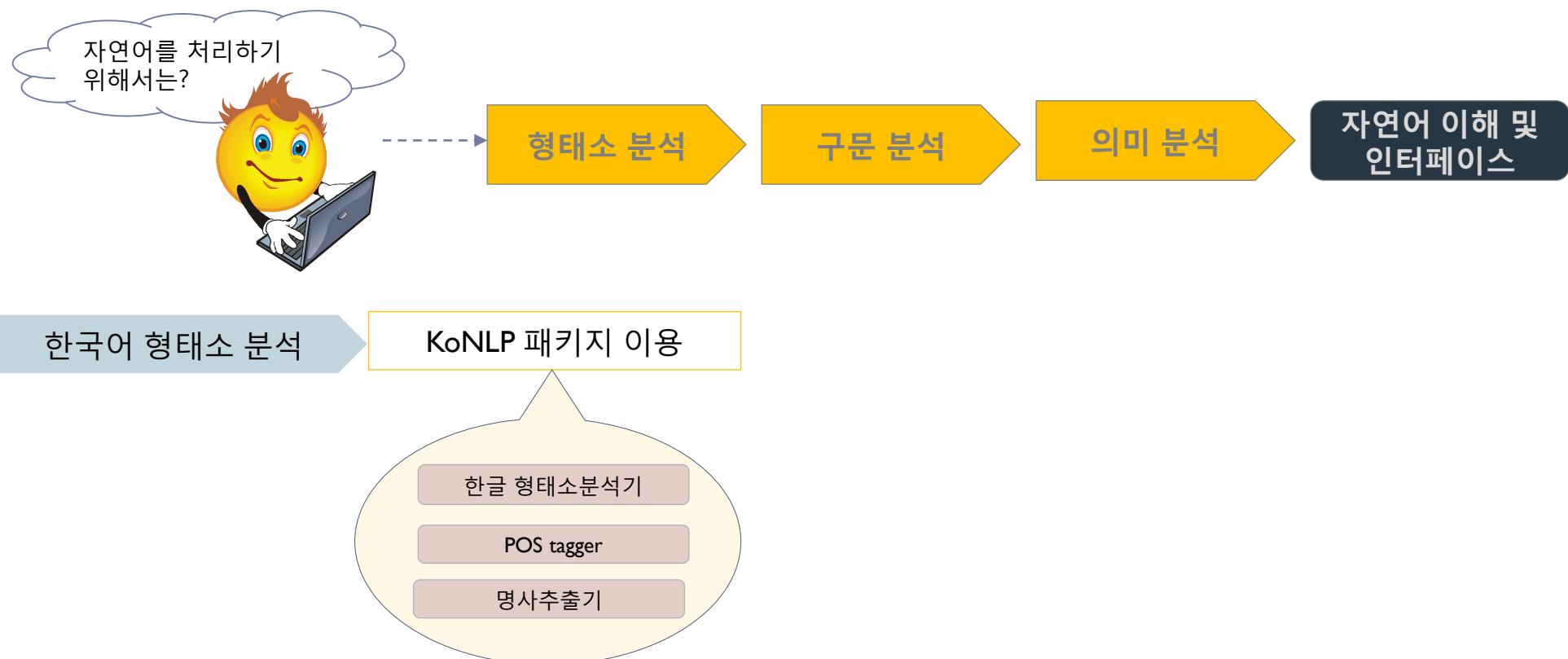


8. 데이터 전처리-3

한글 형태소 분석

I. 형태소 분석

- ◆ 정의 – 텍스트를 입력받아 그것을 형태소 단위로 분석하여 사전에서 해당 품사와 함께 출력하는 것을 의미
- ◆ 하나의 어절에서 분석에 유의미한 최소 단위의 형태소를 찾아내는 과정을 의미



