06

비정형데이터분석

텍스트 데이터 불러오기(2)

통계·데이터과학과 장영재 교수

학습목차

1. 텍스트데이터수집사례
2. 유용한R패키지

01

텍스트 데이터 수집 사례

1. 텍스트 데이터 수집 사례-①텍스트 목록 읽어오기(httr 및 XML 패키지 이용)

- R 프로그래밍 언어 중 간단하지만 널리 활용되는 웹문서 수집 패키지의 사례를 정리
  - 가장 기본적인 패키지는 'httr' 패키지

```
install.packages("httr")
library(httr)
```

1. 텍스트 데이터 수집 사례-①텍스트 목록 읽어오기(httr 및 XML 패키지 이용)

- 한국방송통신대학교 출판문화원 홈페이지 검색

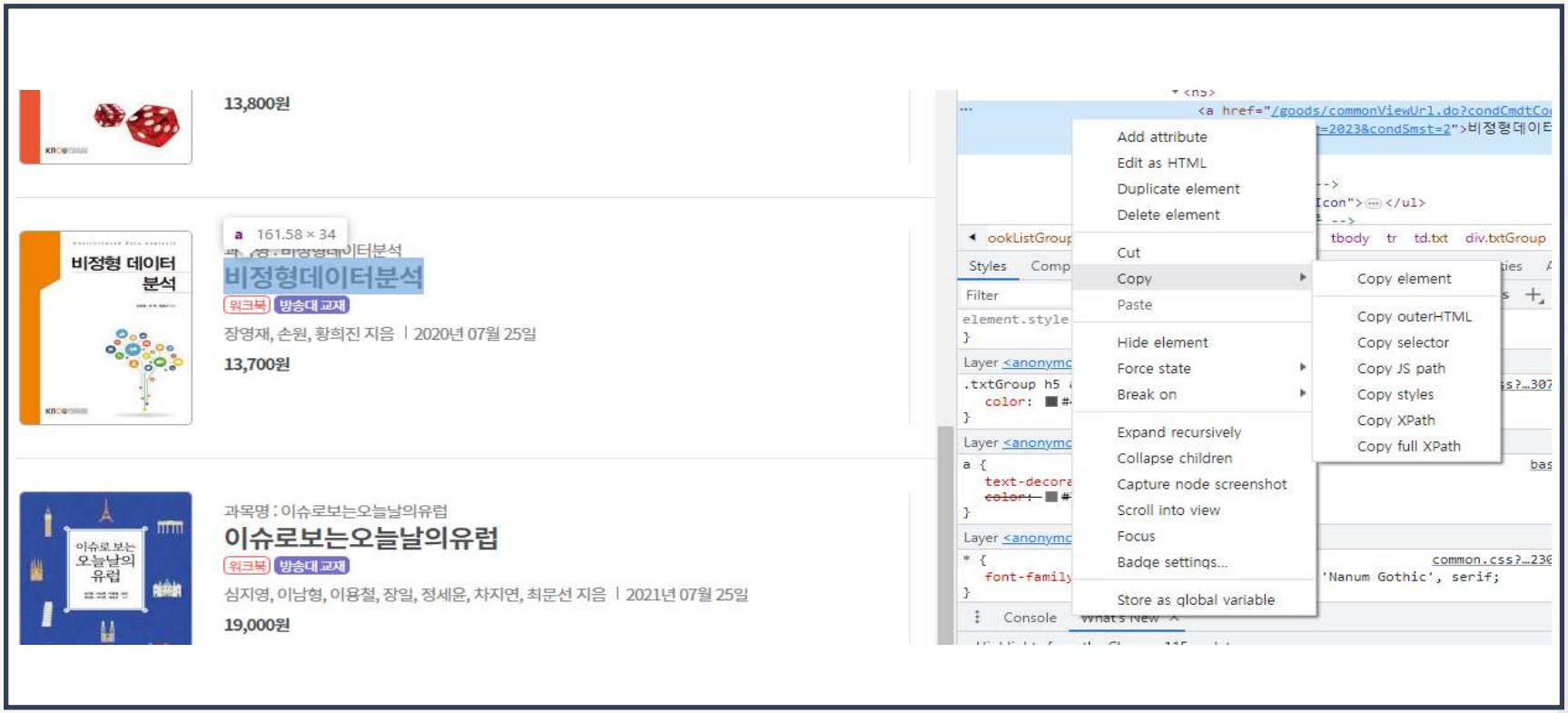
```
webpage=GET('http://press.knou.ac.kr/goods/textBookList.do?condL
scValue=001&condMscValue=003&condSscValue=007&condScyr=4')
```

