R 데이터(텍스트) 마이닝 및 시각화 분석

Seongmin Mun

21st January 2022

아주대학교 AJOU UNIVERSITY

Outline

dplyr을 이용한 데이터 정제

tidyverse설치

```
> install.packages("tidyverse")
Error in install.packages : Updating loaded packages

Restarting R session...

> install.packages("tidyverse")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/tidyverse_1.3.1.tgz'을 시도합니다
Content type 'application/x-gzip' length 421072 bytes (411 KB)
==================================================
downloaded 411 KB


The downloaded binary packages are in
        /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//Rtmp3O7AaM/downloaded_packages
> library(tidyverse)
── Attaching packages ─────────────────────── tidyverse 1.3.1 ──
✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.3     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.1     ✓ forcats 0.5.1
── Conflicts ────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
> # install.packages("dplyr")
> library(dplyr)
```

mpg데이터

mpg {ggplot2} R Documentation

Fuel economy data from 1999 to 2008 for 38 popular models of cars

Description

This dataset contains a subset of the fuel economy data that the EPA makes available on https://fueleconomy.gov/. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

Usage

```
mpg
```

Format

A data frame with 234 rows and 11 variables:

manufacturer

manufacturer name

model

model name

filter()를 사용한 데이터 추출-1

```
> # filter(): 데이터를 사용자가 지정한 기준에 따라 추출하는 방법
> # 제조사중 현대 기업의 차를 추출한다.
> hyundai <- filter(mpg, manufacturer=="hyundai");hyundai
# A tibble: 14 × 11
   manufacturer model   displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>   <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 hyundai      sonata    2.4  1999     4 auto(l4)   f        18    26 r     midsize
 2 hyundai      sonata    2.4  1999     4 manual(m5) f        18    27 r     midsize
 3 hyundai      sonata    2.4  2008     4 auto(l4)   f        21    30 r     midsize
 4 hyundai      sonata    2.4  2008     4 manual(m5) f        21    31 r     midsize
 5 hyundai      sonata    2.5  1999     6 auto(l4)   f        18    26 r     midsize
 6 hyundai      sonata    2.5  1999     6 manual(m5) f        18    26 r     midsize
 7 hyundai      sonata    3.3  2008     6 auto(l5)   f        19    28 r     midsize
 8 hyundai      tiburon   2    1999     4 auto(l4)   f        19    26 r     subcompact
 9 hyundai      tiburon   2    1999     4 manual(m5) f        19    29 r     subcompact
10 hyundai      tiburon   2    2008     4 manual(m5) f        20    28 r     subcompact
11 hyundai      tiburon   2    2008     4 auto(l4)   f        20    27 r     subcompact
12 hyundai      tiburon   2.7  2008     6 auto(l4)   f        17    24 r     subcompact
13 hyundai      tiburon   2.7  2008     6 manual(m6) f        16    24 r     subcompact
14 hyundai      tiburon   2.7  2008     6 manual(m5) f        17    24 r     subcompact
```

filter()를 사용한 데이터 추출-2

```
> # 스포츠카 추출하기
> sportCar <- filter(mpg, cyl >= 8); sportCar  # 머슬카
# A tibble: 70 × 11
   manufacturer model              displ  year   cyl trans    drv     cty   hwy fl    class
   <chr>        <chr>              <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
 1 audi         a6 quattro           4.2  2008     8 auto(s6) 4        16    23 p     midsi…
 2 chevrolet    c1500 suburban 2wd   5.3  2008     8 auto(l4) r        14    20 r     suv
 3 chevrolet    c1500 suburban 2wd   5.3  2008     8 auto(l4) r        11    15 e     suv
 4 chevrolet    c1500 suburban 2wd   5.3  2008     8 auto(l4) r        14    20 r     suv
 5 chevrolet    c1500 suburban 2wd   5.7  1999     8 auto(l4) r        13    17 r     suv
 6 chevrolet    c1500 suburban 2wd   6    2008     8 auto(l4) r        12    17 r     suv
 7 chevrolet    corvette             5.7  1999     8 manual(… r        16    26 p     2seat…
 8 chevrolet    corvette             5.7  1999     8 auto(l4) r        15    23 p     2seat…
 9 chevrolet    corvette             6.2  2008     8 manual(… r        16    26 p     2seat…
10 chevrolet    corvette             6.2  2008     8 auto(s6) r        15    25 p     2seat…
# … with 60 more rows
```

