TM_Lab01_KoNLP

Lab03_KoNLP를 이용한 텍스트 마이닝 분석

학습 목표

- KoNLP는 뭘까?
- 사전이란?
- 형태소란?
- 품사란?

학습 내요

```
03-01 KoNLP 설치
03-02 사전 불러오기
03-03 간단한 명사 추출
03-03 사전에 단어 추가하기
03-04 형태소 분석하기
03-05 품사 달기
```

03-01 KoNLP 설치

- KoNLP에는 한국어를 분석할 수 있는 27개 이상의 함수 존재
- 참고 PDF : https://cran.r-project.org/web/packages/KoNLP/KoNLP.pdf (https://cran.r-project.org/web/packages/KoNLP/KoNLP.pdf)

```
install.packages("KoNLP")
library(KoNLP)
```

```
# install.packages("KoNLP")  # 설치는 한번만 실행되면 library로 불러오기만 실행.
library(KoNLP)
```

```
## Warning: package 'KoNLP' was built under R version 3.3.3
```

```
## Checking user defined dictionary!
```

03-02 사전 불러오기

- 형태소 분석을 위한 참고 사전 필요
- 37만 단어의 사전

```
useSejongDic()
```

```
## Backup was just finished!
## 370957 words dictionary was built.
```

03-03 간단한 명사 추출

- Hannanum analyzer를 사용하여 한국어 문장으로부터 명사를 추출한다.
- extractNoun(sentences, autoSpacing = FALSE)
- sentences : 문장
- autoSpacing : 입력에 대한 자동 공백

```
sentence <- "우리나라 고유의 문화들을 멋진그래픽으로 잘 표현했다. 재미있게 봤다. 스르륵 시간이 지나갔다."
sentence2 <- "님은 갔습니다. 아아, 사랑하는 나의 님은 갔습니다.
푸른 산빛을 깨치고 단풍나무 숲을 향하여 난 작은 길을 걸어서 차마 떨치고 갔습니다.
황금의 꽃같이 굳고 빛나던 옛 맹세는 차디찬 티끌이 되어서 한숨의 미풍에 날아갔습니다.
날카로운 첫 키스의 추억은 나의 운명의 지침을 돌려 놓고 뒷걸음쳐서 사라졌습니다.
나는 향기로운 님의 말소리에 귀먹고 꽃다운 님의 얼굴에 눈멀었습니다."
extractNoun(sentence,  autoSpacing = TRUE)
```

```
## [1] "우리나라"       "유의"          "문화"          "들"
## [5] "멋진그래픽으로" "표현"          "봤"            "스르륵"
## [9] "시간"
```

```
extractNoun(sentence,  autoSpacing = FALSE)
```

```
## [1] "우리나라"       "고유"          "문화"          "들"
## [5] "멋진그래픽으로" "표현"          "스르륵"        "시간"
```

03-04 사전에 단어 추가하기

```
mergeUserDic 함수를 이용하여 sejong 사전에 '스르륵' 단어 '부사'로 추가
데이터 프레임 형태로 추가해야 함. 품사 태그셋 참고하여 추가한다.
품사 태그셋 : https://github.com/haven-jeon/KoNLP/blob/master/etcs/KoNLP-API.md
```

```
mergeUserDic(data.frame(c('스르륵'), c('mag')))
```

```
## Warning: 'mergeUserDic' is deprecated.
## Use 'buidDictionary()' instead.
## See help("Deprecated")
```

```
## 1 words were added to dic_user.txt.
```

03-03 사전에 단어 추가하기

```
명사중에 '스르륵'이 사라졌다.
```

```
extractNoun(sentence)
```

```
## [1] "우리나라"       "고유"          "문화"          "들"
## [5] "멋진그래픽으로" "표현"          "시간"
```

