Python

R studio

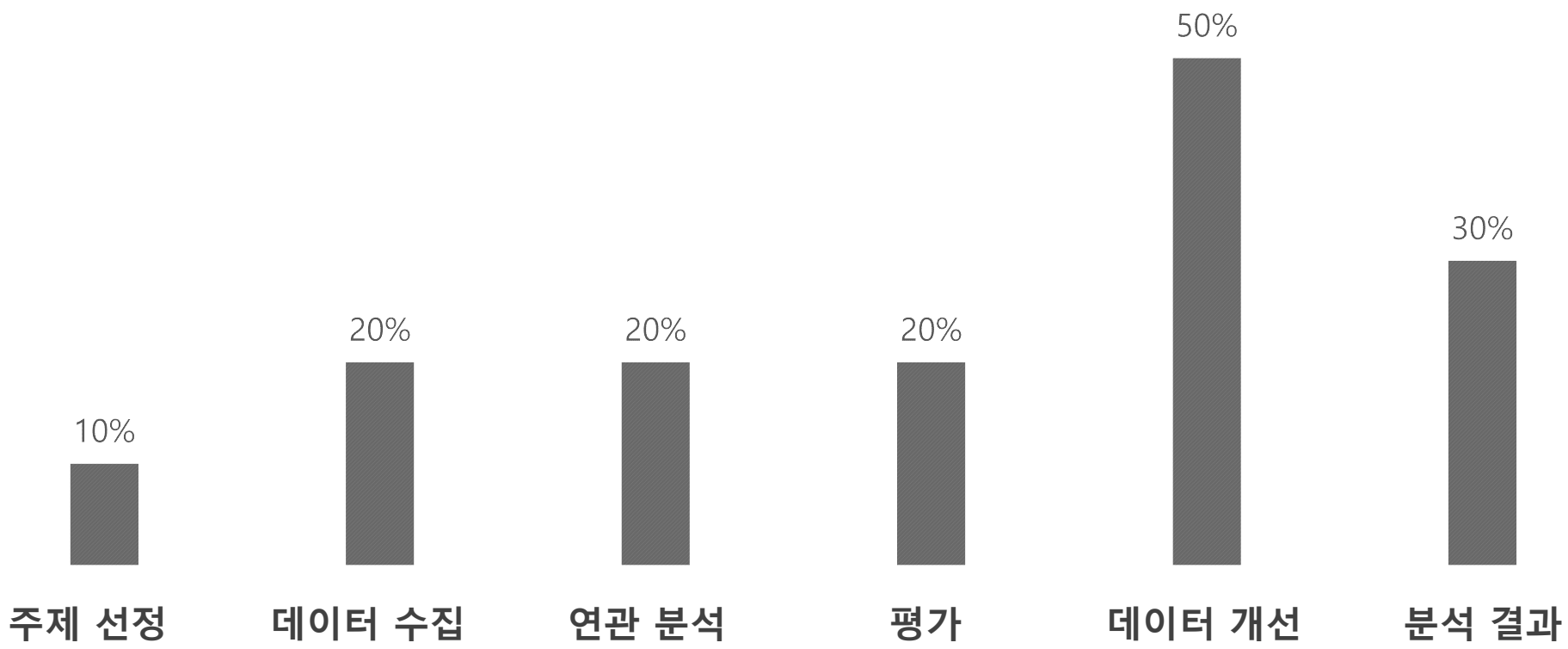
Corono_article.csv

Apriori

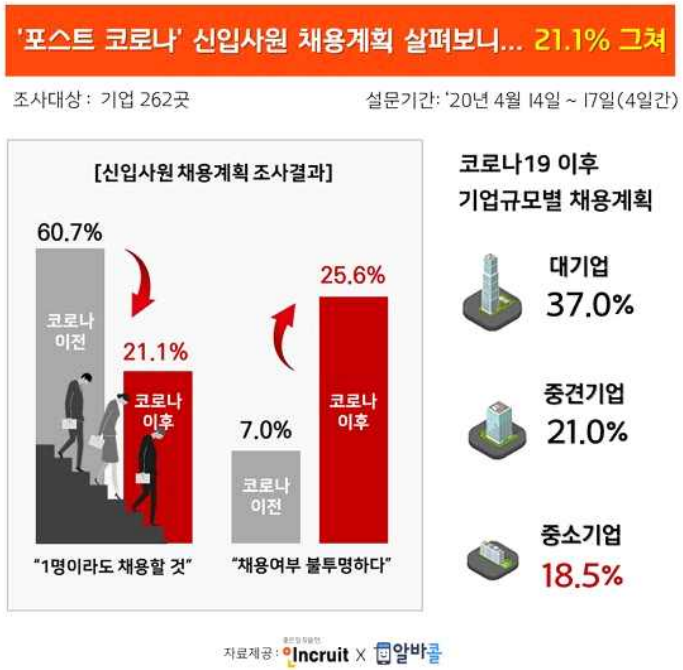
1

Corona 취업 기사 연관 분석

아이티윌 11기 이승혁


10%
20%
20%
20%
50%
30%
주제 선정
데이터 수집
연관 분석
평가
데이터 개선
분석 결과

주제 선정



실업자 ↑

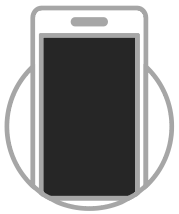
구직자 ↑

채용 계획 ↓

구직 준비 ? ---> 이후 동향 파악 !

데이터 수집

How ?	Keyword ?	Where ?
Web article Crolling	Corona & Job	Naver
Python / R 프로그램 웹 기사를 크롤링을 통해 데이터를 수집한다.	핵심 포인트는 코로나 이후 취업, 구직, 채용	국내 취업 관련 기사 -> 국내 포털 사이트 우리나라 포털 사이트에서 기사를 수집한다.



How ?

Python : BeautifulSoup, Okt

R : KoNLP

Keyword ?

'코로나 이후 취업'

키워드 검색 , 뉴스 카테고리 URL

Where ?

신문사별 검색이 아닌,

< 네이버 뉴스 > 를 통한 통합적 검색

Source Code(Python)

page url

page keyword

기사 내용 url

Source Code(Python)

```python
def naver_scroll(page):
    from konlpy.tag import Okt
    import random
    import urllib.request
    from bs4 import BeautifulSoup
    from newspaper import Article
    okt=Okt()
    f=open('d:\\data\\Project\\articlenoun.txt','w',encoding='utf-8')
    cnt=1
    url_list=[]
    for i in range(page+1):
        list_url="https://search.naver.com/search.naver?&where=news&query=%EC%8F%AC%EC%8A%A4%EC%8B/
        url=urllib.request.Request(list_url)
        result=urllib.request.urlopen(url).read().decode("utf-8")
        soup=BeautifulSoup(result,"html.parser")
        result1=soup.find_all('dd',class_='txt_inline')
        for j in result1:
            for k in j:
                if 'href' in str(k):
                    if k.get('href')!=None:
                        url_list.append(k.get('href'))
    url_list=list(set(url_list))

    for texts in url_list:
        a = Article(texts, language='ko')
        a.download()
        a.parse()
        f.write(str(cnt))
        for noun in a.text:
            f.write(','+noun)
        f.write('\n')
        cnt+=1
    f.close()
```

