뉴스 기사 웹크롤링을 통한 논술 주요 키워드 탐색

경 희 대 학 교
C A I T e c h
손 권 상 구 국 원

CONTENTS

CAITech
(Center for Advanced Information Technology)

1) 비즈니스 환경에 특화된 데이터 과학 연구

2) 데이터 마이닝, 텍스트 마이닝 기법을 활용한 비즈니스 빅데이터 분석

3) JAVA, R, Python 등을 활용한 머신러닝, Deep Learning 연구 및 비즈니스 응용

4) 빅데이터 기반의 트렌드 분석 및 추천 시스템 연구

경희대학교
KYUNG HEE UNIVERSITY

CAITech
Research center for advanced
information technology

R 및 R studio 설치

01

https://cran.r-project.org/bin/windows/base/

R 및 R studio 설치

01

https://www.rstudio.com/products/rstudio/download/

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.1.456 - Windows Vista/7/8/10	85.8 MB	2018-07-19	24ca3fe0dad8187aabd4bfbb9dc2b5ad
RStudio 1.1.456 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-07-19	4fc4f4f70845b142bf96dc1a5b1dc556
RStudio 1.1.456 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-07-19	3493f9d5839e3a3d697f40b7bb1ce961
RStudio 1.1.456 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB	2018-07-19	863ae806120358fa0146e4d14cd75be4
RStudio 1.1.456 - Ubuntu 16.04+/Debian 9+ (64-bit)	64.9 MB	2018-07-19	d96e63548c2add890bac633bdb883f32
RStudio 1.1.456 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2018-07-19	1df56c7cd80e2634f8a9fdd11ca1fb2d
RStudio 1.1.456 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-07-19	5e77094a88fdbdddddb0d35708752462

키워드 검색 및 웹 크롤링을 통한 데이터 수집

02

https://github.com/KGWON/KU/tree/master/talent_donation

README.md

1. 검색할 단어를 컴퓨터에게 알려줍니다.

```
install.packages("rvest") # 크롤링을 위한 패키지 설치, 한번만 실행!!
library(rvest) # 패키지 불러오기
search_query <- readline("검색어를 입력해 주세요: ") # 검색하고 싶은 주제를 입력.
```

2. 네이버 뉴스 기사를 검색하고 100페이지 분량의 제목을 컴퓨터에 저장합니다.

```
temp <- c()

for (i in 1:100)
{
      url <- URLencode(paste0("https://search.naver.com/search.naver?&where=news&query=", search_query, "&sm
  html <- read_html(url, encoding = 'utf-8')
  add <- html %>% html_nodes("._sp_each_title") %>% html_text()
  temp <- c(temp, add)
  print(paste0("[", i, "]", "번째 페이지까지 완료하였습니다. ", "  [예시]: ", add[1]))
  rm(add)
}
```

3. 각 문장에서 '단어'들만 뽑아낸 후 상위 100개의 단어를 살펴봅니다.

```
install.packages("KoNLP") # 문장에서 명사를 꺼내오기 위한 패키지, 한번만 실행 !!
library(KoNLP) # 패키지 불러오기
useNIADic() # 패키지에 내장 된 사전 불러오기

result <- lapply(temp, function(x) extractNoun(x)) # 명사만 꺼내오기
result <- unlist(result)
result <- table(result[nchar(result) > 1]) # 두 단어 이상인 명사만 저장하기
result <- result[order(result, decreasing = T)]

print(result[1:100])
```

4. 워드클라우드를 통해서 단어를 살펴보기

```
install.packages("wordcloud")
install.packages("RColorBrewer")
library(wordcloud)
library(RColorBrewer)

wordcloud(names(result[1:100]), result[1:100], colors = brewer.pal(8, "Dark2"), random.order = T, random.col
```

[주제]

1. 4차 산업혁명
2. 인공지능
3. 빅데이터
4. 블록체인

경희대학교 KYUNG HEE UNIVERSITY

CAITech
Research center for advanced information technology

THANK YOU

경 희 대 학 교

C A I T e c h

손 권 상 구 국 원

miroo1215@khu.ac.kr
qas6125@khu.ac.kr