- 한국방송통신대학교 출판문화원 홈페이지 검색

```
install.packages("XML")
library(XML)
web=htmlParse(webpage)
```

1. 텍스트 데이터 수집 사례-①텍스트 목록 읽어오기(httr 및 XML 패키지 이용)

- '데이터마이닝' 과목명에 마우스를 두고 오른쪽 버튼을 클릭하고 검사를 클릭한 이후 Copy-Copy XPath를 선택



1. 텍스트 데이터 수집 사례-①텍스트 목록 읽어오기(httr 및 XML 패키지 이용)

- //*[@id="listForm"]/div/div[3]/div[4]/div[3]/table/tbody/tr[5]/td[2]/div/h5/a
- 소스 코드의 규칙성을 파악한 뒤, 적당한 반복문을 이용하여 여러 과목명 텍스트를 한꺼번에 수집
  - 교과목이 동일한 코드를 포함하지만 '.../tbody/tr[5]/...'로 나타난 부분에서 대괄호 [ ] 내에 위치한 숫자만 다름을 이용

```
> library(XML)
> web=htmlParse(webpage)
>
>
> crsname=xpathSApply(web,'//*[@id="listForm"]/div/div[3]/div[4]/div[3]/table/tbody/tr[5]/td[2]/div/h5/a', xmlValue)
> crsname=gsub("\r","",crsname)
> crsname=gsub("\n","",crsname)
> crsname=gsub("\t","",crsname)
> crsname
[1] "비정형데이터분석"
>
> crsname=xpathSApply(web,'//*[@id="listForm"]/div/div[3]/div[4]/div[3]/table/tbody/tr[1]/td[2]/div/h5/a', xmlValue)
> crsname=gsub("\r","",crsname)
> crsname=gsub("\n","",crsname)
> crsname=gsub("\t","",crsname)
> crsname
[1] "R데이터분석"
>
>
> ls = rep("",6)
> for(i in 1:6){
+ sub = paste0('//*[@id="listForm"]/div/div[3]/div[4]/div[3]/table/tbody/tr[',i,']/td[2]/div/h5/a')
+ ls[i] = xpathSApply(web,sub,xmlValue)
+ ls[i]=gsub("\r","",ls[i])
+ ls[i]=gsub("\n","",ls[i])
+ ls[i]=gsub("\t","",ls[i])
+ }
>
> ls
[1] "R데이터분석"              "딥러닝의통계적이해"      "마케팅조사"              "베이즈데이터분석"
[5] "비정형데이터분석"         "이슈로보는오늘날의유럽"
```

<그림> xpathSApply 함수를 통한 텍스트 데이터 수집

1. 텍스트 데이터 수집 사례 - ② 웹문서 읽어오기(rvest 및 dplyr 패키지 이용)