03-05 형태소 분석하기

```
MorphAnalyzer : 형태소를 분석해 주는 함수.
```

```
MorphAnalyzer(sentence)
```

```
## $우리나라
## [1] "우리나라/ncn"                  "우리/ncn+나라/ncn"
## [3] "우리/ncn+나/ncn+이/jp+라/ecs" "우리/ncn+나/ncn+이/jp+라/ef"
##
## $고유의
## [1] "고유/ncps+의/jcm" "고유/ncps+의/ncn"
##
## $문화들을
## [1] "문화/ncn+들/ncn+을/jco"   "문화/ncn+들/xsncc+을/jco"
##
## $멋진그래픽으로
## [1] "멋진그래픽/ncn+으로/jca" "멋진그래픽으/ncn+로/jca"
## [3] "멋진그래픽으로/ncn"      "멋진그래픽/nqq+으로/jca"
## [5] "멋진그래픽으/nqq+로/jca" "멋진그래픽으로/nqq"
##
## $잘
## [1] "잘/mag"        "자/pvg+ㄹ/etm"
##
## $표현했다
## [1] "표현/ncpa+하/xsva+었/ep+다/ef"
##
## $.
## [1] "./sf" "./sy"
##
## $재미있게
## [1] "재미있/pvg+게/ecc"       "재미있/pvg+게/ecs"
## [3] "재미있/pvg+게/ecx"       "재미있/pvg+게/ef"
## [5] "재미/ncn+있/xsmn+게/ecc" "재미/ncn+있/xsmn+게/ecs"
## [7] "재미/ncn+있/xsmn+게/ecx" "재미/ncn+있/xsmn+게/ef"
##
## $봤다
## [1] "보/pvg+아/ep+다/ef" "보/px+아/ep+다/ef"
##
## $.
## [1] "./sf" "./sy"
##
## $스르륵
## [1] "스르륵/mag"
##
```

```
## $시간이
## [1] "시간/ncn+이/jcs" "시간/ncn+이/jcc"
##
## $지나갔다
## [1] "지나/pvg+아/ecx+가/px+아/ep+다/ef" "지나가/pvg+아/ep+다/ef"
##
## $.
## [1] "./sf" "./sy"
```

(1) 분석은 띄어쓰기 단위, 즉 어절 단위로 분석이 되고 있다.
(2) 간단한 내용 일부
$고유의
[1] "고유/ncps+의/jcm" "고유/ncps+의/ncn"
의가 jcm(관형격 조사)
의가 ncn(비서술성 명사) 로 두 가지로 분석이 가능하다는 내용이다.

(3) 형태소 분석 결과
MorphAnalyzer라는 함수는 품사 확정보다는 형태소 분석 결과를 산출해 준다.

03-06 품사 달기

SimplePos09 : 09개의 품사 태그를 달아주는 함수
SimplePos22 : 22개의 품사 태그를 달아주는 함수

SimplePos09 : 품사의 상위 분류인 9개의 기준으로만 분류 수행
SimplePos22 : 그 다음 22개의 기준으로 분류를 수행.

```
SimplePos09(sentence)
```

```
## $우리나라
## [1] "우리나라/N"
##
## $고유의
## [1] "고유/N+의/J"
##
## $문화들을
## [1] "문화들/N+을/J"
##
## $멋진그래픽으로
## [1] "멋진그래픽으로/N"
##
## $잘
## [1] "잘/M"
##
## $표현했다
## [1] "표현/N+하/X+었다/E"
##
## $.
## [1] "./S"
##
## $재미있게
## [1] "재미있/P+게/E"
##
## $봤다
## [1] "보/P+아다/E"
##
## $.
## [1] "./S"
##
## $스르륵
## [1] "스르륵/M"
##
## $시간이
## [1] "시간/N+이/J"
##
## $지나갔다
## [1] "지나/P+아/E+가/P+아다/E"
##
```

```
## $.
## [1] "./S"
```

```
SimplePos22(sentence)
```

```
## $우리나라
## [1] "우리나라/NC"
##
## $고유의
## [1] "고유/NC+의/JC"
##
## $문화들을
## [1] "문화들/NC+을/JC"
##
## $멋진그래픽으로
## [1] "멋진그래픽으로/NC"
##
## $잘
## [1] "잘/MA"
##
## $표현했다
## [1] "표현/NC+하/XS+었/EP+다/EF"
##
## $.
## [1] "./SF"
##
## $재미있게
## [1] "재미있/PV+게/EC"
##
## $봤다
## [1] "보/PX+아/EP+다/EF"
##
## $.
## [1] "./SF"
##
## $스르륵
## [1] "스르륵/MA"
##
## $시간이
## [1] "시간/NC+이/JC"
##
## $지나갔다
## [1] "지나/PV+아/EC+가/PX+아/EP+다/EF"
##
```

```
## $.
## [1] "./SF"
```

```
$고유의
[1] "고유/N+의/J" (N: 체언, J:관계언)

$봤다
[1] "보/P+아다/E" (P:용언, E:어미)

-->
$고유의
[1] "고유/NC+의/JC" (NC: 보통명사, JC:격조사)

$봤다
[1] "보/PX+아/EP+다/EF" (PX: 보조용언, EP: 선어말어미, EF:종결어미)
```

03-07 명사 확인하기

```
library(stringr)
txt = SimplePos09(sentence)
txt_n <- str_match(txt, '([A-Z가-힣]+)/N') #명사확인
txt_n
```

```
##       [,1]               [,2]
##  [1,] "우리나라/N"       "우리나라"
##  [2,] "고유/N"           "고유"
##  [3,] "문화들/N"         "문화들"
##  [4,] "멋진그래픽으로/N" "멋진그래픽으로"
##  [5,] NA                 NA
##  [6,] "표현/N"           "표현"
##  [7,] NA                 NA
##  [8,] NA                 NA
##  [9,] NA                 NA
## [10,] NA                 NA
## [11,] NA                 NA
## [12,] "시간/N"           "시간"
## [13,] NA                 NA
## [14,] NA                 NA
```

