[R] R로 하는 비정형 데이터 처리 (facebook 데이터를 통한 긍정/부정 나누기) — 나무늘보의 개발 블로그

노트북: blog
만든 날짜: 2020-10-06 오후 4:43
URL: https://continuous-development.tistory.com/51?category=793392

R

[R] R로 하는 비정형 데이터 처리 (facebook 데이터를 통한 긍정/부정 나누기)

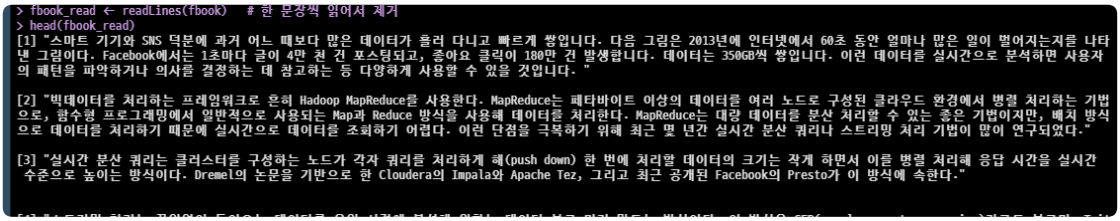
2020. 8. 3. 17:52 수정 삭제 공개

비정형 데이터 처리

일단 기본적인 패키지들을 install 하자

```
# 비정형 데이터 처리(텍스트 마이닝)
# 단어 빈도를 나타내는 시각화(wordcloud, koNLP, tm)


install.packages(c("hash", "tau", "Sejong",
                   "RSQLite", "devtools", "bit",
                   "rex", "lazyeval", "htmlwidgets",
                   "crosstalk", "promises", "later",
                   "sessioninfo", "xopen", "bit64",
                   "blob", "DBI", "memoise", "plogr",
                   "covr", "DT", "rcmdcheck", "rversions"),
                 type = "binary")

# github 버전 설치
install.packages("remotes")
# 64bit 에서만 동작합니다.
remotes::install_github('haven-jeon/KoNLP',
                        upgrade = "never",
                        INSTALL_opts=c("--no-multiarch"))


#감성분석
# service_data_facebook_bigdata.txt

fbook <- file(file.choose(),encoding="UTF-8")
fbook_read <- readLines(fbook)
head(fbook_read)
str(fbook_read)
```

여기서는 페이스북의 데이터를 가져왔다. 데이터의 내용은 아래와 같다. 이런 식으로 각 행에 대해서 문장형 데이터가 들어가 있었다.



이 문장들을 전처리할 필요가 있었다. 그래서 정규표현식을 통해 전처리를 하였다.

```
336 #2. 전처리(정규표현식이 필요하다)
337 #문장 부호 제거[[:punct:]]하는 정규표현식 활용
338 #특수문자 제거[[:cntrl:]]
339 #숫자 제거 [[0-9]] \\d+(숫자) , \\w(단어) , \\s+ (공백) , \n , \t
340 #gsub() 함수를 이용해서 전처리를 한다.
341
342
343 s1 ← gsub('[[:punct:]]','',fbook_read) #문장부호 제거
344 s1
345
346 s2 ← gsub('[[:cntrl:]]','',s1) # 특수문자 제거
347 s2
348
349 s3 ← gsub('\\d+','',s2) # 숫자제거
350 s3
351
352 s4 ← tolower(s3) #소문자로 변환
353 s4[1]
```

gsub을 통해서 정규표현식에 해당하는 데이터를 ''로 변환해주었다.
처음에는 문장부호를 제거하고 그다음에는 특수문자 , 숫자제거 이렇게 하였고 마지막에는 모든 대문자를 소문자로 바꿔주는 tolower를 사용하였다.

```
355
356 wordList ← str_split(s4,"\\s+") #공백으로 분류
357 wordVec←unlist(wordList) # vector로 만든다.
```