- ◆ KoNLP에서는 형태소분석을 위한 사용자 사전을 제공 – 기본 SystemDic, SejongDic, NIADic

SystemDictionary를 이용

```
> Sys.setenv(JAVA_HOME="c:\\program files\\java\\jdk1.8.0_141")
> install.packages("KoNLP")
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---
URL 'http://cran.nexr.com/bin/windows/contrib/3.4/KoNLP_0.80.1.zip'을 시도합니다
Content type 'application/zip' length 5867634 bytes (5.6 MB)
downloaded 5.6 MB
```

패키지 'KoNLP'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다

다운로드된 바이너리 패키지를 다음의 위치에 있습니다

C:\Users\Administrator\AppData\Local\Temp\Rtmpq6M9gm\downloaded_packages

```
> library(KoNLP)
```

Checking user defined dictionary!

```
> useSystemDic()
```

Backup was just finished!

283949 words dictionary was built.

```
> tx_1<- "내 트친들과 깨방정, 페이스북으로 링크걸어서 총총 이동했다. 카톡도 프사를 어제 바꿨는$
```

```
> extractNoun(tx_1)
```

[1] "내"	"트친들과"	"깨방정"	"페이스"	"북"
[6] "링크걸어서"	"이동"	"카톡도"	"프사를"	"어제"
[11] "피곤"	"타"			

```
>
```

“트친”, “페이스북”, “프사” 와 같은
SNS 신조어가 사전에 등록되지 않아서
이들을 명사로 인식하지 못하는 문제가
발생했네요^^

SejongDictionary를 이용

```
> useSejongDic()
Backup was just finished!
370957 words dictionary was built.
```

```
> tx_1<- "내 트친들과 깨방정, 페이스북으로 링크걸어서 총총 이동했다. 카톡도 프사를 어제 바꿨는$
```

```
> extractNoun(tx_1)
```

[1] "내"	"트친들과"	"깨방정"	"페이스"	"북"	"링크걸어서"	"이동"
[8] "카톡도"	"프사를"	"어제"	"피곤"	"타"		

```
>
```

NIADictionary를 이용

```
> useNIADic()
Backup was just finished!
983012 words dictionary was built.
```

```
> tx_1<- "내 트친들과 깨방정, 페이스북으로 링크걸어서 총총 이동했다. 카톡도 프사를 어제 바꿨는$
```

```
> extractNoun(tx_1)
```

[1] "내"	"트친"	"들"	"깨방"	"정"	"페이스북"	"링크걸어서"
[8] "이동"	"카톡"	"프사"	"어제"	"피곤"	"타"	

```
>
```

“트친”, “페이스북”, “프사”, “카톡”과 같은
SNS 신조어가 모두 명사로 인식되는 것을 알 수 있어요.

형태소 태깅 작업

- ◆ 입력된 문자열을 각 품사별로 태깅한 결과를 출력 – 카이스트 태그 셋을 바탕으로 함
- ◆ SimplePos09(), SimplePos22() 함수 이용

		7. su	단위기호	8. sy	기타 기호
외국어(f)		9. f	외국어		
체언(n)	보통명사(nc)				
서술성 명사(ncp)	10. ncpa	동작성 명사	11. ncps	상태성 명사	
비서술성 명사(ncn)	12. ncn	비서술성 명사			
고유명사(nq)	13. nq	고유명사			
의존명사(nb)	14. nbu	단위성 의존 명사	15. nbn	비단위성 의존명사	
대명사(np)	16. npp	인칭대명사	17. npd	지시대명사	
수사(nn)	18. nnc	양수사	19. nno	서수사	
용언(p)	동사(pv)	20. pvd	지시동사	21. pvg	일반동사
혈용사(pa)	22. pad	지시혈용사	23. paa	성상혈용사	
보조용언(px)	24. px	보조용언			
수식언(m)	관형사(mm)	25. mmd	지시관형사	26. mma	성상관형사
부사(ma)	28. mad	지시부사	29. maj	접속부사	
	30. mag	일반부사			
독립언(l)	감탄사(li)	31. ii	감탄사		
관계언(j)	격조사(jc)	32. jcs	주격조사	33. jco	목격격조사
	34. jcc	보격조사	35. jcm	관형격조사	
	36. jcj	호격조사	37. jca	부사격조사	
	38. jcj	접속격조사	39. jct	공동격조사	
	40. jcr	인용격조사			
보조사(jx)	41. jxc	통용보조사	42. jxf	종결보조사	
서술격조사(jcp)	43. jcp	서술격조사			
어미(e)	선어말어(ep)	44. ep	선어말어미		
연결어미(ec)	45. ecc	대등적 연결어미	46. ecs	증속적 연결어미	
	47. ecx	보조적 연결어미			
전성어미(et)	48. etn	명사형어미	49. etm	관형사여어미	
증결어미(ef)	50. ef	증결어미			
접사(x)	접두사(xp)	51. xp	접두사		
접미사(xs)	52. xsn	명사파생접미사	53. xsv	동사파생접미사	
	54. xsm	형용·사파생접미사	55. xsa	부사파생접미사	