실습 1. 주어진 아래 내용을 extractNoun과 형태소 분석을 통해 명사를추출해 보자.

```
v1 = '오늘은 좋은 날. 스스륵 하고 오늘 하루가 지나갔다.'
```

03-08 여러줄 텍스트 파일 읽기

```
movie_review <- readLines("./data/15_TheExtreme_min.txt", encoding="UTF-8")  # 한줄 한줄 읽어 각 줄이 벡터에서 하나의 원소가 된다.
```

```
## Warning in readLines("./data/15_TheExtreme_min.txt", encoding = "UTF-8"):
## './data/15_TheExtreme_min.txt'에서 불완전한 마지막 행이 발견되었습니다
```

```
movie_review
```

```
##  [1] "<U+FEFF>₩"x₩""
##  [2] "₩"1₩" ₩" 분노의 질주 시리즈중에서 제일 별루₩""
##  [3] "₩"2₩" ₩" 스케일 큰 시끄러운 액션이 난무하는데도 이렇게까지 지루할수 있다니.....₩""
##  [4] "₩"3₩" ₩" 시~원 하게 잘 본 영화. 다음 시리즈에서는 여자 주인공의 비중이 더 높아졌으면 하는 바램!₩""
##  [5] "₩"4₩" ₩" 반지닦기, 자살닦이, 고무닦이, 정의닦이...로 이어지는 한심한 DC 시리즈 "
##  [6] "레지던트 이블 시리즈 "
##  [7] "그리고 이 영화 분노의 질주 시리즈 "
##  [8] "공통점은 시리즈가 거듭될수록 돈은 많이 들지만 재미는 없어지고 "
##  [9] "CG는 떡질되지만 실감나는 장면은 더 없어지도 뻔히 가짜라는게 드러나는 영화들 "
## [10] "그러나"
## [11] "익스트림 익스트림"
## [12] "아무리 엉터리로 만들고, 자국에서 망해도 "
## [13] "미국 블록버스터라면 맹목적으로 보는 중국애들 땜에 "
## [14] "아무리 쓰레기 영화라도 본전 건지는 것은 물론 상당히 많은 돈을 버니... "
## [15] "이런 쓰레기들이 매년 양산된다. "
## [16] "물론, 중국애들도 할말은 있을 거다 "
## [17] "공산당이 검열하는 자국영화보다는 낫다고... "
## [18] "하지만 우리들은 다른 전세계의 재미있는 영화를 볼 선택의 자유가 있잖아! "
## [19] "왜 이런 쓰레기 영화를 보는 거지?₩""
## [20] "₩"5₩" ₩" ₩""
## [21] "₩"6₩" ₩" 그냥 액션만 보면 멋진데"
## [22] "스토리는 주인공이 전여친한테 싸지른"
## [23] "애새끼 구하러 간다며 아빠행세하면서"
## [24] "그 덕분에 지동료들 다 버리고 미쳐 날뛰는 내용₩""
## [25] "₩"7₩" ₩" ₩""
## [26] "₩"8₩" ₩" 아래는 다들 평점 알바들인가부네.. 이런 개 쓰레기 영화가 평점이 이리 높다니₩""
## [27] "₩"9₩" ₩" ₩""
## [28] "₩"10₩" ₩" ₩""
## [29] "₩"11₩" ₩" ₩""
## [30] "₩"12₩" ₩" ₩""
## [31] "₩"13₩" ₩" ₩""
## [32] "₩"14₩" ₩" ₩""
## [33] "₩"15₩" ₩" 스케일은 점점 더 커지지만, 액션은 멍청할정도로 어이가없음 과유불급₩""
## [34] "₩"16₩" ₩" 이 시리즈로 이렇게 길게 간다는게 신기.. 새로울 건 없지만 달리는 걸 좋아하시는 분이라면 ₩""
## [35] "₩"17₩" ₩" ₩""
## [36] "₩"18₩" ₩" ₩""
## [37] "₩"19₩" ₩" ₩""
## [38] "₩"20₩" ₩" 대머리들은 TV물로 찍고,"
## [39] "감독은 CG실에서 이어붙히고.₩""
```

실습 2. 주어진 txt 파일을 명사 추출을 해 보자.

data file : ohzun.txt

도전 1.

- 뉴스 키워드를 하나 골라 신문사 2개 이상의 사이트에 대해 웹 크롤링을 한 후에, 이에 대한 명사 분석을 수행해 보자.