Yok Dictionary

GI, SEUNG

2019 년 4 월 17 일

#욕사전

##1. 크롤링

- **필요한 패키지 요청**

```
require(stringr)
```

```
## Loading required package: stringr
```

```
## Warning: package 'stringr' was built under R version 3.5.3
```

```
require(rvest)
```

```
## Loading required package: rvest
```

```
## Warning: package 'rvest' was built under R version 3.5.3
```

```
## Loading required package: xml2
```

```
## Warning: package 'xml2' was built under R version 3.5.2
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(KoNLP)
```

```
## Loading required package: KoNLP
```

```
## Warning: package 'KoNLP' was built under R version 3.5.3

## Checking user defined dictionary!
```

- **욕사전 url 을 받아온다.**

```
yok_url='https://namu.wiki/w/%EC%9A%95%EC%84%A4/%ED%95%9C%EA%B5%AD%EC%96%B4'
```

나무위키

- **텍스트 가져오기**

```
html=read_html(yok_url)
yok_crawling=html %>%
  html_nodes('.wiki-heading-content') %>%
  html_nodes('.wiki-list') %>%
  html_nodes('.wiki-paragraph') %>%
  html_text()
head(yok_crawling)

## [1] "간나새끼, 개간나, 쌍간나, 종간나, 좆간나. 대표적인 북한의 욕."

## [2] "갈보[1] : 걸레, 창녀와 비슷한 뜻."

## [3] "개, 개-(두 '개'의 링크가 다르다) : '개'라는 동물 자체를 일컫는 말이라고 하기에는 '비하' 또는 '격하'의 의미가 강하다. 네이버 국어사전에 따르면 동물 '개'가 아
```

닌 다른 의미로 '개'는 '행실이 형편없는 사람을 비속하게 이르는 말'이라는 풀이가 있는데, 이 경우가 욕설로서의 '개'의 어원이 된다고 할 수 있다. '너는 개다'라는 식으로 '개' 그 자체로 욕설이 되기도 하지만, '개-'와 같이 접두사로 사용되면서 비하 또는 격하의 의미(개년, 개새끼, 개판, 개씨발, 개꼴통, 개똥차, 개쌍판, 개꿈, 개수작, 개망나니 등)가 되기도 하고, 2000 년대 들어서는 청소년과 온라인을 중심으로 '강조'의 의미(개맛있다, 개재미있다(개꿀잼)/없다, 개덥다, 개무겁다, 개비싸다, 개 존나 잘생겼다(개존잘) 등) 를 띠기도 하는 추세이다. 후자의 경우 어감은 좀 그렇지만 '욕설' 또는 '비하'의 의미는 없는 것으로 이해하는 편이다. 하지만, 기본적으로 '개-'를 접두어로 사용하는 경우 욕이 아닌 것도 웬만하면 욕인 듯한 느낌으로 만들 수 있고, 특히 듣는 사람들 입장에서는 욕설처럼 들릴 수 있으니 사용에 유의하자."
[4] "개간년 : 개같은 여자의 약자"

[5] "개나리"

[6] "개년"

##2. 사전 만들기 + #### 단어와 설명나누기

```
df1=data.frame(word=NA,description=NA)
for(i in 1:length(yok_crawling)){
```

```
  df1[i,1]=str_split(yok_crawling,":")[[i]][1]
  df1[i,2]=str_split(yok_crawling,":")[[i]][2]
}
#Rstudio
```

- “,”로 되어있는 단어들 나누기

```
word_split=list()
descript_split=list()
for(i in 1:nrow(df1)){
  word_split[[i]]=if(length(unlist(str_split(df1[i,"word"],","))) >=1){unlist
(str_split(df1[i,"word"],","))}
  word_split[[i]]=gsub('\\s','',word_split[[i]])
  descript_split[[i]]=rep(df1[i,2],length(word_split[[i]]))
}
df2=data.frame(word=unlist(word_split),description=unlist(descript_split),str
ingsAsFactors = F)
head(df2[,1],20)

##  [1] "간나새끼"                  "개간나"
##  [3] "쌍간나"                    "종간나"
##  [5] "좆간나.대표적인북한의욕."  "갈보[1]"
##  [7] "개"                        "개-(두'개'의링크가다르다)"
##  [9] "개간년"                    "개나리"
## [11] "개년"                      "개돼지"
## [13] "개새끼(견공자제분)"        "개소리"
## [15] "개씨발(개씹할)"            "개씹"
## [17] "개좆"                      "개지랄"
## [19] "개족새"                    "개차반"
```