articlenoun.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

1,무안,배상현,진도,비치,리조트,내일,힐링캠프,Healing,Camp,영화배우,박진주,참여,취
2,오마이,실시,오마이,현재,가입자,의무,이유,실업,급여,취업,단계,추진,확대,오마이,문재
3,권현수,한민족,여름,과정,원격,프로그램,국제,사업,공주대,프로그램,한민족,이번,프로
4,기획재정부,지원,접수,접수,대화,기재부,진철,시장,얼마나,충격,통계,경기,위축,은행,기
5,확대,국가,역할,시사,IN,시사,IN,시사,IN,시사,IN,취업,현실,안주,실제,과제,사람,취업,시
6,문재인,김정현,이후,문재인,다음주,비상,회의,청와대,회의,회의,청와대,다음주,논의,지
7,민주당,국난,극복,비상,대책,정부,극복,회안,전망,확충,민주당,국난,복위,대변인,오늘,긴
8,문재인,청와대,춘추관,취임,주년,대국민,특별,취재,진의,질문,문재인,청와대,춘추관,취
9,부산,LINC,AI,취업,컨설팅,개발,운영,업무,협력,협약,체결,부산,LINC,AI,취업,컨설팅,개
10,김경환,정책,사회,부장,여유,커피,여유,카페,아르바이트생,카페,사람,아르바이트생,카
11,서미영,실전,필기,삼성,적성검사,GSAT,필기,대해,실전,감각,구직,자의,준비,필기,실전
12,게임,재단,서울특별시,교육청,직업,재단,직업,대회,대회,행사,직업,학생,직업,서울특별
13,사회,가치,혁신,실장,주목,마산,여성,인력,관장,취업,취약,계층,체결,포즈,취하,제공,시
14,언론사,언론사,청와대,주재,단계,한파,눈앞,제시,문재인,청와대,주재,청와대,출자,보충
대란,차단,면서,안전,촘촘하,첫째,수당,보전,원금,지속,무급,대상,무급,신속,프로그램,항
15,유은혜,사회,부총리,교육부,서울,종로구,정부,서울,청사,회관,회의,발언,호남,munona
16,정부,민간,투자,투자,부총리,정부,경제,민간,투자,관련,투자,부총리,지난달,고용,관련,
17,서울,박영태,문재인,청와대,본관,참석,발언,서울,박영태,문재인,청와대,본관,참석,발언
18,이재,서울,삼일대로,지방,노동청,노동,위기,대응,태스크포스,TF,회의,참석,발언,제공,
19,경제,임혜선,이선,이전,준비,발생,이후,영업,생태계,유통,시장,오프라인,온라인,커머스
20,지난달,서울,여의도,국회,실업,부조,도입,확대,촉구,기자회견,참석자,구호,외치,지난달
21,문재인,유승,문재인,유승,서울,문재인,주재,시작,생활,거리,두기,생활,방역,청와대,청와
22,청와대,주년,특별,적용,취업,단계,강조,대상,보험료,보험료,정은,대상,취업,입장,앞서,
23,전남,교육청,제공,전남,교육청,제공,전라남도교육청,취업,교육,정책,설명,포스트,교육

1227개의 기사, 약 80만개의 단어 추출

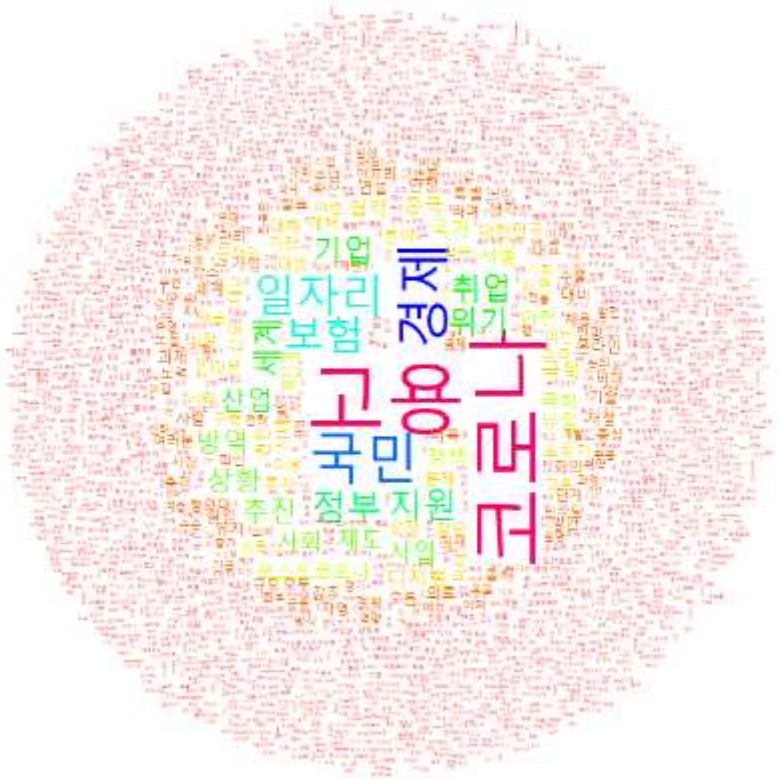
Source Code(R)