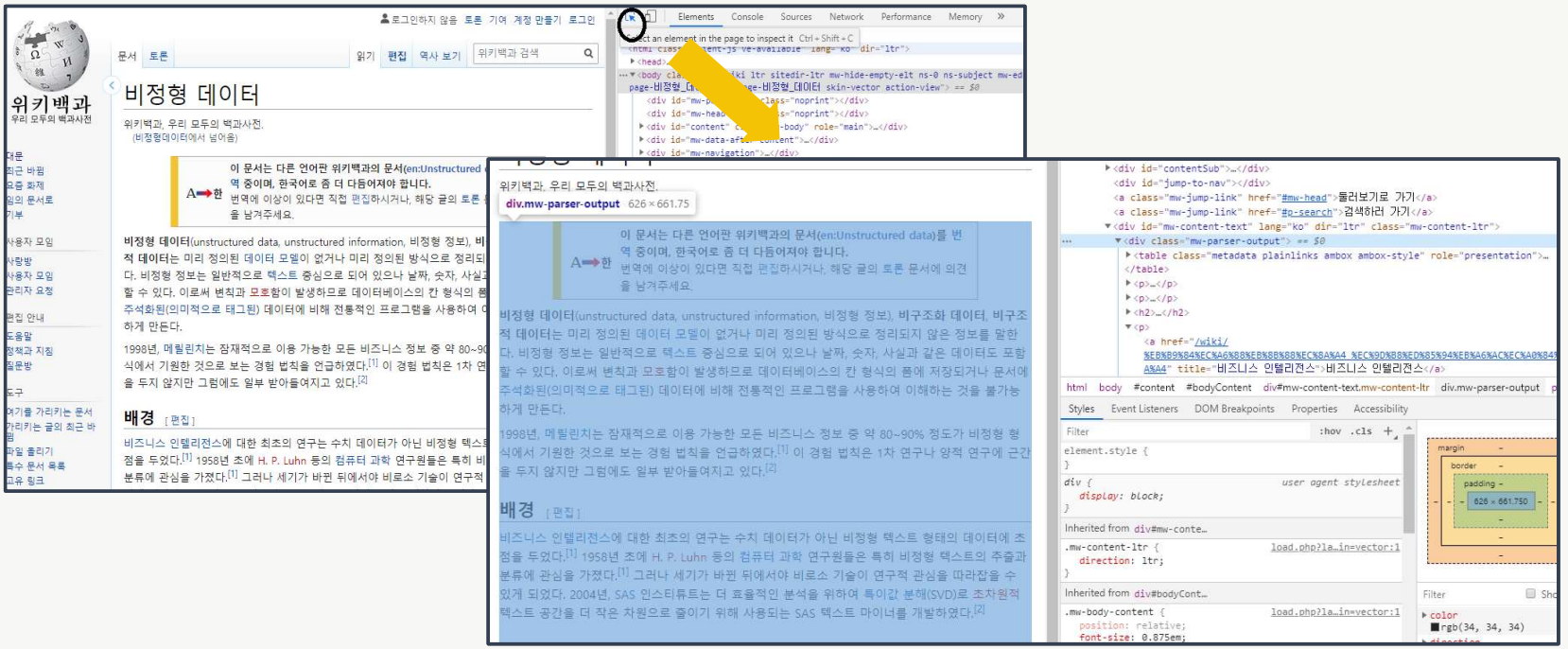
- R의 'rvest'패키지를 이용하면 웹문서의 내용을 쉽게 가져올 수 있음
  - dplyr' 패키지를 활성화하면 파이프(pipe)라고 하는 도구를 사용하면 문서를 읽어 들이는 기능을 보다 원활하게 할 수 있음

```
함수3(함수2(함수1(데이터)))

데이터%>%함수1%>%함수2%>%함수3
```

1. 텍스트 데이터 수집 사례 - ② 웹문서 읽어오기(rvest 및 dplyr 패키지 이용)

- Chrome 웹 브라우저 기준으로 F12 키를 눌러 창이 열리면 화살표 그림을 클릭하고 정보를 얻고 read_html(), html_nodes() 및 html_text() 함수를 이용하여 문서의 내용을 읽어올 수 있음

