교육 빅데이터의 이해와 활용(11주차)

흥도초등학교
교사 박 정

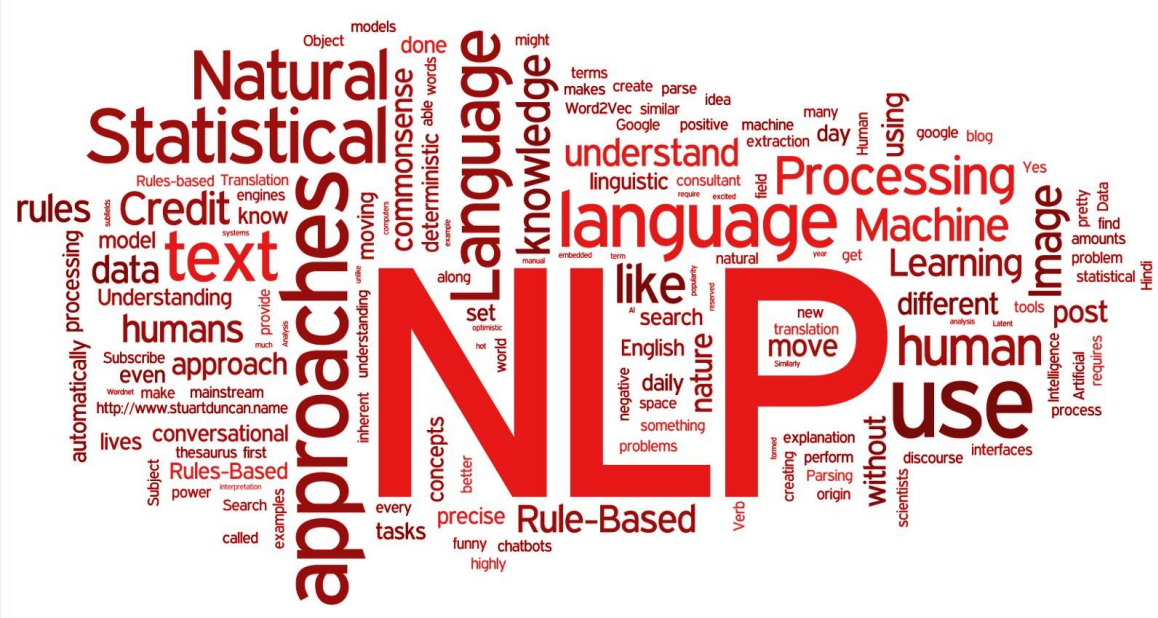
KoNLP 패키지 설치

Source on Save

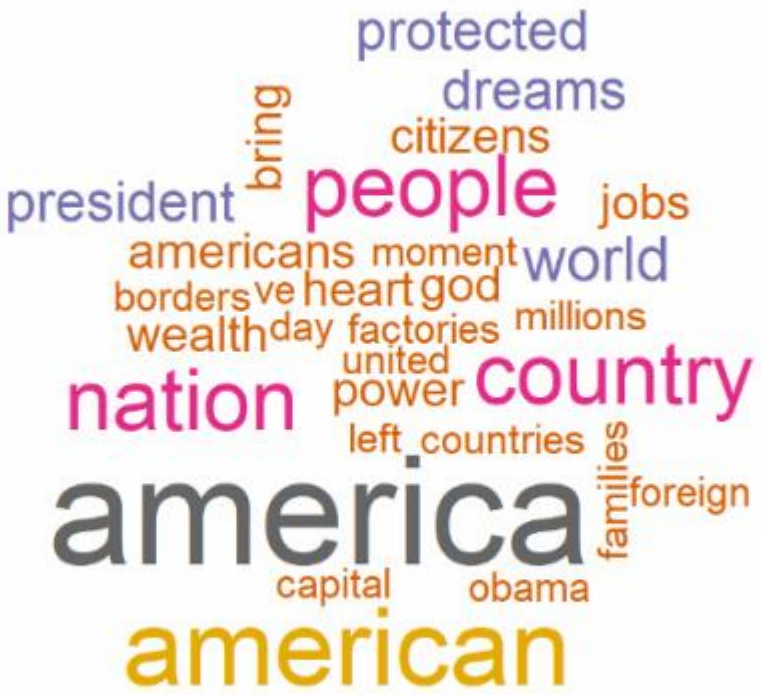
```
#### STEP0 ####
# KoNLP 설치
# https://www.facebook.com/notes/r-korea-krugkorean-r-user-group/konlp-%EC%84%A4%EC%B9%98-%EC%9D%B4%EC%8A%88-%EA%B3%B5%EC%9C%A0/1847510068715020/

install.packages("multilinguer")
library(multilinguer)

install_jdk()

install.packages(c('stringr', 'hash', 'tau', 'Sejong', 'RSQLite', 'devtools'), type = "binary")
install.packages("remotes")
remotes::install_github('haven-jeon/KoNLP', upgrade = "never", INSTALL_opts=c("--no-multiarch"))


library(KoNLP)