그다음은 str_split를 사용하여 공백을 통한 단어 분리를 하였다. 그다음 데이터 프레임 형식으로 된 데이터를

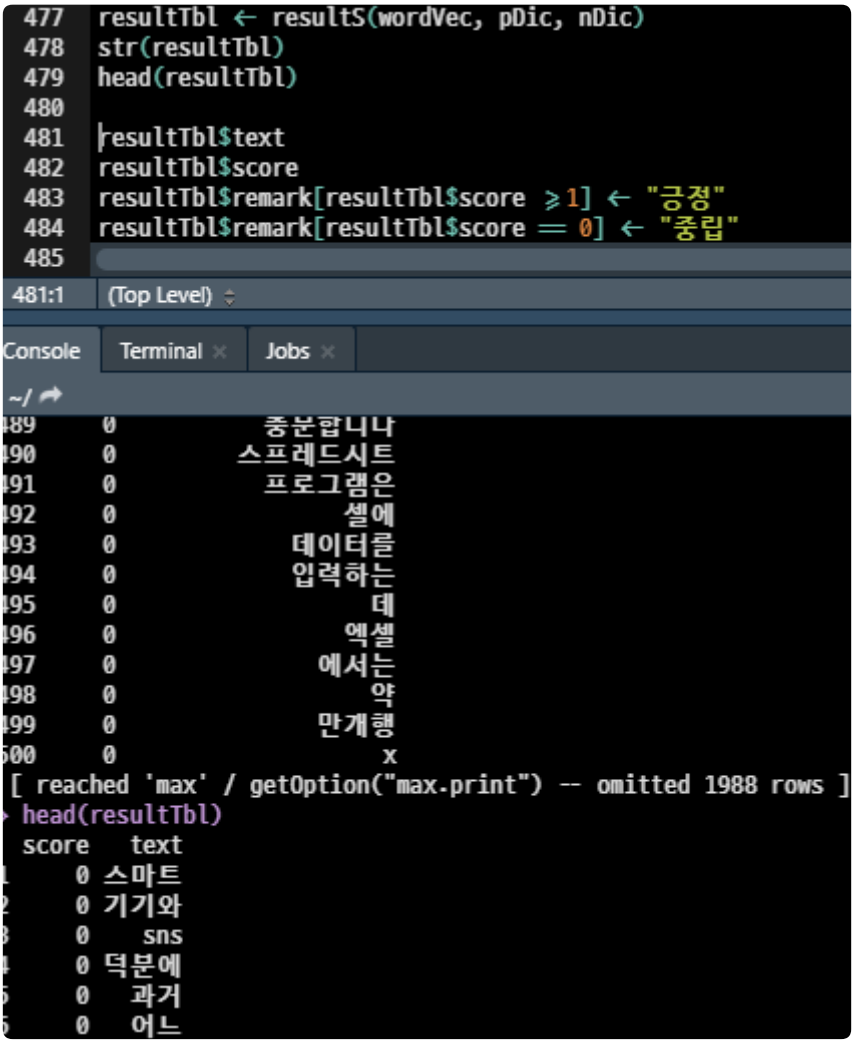
```
[[74]]
 [1] "빅데이터란"   "얼차적으로"   "데이터의"   "양이"     "방대해"   "종래의"   "방법으로는"   "수집"
[12] "어려운"     "것을"       "말한다"     "이차적으로는" "그런"     "큰"       "데이터를"     "여러"
[23] "정보로"     "만들어내는"   "과정까지를"  "포함한다"   "년"       "하반기부터" "우리나라에서도" "무한경쟁의"
[34] "새로운"     "블루오션으로" "빅데이터란"  "말이"     "대유행하기" "시작했다"  "삼성경제연구소는" "년"
[45] "빅데이터를"  "지목하며"   "이것이"     "미래의"   "성장"     "동력이"   "될"          "거라고"
[56] ""

[[75]]
 [1] "그런데"   "사실상"   "빅데이터는" "오래전부터" "우리"     "삶에"     "이미"     "들어와"   "있는"     "매우"
[14] "책의"     "저자는"   "말한다"     "또한"     "빅데이터"  "시대에"   "해독능력을" "위한"     "통계적"   "사고를"
[27] "데이터를"  "모아"     "분석해"     "가장"     "올바르고"  "빠른"     "답을"     "얻려주는"  "실용적인" "학문인"
[40] "근거가"   "되기에"   "현대"       "비즈니스맨이" "지녀야"   "할"       "최강의"   "무기라는"  "것이다"

[[76]]
 [1] "이"       "책은"     "일본에서"   "통계"     "관련"     "서적으로는" "이례적으로" "출간"     "계월"     "만에"
[14] "이례적인"  "현상을"   "불러일으키며" "상반기"   "경제경영"  "분야"     "베스트셀러" "위에"     "올랐다"   "문제"
[27] "척도로"   "통계의"   "역할을"     "새롭게"   "인식한"   "이"       "책은"     "통계학을"  "공부하려는" "사람을"
[40] "지금"     "이"       "순간"       "당신의"   "업무에"   "기업에"   "속한"     "공동체에"  "업무"     "비용을"
[53] "계획할"   "수"       "있게"       "하는"     "최고의"   "활용서이다"
```