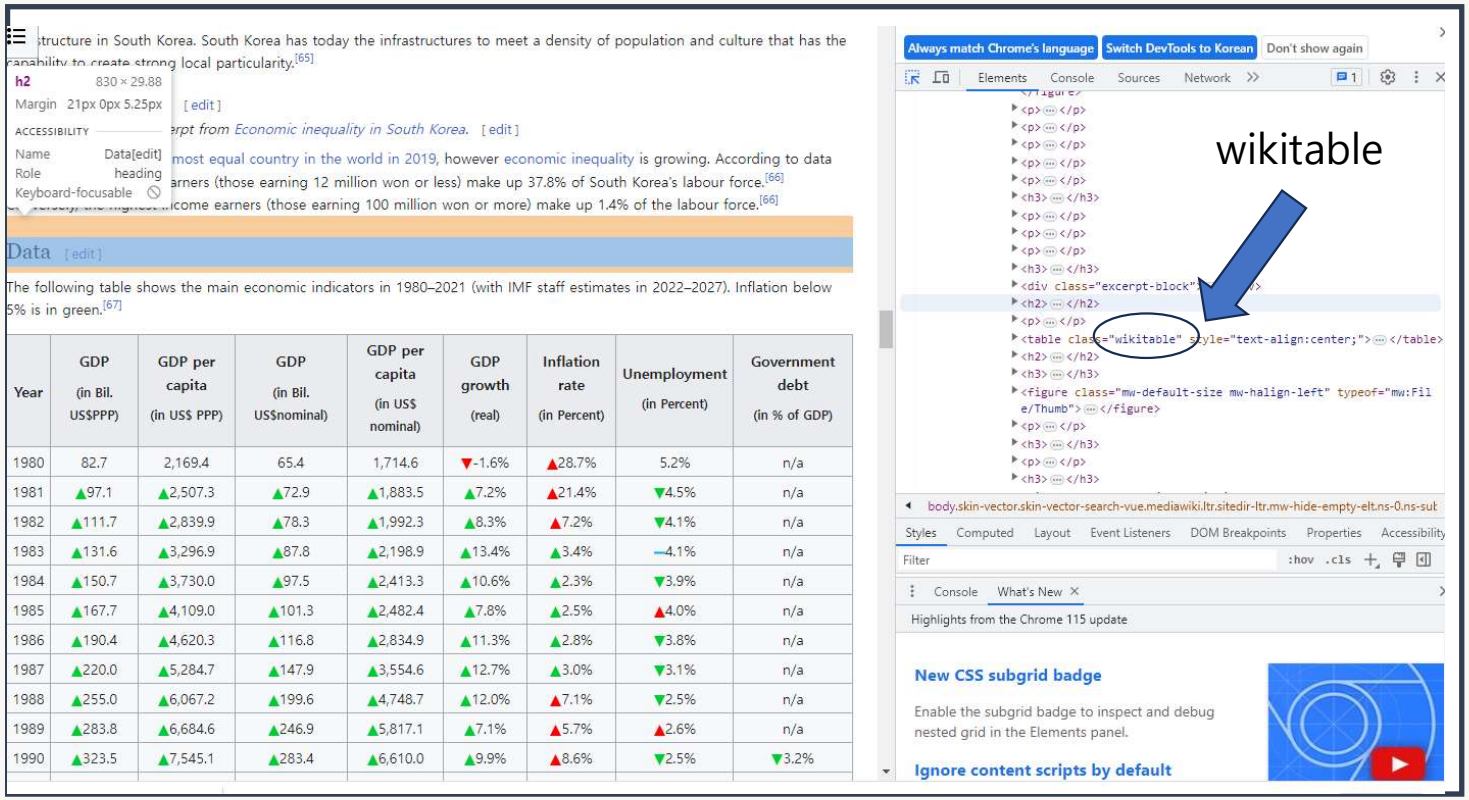


1. 텍스트 데이터 수집 사례 - ② 웹문서 읽어오기(rvest 및 dplyr 패키지 이용)

```
library(rvest)
library(dplyr)
exurl <- "https://ko.wikipedia.org/wiki/%EB%B9%84%EC%A0%95%ED%98%95
_%EB%8D%B0%EC%9D%B4%ED%84%B0"
html_ex <- read_html(exurl,encoding="UTF-8")
html_ex%>%html_nodes(".mw-parser-output p")%>%html_text()
```

1. 텍스트 데이터 수집 사례 - ③ 웹문서의 표 읽어오기 (rvest 및 dplyr 패키지 이용)