#### JAVA 설치 및 경로 지정 ####
#### STEP1 ####
# https://blog.naver.com/hss2864/220980568640
# 자바 다운로드 https://www.oracle.com/java/technologies/javase-jdk13-downloads.html

#### STEP2 ####
# http://blog.naver.com/PostView.nhn?blogId=hss2864&logNo=221378606062

#### STEP3 ####
# http://shorturl.at/rBCS7
```

NLP(Natural Language Processing: 자연어 처리)

워드클라우드



NLP(Natural Language Processing: 자연어 처리)



워드클라우드:

미국 트럼프 대통령 취임 연설

NLP(Natural Language Processing: 자연어 처리)

워드클라우드

○ 연수 만족도 조사 서술형 내용 분석 결과

〈표 12〉 연수 만족도 조사 결과

온라인 원격연수에서 가장 좋았던 점은?

온라인 연수 콘텐츠 가장 유용했던 내용은?

더 심화해서 배우고 싶은 내용은?

연수 프로그램 관련 기타 의견은?

NLP(Natural Language Processing: 자연어 처리)

언어 네트워크 분석

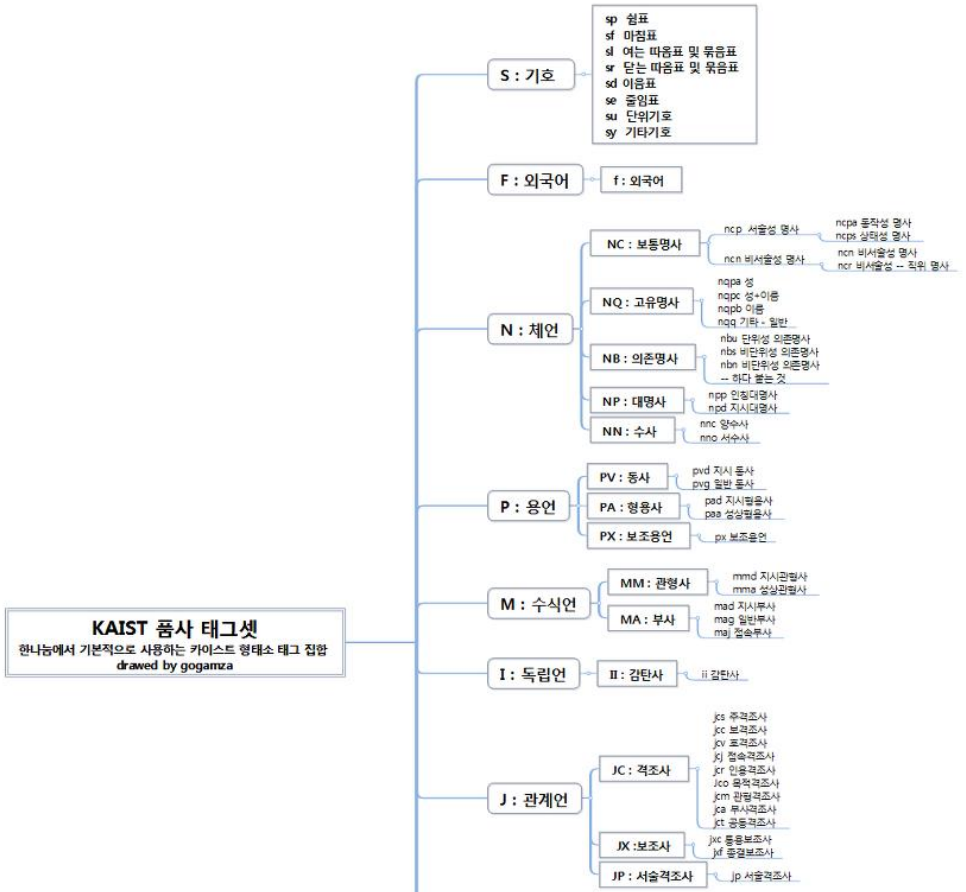
데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 공출현 행렬 연산 → 네트워크 시각화