분석에서 가장 핵심이 되는 주제어 추출시 체언의 "명사" 가 가장 많이 해당됨!!

> SimplePos09(tx_1)

\$`내`
[1] "내/N"

\$트친들과
[1] "트친들/N+과/JC"

\$`깨방정`
[1] "깨방정/N+, /SP"

\$페이스북으로
[1] "페이스북/N+으로/JC"

\$링크걸어서
[1] "링크걸어서/N"

\$총총
[1] "총총/M"

\$이동했다
[1] "이동/N+하/X+었다/E"

\$.
[1] "./S"

\$카톡도
[1] "카톡/N+도/JX"

\$프사률
[1] "프사/NC+률/JC"

\$어제
[1] "어제/NC"

9개 품사태그로 분류하여 결과보여줌

> SimplePos22(tx_1)

\$`내`
[1] "내/NP"

\$트친들과
[1] "트친들/NC+과/JC"

\$`깨방정`
[1] "깨방정/NC+, /SP"

\$페이스북으로
[1] "페이스북/NC+으로/JC"

\$링크걸어서
[1] "링크걸어서/NC"

\$총총
[1] "총총/MA"

\$이동했다
[1] "이동/NC+하/X+었다/EP+다/EF"

\$.
[1] "./SF"

\$카톡도
[1] "카톡/NC+도/JX"

\$프사률
[1] "프사/NC+률/JC"

\$바꿨는데
[1] "바꾸/PV+었/EP+는데/EC"

22개 품사태그로 분류하여 결과보여줌

사전에 단어 추가하기

- ◆ 사전에 등록되지 않은 단어인 경우 형태소 분석의 결과가 달라질 수 있음.
- ◆ KoNLP에서는 사용자가 직접 사전에 단어를 등록할 수 있음 → 사용자 사전
- ◆ buildDictionary() 함수 이용

```
> t_1<- "스마트폰 엘지V30 화소 너무좋다. 갤럭시note8의 기능도 좋고 가상화폐 이더리움과 리플도 비트코인 못지않게 가격이 폭등했"
```

```
> extractNoun(t_1)
```

```
[1] "스마트폰" "엘지V30" "화" "너무" "좋" "갤럭시note8" "기능" "가상"  
[9] "화폐" "더리" "옴" "리플" "비트코인" "가격" "폭등" "며"
```

```
>
```

"화소", "이더리움" 등의 단어를 명사로 인식하지 못하는 문제가 있죠? 사전에 해당 단어가 등록되지 않아서 발생하는 문제예요.

```
> buildDictionary(ext_dic="NIADic", user_dic=data.frame(c("화소", "이더리움"), c("ncn")))
```

```
283955 words dictionary was built.
```

현재 NIADic에 화소/이더리움 단어를 명사로 등록하라는 의미

```
> extractNoun(t_1)
```

```
[1] "스마트폰" "엘지V30" "화소" "너무좋다" "갤럭시note8" "기능" "가상" "화폐"  
[9] "이더리움과" "리플도" "비트" "코" "가격" "폭등" "며"
```

```
>
```