- R의 'rvest'패키지를 이용하고 html_table() 함수를 이용하면 표를 읽게 됨
  - https://en.wikipedia.org/wiki/Economy_of_South_Korea



...tructure in South Korea. South Korea has today the infrastructures to meet a density of population and culture that has the capability to create strong local particularity.[65]

h2 830 × 29.88
Margin 21px 0px 5.25px
ACCESSIBILITY
Name Data[edit]
Role heading
Keyboard-focusable

[edit]

...erpt from *Economic inequality in South Korea*. [edit]

...most equal country in the world in 2019, however economic inequality is growing. According to data ...arners (those earning 12 million won or less) make up 37.8% of South Korea's labour force.[66] ...ncome earners (those earning 100 million won or more) make up 1.4% of the labour force.[66]

Data [edit]

The following table shows the main economic indicators in 1980–2021 (with IMF staff estimates in 2022–2027). Inflation below 5% is in green.[67]

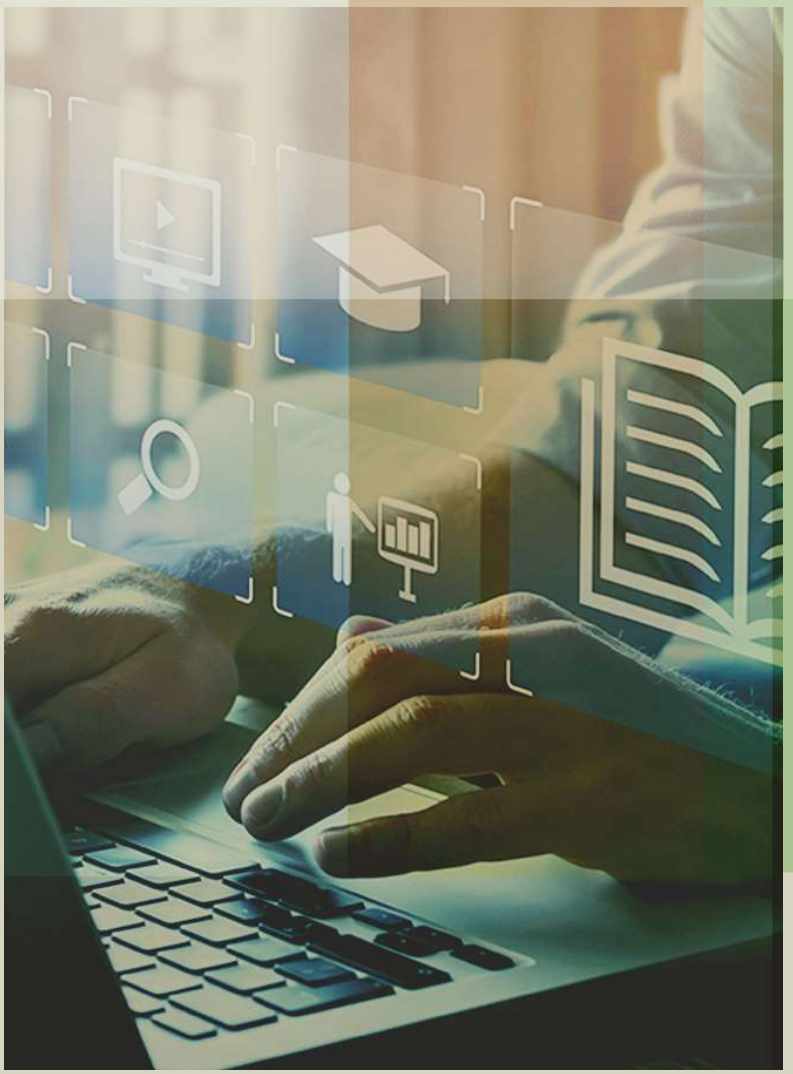
Year	GDP (in Bil. US$PPP)	GDP per capita (in US$ PPP)	GDP (in Bil. US$nominal)	GDP per capita (in US$ nominal)	GDP growth (real)	Inflation rate (in Percent)	Unemployment (in Percent)	Government debt (in % of GDP)
1980	82.7	2,169.4	65.4	1,714.6	▼-1.6%	▲28.7%	5.2%	n/a
1981	▲97.1	▲2,507.3	▲72.9	▲1,883.5	▲7.2%	▲21.4%	▼4.5%	n/a
1982	▲111.7	▲2,839.9	▲78.3	▲1,992.3	▲8.3%	▲7.2%	▼4.1%	n/a
1983	▲131.6	▲3,296.9	▲87.8	▲2,198.9	▲13.4%	▲3.4%	–4.1%	n/a
1984	▲150.7	▲3,730.0	▲97.5	▲2,413.3	▲10.6%	▲2.3%	▼3.9%	n/a
1985	▲167.7	▲4,109.0	▲101.3	▲2,482.4	▲7.8%	▲2.5%	▲4.0%	n/a
1986	▲190.4	▲4,620.3	▲116.8	▲2,834.9	▲11.3%	▲2.8%	▼3.8%	n/a
1987	▲220.0	▲5,284.7	▲147.9	▲3,554.6	▲12.7%	▲3.0%	▼3.1%	n/a
1988	▲255.0	▲6,067.2	▲199.6	▲4,748.7	▲12.0%	▲7.1%	▼2.5%	n/a
1989	▲283.8	▲6,684.6	▲246.9	▲5,817.1	▲7.1%	▲5.7%	▲2.6%	n/a
1990	▲323.5	▲7,545.1	▲283.4	▲6,610.0	▲9.9%	▲8.6%	▼2.5%	▼3.2%