불용어 처리

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → **형태소 분석** → 핵심 키워드 추출 및 공출현 행렬 연산 → 네트워크 시각화

https://lightblog.tistory.com/55



https://statkclee.github.io/text/nlp-bag-of-words.html

단어문서행렬(Term Document Matrix)을 전치(Transpose)하게 되면 문서단어행렬(DTM)이 된다. 단어문서행렬은 다음과 같은 형태를 갖는다.

	문서$_1$	문서$_1$	문서$_1$	$\cdots$	문서$_n$
단어$_1$	0	0	0	0	0
단어$_2$	1	1	0	0	0
단어$_3$	1	0	0	0	0
$\cdots$	0	0	2	1	1
단어$_m$	0	0	0	1	0

문서단어행렬은 단서문서행렬을 전치하여 다음과 같은 형태를 갖는다.

	단어$_1$	단어$_1$	단어$_1$	$\cdots$	단어$_n$
문서$_1$	0	1	1	0	0
문서$_2$	0	1	0	0	0
문서$_3$	0	0	0	2	0
$\cdots$	0	0	0	1	1
문서$_m$	0	0	0	1	0

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → **핵심 키워드 추출** 및 공출현 행렬 연산 → 네트워크 시각화

〈표 1〉 6차~2015 개정 고등학교 「진로와 직업」 교육과정 핵심 키워드 목록

6차		7차		2007 개정		2009 개정		2015 개정	
단어	빈도수	단어	빈도수	단어	빈도수	단어	빈도수	단어	빈도수
직업	44	직업	46	진로	42	진로	44	진로	133
진로	32	진로	28	직업	30	직업	32	직업	93
이해	17	자신	21	이해	21	이해	21	자신	74
세계	14	이해	17	자신	21	자신	21	학생	39
계획	14	학습	14	탐색	17	탐색	17	역량	36
산업	12	학생	13	학생	14	학생	14	탐색	35
발전	11	세계	12	결정	12	결정	13	이해	32
자신	11	계획	11	과목	10	학습	11	다양	25
특성	10	방법	9	방법	10	과목	10	변화	23
평가	9	능력	8	활용	10	방법	10	관련	21
선택	8	생활	8	활동	9	활용	10	고등학교	20
사회	8	과목	7	세계	8	활동	9	세계	20
방법	8	교과	7	준비	8	세계	8	선택	19
능력	7	태도	7	계획	8	준비	8	학습	19
학습	7	사회	7	특성	8	계획	8	결정	18
		행복	7	사회	8	특성	8	활동	17
				학습	8	사회	8	방법	17
				과정	8	과정	8	정보	17
								평가	17

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 **공출현 행렬 연산** → 네트워크 시각화

1) 전치 행렬 (Transposed Matrix)

전치 행렬 (Transposed Matrix)은 원래의 행렬에서 행과 열을 바꾼 행렬이다. 즉, 주대각선을 축으로 반사대칭을 하여 얻는 행렬이다. 기호는 기존 행렬 표현의 우측에 T를 붙인다.

