



실습 6강 웹 크롤링

텍스트마이닝

웹크롤링

텍스트 전처리

토픽분석

Tokeniz
ation

Normali
zation

Stemmi
ng

TF-IDF

LSA

LDA



❖ Product review data

❖ 아마존 Alexa 4세대 Eco Dot



All-new Echo Dot (4th Gen) | Smart speaker with clock and Alexa | Glacier White

Brand: Amazon

★★★★★ 64,233 ratings | 682 answered questions

Climate Pledge Friendly

#1 Best Seller in Home Audio Speakers

Temporarily out of stock.

We are working hard to be back in stock as soon as possible.

Echo Dot 4th Gens [See the differences](#)



Color: **Glacier White**



Configuration: **Echo Dot with clock**

Echo Dot with clock

with \$10 Smart Plug

with Echo Auto

- Meet the all-new Echo Dot with clock - Our most popular smart speaker with Alexa. The sleek, compact design delivers crisp vocals and balanced bass for full sound.
- Perfect for your nightstand - See the time, alarms, and timers on the LED display. Tap the top to snooze an alarm.

라이브러리 불러오기

```
# 라이브러리 불러오기
# tidyverse : 텍스트 전처리 관련
# rvest : html 크롤링
library(tidyverse)
library(rvest)
```

```
> library(tidyverse)
-- Attaching packages ----- tidyverse 1.3.0 --
✓ ggplot2 3.3.0    ✓ purrr   0.3.4
✓ tibble  3.0.1    ✓ dplyr   0.8.3
✓ tidyr   1.0.0    ✓ stringr 1.4.0
✓ readr   1.3.1    ✓ forcats 0.5.0
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
```

```
> library(rvest)
```

필요한 패키지를 로딩중입니다: xml2

다음의 패키지를 부착합니다: 'rvest'

The following object is masked from 'package:purrr':

pluck

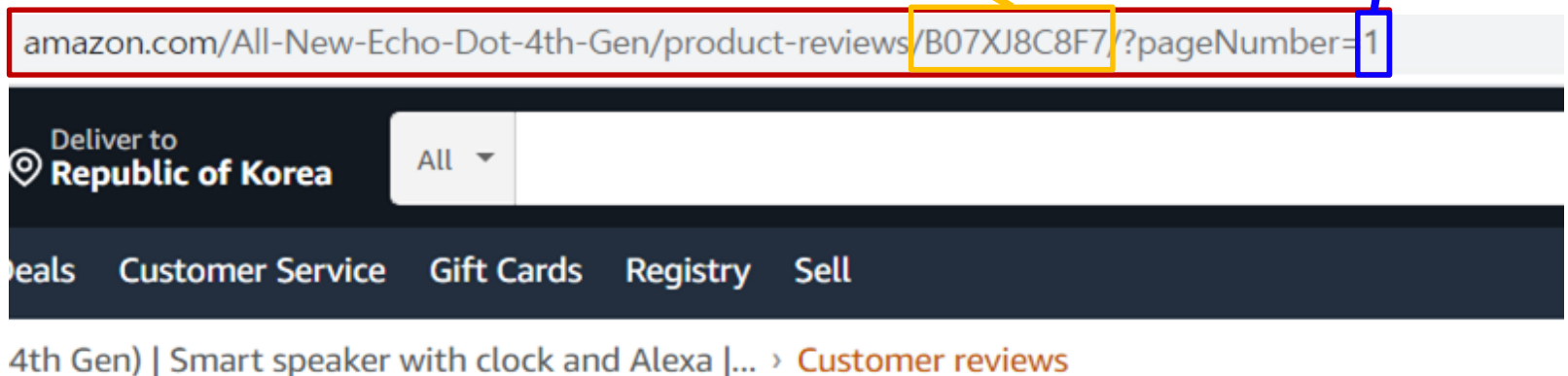
The following object is masked from 'package:readr':

guess_encoding

크롤링 준비

```
scrap_amazon <- function(ASIN, page_num){
```

```
  url_reviews <- paste0("https://www.amazon.com/All-New-Echo-Dot-4th-Gen/pro  
                        duct-reviews/", ASIN, "/?pageNumber=", page_num)
```