1. 텍스트 데이터 수집 사례 - ③ 웹문서의 표 읽어오기 (rvest 및 dplyr 패키지 이용)

```
library(rvest)
library(dplyr)
ex2url<- "https://en.wikipedia.org/wiki/Economy_of_South_Korea"
html_ex2 <- read_html(ex2url,encoding="UTF-8")
html_ex2%>%html_nodes(".wikitable")%>%html_table()
```

02

유용한 R 패키지



2. 유용한 R 패키지 - ① textstem 패키지

- 어간추출(stemming)을 위해 사용되는 패키지
  - textstem 패키지의 가장 기본이 되는 함수는 stem_strings()

```
library(textstem)
stem_strings(x, language = "porter", ...) # x는 텍스트 벡터
```

2. 유용한 R 패키지 - ② stopwords 패키지

- 불용어를 파악하고 제거하기 위한 목적으로 만들어진 패키지
  - stopwords()와 같은 함수를 통해 불용어리스트를 출력

```
library(stopwords)
stopwords(language = "en", source = "snowball")
# example
stopwords("en")
stopwords("de")
```

2. 유용한 R 패키지 - ③ tidytext 패키지

- R에서는 tidytext 패키지에서 제공하는 감성어 사전을 활용할 수 있음
  - get_sentiments( ) 함수를 이용하고 옵션을 "bing"으로 지정하면 단어들을 negative와 positive로 분류

```
library(tidytext)
get_sentiments(lexicon = c("afinn", "bing", "loughran", "nrc"))
```

2. 유용한 R 패키지 - ③ tidytext 패키지

- lexicon 옵션을 "afinn"로 지정하면 단어들에 -5점에서 5점 사이의 점수를 부여하고 "nrc"로 지정하면 단어들을 10가지 감정의 범주로 분류

```
library(tidytext)
library(textdata)
get_sentiments("afinn")
get_sentiments("nrc")
```

2. 유용한 R 패키지 - ④ wordcloud 패키지

- 텍스트 데이터의 시각화의 가장 기본이 되는 워드클라우드를 생성
  - wordcloud( ) 함수는 텍스트를 읽어 들인 뒤 단어의 빈도 등 통계량을 바탕으로 단어의 집합을 출력

```
library(wordcloud)
wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,
          random.order=TRUE, random.color=FALSE, rot.per=.1,
          colors="black",ordered.colors=FALSE,use.r.layout=FALSE,
          fixed.asp=TRUE, ...)
```

실습하기

다음시간안내

07

텍스트 데이터의 전처리(1)