$$M = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \xrightarrow{\text{Transpose}} M^{\mathrm{T}} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 5 & 11 & 17 \\ 11 & 25 & 39 \\ 17 & 39 & 61 \end{pmatrix}$$

▼ 세부 (행렬 곱셈)

행렬 곱셈 : 첫 번째 행렬의 행에 두 번째 행렬의 열을 곱합니다.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 1 \cdot 1 + 2 \cdot 2 & 1 \cdot 3 + 2 \cdot 4 & 1 \cdot 5 + 2 \cdot 6 \\ 3 \cdot 1 + 4 \cdot 2 & 3 \cdot 3 + 4 \cdot 4 & 3 \cdot 5 + 4 \cdot 6 \\ 5 \cdot 1 + 6 \cdot 2 & 5 \cdot 3 + 6 \cdot 4 & 5 \cdot 5 + 6 \cdot 6 \end{pmatrix} = \begin{pmatrix} 5 & 11 & 17 \\ 11 & 25 & 39 \\ 17 & 39 & 61 \end{pmatrix}$$

http://asq.kr/DclvXbdetiPbC

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 **공출현 행렬 연산** → 네트워크 시각화

단어문서행렬(Term Document Matrix)을 전치(Transpose)하게 되면 문서단어행렬(DTM)이 된다. 단어
를 갖는다.

	문서$_1$	문서$_1$	문서$_1$	$\cdots$
단어$_1$	0	0	0	0
단어$_2$	1	1	0	0
단어$_3$	1	0	0	0
$\cdots$	0	0	2	1
단어$_m$	0	0	0	1

문서단어행렬은 단서문서행렬을 전치하여 다음과 같은 형태를 갖는다.

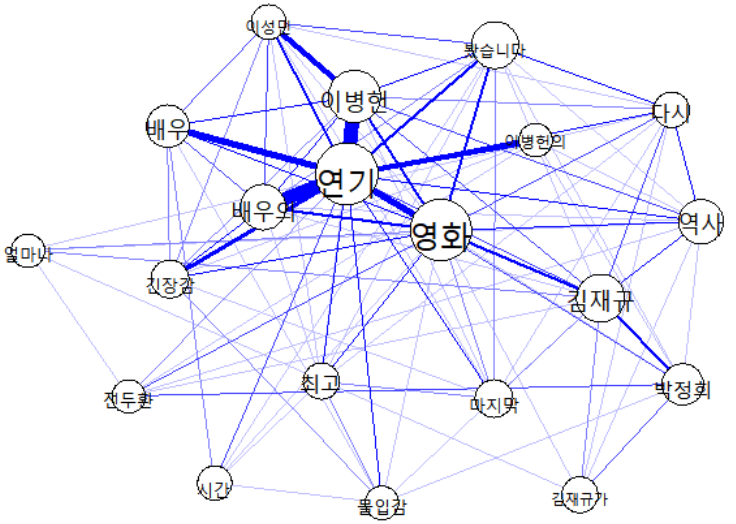
	단어$_1$	단어$_1$	단어$_1$	$\cdots$
문서$_1$	0	1	1	0
문서$_2$	0	1	0	0
문서$_3$	0	0	0	2
$\cdots$	0	0	0	1
문서$_m$	0	0	0	1