1페이지 당 10개씩의 리뷰가 있음

- paste0 : 문자열 결합 함수
- ASIN : 아마존이 만든 10자리의 고유 식별 번호

크롤링 준비

```
doc <- read_html(url_reviews)
```

The screenshot shows the Amazon product page for the Echo Dot (4th Gen). The page includes the Amazon logo, delivery location (Republic of Korea), and navigation links. The product title is "All-new Echo Dot (4th Gen) ... by Amazon". It has a 4.8 out of 5 star rating from 64,233 global ratings. The color is "Glacier White" and the configuration is "Echo Dot with clock". There are two review sections: "Top positive review" by Honest Reviewer (TOP 1000 REVIEWER) and "Top critical review" by RedOneStandingBy. The positive review is 5 stars and mentions the globe design and fabric speaker. The critical review is 3 stars and mentions that it doesn't hear well. A red box highlights the "Customer reviews" section.

The screenshot shows the Chrome DevTools interface. The "Elements" panel on the left shows the HTML structure of the page, with a red box highlighting the "body" element. The "Styles" panel on the right shows the CSS styles for the "body" element, including "font-family: 'Amazon Ember', Arial, sans-serif;". A red box also highlights the "body" element in the "Styles" panel. The "Console" panel at the bottom shows the "What's New" section with links to "New CSS Flexbox debugging tools" and "New Core Web Vitals overlay".

- read_html : 해당하는 벡터의 html을 read

크롤링 준비

Review Date

doc %>%

html_nodes("[data-hook='review-date']")%>%

html_text() -> Data

```
<span data-hook="review-title" class="a-size-base review-  
title a-text-bold">Overall improvement, but lost its  
flexibility</span> == $0
```

Review Title

doc %>%

html_nodes("[class='a-size-base a-link-normal review-title a-color-base
review-title-content a-text-bold']")%>%

html_text() -> Title

```
▼<div class="a-row">  
  ::before  
  <span class="a-size-base a-color-secondary review-date">  
    Reviewed in the United States on October 15, 2018</span>  
  ::after  
</div>
```

Html의 Text 부분

```
▼<a data-hook="review-title" class="a-size-base a-link-normal  
review-title a-color-base review-title-content a-text-bold"  
href="/gp/customer-reviews/RVDNM4DW0ELQE/  
ref=cm_cr_getr_d_rvw_ttl?ie=UTF8&ASIN=B082TJT44G">  
  <span>Best Echo dot ever!</span>  
</a>  
  ::after
```

Html의 Text 부분

- `html_nodes` : html 문서에서 노드 찾는 함수
- `A%>%B` : $A \supset B$ / `A%<%B` : $A \subset B$
- `html_text` : html에서 해당하는 텍스트

크롤링 준비

Review Text

doc %>%

html_nodes("[class='a-size-base review-text review-text-content']")%>%

html_text() -> Review

Number of Stars in Review

doc %>%

html_nodes("[data-hook='review-star-rating']")%>%

html_text() -> Rating

Return a tibble

tibble(Data, Title, Review, Rating, Page = page_num)%>%

return()

}

```
▼<span data-hook="review-body" class="a-size-base  
review-text review-text-content"> == $0  
  ▼<span>  
    "They're pretty responsive. Sounds not the bad and  
    not super great. But for the price? Why not."  
  </span>  
</span>
```

Html의 Text 부분

```
▼<i data-hook="review-star-rating" class="a-icon a-  
icon-star a-star-4 review-rating">  
  <span class="a-icon-alt">4.0 out of 5 stars</span>  
</i>
```

Html의 Text 부분

- tibble : dataframe와 같은 역할로, 열 이름만 사용해서 편하게 만들 수 있음

Page별 크롤링 시작

Product name = All-New-Echo-Dot-4th-Gen

ASIN = B07XJ8C8F7

ASIN : 아마존이 만든 10자리의 고유 식별 번호

review_all <- vector("list", length = 10)

1페이지 당 10개씩의 리뷰가 있음

Showing 11-20 of 115,177 reviews

스크랩 시작

```
for (i in 1:10){
```

```
  review_all[[i]] <- scrap_amazon(ASIN = "B07XJ8C8F7", page_num = i)
```

```
}
```

review_all 내용을 rbind를 이용하여 한줄씩 리뷰를 저장

```
amazon <- do.call(rbind, review_all)
```

