웹에서 텍스트 자료의 수집

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-03-05

Contents

1 텍스트 자료의 처리

1.1 네이버 뉴스 수집 및 처리

- 검색어: 빅데이터
- URL
  - https://search.naver.com/search.naver?&where=news&query=%EB%B9%85%EB%8D%B0%EC%
- 웹스크래핑을 위한 R 코드

```
library(rvest)
n_news <- lapply(1:10, function(page){
  url <- paste0("https://search.naver.com/search.naver?&where=news&query=%EB%B9%85%EB%8D%B0%EC
  nr_table  <- read_html(url)
  title <- html_nodes(nr_table, xpath='//*[@class="type01"]/li/dl/dt/a')
  title <- html_text(title)
  sources <- html_nodes(nr_table, xpath='//*[@class="type01"]/li/dl/dd[1]/span[1]')
  sources <- html_text(sources)
  contents <- html_nodes(nr_table, xpath='//*[@class="type01"]/li/dl/dd[2]')
  contents <- html_text(contents)
```

```
  data.frame(title, sources, contents)
  })
n_news <- do.call(rbind, n_news)
```

```
load(file="naver_news.RData")
head(n_news)
```

```
##                                                                    title
## 1                                심평원, 건강보험 빅데이터 분석 협업 과제 공모
## 2 [브랜드평판] 건설회사 브랜드 2018년 3월 빅데이터 분석...1위 현대건설, 2위...
## 3 [뉴스투데이 E] 현대산업개발, 부동산114 빅데이터 활용한 신규사업 추진 본격...
## 4                                  통계청, '빅데이터 활용' 통계법 논문 공모
## 5                      [빅데이터MSI]5일 오전 주식시장 심리 1단계 '매우 나쁨'
## 6                       마이23헬스케어, 헬스케어 빅데이터 유통 위한 ICO 추진
##             sources
## 1          연합뉴스
## 2          미래한국
## 3        뉴스투데이
## 4        아시아경제
## 5            뉴시스
## 6 블로터언론사 선정
```

```
##
## 1                  건강보험심사평가원(심평원)은 오는 25일까지 '보건의료빅데이터 분석 협업 과제
## 2                        건설회사 브랜드 2018년 3월 빅데이터 분석 건설회사 브랜드 2018년 3월
## 3                 부동산114의 부동산 빅데이터를 활용한 복합개발의 효과성 제고, 지역 수요에 특
## 4               지난해까지 별개로 개최해 온 '대학(원)생 논문 공모'와 '마이크로데이터 우수활용
## 5 = 5일 빅데이터로 분석한 주식시장 코스피2500 종목의 시장심리지수(Market Sentiment Index·MS
## 6                  헬스케어 기업인 마이23헬스케어가 데이터 유통 및 가치 극대화를 위해 암호화
```

1.2 네이버 뉴스의 처리

1. 네이버 뉴스의 내용(contents)처리

    1. 문장부호 제거

```
contents <- gsub("[[:punct:]]", " ", n_news$contents)
```

    2. 띄어쓰기 교정: Web API이용, 한번에 200자까지만 제한, 오래 걸림

```
library(httr)
out <- character(length(contents))
for(i in seq_along(contents)){
    body <- list(sent=contents[i])
    res  <- PUT(url='http://35.201.156.140:8080/spacing', body=body)
    out[i] <- content(res)$sent
}
```

    3. 숫자 제거

```
contents <- gsub("[[:digit:]]", "", out)
```

2. 어근/명사 추출 및 사용자 단어의 추가

    1. 명사추출

```
library(KoNLP)
library(NIADic)
useNIADic()
```

```
## Backup was just finished!
## 983012 words dictionary was built.
```

```
out1 <- lapply(contents, extractNoun)
out1[1:2]
```

    2. 사용자 단어 추가

```r
new_term <- c("열린데이터광장", "헬스케어", "빅카인즈", "서울특별시", "박원순",
              "원주의료기기테크노밸리", "건강보험심사평가원", "심평원", "삼성물산",
              "코스피", "뉴시스", "코스콤", "암호화폐", "하나카드")
new_dic <- data.frame(new_term , "ncn")
buildDictionary(ext_dic = c('sejong', 'woorimalsam', 'insighter'),user_dic = new_dic)
#                 , category_dic_nms=c('political'))

out2 <- lapply(contents, extractNoun)
out2[1:2]
```

1.3 네이버 뉴스의 기초분석

1. 단어 빈도

```
#out2 : list type ==> vector
out_v <- do.call(c, out2)
wt <- table(out_v)
wt <- wt[nchar(names(wt)) > 1]
sort(wt, decreasing = T)[1:10]
```

```
## out_v
## 빅데이터     분석   데이터 아키텍처     활용 인공지능     개발     출시
##      130       76       67       30       29       27       25       23
##   부동산 비즈니스
##       21       20
```

2. 워드 클라우드

```
library(wordcloud)
pdf.options(family = "Korea1deb")
pal <- brewer.pal(8,"Dark2")
```

```
wordcloud(words=names(wt), freq=wt, min.freq=5,
          random.order=F, random.color=T, colors=pal)
```


빅데이터
분석
데이터
아키텍처
인공지능
출시
활용
서비스
비즈니스
하나카드
블록체인
부동산
클라우드
헬스케어
서울시
기업
특집
통합
개발
산업
저장
수집
혜택
강화
브랜드
창간
뉴스
시장
98
114
기반
1Q
베이징
암호화폐
산업혁명
삼성SDS
빅카인즈
플러스
거점
조정
의미
운용
서버
400
특파원
생활
계획
주석
바탕
사업영역
유력
정책
게임
종목
기업용
핵심기술
전인대
시스템
국가주석
매출
건설회사
수단
특화
html
마이23헬스케어
마이23헬스케어가
열린데이터광장
효과
공룡
하게
ICO
가치
지난해
Index
newsis
CNS
AI
LG
경제론
복합
마원
심평원
위변조
지각화
알리바바
목적
10카드
평균
넥슨
정보
wid
지수
안전
포함
수행
통계
코스피2500
축적
진출
1단계
IoT
공략
해커
국가
공동
보관
극대화
시진핑
필요
과정
배제
MSI
서울시
KT
2018
유통
지역
완료
Redundancy
Integration
일련
공개
포털
씨앤비뉴스
수요
사업
신사업
생태계
텔로모바일
서울특별시
CNBNEWS
추구
주식시장
심리
삼성SDS가
출처
마이23헬스케어
인구
코스콤
하루
세상
돌파
일반
센터
Enterprise
변환
stock
CNB
알파
디지털
해킹
종합적
http
기간
DAP
Sentiment
Market