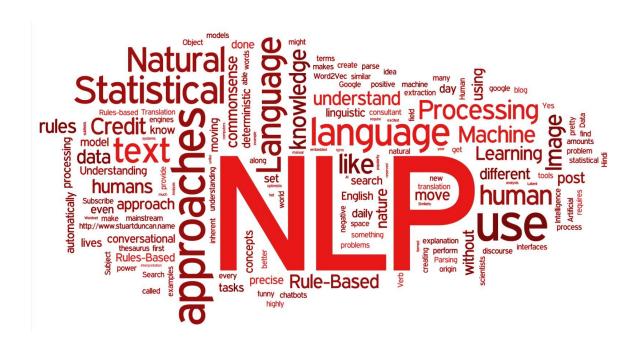
# 교육 빅데이터의 이해와 활용(11주차)

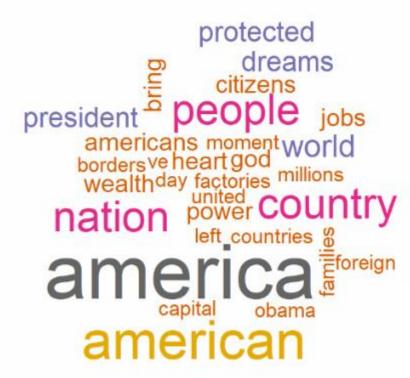
흥도초등학교 교사 박 정

#### KoNLP 패키지 설치

```
📠 📗 🖩 Source on Save 🔍 🏸 ▾ 📗
 2 * #### STEPO ####
    # KONLP 설치
    # https://www.facebook.com/notes/r-korea-krugkorean-r-user-group/kon]p-%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()
10
    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"))
13
15
    library(KoNLP)
19 * #### JAVA 설치 및 경로 지정 ####
20 - #### STEP1 ####
   # 자바 다운로드 https://www.oracle.com/java/technologies/javase-jdk13-downloads.html
24 v #### STEP2 ####
25 # http://blog.naver.com/PostView.nhn?blogId=hss2864&logNo=221378606062
    # http://shorturl.at/rBC57
29
```

#### 워드클라우드





워드클라우드:

미국 트럼프 대통령 취임 연설

#### 워드클라우드

연수 만족도 조사 서술형 내용 분석 결과
 〈표 12〉연수 만족도 조사 결과



#### 언어 네트워크 분석

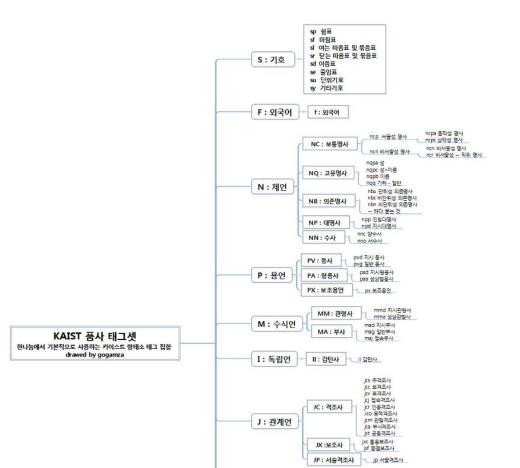
<u>데이터 전처리</u> → 형태소 분석 → 핵심 키워드 추출 및 공출현 행렬 연산 → 네트워크 시각화

불용어 처리

```
"1" "원더 우먼 1984
              별점 - 총 10점 중8
          영화 스토리텔링의 모든 요소가 다 들어가 있다.가진 자의 책임감, 없는 자의 열등감, 인간의 끝없는 욕망, 가족에 대한 사랑, 연인에 대한 사랑, 편법에 따르는 대가, 정치적 분열이 가
"2" "원더 우먼 1984
             별점 - 총 10점 중1
                                         잼없다 ㅜㅜ 3탄까지는 아닌것 같은데
          연출도 아니고 액션도 아니고 연기도 아니고
"3" "원더 우먼 1984
              별점 - 총 10점 중4
           회상씬 유치한건 어릴때라는 설정이라 넘어간다. 현세 등장씬에서 강도를 연기실화냐..가방들고 덤벙대다가 총을 ''실수로\"떨궈서강도인거 들키는거 실화냐... DC히어로증에 그나「
```

#### 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 $_1$	• • •	문서 $_n$
단어 $_1$	0	0	0	0	0
단어 $_2$	1	1	0	0	0
단어 $_3$	1	0	0	0	0
• • •	0	0	2	1	1
단어 $_m$	0	0	0	1	0

문서단어행렬은 단서문서행렬을 전치하여 다음과 같은 형태를 갖는다.