- do.call : 이름이나 함수 및 전달할 인수 목록을 위하여, 함수 호출 구성 및 실행

텍스트 전처리

❖ 현재 데이터 프레임 상태

	Data	Title	Review	Rating	Page
1	Reviewed in the United States on November 5, 2020	The Pros, Cons and Ok's for Echo Dot 4th Gen (HR).	Pros: _____ • The globe design with the fabric spe...	4.0 out of 5 stars	1
2	Reviewed in the United States on November 6, 2020	4th Gen. Twilight Blue w/ clock A great upgrade	Your browser does not support HTML5 video. I preordere...	5.0 out of 5 stars	1
3	Reviewed in the United States on November 5, 2020	A big improvement. Worth the wait, and money!	Your browser does not support HTML5 video. Just receiv...	5.0 out of 5 stars	1
4	Reviewed in the United States on November 7, 2020	Great echo dot stereo system !!!!	I purchased two of these to achieve true stereo. The...	5.0 out of 5 stars	1
5	Reviewed in the United States on November 8, 2020	Doesn't hear well	I got this to replace a 3rd generation Echo Dot and ...	1.0 out of 5 stars	1
6	Reviewed in the United States on November 26, 2020	Clock Display Too Bright	Alexa is awesome. We have 8 echos including our 2 ...	4.0 out of 5 stars	1
7	Reviewed in the United States on November 5, 2020	Worth the money 🍋.	Good looking 🍋. Perfect size and sound quality is g...	5.0 out of 5 stars	1
8	Reviewed in the United States on November 25, 2020	Echo dot 4th gen	Echo dot 4th gen. I've had all the versions of the Ec...	1.0 out of 5 stars	1
9	Reviewed in the United States on December 21, 2020	Shares your internet	Just got this, and then I come to find out that it is s...	1.0 out of 5 stars	1
10	Reviewed in the United States on November 5, 2020	Modern look with clock and alarm	Love it, small has a clock. Very modern. Replaced o...	5.0 out of 5 stars	1
11	Reviewed in the United States on November 5, 2020	Easy set up sounds so good	It's was so easy to set up replaced my google home...	5.0 out of 5 stars	2
12	Reviewed in the United States on November 30, 2020	Alexa, I think we should see other people	Bought product because it was a newer, more stylis...	1.0 out of 5 stars	2
13	Reviewed in the United States on December 19, 2020	Definite improvements, replacing a gen 2 Dot	Your browser does not support HTML5 video. My 2nd ge...	5.0 out of 5 stars	2
14	Reviewed in the United States on December 16, 2020	One star due to being a paper weight without interne...	The 4th gen echo dot with clock was my first Alexa ...	1.0 out of 5 stars	2
15	Reviewed in the United States on December 12, 2020	Best One Yet Because of More Than Its Shape .	I've got Echoes and Dots from 1st to 3rd generation...	5.0 out of 5 stars	2
16	Reviewed in the United States on December 6, 2020	Maybe mine was a lemon? BUT I'LL NEVER KNOW!	Super bummed. Bought this for my birthday. It wou...	1.0 out of 5 stars	2