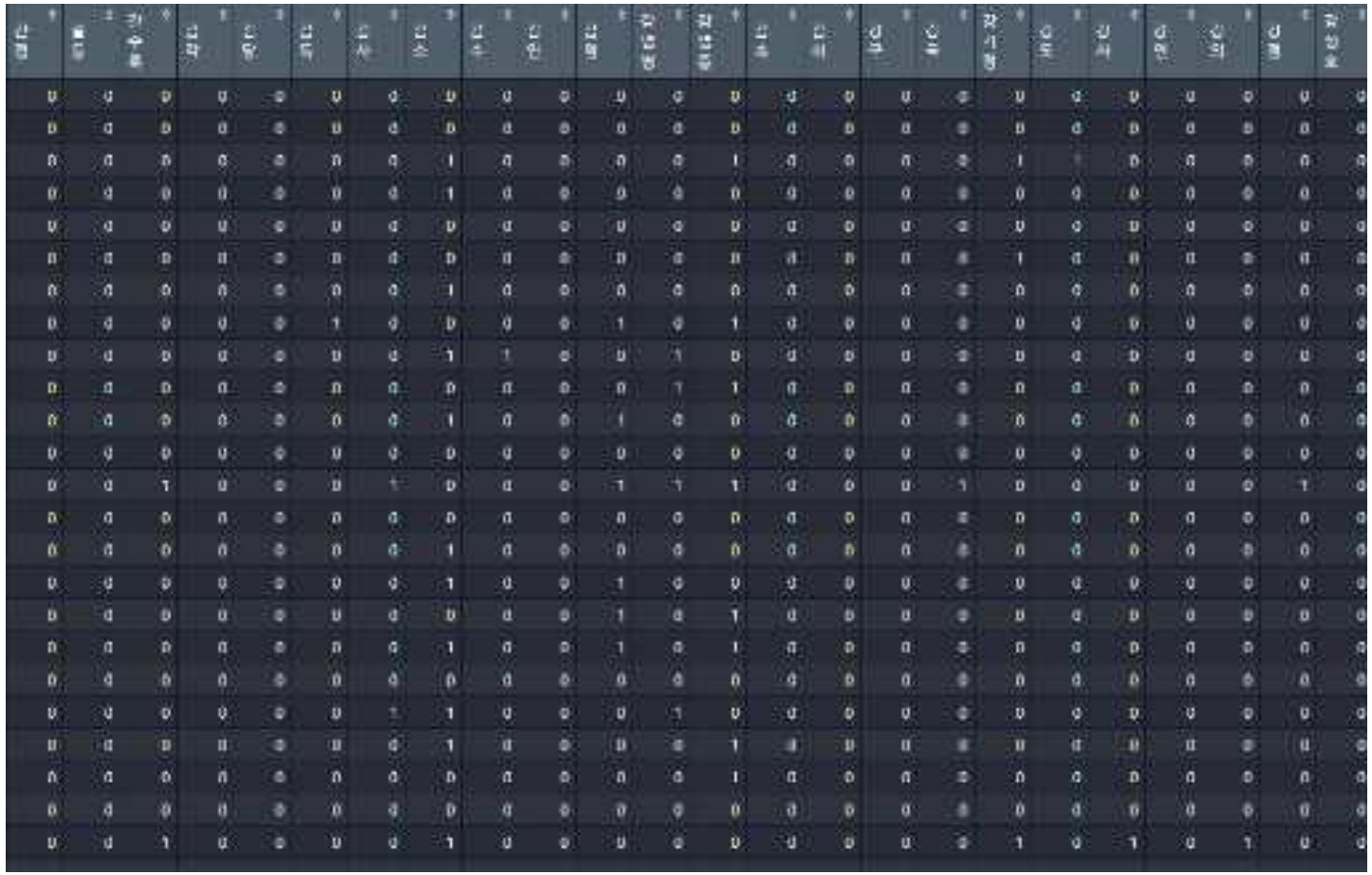
```
a<-read.csv('d:\\data\\Project\\articlenoun.csv',header=F)
# 칼럼 찾기
l<-c()
for(i in 1:dim(a)[1]){
  for (j in 1:dim(a)[2]){
    if(a[i,j]!=''){
      if(nchar(a[i,j])>=2){l<-append(l,a[i,j])}
    }else{
      break
    }
  }
}
tab<-sort(table(l),decreasing=T)
tab<-tab[tab>=3]
tab<-tab[names(tab)!='']

wordcloud(tab,freq=tab, min.freq=10, scale=c(5,0.2), rot.per=0.1,
          random.order = F,random.color=F,col=rainbow(15))
```

```
# 칼럼명 지정
col_list<-names(tab)
# 데이터 프레임 생성 위해 칼럼 갯수만큼 0리스트 생성
rr<-rep(0,length(col_list))
# 데이터 프레임 생성
df<-data.frame(rr)
# 전치 및 데이터 프레임으로 설정
df<-t(df)
df<-as.data.frame(df)
# 데이터 프레임 칼럼명 지정
names(df)<-col_list
target<-c()
for (i in 1:dim(a)[1])
{
  if (!is.na(as.numeric(a[i,1])) & length(target)!=0){ # 첫 데이터가 숫자고, 타겟 리스트의 갯수가 0보다 크면
    al<-c() # append list                              # 갯수가 0보다 크다( 처음이 아니다 ). 즉, 새로운 기사의 시작
    #print(a[i,1])
    #print(al)
    for (k in sort(col_list)){
      #print(k)
      if (k %in% target){
        al<-append(al,1)
      }else{al<-append(al,0)}
    }
    df<-rbind(df,al)
    #print(df)
    target<-c()
  }
  # i 번째 기사 단어들을 target에 담음
  for (j in 1:dim(a)[2])
  {
    if (a[i,j]!=''){
      target<-append(target,a[i,j])
    }else{
      break
    }
  }
  target<-unique(target)
}
df<-df[-1,]
```