arrange()를 사용한 데이터 정렬-1

```
> #arrange(): 선택된 변수 값을 기준으로 데이터를 정렬하는 방법
> # 현대차 추출
> hyundai <- filter(mpg, manufacturer=="hyundai");hyundai
# A tibble: 14 × 11
   manufacturer model   displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>   <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 hyundai      sonata    2.4  1999     4 auto(l4)   f        18    26 r     midsize
 2 hyundai      sonata    2.4  1999     4 manual(m5) f        18    27 r     midsize
 3 hyundai      sonata    2.4  2008     4 auto(l4)   f        21    30 r     midsize
 4 hyundai      sonata    2.4  2008     4 manual(m5) f        21    31 r     midsize
 5 hyundai      sonata    2.5  1999     6 auto(l4)   f        18    26 r     midsize
 6 hyundai      sonata    2.5  1999     6 manual(m5) f        18    26 r     midsize
 7 hyundai      sonata    3.3  2008     6 auto(l5)   f        19    28 r     midsize
 8 hyundai      tiburon   2    1999     4 auto(l4)   f        19    26 r     subcompact
 9 hyundai      tiburon   2    1999     4 manual(m5) f        19    29 r     subcompact
10 hyundai      tiburon   2    2008     4 manual(m5) f        20    28 r     subcompact
11 hyundai      tiburon   2    2008     4 auto(l4)   f        20    27 r     subcompact
12 hyundai      tiburon   2.7  2008     6 auto(l4)   f        17    24 r     subcompact
13 hyundai      tiburon   2.7  2008     6 manual(m6) f        16    24 r     subcompact
14 hyundai      tiburon   2.7  2008     6 manual(m5) f        17    24 r     subcompact
```

arrange()를 사용한 데이터 정렬-2

```
> #실린더의 수를 기준으로 정렬 (오름차순이다.)
> arrange(hyundai, cyl)
# A tibble: 14 x 11
   manufacturer model   displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>   <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 hyundai      sonata    2.4  1999     4 auto(l4)   f        18    26 r     midsize
 2 hyundai      sonata    2.4  1999     4 manual(m5) f        18    27 r     midsize
 3 hyundai      sonata    2.4  2008     4 auto(l4)   f        21    30 r     midsize
 4 hyundai      sonata    2.4  2008     4 manual(m5) f        21    31 r     midsize
 5 hyundai      tiburon   2    1999     4 auto(l4)   f        19    26 r     subcompact
 6 hyundai      tiburon   2    1999     4 manual(m5) f        19    29 r     subcompact
 7 hyundai      tiburon   2    2008     4 manual(m5) f        20    28 r     subcompact
 8 hyundai      tiburon   2    2008     4 auto(l4)   f        20    27 r     subcompact
 9 hyundai      sonata    2.5  1999     6 auto(l4)   f        18    26 r     midsize
10 hyundai      sonata    2.5  1999     6 manual(m5) f        18    26 r     midsize
11 hyundai      sonata    3.3  2008     6 auto(l5)   f        19    28 r     midsize
12 hyundai      tiburon   2.7  2008     6 auto(l4)   f        17    24 r     subcompact
13 hyundai      tiburon   2.7  2008     6 manual(m6) f        16    24 r     subcompact
14 hyundai      tiburon   2.7  2008     6 manual(m5) f        17    24 r     subcompact
```