	단어 $_1$	단어 $_1$	단어 $_1$	• • •	단어 $_n$
문서 $_1$	0	1	1	0	0
문서 $_2$	0	1	0	0	0
문서 $_3$	0	0	0	2	0
	0	0	0	1	1
문서 $_m$	0	0	0	1	0

## 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

	7:	2].		12 mg		10 mg 2				
6차		7차		개정	2009	개정	2015 개정			
빈도수	단어	빈도수	단어	빈도수	단어	빈도수	단어	빈도수		
44	직업	46	진로	42	진로	44	진로	133		
32	진로	28	직업	30	직업	32	직업	93		
17	자신	21	이해	21	이해	21	자신	74		
14	이해	17	자신	21	자신	21	학생	39		
14	학습	14	탐색	17	탐색	17	역량	36		
12	학생	13	학생	14	학생	14	탐색	35		
11	세계	12	결정	12	결정	13	이해	32		
11	계획	11	과목	10	학습	11	다양	25		
10	방법	9	방법	10	과목	10	변화	23		
9	능력	8	활용	10	방법	10	관련	21		
8	생활	8	활동	9	활용	10	고등학교	20		
8	과목	7	세계	8	활동	9	세계	20		
8	교과	7	준비	8	세계	8	선택	19		
7	태도	7	계획	8	준비	8	학습	19		
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		
	44 32 17 14 14 12 11 10 9 8 8 8 7	44     직업       32     진로       17     자신       14     이해       14     학생       12     학생       11     세계       10     방법       9     능력       8     과목       8     교과       7     대도       7     사회	44     직업     46       32     진로     28       17     자신     21       14     이해     17       14     학생     13       11     세계     12       11     계획     11       10     방법     9       9     능력     8       8     광목     7       8     교과     7       7     대도     7       7     사회     7	44     직업     46     진로       32     진로     28     직업       17     자신     21     이해       14     이해     17     자신       14     학습     14     탐색       12     학생     13     학생       11     세계     12     결정       11     계획     11     과목       10     방법     9     방법       9     능력     8     활용       8     생활     8     활동       8     과목     7     세계       8     교과     7     준비       7     대도     7     사회       7     사회     7     특성       행복     7     사회     학습	44     직업     46     진로     42       32     진로     28     직업     30       17     자신     21     이해     21       14     이해     17     자신     21       14     학생     14     탐색     17       12     학생     13     학생     14       11     세계     12     결정     12       11     계획     11     과목     10       10     방법     9     방법     10       9     능력     8     활용     10       8     생활     8     활동     9       8     과목     7     세계     8       8     교과     7     준비     8       7     사회     7     특성     8       행복     7     사회     8       학습     8     학습     8	44     직업     46     진로     42     진로       32     진로     28     직업     30     직업       17     자신     21     이해     21     이해       14     이해     17     자신     21     자신       14     학습     14     탐색     17     탐색       12     학생     13     학생     14     학생       11     세계     12     결정     12     결정       11     계획     11     과목     10     학습       10     방법     9     방법     10     과목       9     능력     8     활용     10     방법       8     생활     8     활동     9     활용       8     과목     7     세계     8     활동       8     교과     7     준비     8     세계       7     사회     7     사회     8     특성       행복     7     사회     8     두성       학습     8     사회	44     직업     46     진로     42     진로     44       32     진로     28     직업     30     직업     32       17     자신     21     이해     21     이해     21       14     이해     17     자신     21     자신     21       14     학습     14     탐색     17     탐색     17       12     학생     13     학생     14     학생     14       11     세계     12     결정     12     결정     13       11     계획     11     과목     10     학습     11       10     방법     9     방법     10     과목     10       9     능력     8     활용     10     방법     10       8     생활     8     활동     9     활용     10       8     과목     7     세계     8     활동     9       8     교과     7     준비     8     세계     8       7     사회     7     두성     8     계획     8       행복     7     사회     8     특성     8       학생     10     10     10     10     10     10       8     과목     7<	44     직업     46     진로     42     진로     44     진로       32     진로     28     직업     30     직업     32     직업       17     자신     21     이해     21     아신       14     이해     17     자신     21     자신     21     학생       14     학습     14     탐색     17     탐색     17     역량       12     학생     13     학생     14     학생     14     탐색       11     세계     12     결정     12     결정     13     이해       11     계획     11     과목     10     학습     11     다양       10     방법     9     방법     10     과목     10     변화       9     능력     8     활용     10     방법     10     관련       8     생활     8     활동     9     활용     10     고등학교       8     과목     7     세계     8     활동     9     세계       8     교과     7     준비     8     세계     8     전ط       7     사회     7     사회     8     활동       9     학업     8     사회     8     발생       <		