Source Code(R)



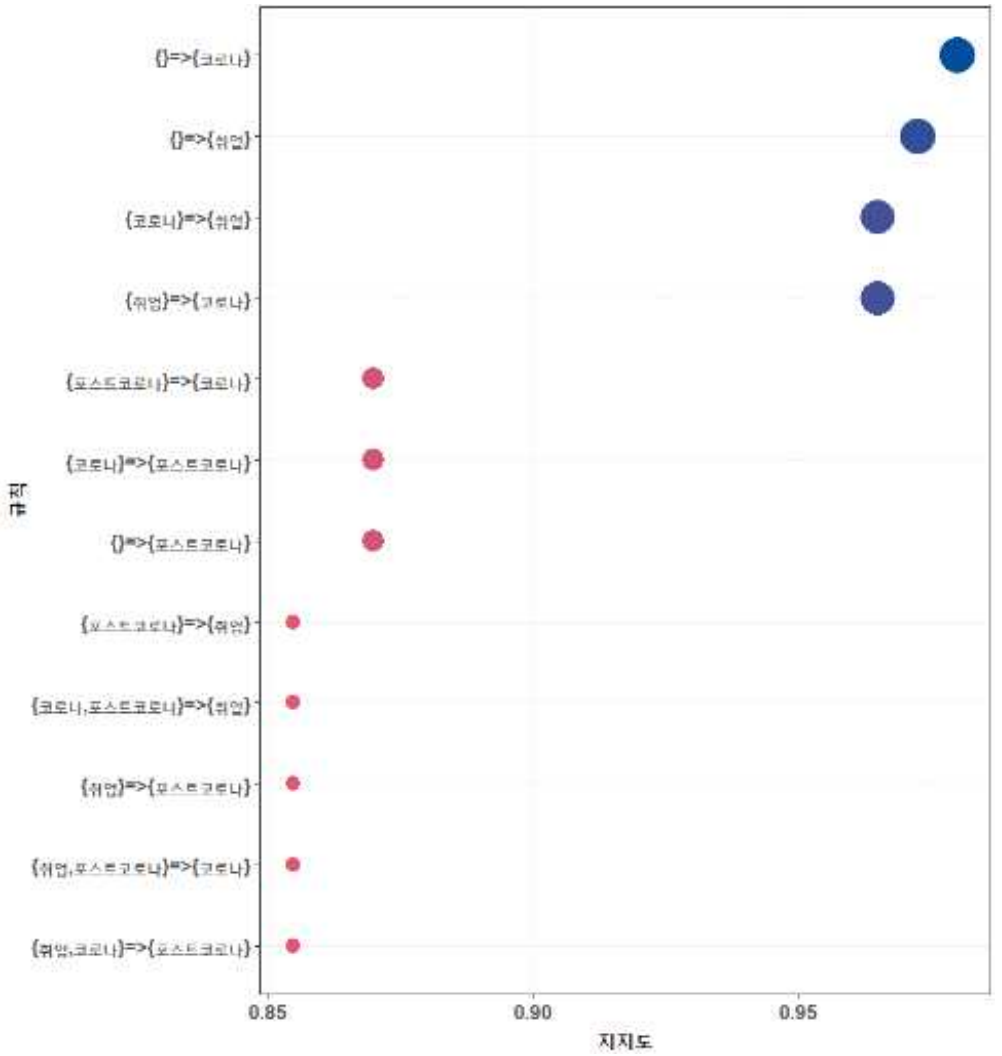


너무 많은 불용어

연관 분석

```
trans<-as.matrix(df,"transaction")
article_rules <- apriori(trans, parameter=list(supp=0.5, conf=0.6,target="rules") )
inspect_article<-inspect(sort(article_rules,decreasing = T))
inspect_article
```

```
> inspect_article
     lhs                          rhs    support confidence  coverage     lift count
[1]  {}                        => {코로나} 0.9799499  0.9799499 1.0000000 1.000000   391
[2]  {}                        => {취업}   0.9724311  0.9724311 1.0000000 1.000000   388
[3]  {취업}                    => {코로나} 0.9649123  0.9922680 0.9724311 1.012570   385
[4]  {코로나}                  => {취업}   0.9649123  0.9846547 0.9799499 1.012570   385
[5]  {}                        => {포스트코로나} 0.8696742  0.8696742 1.0000000 1.000000   347
[6]  {포스트코로나}            => {코로나} 0.8696742  1.0000000 0.8696742 1.020460   347
[7]  {코로나}                  => {포스트코로나} 0.8696742  0.8874680 0.9799499 1.020460   347
[8]  {포스트코로나}            => {취업}   0.8546366  0.9827089 0.8696742 1.010569   341
[9]  {취업}                    => {포스트코로나} 0.8546366  0.8788660 0.9724311 1.010569   341
[10] {취업,포스트코로나}       => {코로나} 0.8546366  1.0000000 0.8546366 1.020460   341
[11] {코로나,포스트코로나}     => {취업}   0.8546366  0.9827089 0.8696742 1.010569   341
[12] {취업,코로나}             => {포스트코로나} 0.8546366  0.8857143 0.9649123 1.018444   341
[13] {}                        => {경제}   0.6691729  0.6691729 1.0000000 1.000000   267
[14] {경제}                    => {코로나} 0.6666667  0.9962547 0.6691729 1.016638   266
[15] {코로나}                  => {경제}   0.6666667  0.6803069 0.9799499 1.016638   266
[16] {경제}                    => {취업}   0.6641604  0.9925094 0.6691729 1.020648   265
[17] {취업}                    => {경제}   0.6641604  0.6829897 0.9724311 1.020648   265
[18] {경제,취업}               => {코로나} 0.6616541  0.9962264 0.6641604 1.016610   264