텍스트 전처리

❖ 전처리 할 부분들

속성의 순서 변경

	Data	Title	Review	Rating	Page
1	Reviewed in the United States on November 5, 2020	The Pros, Cons and Oks for Echo Dot 4th Gen (HR).	Pros:_____ • The globe design with the fabric spe...	4.0 out of 5 stars	1
2	Reviewed in the United States on November 6, 2020	4th Gen. Twilight Blue w/ clock A great upgrade	Your browser does not support HTML5 video. I preordere...	5.0 out of 5 stars	1
3	Reviewed in the United States on November 5, 2020	A big improvement. Worth the wait, and money!	Your browser does not support HTML5 video. Just receiv...	5.0 out of 5 stars	1
4	Reviewed in the United States on November 7, 2020	Great echo dot stereo system !!!!	I purchased two of these to achieve true stereo. The...	5.0 out of 5 stars	1
5	Reviewed in the United States on November 8, 2020	Doesn't hear well	I got this to replace a 3rd generation Echo Dot and ...	1.0 out of 5 stars	1
6	Reviewed in the United States on November 26, 2020	Clock Display Too Bright	Alexa is awesome. We have 8 echos including our 2 ...	4.0 out of 5 stars	1
7	Reviewed in the United States on November 5, 2020	Worth the money 🍋.	Good looking 🍋. Perfect size and sound quality is g...	5.0 out of 5 stars	1
8	Reviewed in the United States on November 25, 2020	Echo dot 4th gen	Echo dot 4th gen. I've had all the versions of the Ec...	1.0 out of 5 stars	1
9	Reviewed in the United States on December 21, 2020	Shares your internet	Just got this, and then I come to find out that it is s...	1.0 out of 5 stars	1
10	Reviewed in the United States on November 5, 2020	Modern look with clock and alarm	Love it, small has a clock. Very modern. Replaced o...	5.0 out of 5 stars	1
11	Reviewed in the United States on November 5, 2020	Easy set up sounds so good	It's was so easy to set up replaced my google home...	5.0 out of 5 stars	2
12	Reviewed in the United States on November 30, 2020	Alexa, I think we should see other people	Bought product because it was a newer, more stylis...	1.0 out of 5 stars	2
13	Reviewed in the United States on December 19, 2020	Definite improvements, replacing a gen 2 Dot	Your browser does not support HTML5 video. My 2nd ge...	5.0 out of 5 stars	2
14	Reviewed in the United States on December 16, 2020	One star due to being a paper weight without interne...	The 4th gen echo dot with clock was my first Alexa ...	1.0 out of 5 stars	2
15	Reviewed in the United States on December 12, 2020	Best One Yet Because of More Than Its Shape.	I've got Echoes and Dots from 1st to 3rd generation...	5.0 out of 5 stars	2
16	Reviewed in the United States on December 6, 2020	Maybe mine was a lemon? BUT I'LL NEVER KNOW!	Super bummed. Bought this for my birthday. It wou...	1.0 out of 5 stars	2

국가이름

날짜

빈칸

빈칸

반복 내용

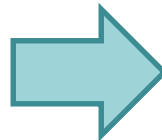
텍스트 전처리

Rating에서 ".0 out of 5 stars" 지우기

점수만 확인

```
amazon$Rating <- gsub(".0 out of 5 stars", "", amazon$Rating )
```

Rating
4.0 out of 5 stars
5.0 out of 5 stars
3.0 out of 5 stars
5.0 out of 5 stars
5.0 out of 5 stars
1.0 out of 5 stars



Rating
4
5
3
5
5
1

- `gsub` : 텍스트를 지우거나 변경할 수 있는 함수

텍스트 전처리

Data에서 국가 및 날짜 나누기

“ on”과 “on ”을 기준으로 문장을 나누기

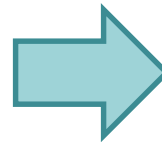
Country와 Date 생성

```
amazon$Country <- strsplit(amazon$Data, " on")[[1]][1]
```

```
amazon$Country <- gsub("Reviewed in ", "", amazon$Country)
```

```
amazon$Date <- strsplit(amazon$Data, "on ")[[1]][2]
```

	Data
1	Reviewed in the United States on October 15, 2018
2	Reviewed in the United States on January 8, 2020
3	Reviewed in the United States on October 15, 2018
4	Reviewed in the United States on October 17, 2018
5	Reviewed in the United States on January 21, 2019



Country	Date
the United States	October 15, 2018
the United States	October 15, 2018
the United States	October 15, 2018
the United States	October 15, 2018
the United States	October 15, 2018

- Strsplit : 문장을 나누는 함수

텍스트 전처리

Data 속성 지우기

```
amazon <- amazon[, -1]
```

	Data	Title
1	Reviewed in the United States on October 15, 2018	Love the Do
2	Reviewed in the United States on January 8, 2020	Love my ech
3	Reviewed in the United States on October 15, 2018	Overall impr
4	Reviewed in the United States on October 17, 2018	Actually imp



	Title	Review
1	Love the Dot and its new look, but the sound is still di...	I've
2	Love my echo dot! So helpful and the GGMM D3batte...	I k
3	Overall improvement, but lost its flexibility	I h
4	Actually impressed.	I w
5	Nueva amiga!	Es ur