unlist를 사용하여 vector 형식으로 바꿔주었다.

```
[793] "원해서는"  "데이터를"  "체계적으로" "정리하고"  "수집보관한"  "뒤"       "필요한"   "데이터를"  "질질하게"
[802] "뽑아서"   "사용할"   "수"       "있어야"   "합니다"   "그래서"   "데이터베이스의" "테이블을"  "설계하고"
[811] "쿼리를"   "저유은계"  "사용하는"  "능력이"   "필요합니다" "테이블"   "설계과"   "쿼리를"   "사용하기"
[820] "위한"     "도구도"   "다양하지만" "저는"     "exerd와"  "sql"      "developer을" "사용합니다" "데이터베이스"
[829] "설계"     "나에게"   "맡겨"     "exerd"    "exerd는"  "토마토"   "시스템즈에서" "개발한"   "이클립스"
[838] "거반의"   "지능형"   "er"       "도구입니다" "개인사용자는" "무료로"   "이용할"   "수"       "있습니다"
[847] "현재"     "oracle"   "microsoft" "sql"      "server"   "ibm"      "db"       "mysql을"  "대성으로"
[856] "리비스포워드" "엔지니어링과" "물리적"   "특성"     "필십을"   "지켜하고"  "있습니다"  "초보자도"  "데이터베이스"
[865] "설계를"   "적관적이고"  "쉽고"     "빠르게"   "알"       "수"       "있습니다"  ""         "쿼리"
[874] "이거"     "하나면"   "충분"     "oracle"   "sql"      "developer"  "sql"      "developer는" "oracle에서"
[883] "제공하는"  "sql"      "개발도구입니다" "오라클"   "데이터베이스와" "ms"       "access에"  "대한"     "데이터베이스"
[892] "검속을"   "지원합니다" "데이터베이스에서" "여러가지"  "쿼리를"   "자유자재로"  "사실간"   "추출하고"
[901] "가공할"   "수"       "있도록"   "도와줍니다"  ""         "대용량"   "데이터와"  "배치"
[910] "데이터를"  "다양한"   "분석"     "도구를"   "이용하여"  "쉽고"     "빠르게"   "분석할"   "수"
[919] "있는"     "inmomory" "computing" "기반의"   "고성능"   "빅데이터"  ""         "일반적으로" "빅데이터"
[928] "시스템을"  "얘기하자면" "다음과"   "같이"     "구성을"   "수"       "얻을"     "듯"       "싶다"
[937] "다양한"   "채널을"   "통한"     "대량의"   "데이터"   "수집"     "작업"     "그리고"   "수집된"
[946] "데이터를"  "분석하여"  "연관된"   "데이터끼리"  "분류하는"  "작업"     "분류된"   "데이터를"  "기반으로"
[955] "분석"     "정리하여"  "차"       "데이터를"  "만드는"   "작업"     "거"       "데이터를"  "분석하여"
[964] "최종"     "결과물을"  "만드는"   "작업으로"  "구성된"   "수"       "있을"     "것이다"   "믿은"
[973] "자"       "데이터를"  "기반으로"  "차"       "데이터를"  "만들"     "수도"     "있고"     "자"
[982] "데이터를"  "만드기"   "위해"     "복로된"   "데이터를"  "적적"     "분석하여차" "데이터의"  "생성"
```