select()를 사용한 데이터 선택-1

```
> #select(): 주어진 데이터에서 입력된 이름을 기준으로 선택하는 방법
> # 폭스바겐차 추출
> volkswagen <- filter(mpg, manufacturer=="volkswagen");volkswagen
# A tibble: 27 × 11
   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 volkswagen   gti     2    1999     4 manual(m5) f        21    29 r     compact
 2 volkswagen   gti     2    1999     4 auto(l4)   f        19    26 r     compact
 3 volkswagen   gti     2    2008     4 manual(m6) f        21    29 p     compact
 4 volkswagen   gti     2    2008     4 auto(s6)   f        22    29 p     compact
 5 volkswagen   gti     2.8  1999     6 manual(m5) f        17    24 r     compact
 6 volkswagen   jetta   1.9  1999     4 manual(m5) f        33    44 d     compact
 7 volkswagen   jetta   2    1999     4 manual(m5) f        21    29 r     compact
 8 volkswagen   jetta   2    1999     4 auto(l4)   f        19    26 r     compact
 9 volkswagen   jetta   2    2008     4 auto(s6)   f        22    29 p     compact
10 volkswagen   jetta   2    2008     4 manual(m6) f        21    29 p     compact
# … with 17 more rows
```

select()를 사용한 데이터 선택-2

```
> # 폭스바겐에서 원하는 변수들만을 출력
> volkswagenSelected <- select(volkswagen, model, year, cyl); volkswagenSelected
# A tibble: 27 × 3
   model  year   cyl
   <chr> <int> <int>
 1 gti    1999     4
 2 gti    1999     4
 3 gti    2008     4
 4 gti    2008     4
 5 gti    1999     6
 6 jetta  1999     4
 7 jetta  1999     4
 8 jetta  1999     4
 9 jetta  2008     4
10 jetta  2008     4
# … with 17 more rows
```

mutate()를 사용한 데이터 추가-1

```
> #mutate(): 기존 데이터에 새로운 변수를 삽입하는 방법
> # 아우디 차량 추출
> audi <- filter(mpg, manufacturer=="audi");audi
# A tibble: 18 × 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual(m6) 4        20    28 p     compact
11 audi         a4 quattro   2    2008     4 auto(s6)   4        19    27 p     compact
12 audi         a4 quattro   2.8  1999     6 auto(l5)   4        15    25 p     compact
13 audi         a4 quattro   2.8  1999     6 manual(m5) 4        17    25 p     compact
14 audi         a4 quattro   3.1  2008     6 auto(s6)   4        17    25 p     compact
15 audi         a4 quattro   3.1  2008     6 manual(m6) 4        15    25 p     compact
16 audi         a6 quattro   2.8  1999     6 auto(l5)   4        15    24 p     midsize
17 audi         a6 quattro   3.1  2008     6 auto(s6)   4        17    25 p     midsize
18 audi         a6 quattro   4.2  2008     8 auto(s6)   4        16    23 p     midsize
```

mutate()를 사용한 데이터 추가-2

```
> # 아우디에서 원하는 변수들만을 출력
> audiSelected <- select(volkswagen, model, year, cyl, cty, hwy); audiSelected
# A tibble: 27 × 5
   model  year   cyl   cty   hwy
   <chr> <int> <int> <int> <int>
 1 gti    1999     4    21    29
 2 gti    1999     4    19    26
 3 gti    2008     4    21    29
 4 gti    2008     4    22    29
 5 gti    1999     6    17    24
 6 jetta  1999     4    33    44
 7 jetta  1999     4    21    29
 8 jetta  1999     4    19    26
 9 jetta  2008     4    22    29
10 jetta  2008     4    21    29
# … with 17 more rows
```

mutate()를 사용한 데이터 추가-3

```
> # 도시 및 고속도로에서의 연비의 합
> audiSelectedAdd <- mutate(audiSelected, sum = cty + hwy); audiSelectedAdd
# A tibble: 27 × 6
   model  year   cyl   cty   hwy   sum
   <chr> <int> <int> <int> <int> <int>
 1 gti    1999     4    21    29    50
 2 gti    1999     4    19    26    45
 3 gti    2008     4    21    29    50
 4 gti    2008     4    22    29    51
 5 gti    1999     6    17    24    41
 6 jetta  1999     4    33    44    77
 7 jetta  1999     4    21    29    50
 8 jetta  1999     4    19    26    45
 9 jetta  2008     4    22    29    51
10 jetta  2008     4    21    29    50
# … with 17 more rows
```