#### 데이터 전처리 -> 형태소 분석 -> 핵심 키워드 추출 및 <mark>공출현 행렬 연산</mark> -> 네트워크 시각화

#### 1) 전치 행렬(Transposed Matrix)

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

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$
Transpose 
$$\begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

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

http://asq.kr/DclvXbdetiPbC

## 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. 단어를 갖는다.

	문서 $_1$	문서 $_1$	문서 $_1$	• • •
단어 $_1$	0	0	0	0
단어 <sub>1</sub> 단어 <sub>2</sub> 단어 <sub>3</sub>	1	1	0	0
단어 $_3$	1	0	0	0
•••	0	0	2	1
단어 $_m$	0	0	0	1

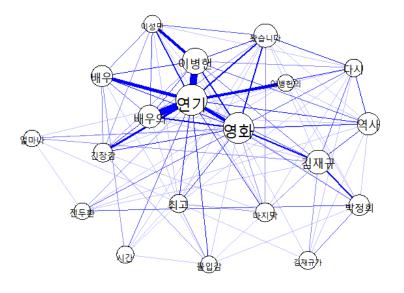
문서단어행렬은 단서문서행렬을 전치하여 다음과 같은 형태를 갖는다.

	단어 $_1$	단어 $_1$	단어 $_1$	• • •
문서 <sub>1</sub>	0	1	1	0
문서 <sub>1</sub> 문서 <sub>2</sub>	0	1	0	0
문서 <sub>3</sub>	0	0	0	2
• • •	0	0	0	1
문서 $_m$	0	0	0	1

	보복조치	행이안	롯 데 마 트	구한	희	경제	\$6 OT	संग्रा तक रहा	기 업	10 O	당	유	항자인	액	면 세 점	수 출	함전	go aa	어 숲	20 St St	WTO
보복 조치	0	292	78	243	55	50	134	82	34	90	88	59	39	59	20	18	17	14	24	25	11
결의 안	292	0	4	501	0	9	367	0	7	84	4	0	0	0	1	0	0	7	0	0	0
롯데 마트	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
II	55	0	35	0	0	31	0	17	35	7	18	12	4	15	13	27	2	5	0	5	0
경계	50	9	20	6	31	0	3	3	10	4	18	17	3	19	6	19	12	10	3	1	5
딸의	134	367	1	260	0	3	0	0	1	41	4	0	0	0	0	0	0	3	0	0	0
근광 객	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
충단	90	84	16	30	7	4	41	7	4	0	12	5	1	10	2	1	2	1	1	2	2
다용	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
중국 인	39	0	25	0	4	3	0	60	3	1	0	5	0	9	37	0	17	3	15	6	0
다박	59	0	0	0	15	19	0	56	4	10	61	0	9	0	1	2	2	3	1	5	0
면서 점	20	1	9	1	13	6	0	46	0	2	5	19	37	1	0	2	16	4	53	8	0
수송	18	0	2	0	27	19	0	1	17	1	6	4	0	2	2	0	7	17	1	6	0

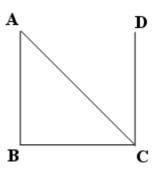
데이터 전처리 > 형태소 분석 > 핵심 키워드 추출 및 공출현 행렬 연산 > 네트워크 시각화

##		Terms									
##	Terms	연기	영화	이병헌	김재규	봤습니다 년	배우의 의	4사 배	우 박정	형희 마지역	<u> </u>
##	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	e	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



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

https://mathbang.net/582



그래프를 표로 나타내기

	А	В	С	D
А	0	1	1	0
В	1	0	1	0
С	1	1	0	1
D	0	0	1	0

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

#### https://mathbang.net/582

그래프를 표로 나타내기

	А	В	С	D
А	0	1	1	0
В	1	0	1	0
С	1	1	0	1
D	0	0	1	0

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

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

##		Terms									
##	Terms	연기	영화	이병헌	김재규	봤습니다 년	배우의 의	4사 배	우 박정	형희 마지역	<u> </u>
##	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	e	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