- 욕설을 제외한 부가설명 지워주기

```
for(i in 1:nrow(df2)){
  df2[i,1]=str_split(df2[,1],'\\(')[[i]][1]
  df2[i,1]=str_split(df2[,1],'\\[')[[i]][1]
  df2[i,1]=str_split(df2[,1],'\'')[[i]][1]
  df2[i,1]=str_split(df2[,1],'=')[[i]][1]
  df2[i,1]=str_split(df2[,1],'\\.')[[i]][1]
  df2[i,1]=gsub('-','',df2[i,1])
  df2[i,1]=gsub('\\)','',df2[i,1])
}
head(df2[,1],20)

##  [1] "간나새끼" "개간나"   "쌍간나"   "종간나"   "좆간나"   "갈보"
##  [7] "개"       "개"       "개간년"   "개나리"   "개년"     "개돼지"
```

```
## [13] "개새끼"   "개소리"   "개씨발"   "개씹"     "개좆"     "개지랄"
## [19] "개족새"   "개차반"
```

###2.1 초성 사전 만들기 + #### 욕사전 초성으로 바꾸기

```
for(i in 1:nrow(df2)){
  initial=vector()
  jamos=convertHangulStringToJamos(df2[i,1])
  for(j in 1:length(jamos)){
    jamos2=convertHangulStringToJamos(jamos[j])
    initial[j]=jamos2[1]
  }
  initial_sum=paste(initial,collapse='')
  df2[i,'initial']=initial_sum
}
head(df2$initial)
```

```
## [1] "ㄱㄴㅅㄲ" "ㄱㄱㄴ"   "ㅆㄱㄴ"   "ㅈㄱㄴ"   "ㅈㄱㄴ"   "ㄱㅂ"
```

###2.2 자음,모음으로 분리된 사전 만들기

```
df_jamo=list()
for (i in 1:nrow(df2)){
  hangul=convertHangulStringToJamos(df2[i,1])
  df_jamo[[i]]=hangul
}
head(df_jamo)
```

```
## [[1]]
## [1] "ㄱㅏㄴ" "ㄴㅏ"   "ㅅㅐ"   "ㄲㅣ"
##
## [[2]]
## [1] "ㄱㅐ"   "ㄱㅏㄴ" "ㄴㅏ"
##
## [[3]]
## [1] "ㅆㅏㅇ" "ㄱㅏㄴ" "ㄴㅏ"
##
## [[4]]
## [1] "ㅈㅗㅇ" "ㄱㅏㄴ" "ㄴㅏ"
##
## [[5]]
## [1] "ㅈㅗㅈ" "ㄱㅏㄴ" "ㄴㅏ"
##
## [[6]]
## [1] "ㄱㅏㄹ" "ㅂㅗ"
```

##3.욕판별 하는 함수 만들기 + #### 3.1 숫자, 영어 특수문자 제거

```
yok_num=function(word){
  word=str_replace_all(word,'[A-z0-9]','')
  if (word %in% df2[,1]){
    word2=df2[which(df2[,1]==word),2]
    return(word2)
  }
}
yok_num('시1발')
```

```
## [1] " 성행위를 뜻하는 '씹할'에서 기원한 욕설.[5] 씨발놈, 씨발년, 씨발새끼 등과 같이 쓰인다. 식빵"
```

- **3.2 자음으로 들어온 단어**

```
yok_jaum=function(word){
  word=str_replace_all(word,'[A-z0-9]','')
  if(word %in% df2[,3]){
    word2=df2[which(df2[,3] == word),2]
    return(word2)
  }
}
yok_jaum('ㅅㅂ')
```

```
## [1] " 성행위를 뜻하는 '씹할'에서 기원한 욕설.[5] 씨발놈, 씨발년, 씨발새끼 등과 같이 쓰인다. 식빵"
```

- **3.3 이상한 단어 바꿔주기**

```
yok_strange=function(word){
  word=str_replace_all(word,'[A-z0-9]','')
  word2=convertHangulStringToJamos(word)
  word3=Filter(function(x){nchar(x)>=2},word2)
  for (i in 1:length(df_jamo)){
    word4=word3[word3 %in% df_jamo[[i]]]
    if(length(word4) == length(df_jamo[[i]])){
      word5=paste(word4,collapse='')
      word6=HangulAutomata(word5)
      return(df2[which(df2[,1]==word6),2])
    }
  }
}
yok_strange('시이이이이이이ㅣ1발')
```

```
## [1] " 성행위를 뜻하는 '씹할'에서 기원한 욕설.[5] 씨발놈, 씨발년, 씨발새끼 등과 같이 쓰인다. 식빵"
```

##4.욕변환

```
yok_convert=function(word){
  word=str_replace_all(word,'[A-z0-9]','')
  if (word %in% df2[,1]){
    return(df2[which(df2[,1]==word),2])
  }
  else{
    word2=convertHangulStringToJamos(word)
    word3=Filter(function(x){nchar(x)>=2},word2)
    for (i in 1:length(df_jamo)){
      word4=word3[word3 %in% df_jamo[[i]]]
      if(length(word4) == length(df_jamo[[i]])){
        word5=paste(word4,collapse='')
        word6=HangulAutomata(word5)
        return(df2[which(df2[,1]==word6),2])
      }
    }
  }
}
yok_convert('시이이이이이이ㅣ1발')
```

```
## [1] " 성행위를 뜻하는 '씹할'에서 기원한 욕설.[5] 씨발놈, 씨발년, 씨발새끼 등과 같이 쓰인다. 식빵"
```