R 데이터(텍스트) 마이닝 및 시각화 분석

Seongmin Mun

27th January 2022

아주대학교
AJOU UNIVERSITY

Outline

데이터 전처리(i.e., pre-processing)

데이터 처리(i.e., processing)

데이터 분석하기

데이터 전처리(i.e., pre-processing)

초기 설정하기

```
> #메모리정리하기
> gc()
         used (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
Ncells  597326 32.0    1374068 73.4         NA   808471 43.2
Vcells 1608977 12.3    8388608 64.0      16384  1823793 14.0
> rm(list=ls())
> #한글깨짐현상
> #한글 인코딩 문제 해결
> #맥
> Sys.setlocale(category = "LC_CTYPE", locale = "ko_KR.UTF-8")
[1] "ko_KR.UTF-8"
> #윈도우
> options(encoding = "UTF-8")
>
> #경로확인하기
>
> getwd()
[1] "/Users/seongminmun/Desktop/2022/Ajou/Data/Chosun"
> setwd("/Users/seongminmun/Desktop/2022/Ajou/Data/Chosun")
> getwd()
[1] "/Users/seongminmun/Desktop/2022/Ajou/Data/Chosun"
> dir()
 [1] "Cleaned"                        "historyRecords_경종(20).txt"
 [3] "historyRecords_고종(26).txt"    "historyRecords_광해군(15).txt"
 [5] "historyRecords_단종(6).txt"     "historyRecords_명종(13).txt"
```

관련 패키지 불러오기

```
> # 데이터 정제하기
> library(remotes)
> library(rJava)
> library(KoNLP)
Checking user defined dictionary!

> library(stringr)
>
```

데이터 불러오기

```
> #데이터 불러오기
> King1 <- readLines("historyRecords_태조(1).txt"); King1
  [1] "\"x\""
```

데이터 정제하기-1

```
> # 한글만 남기고 제거하기
> King1Korean <- NULL
> for (i in 1:length(King1)){
+   cleanedText <- gsub("[^가-힣]", " ", King1[i])
+   cleanedText <- gsub("\\s+", " ", cleanedText)
+   if (cleanedText==" "){
+     next
+   } else {
+     cleanedText <- gsub("^\\s","",cleanedText)
+     cleanedText <- gsub("\\s$","",cleanedText)
+     King1Korean <- c(King1Korean,cleanedText)
+   }
+ }
> head(King1Korean)
[1] "태조가 백관의 추대를 받아 수창궁에서 왕위에 오르다"
```

데이터 정제하기-2

```
> King1KoreanPOS <- NULL
> for (i in 1: length(King1Korean)){
+   input <- King1Korean[i]
+
+   textPos22 <- SimplePos22(input); textPos22
+   textPOS <- NULL
+   for (j in 1:length(textPos22)){
+     process1<-textPos22[j];process1
+     process2<-as.character(process1);process2
+     process3<-gsub("\\+"," ",process2);process3
+     textPOS <- paste(textPOS,process3,sep=" ")
+   }
+   textPOS <- gsub("^\\s","",textPOS)
+   textPOS <- gsub("\\s$","",textPOS)
+   King1KoreanPOS <- c(King1KoreanPOS,textPOS)
+ }
> head(King1KoreanPOS)
[1] "태조가/NC 백관/NC 의/JC 추대/NC 를/JC 받/PV 아/EC 수창궁/NC 에서/JC 왕위/NC 에/JC 오르/NC 다/MA"
```

데이터 정제하기-3

```
> King1KoreanN_P <- NULL
> for (i in 1: length(King1KoreanPOS)){
+   input <- King1KoreanPOS[i]
+   if (input!=""){
+     inputWords <- strsplit(input, split=" ")
+     inputWords <- unlist(inputWords)
+     cleanedInput <- NULL
+     for (j in 1:length(inputWords)){
+       if(inputWords[j]!=""){
+         if(str_detect(inputWords[j],'([A-Z가-힣]+)/N')==TRUE | str_detect(inputWords[j],'([A-Z가-힣]
+)/P')==TRUE){
+           cleanedInput <- paste(cleanedInput, inputWords[j], sep=" ")
+         }
+       }
+     }
+   }
+   King1KoreanN_P <- c(King1KoreanN_P, cleanedInput)
+ }
> head(King1KoreanN_P)
[1] " 태조가/NC 백관/NC 추대/NC 받/PV 수창궁/NC 왕위/NC 오르/NC"
```

데이터 정제하기-4

```
> # 기호 제거하기
> King1KoreanN_P <- gsub("[^가-힣]", " ",King1KoreanN_P)
> King1KoreanN_P <- gsub("\\s+", " ",King1KoreanN_P)
> King1KoreanN_P <- gsub("^\\s", " ",King1KoreanN_P)
> King1KoreanN_P <- gsub("\\s$", " ",King1KoreanN_P); head(King1KoreanN_P)
[1] " 태조가 백관 추대 받 수창궁 왕위 오르 "
```