[19] {경제,코로나}             => {취업}   0.6616541  0.9924812 0.6666667 1.020619   264
[20] {취업,코로나}             => {경제}   0.6616541  0.6857143 0.9649123 1.024719   264
[21] {}                        => {정부}   0.6365915  0.6365915 1.0000000 1.000000   254
[22] {}                        => {고등}   0.6340852  0.6340852 1.0000000 1.000000   253
[23] {고등}                    => {코로나} 0.6340852  1.0000000 0.6340852 1.020460   253
[24] {코로나}                  => {고등}   0.6340852  0.6470588 0.9799499 1.020460   253
[25] {정부}                    => {코로나} 0.6315789  0.9921260 0.6365915 1.012425   252
[26] {코로나}                  => {정부}   0.6315789  0.6445013 0.9799499 1.012425   252
[27] {}                        => {지원}   0.6290727  0.6290727 1.0000000 1.000000   251
[28] {정부}                    => {취업}   0.6265664  0.9842520 0.6365915 1.012156   250
[29] {취업}                    => {정부}   0.6265664  0.6443299 0.9724311 1.012156   250
[30] {지원}                    => {취업}   0.6240602  0.9920319 0.6290727 1.020156   249
[31] {취업}                    => {지원}   0.6240602  0.6417526 0.9724311 1.020156   249
[32] {고등}                    => {취업}   0.6240602  0.9841897 0.6340852 1.012092   249
[33] {취업}                    => {고등}   0.6240602  0.6417526 0.9724311 1.012092   249
[34] {고등,취업}               => {코로나} 0.6240602  1.0000000 0.6240602 1.020460   249
[35] {고등,코로나}             => {취업}   0.6240602  0.9841897 0.6340852 1.012092   249
[36] {취업,코로나}             => {고등}   0.6240602  0.6467532 0.9649123 1.019978   249
[37] {}                        => {일자리} 0.6215539  0.6215539 1.0000000 1.000000   248
[38] {지원}                    => {코로나} 0.6215539  0.9880478 0.6290727 1.008264   248
[39] {코로나}                  => {지원}   0.6215539  0.6342711 0.9799499 1.008264   248
[40] {정부,취업}               => {코로나} 0.6215539  0.9920000 0.6265664 1.012297   248
[41] {정부,코로나}             => {취업}   0.6215539  0.9841270 0.6315789 1.012027   248
```