KoNLP를 이용한 형태소 분석

기본 설정

```
> #메모리정리하기
> gc()
         used (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
Ncells 1228617 65.7    1945355 103.9         NA  1945355 103.9
Vcells 2075557 15.9    8388608  64.0      16384  8195034  62.6
> rm(list=ls())
> #한글깨짐현상
> #한글 인코딩 문제 해결
> #맥
> Sys.setlocale(category = "LC_CTYPE", locale = "ko_KR.UTF-8")
[1] "ko_KR.UTF-8"
> #윈도우
> options(encoding = "UTF-8")
> getwd()
[1] "/Users/seongminmun"
> setwd("/Users/seongminmun/Desktop/2022/Ajou/Data")
```

JDK설치

- https://www.oracle.com/java/technologies/downloads/

```
> dyn.load('/Library/Java/JavaVirtualMachines/jdk1.8.0_241.jdk/Contents/Home/jre/lib/server/libjvm.dylib')
```

remotes설치

```
> install.packages("remotes")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/remotes_2.4.2.tgz'을 시도합니다
Content type 'application/x-gzip' length 395393 bytes (386 KB)
==================================================
downloaded 386 KB


The downloaded binary packages are in
        /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//RtmpnezOSM/downloaded_packages
> library(remotes)
```

rJava설치

```
> install.packages("rJava")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/rJava_1.0-6.tgz'을 시도합니다
Content type 'application/x-gzip' length 1138672 bytes (1.1 MB)
==================================================
downloaded 1.1 MB


The downloaded binary packages are in
        /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//RtmpnezOSM/downloaded_packages
> library(rJava)
```

KoNLP설치

```
> remotes::install_github('haven-jeon/KoNLP', upgrade = "never", INSTALL_opts=c("--no-multiarc
h"))
Skipping install of 'KoNLP' from a github remote, the SHA1 (960fbbcf) has not changed since last
 install.
  Use `force = TRUE` to force installation
> library(KoNLP)
Checking user defined dictionary!
```

형태소 분석기 정보

- 한나움 형태소 분석기
- POS태그 정보

명사 추출하기

```
> message <- "아주대학교에서 겨울 특강으로 진행된 R 데이터(텍스트) 마이닝 및 시각화 분석에서 학생분들이 보여준 학구열로 인해
 힘이 납니다."
> #명사 추출하기
> extractNoun(message)
 [1] "아주"          "대학교"          "겨울"          "특강"          "진행"
 [6] "R"             "데이터(텍스트)" "마이닝"        "시각화"        "분석"
[11] "학생"          "분"              "학구열"        "힘"
```

형태소 분석하기-1

```
> #형태소 분석하기
> MorphAnalyzer(message)
$아주대학교에서
 [1] "아주/ncn+대학교/ncn+에서/jca"           "아주/ncn+대학교/ncn+에서/jcs"
 [3] "아주/ncn+대학교/ncn+에/ncn+서/jca"      "아주/ncn+대학교/ncn+에/ncn+서/jcs"
 [5] "아주/ncn+대학/ncn+교/ncn+에서/jca"      "아주/ncn+대학/ncn+교/ncn+에서/jcs"
 [7] "아/xp+주대/ncn+학교/ncn+에서/jca"       "아/xp+주대/ncn+학교/ncn+에서/jcs"
 [9] "아/xp+주대/ncn+학교/ncn+에/ncn+서/jca" "아/xp+주대/ncn+학교/ncn+에/ncn+서/jcs"

$겨울
[1] "겨울/ncn"
```

형태소 분석하기-2

```
> #22가지 분류코드로 분류하기
> SimplePos22(message)
$아주대학교에서
[1] "아주대학교/NC+에서/JC"

$겨울
[1] "겨울/NC"

$특강으로
[1] "특강/NC+으로/JC"
```

형태소 분석하기-3

```
> #9가지 분류코드로 분류하기
> SimplePos09(message)
$아주대학교에서
[1] "아주대학교/N+에서/J"

$겨울
[1] "겨울/N"

$특강으로
[1] "특강/N+으로/J"
```