텍스트 전처리

Pattern을 이용하여 newline을 의미하는 “\n” 지우고,

두 칸 white space 없애기

```
amazon$Title <- gsub(pattern="\n", "", amazon$Title)
```

```
amazon$Review <- gsub(pattern="\n", "", amazon$Review)
```

```
amazon$Title <- gsub("  ", "", amazon$Title)
```

```
amazon$Review <- gsub("  ", "", amazon$Review)
```

Data	
amazon	100 obs. of 6 variables
Title :	chr "\n\n\n\n\n\n\n\n\n\n"
Review :	chr "\n\n\n\n\n\n\n\n\n\n"



Data	
amazon	100 obs. of 6 variables
Title :	chr " Love the Dot and
Review :	chr " I've been a ha



Data	
amazon	100 obs. of 6 variables
Title :	chr "Love the Dot and
Review :	chr "I've been a happ

텍스트 전처리

속성 중요도에 따라 reorder

```
amazon <- amazon[ , c(6,5,3,1,2,4)]
```

	Title	Review	Rating	Page	Country	Date
1	The Pros, Cons and Oks for Echo Dot 4th Gen (HR).	Pros:_____ • The globe design with the fabric speaker lay...	4	1	the United States	November 5, 2020
2	4th Gen. Twilight Blue w/ clock A great upgrade	Your browser does not support HTML5 video. I preordered t...	5	1	the United States	November 5, 2020
3	A big improvement. Worth the wait, and money!	Your browser does not support HTML5 video. Just received i...	5	1	the United States	November 5, 2020
4	Great echo dot stereo system !!!!	I purchased two of these to achieve true stereo. The update...	5	1	the United States	November 5, 2020
5	Doesn't hear well	I got this to replace a 3rd generation Echo Dot and quite fra...	1	1	the United States	November 5, 2020
6	Clock Display Too Bright	Alexa is awesome. We have 8 echos including our 2 echo-au...	4	1	the United States	November 5, 2020
7	Worth the money 🍷.	Good looking 🍷. Perfect size and sound quality is good.	5	1	the United States	November 5, 2020

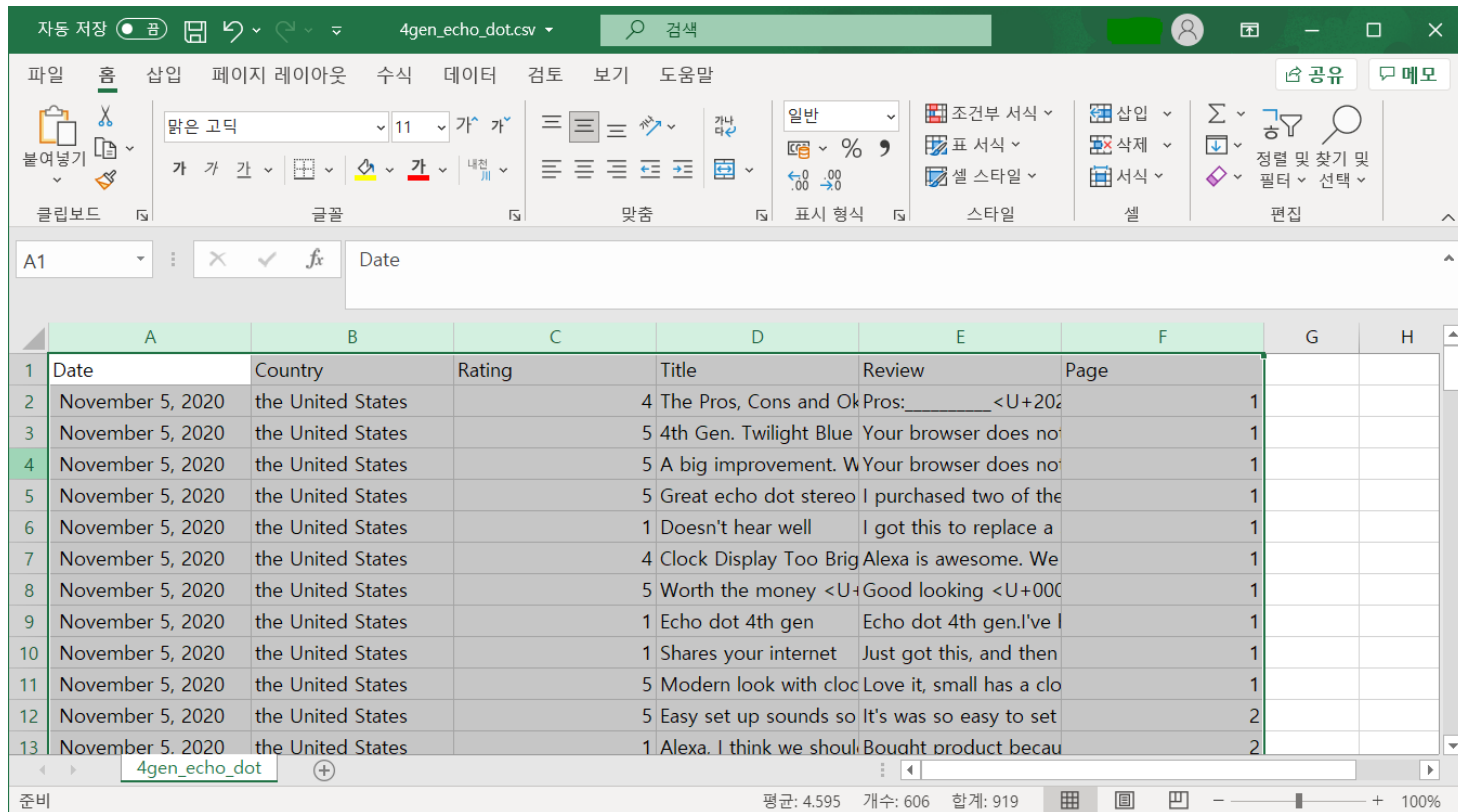


	Date	Country	Rating	Title	Review	Page
1	November 5, 2020	the United States	4	The Pros, Cons and Oks for Echo Dot 4th Gen (HR).	Pros:_____ • The globe design with the fabric speaker lay...	1
2	November 5, 2020	the United States	5	4th Gen. Twilight Blue w/ clock A great upgrade	Your browser does not support HTML5 video. I preordered t...	1
3	November 5, 2020	the United States	5	A big improvement. Worth the wait, and money!	Your browser does not support HTML5 video. Just received i...	1
4	November 5, 2020	the United States	5	Great echo dot stereo system !!!!	I purchased two of these to achieve true stereo. The update...	1
5	November 5, 2020	the United States	1	Doesn't hear well	I got this to replace a 3rd generation Echo Dot and quite fra...	1
6	November 5, 2020	the United States	4	Clock Display Too Bright	Alexa is awesome. We have 8 echos including our 2 echo-au...	1
7	November 5, 2020	the United States	5	Worth the money 🍷.	Good looking 🍷. Perfect size and sound quality is good.	1

CSV 저장하기

CSV 저장하기

```
write.csv(amazon, file= "4gen_echo_dot.csv", row.names = FALSE)
```



	A	B	C	D	E	F	G	H
1	Date	Country	Rating	Title	Review	Page		
2	November 5, 2020	the United States		4 The Pros, Cons and Ok	Pros: <U+202	1		
3	November 5, 2020	the United States		5 4th Gen. Twilight Blue	Your browser does no	1		
4	November 5, 2020	the United States		5 A big improvement. W	Your browser does no	1		
5	November 5, 2020	the United States		5 Great echo dot stereo	I purchased two of the	1		
6	November 5, 2020	the United States		1 Doesn't hear well	I got this to replace a	1		
7	November 5, 2020	the United States		4 Clock Display Too Brig	Alexa is awesome. We	1		
8	November 5, 2020	the United States		5 Worth the money <U+	Good looking <U+00C	1		
9	November 5, 2020	the United States		1 Echo dot 4th gen	Echo dot 4th gen.I've l	1		
10	November 5, 2020	the United States		1 Shares your internet	Just got this, and then	1		
11	November 5, 2020	the United States		5 Modern look with cloc	Love it, small has a clo	1		
12	November 5, 2020	the United States		5 Easy set up sounds so	It's was so easy to set	2		
13	November 5, 2020	the United States		1 Alexa. I think we shoul	Bought product becau	2		