의미 있는 결과를 찾기 힘들다.

1. 특정 단어 편향

코로나 , 취업 , 포스트코로나, 고용 등 특정 단어의 빈도가 너무 높다.
데이터를 정제한다.

2. 특정 기사의 연관 분석에서 지지도의 의미 ?

기사에서 데이터 수집시 무작위가 아닌 **특정 단어에 대한 조건**을 부여한 상태로 수집한다.
지지도 보다는 신뢰도 , 향상도를 기준으로 시각화 및 분석을 수행한다.

3. 데이터의 개수 ?

연관 분석은 모델을 훈련시키는 것이 아니다.
데이터의 개수 보다는 데이터 속에서 연관성을 잘 찾는것이 중요하다.

4. 개선 방향 ?

특정 단어는 포함을 전제하여 분석시 제거한다.
특정 주제 → 다양한 내용의 기사 수집 → 지지도 보다는 신뢰도, 향상도 유의하며 분석을 수행
해당 주제에서 너무 많은 데이터는 불필요 단어, 관심 없는 연관 규칙이 생길 수 있다.

데이터 개선

☞ 제거 리스트 생성

```
sublist=['기사','두단','배포','뉴시스','중앙','일보','기자','대통령','일보','금지','아랫목','부울경','이
         '파이낸셜뉴스','연합뉴스','뉴스','동아일보','가금류','강서','강서구','강미경','강민석',
        '사진','씩션','지난','지금','오후','오전','왼쪽','YBLN','CJ','OK','이사장','프레','테이','빗장'
        '인크루트','한국','통해','위해','우리','가운데','가지','경우','통해','위해','대한','네이버','노컷
         '뉴스레터','뉴스스탠드','구독','바로가기','장관','충북','오른쪽','코리아','포털','News',\
        'SBS','click','강영','결과','애문','국민','교수','관리','대표','더욱','라며','도두',\
         '모든','마련','부분','사실','생각','연설','코로나','포스트 코로나','포스트코로나','데일리안']
```