데이터 처리(i.e., processing)

패키지 설치하기

```
> install.packages("tm")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/tm_0.7-8.tgz'을 시도합니다
Content type 'application/x-gzip' length 1102134 bytes (1.1 MB)
==================================================
downloaded 1.1 MB


The downloaded binary packages are in
        /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//Rtmpl9Sa8b/downloaded_packages
> # word matrix생성하기
> #tm패키지 설치
> #install.packages("tm")
> library(tm)
필요한 패키지를 로딩중입니다: NLP
```

문서-단어 행렬 생성

```
> #코퍼스데이터 생성(말뭉치)
> King1KoreanN_P_corpus <- Corpus(VectorSource(King1KoreanN_P))
> King1KoreanN_P_corpus
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 173
> #TermDocumentMatrix를 활용하여 수치형 데이터로 형변환
> King1KoreanN_P_Tdm <- TermDocumentMatrix(King1KoreanN_P_corpus, control = list(wordLengths = c(2,
 Inf)))
> King1KoreanN_P_Tdm
<<TermDocumentMatrix (terms: 2057, documents: 173)>>
Non-/sparse entries: 5157/350704
Sparsity           : 99%
Maximal term length: 8
Weighting          : term frequency (tf)
```

데이터 분석하기

연관 단어 분석

```
> # 연관 단어 분석
> #10번 이상 출현한 명사
> findFreqTerms(King1KoreanN_P_Tdm, lowfreq = 10)
 [1] "오르"    "왕위"    "태조가" "등"      "로"      "에"      "이"      "일"      "하"      "가"
[11] "과"      "되"      "백성"   "사직"    "수"      "없"      "임금"    "청하"    "가지"    "를"
[21] "말하기" "문하"    "오"      "의"      "이르"    "나"      "못하"    "사"      "세우"    "않"
[31] "알"      "두"      "사람"   "삼"      "있"      "전하"    "정도전" "주"      "두렵"    "들"
[41] "말"      "으로"    "지"      "같"      "것"      "고려"    "관"      "나라"    "늘"      "대"
[51] "라"      "마음"    "법"      "서"      "씨"      "아들"    "은"      "을"      "이것"    "이색"
[61] "일으키" "죄"      "지방"   "하늘"    "데"      "뜻"      "신"      "덕"      "명"      "아니하"
[71] "이상"    "전"      "경"      "보"      "적"      "타"      "비"      "한"      "나머지" "승"
[81] "자"      "사리"    "따르"   "풀"      "관장"    "대부"    "종"      "거관"    "녹사"    "부사"
[91] "판사"    "주부"    "직장"   "장군"    "공신"
> #'태조'과 20% 연관성있는 명사
> findAssocs(King1KoreanN_P_Tdm, "태조", 0.2)
$태조
    더불     동맹    술자리   아뢰었 감록국사     고이     골목     기쁘 기울이     김균
    0.44     0.44     0.44     0.44     0.44     0.44     0.44     0.44     0.44     0.44
  나타내     눈물 대소신료   두서너     마을       메 민여익 배극렴과     상경     선포
```

워드클라우드-1

```
> install.packages("wordcloud")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/wordcloud_2.6.tgz'을 시도합니다
Content type 'application/x-gzip' length 278557 bytes (272 KB)
==================================================
downloaded 272 KB


The downloaded binary packages are in
        /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//Rtmpl9Sa8b/downloaded_packages
> #불러오기
> library(wordcloud)
필요한 패키지를 로딩중입니다: RColorBrewer
```

워드클라우드-2

```
> #단어의 출현빈도를 기반으로 생성된 Tdm데이터를 매트릭스형으로 변환
> King1KoreanN_P_Tdm_M <- as.matrix(King1KoreanN_P_Tdm)
> #단어들의 출현 빈도를 합한다.
> King1KoreanN_P_wordFreq <- sort(rowSums(King1KoreanN_P_Tdm_M), decreasing = TRUE)
> King1KoreanN_P_wordFreq
  명   하   풀   종   것   을   의 
377 347 303 156 149 135 127 
  를   일   이   되 사람   늘   있 
106 104  98  83  82  74  73 
  등   은 관장   가   오   말   없 
 69  60  55  53  50  49  46 
  한   로   과 으로   나   않   수 
 43  41  41  39  36  35  34 
```

워드클라우드-3

```
> #워드클라우드 색상지정
> pal <- brewer.pal(8, "Dark2")
> #그래프 폰트 깨짐 수정
> par(family = "AppleGothic")
> wordcloud(words = names(King1KoreanN_P_wordFreq), freq = King1KoreanN_P_wordFreq, min.freq = 4, ra
ndom.order = F, rot.per = 0.1, colors = pal)
>
```

워드클라우드-4