	보복조치	결의안	롯데마트	규탄	피해	경제	발의	관광객	기업	중단	대응	우려	중국인	대책	면세점	수술	타격	영향	대출	직격탄	WTO
보복조치	0	292	78	243	55	50	134	82	34	90	88	59	39	59	20	18	17	14	24	25	11
결의안	292	0	4	501	0	9	367	0	7	84	4	0	0	0	1	0	0	7	0	0	0
롯데마트	78	4	0	15	35	20	1	22	12	16	2	5	25	0	9	2	38	3	16	21	2
규탄	243	501	15	0	0	6	260	0	5	30	4	0	0	0	1	0	0	4	0	0	0
피해	55	0	35	0	0	31	0	17	35	7	18	12	4	15	13	27	2	5	0	5	0
경제	50	9	20	6	31	0	3	3	10	4	18	17	3	19	6	19	12	10	3	1	5
발의	134	367	1	260	0	3	0	0	1	41	4	0	0	0	0	0	0	0	3	0	0
관광객	82	0	22	0	17	3	0	0	1	7	54	11	60	56	46	1	29	4	22	11	0
기업	34	7	12	5	35	10	1	1	0	4	9	3	3	4	0	17	2	4	2	6	20
중단	90	84	16	30	7	4	41	7	4	0	12	5	1	10	2	1	2	1	1	2	2
대응	88	4	2	4	18	18	4	54	9	12	0	3	0	61	5	6	4	0	0	0	3
우려	59	0	5	0	12	17	0	11	3	5	3	0	5	0	19	4	1	5	12	3	2
중국인	39	0	25	0	4	3	0	60	3	1	0	5	0	9	37	0	17	3	15	6	0
대책	59	0	0	0	15	19	0	56	4	10	61	0	9	0	1	2	2	3	1	5	0
면세점	20	1	9	1	13	6	0	46	0	2	5	19	37	1	0	2	16	4	53	8	0
수술	18	0	2	0	27	19	0	1	17	1	6	4	0	2	2	0	7	17	1	6	0

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 **공출현 행렬 연산** → 네트워크 시각화

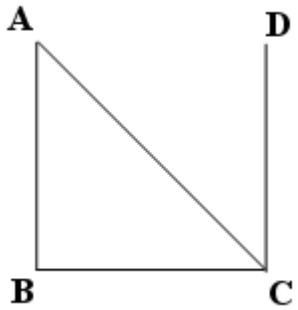
```
##          Terms
## Terms     연기 영화 이병헌 김재규 봤습니다 배우의 역사 배우 박정희 마지막
##   연기     174   14     29      3         7     29    5   15      3      4
##   영화      14  161      7      8         9      8    5    5      1      2
##   이병헌    29    7     76      3         5      4    3    5      0      1
##   김재규     3    8      3     54         1      1    4    0      7      2
##   봤습니다   7    9      5      1        49      2    1    0      1      2
##   배우의    29    8      4      1         2     44    1    3      0      0
##   역사       5    5      3      4         1      1   40    0      1      0
##   배우      15    5      5      0         0      3    0   34      0      0
##   박정희     3    1      0      7         1      0    1    0     31      0
##   마지막     4    2      1      2         2      0    0    0      0     23
```



NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 공출현 행렬 연산 → **네트워크 시각화**

https://mathbang.net/582



그래프를 표로 나타내기

	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	1	1	0	1
D	0	0	1	0

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 공출현 행렬 연산 → **네트워크 시각화**

https://mathbang.net/582

그래프를 표로 나타내기

	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	1	1	0	1
D	0	0	1	0

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

NLP(Natural Language Processing: 자연어 처리)

데이터 전처리 → 형태소 분석 → 핵심 키워드 추출 및 **공출현 행렬 연산** → 네트워크 시각화

```
##          Terms
## Terms     연기 영화 이병헌 김재규 봤습니다 배우의 역사 배우 박정희 마지막
##   연기     174   14     29      3         7     29    5   15      3      4
##   영화      14  161      7      8         9      8    5    5      1      2
##   이병헌    29    7     76      3         5      4    3    5      0      1
##   김재규     3    8      3     54         1      1    4    0      7      2
##   봤습니다   7    9      5      1        49      2    1    0      1      2
##   배우의    29    8      4      1         2     44    1    3      0      0
##   역사       5    5      3      4         1      1   40    0      1      0
##   배우      15    5      5      0         0      3    0   34      0      0
##   박정희     3    1      0      7         1      0    1    0     31      0
##   마지막     4    2      1      2         2      0    0    0      0     23
```