이 다음에는 긍정 단어와 부정 단어의 데이터를 통해 내가 가지고 있는 데이터와 매칭 하는 작업을 하였다.
아래 함수를 생성했다. 이 함수를 통해 긍정 / 부정 / 중립을 나눴다.

```r
library(stringr)
library(plyr)
?laply

#여기서 생략된게 1. list일때 for문처럼 계속 도는것 2.sum(true) 1의 값을 나타내고 3.laply에서 function을 쓸경우에 매개변수를 밖에서 한번 정의줘야 한다.

# 이 함수를 정의하세요
resultS <- function(words , positive , negative) { #여기서 애초에 리스트로 돌아가기 때문에 for 문처럼 계속 돈다.

  scores = laply(words, function(words, positive, negative) {
    pMatch = match(words, positive)
    nMatch = match(words, negative)

    pMatch = !is.na(pMatch)  # true false는 1과 0을 나타낸다. 그래서 이 true를 sum으로 하면 1을 나타내고 이렇게 계산식을 나타낸다.
    nMatch = !is.na(nMatch)

    score = sum(pMatch) - sum(nMatch) #값이 양수이면 긍정의 단어로 -일경우 부정의 단어로 0 일경우 중립단어로서 나타낸다.
    return(score)
  }, positive, negative) # 이거 두개는 문법이다. 안에 function(words, positive, negative) 에 쓴거는 밖에서 이런식으로 적어준다. 결과적으로 retrun하는건 scroe 하나



  scores.df = data.frame(score=scores , text=words)
  return(scores.df)
}


resultTbl <- results(wordVec, pDic, nDic)
head(resultTbl)
```



```
477 resultTbl ← resultS(wordVec, pDic, nDic)
478 str(resultTbl)
479 head(resultTbl)
480
481 resultTbl$text
482 resultTbl$score
483 resultTbl$remark[resultTbl$score ⩾1] ← "긍정"
484 resultTbl$remark[resultTbl$score == 0] ← "중립"
485
481:1  (Top Level)

Console  Terminal  Jobs
~/
489     0          중분합니다
490     0         스프레드시트
491     0          프로그램은
492     0                셀에
493     0           데이터를
494     0           입력하는
495     0                데
496     0              엑셀
497     0             에서는
498     0                약
499     0             만개행
500     0                 x
[ reached 'max' / getOption("max.print") -- omitted 1988 rows ]
> head(resultTbl)
  score   text
1     0 스마트
2     0 기기와
3     0    sns
4     0 덕분에
5     0   과거
6     0   어느
```

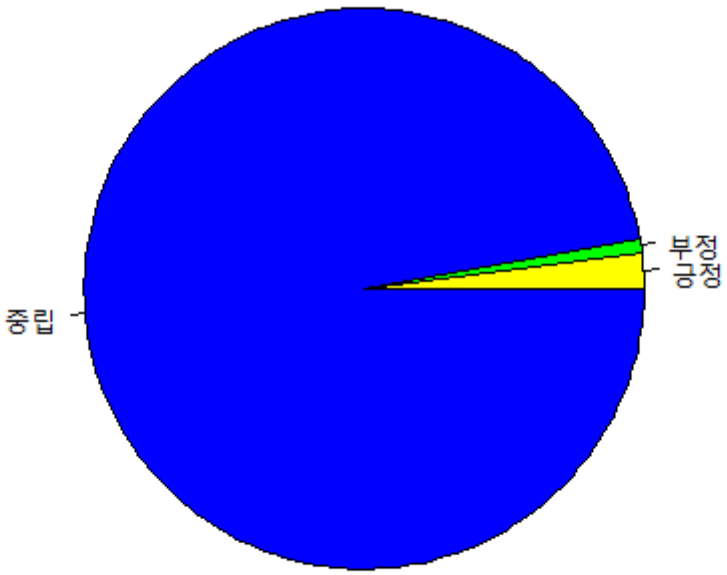
이런식으로 각 단어에 대한 긍정 부정을 볼 수 있다.

```
472
473 # 긍정부정에 따라서 파이차트 만들기
474 resultTbl ← resultS(wordVec, pDic, nDic)
475 str(resultTbl)
476 head(resultTbl)
477
478 resultTbl$text
479 resultTbl$score
480 resultTbl$remark[resultTbl$score ⩾1] ← "긍정"
481 resultTbl$remark[resultTbl$score == 0] ← "중립"
482 resultTbl$remark[resultTbl$score < 0] ← "부정"
483
484 resultTbl$remark
485
486 table(resultTbl$remark)
487
488 pieResult ← table(resultTbl$remark)
489 pieResult ← table(resultTbl$remark)
490
491 ?pie
492 pie(pieResult,
493     labels=names(pieResult),
494     col = c('yellow','green','blue'))
```

```
> table(resultTbl$remark)

긍정 부정 중립
  51   20 2417
```

이렇게 긍정 부정중립의 개ㅐ수를 셀수 있다.

이걸 파이차트로 나타내면 아래와 같다.



'R' 카테고리의 다른 글

[R] R을 활용한 상관분석과 회귀분석 - 1

[R] R을 통한 텍스트마이닝에서 워드클라우드 까지

[R] R로 하는 비정형 데이터 처리 (facebook 데이터를 통한 긍정/부정 나누기)

[R] R에서 Database 사용하기 / DB 기본적인 구문 사용하기

[R] 예제를 통한 데이터 전처리 작업

[R] R을 통해 시계열 그래프 만들기 (자료 분석을 위한 시각화와 실습 예제)

R 비정형 데이터 처리 R로 하는 비정형 데이터 처리 비정형 데이터 비정형 데이터 처리

꾸까꾸

혼자 끄적끄적하는 블로그 입니다.