25  0.00  0.75  0.00  0.00  0.00 -0.50  0.00  1.00  0.00  0.85
## [13]  0.00  0.25  0.00 -0.70  0.00  0.25  0.00  1.60  0.00  0.50  0.00  2.15
## [25]  0.00  1.40  0.00  1.85 -0.25  0.75  1.00  2.05  2.30  0.75  0.75  0.75
## [37]  0.00  0.00  0.75  0.75  0.75  0.75  2.10  0.50  0.00  2.00
```

```
summary(sentiment_text)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.7000  0.0000  0.2500  0.5402  0.7500  2.3000
```

```
hist(sentiment_text)
```



Klasifikasi Emosi

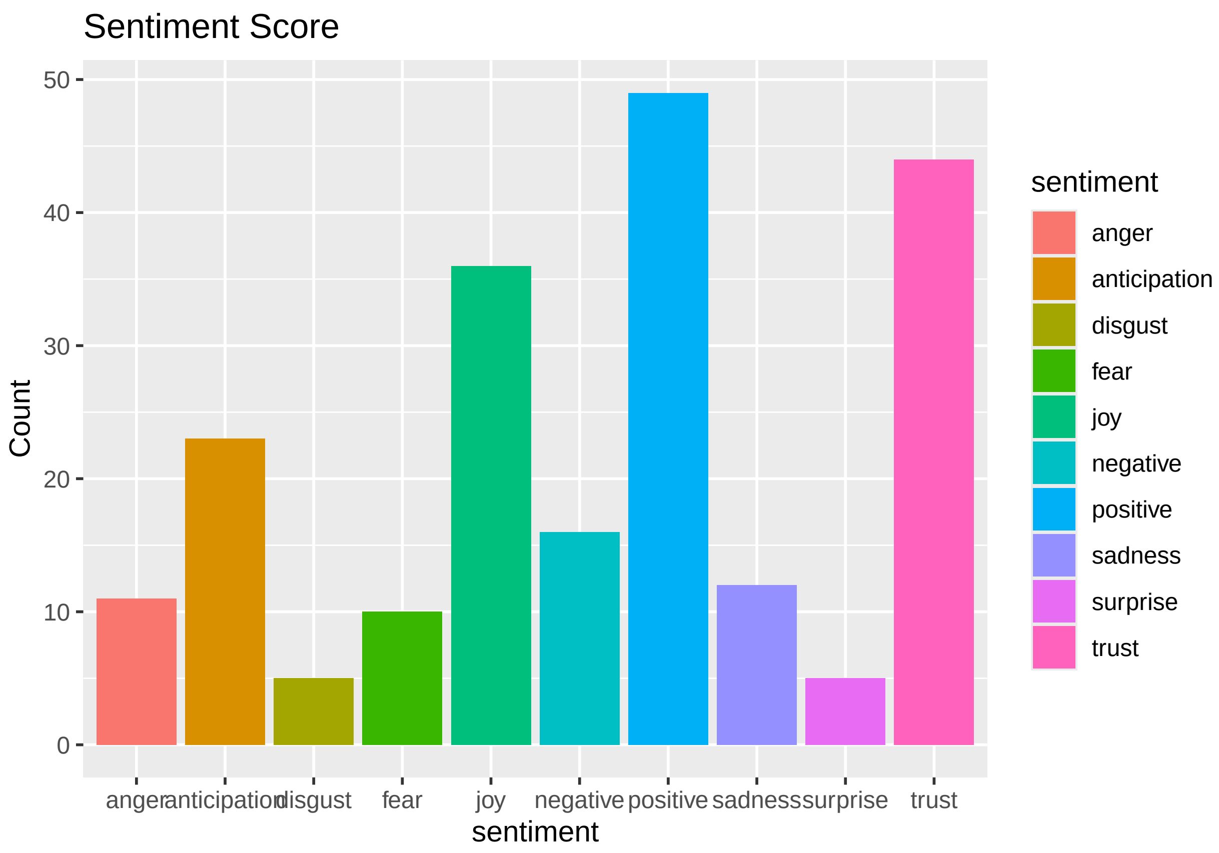
```
d2 = get_nrc_sentiment(text)
td = data.frame(t(d2))
```

Pengvisualan

```
td_new = data.frame(rowSums(td))
names(td_new)[1] = 'Count'
td_new = cbind("sentiment" = rownames(td_new), td_new)
rownames(td_new) = NULL

qplot(sentiment, weight=Count, data=td_new,
      geom='bar', fill=sentiment, ylab='Count') + ggtitle("Sentiment Score")
```

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```


Sentiment Score
Count
50
40
30
20
10
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust