웹에서 텍스트 자료의 수집

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-03-14

Contents

1 웹에서 텍스트 자료의 수집

1.1 웹스크랩이란?

1. 웹스크래핑(Web scraping; web harvesting; web data extraction): 웹사이트에 있는 정보를 추출하는 컴퓨팅 기술
2. 웹문서(사이트)는 통상 텍스트와 이미지가 혼합되어 있는 HTML형식으로 구성됨
3. 웹스크래핑은 비구조화된 웹문서 자료를 정형화된(구조화된) 형태로 변환하여 데이터베이스나 스프레드시트에 저장, 분석할 수 있도록 하는 것

1.2 HTML/XML

- HTML (HyperText Markup Language)
  - 팀 버너스리가 개발한 마크업 요소(tag)와 속성등을 이용하여 웹 페이지를 쉽게 작성할 수 있도록 하는 마크업 언어
- XML(Extensible Markup Language)
  - XML은 서로 다른 유형의 데이터를 기술하는 마크업 언어
  - 다른 종류의 시스템간 (특히 인터넷에 연결된 시스템)끼리 데이터를 쉽게 주고 받을 수 있도록 고안
  - HTML의 한계에 대한 대안

1.3 HTML/XML의 예시

- HTML

```
<!DOCTYPE html>
<html>
  <body>
    <p>This is a paragraph.</p>
    <p>This is another paragraph.</p>
  </body>
</html>
```

- XML

```
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

  - https://www.w3schools.com/xml/cd_catalog.xml

1.4 웹문서 가져오기 - 웹스크랩/웹크롤링

- 웹 크롤러(web crawler): 조직적, 자동화된 방법으로 웹을 탐색하는 컴퓨터 프로그램
- 웹 크롤링(web crawling): 웹 크롤러가 하는 작업

1.5 R에서 웹문서 가져오기 - 웹스크랩/웹크롤링

- 웹에 있는 데이터를 가져오는 단계
  1. 요청: GET과 POST 방식
  2. 추출 및 저장
- 관련 R 패키지
  - XML, RCurl, httr, rvest, ⋯

1.6 XML, RCurl

```
library(XML)
library(RCurl)
url <- "http://survey.joins.com/detail.asp?tp=1&cn=19853"
```

```
doc <- getURL(url, .encoding="UTF-8")
d <- htmlParse(doc)
oo <- xpathSApply(d, '//*[@id="body"]', xmlValue)
oo
# 윈도우즈에서 코딩할 경우에만 아래코드를 실행
#xx <- iconv(oo, "latin1", "CP949")
#
oo <- gsub("\n", " ", oo)
oo
```

1.7 rvest, httr

- rvest의 동작 순서

1. html 문서 데이터 가져오기
2. 필요한 노드 선택하기
3. 노드에서 필요한 작업
    - 노드에서 text를 가져오기
    - 노드내에 특정 속성(attr)을 추출

- 사용법

```
read_html(url)
read_html(url) %>% html_nodes("tag.class")
read_html(url) %>% html_nodes("tag.class") %>% html_text
read_html(url) %>% html_nodes("tag.class") %>% html_attr("attr1")
```

2 웹스크래핑 R 예제

2.1 경주캠퍼스 민원 게시판 예제

- 웹브라우저 화면

1/72페이지 총3186 공개: 1046건, 비공개: 2140건 제목 ▼ Search

번호	상태	제목	업무	담당부서	등록일	조회수
1066	공개	장학금을 돌려주세요!	학생복지/장학	학생서비스팀(경)	2016.02.29	999
1065	공개	토익	교양/외국어…	파라미타칼리…	2016.02.24	631
1064	공개	재학증명서 Fax 부탁드립니다.	학생복지/장학	학생서비스팀(경)	2016.02.18	999
1063	공개	교양필수과목은 다 듣게 해줘야되는 거 아..	취업/경력개…	취업지원센터(경)	2016.02.12	929
1062	공개	교양필수에 대해서 문의드립니다	교양/외국어…	파라미타칼리…	2016.02.05	678
1061	공개	기숙사 입관 관련	기숙사	관리팀(경)	2016.02.01	628
1060	공개	근로장학과 관련된 글을 올린지 꽤 됐는데..	학생복지/장학	학생서비스팀(경)	2016.01.11	410
1059	공개	답변이 뭐이리 느리죠??	기획/예산/혁…	경영평가실(경)	2016.01.09	453
1058	공개	성적증명서	학생복지/장학	학생서비스팀(경)	2016.01.08	372
1057	공개	수강교과목 삭제	수업/학사/학…	교무팀(경)	2016.01.07	370

Figure 1: 동국대 경주캠퍼스 민원 게시판

 R 코드

```
library(rvest)
url <- "https://web.dongguk.ac.kr/mbs/kr/jsp/community/bbsList3.jsp?page=1&table=VOC.TB_VOC&bb
nv  <- read_html(url)
html_table(nv)[[1]][, c(3,5)]
```

```
##                                      제목                 담당부서
## 1                          장학금을 돌려주세요!          학생서비스팀(경)
## 2                                  토익 파라미타칼리지 학사운영실(경)
## 3                  재학증명서 Fax 부탁드립니다.          학생서비스팀(경)
## 4  교양필수과목은 다 듣게 해줘야되는 거 아..          취업지원센터(경)
## 5              교양필수에 대해서 문의드립니다 파라미타칼리지 학사운영실(경)
## 6                            기숙사 입관 관련             관리팀(경)
## 7  근로장학과 관련된 글을 올린지 꽤 됐는데..          학생서비스팀(경)
## 8                      답변이 뭐이리 느리죠??           경영평가실(경)
## 9                                 성적증명서          학생서비스팀(경)
## 10                          수강교과목 삭제             교무팀(경)
## 11                 주소변경 신청하고 싶습니다           경영평가실(경)
## 12                   근로장학을 신청했습니다.          학생서비스팀(경)
## 13 안녕하십니까 농구코트 불 대관 관련 건의..          학생서비스팀(경)
## 14             국가근로장학을 신청했는데요,,,          학생서비스팀(경)
```

15 4학년 2학기의 취득학점 포기도 가능하게.. 교무팀(경)

2.2 쿠팡 - 등산화

 R 코드

```
url <- "http://www.coupang.com/np/search?q=%EB%93%B1%EC%82%B0%ED%99%94&channel=user&component=
shoes  <- read_html(url)
x <- html_nodes(shoes, xpath='//*[@id="productList"]/li')

#상품명
p_name <- html_nodes(shoes, xpath='//*[@class="name"]')
p_name <- html_text(p_name)
#상품 옵션
p_name_opt <- html_nodes(shoes, xpath='//*[@class="more-options"]')
p_name_opt <- html_text(p_name_opt)
#상품 가격
p_price <- html_nodes(shoes, xpath='//*[@class="price-value"]')
p_price <- html_text(p_price)
#상품평점
p_rating <- html_nodes(shoes, xpath='//*[@class="rating"]')
p_rating <- html_text(p_rating)

head(data.frame(p_name, p_price), 10)
```

Shop
PGA Stats
Customize results by filtering below
View by World Rank ▾ Year 2017 ▾ Filter
World Rank

Rank	Name	Events	Avg Points	Total Points	Points Lost	Points Gained
1	Dustin Johnson	45	10.6	478.0	-420.0	450.0
2	Jordan Spieth	48	9.4	450.0	-411.0	450.0
3	Justin Thomas	52	8.5	441.0	-213.0	446.0
4	Jon Rahm	40	8.2	326.0	-100.0	379.0
5	Hideki Matsuyama	49	8.1	395.0	-309.0	315.0
6	Justin Rose	45	7.9	357.0	-224.0	367.0
7	Rickie Fowler	48	6.9	331.0	-268.0	346.0
8	Brooks Koepka	47	6.4	302.0	-191.0	276.0
9	Henrik Stenson	44	6.0	265.0	-322.0	179.0
10	Rory McIlroy	40	5.8	231.0	-327.0	145.0
11	Sergio Garcia	44	5.8	254.0	-212.0	262.0
12	Marc Leishman	52	5.2	270.0	-133.0	285.0
13	Jason Day	42	5.2	217.0	-369.0	149.0
14	Paul Casey	47	5.0	234.0	-190.0	189.0

Figure 2: Yahoo PGA골프 선수 세계랭킹

2.3 PGA골프 선수별 통계

- URL : http://sports.yahoo.com/golf/pga/stats/bycategory?cat=WORLD_RANK&season=2017

- R code

```r
url  <- "http://sports.yahoo.com/golf/pga/stats/bycategory?cat=WORLD_RANK&season=2017"
world_ranking  <- read_html(url)
x <- html_table(world_ranking)[[2]] # 두번째 TABLE
head(x)
```

```
##   Rank             Name Events Avg Points Total Points Points Lost
## 1    1   Dustin Johnson     45       10.6          478        -420
## 2    2    Jordan Spieth     48        9.4          450        -411
## 3    3    Justin Thomas     52        8.5          441        -213
## 4    4         Jon Rahm     40        8.2          326        -100
## 5    5 Hideki Matsuyama     49        8.1          395        -309
## 6    6      Justin Rose     45        7.9          357        -224
##   Points Gained
## 1           450
## 2           450
## 3           446
## 4           379
## 5           315
## 6           367
```

3 부록 - R 코드

```
url0 <- "https://web.dongguk.ac.kr/mbs/kr/jsp/community/bbsList3.jsp?"
url1 <- "&table=VOC.TB_VOC&bbs=05&mode=open&keyword=&keyfield=&id=kr_070303030000"
out <- lapply(1:20, function(i){
  url <- paste0(url0, "page=",i, url1)
  nv  <- read_html(url)
  html_table(nv)[[1]][, c(3,5)]
})
save(out, file="campus_qna.RData")
```

```
load(file="campus_qna.RData")
oo <- do.call("rbind", out)
text <- oo$제목
library(KoNLP)
library(wordcloud)
x1 <- lapply(text, extractNoun)
x2 <- lapply(x1, function(x) x[nchar(x)>1])
x3 <- do.call(c, x2)
o <- table(x3)
```

```
pal <- brewer.pal(8,"Dark2")
wordcloud(names(o), o, min.freq=3, random.order=F,random.color=T,colors=pal)
```