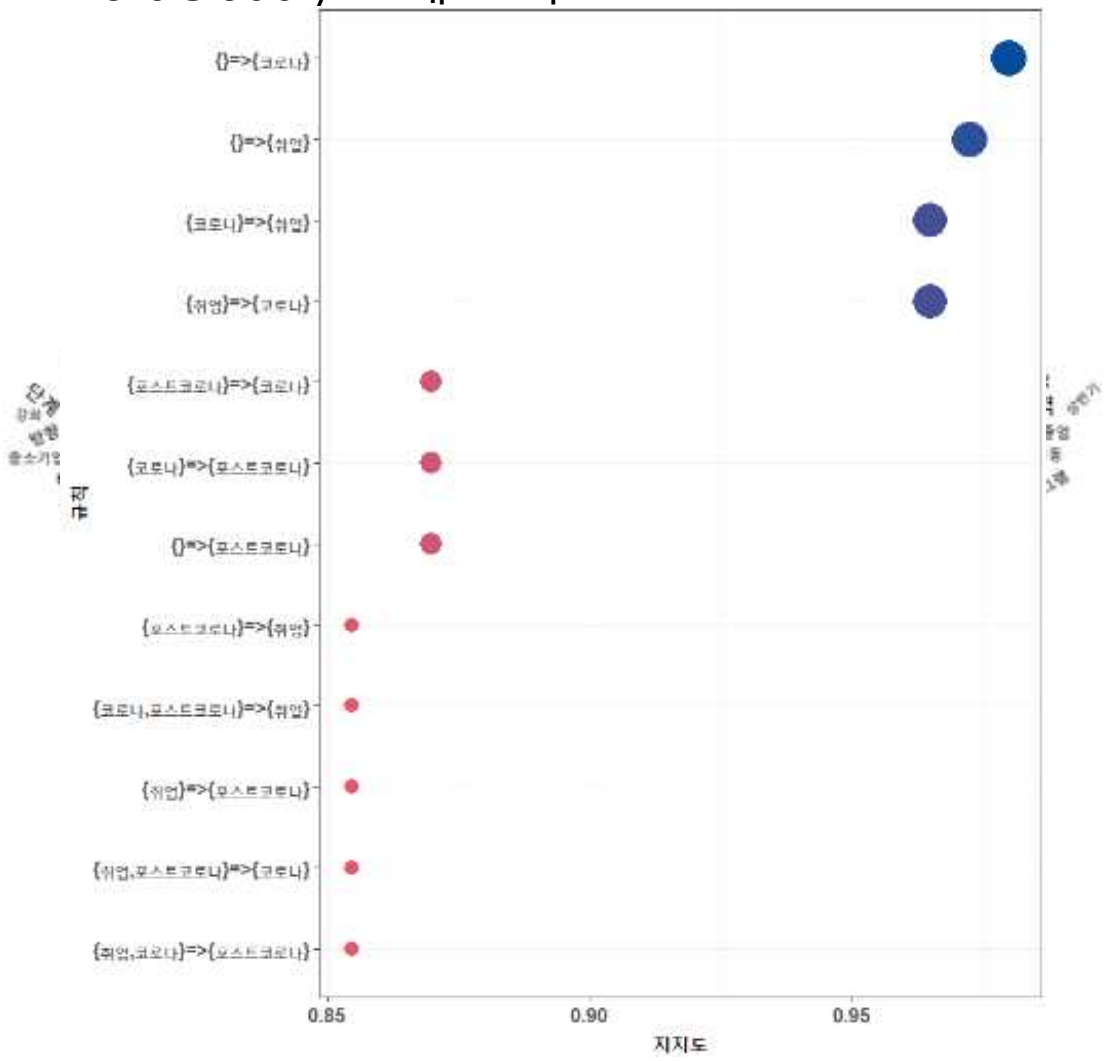
☞ 기사 제목으로 선정
☞ 기사 내 등장 횟수 2회 이상 5회 이하 단어 제한
☞ 제거 리스트로 마지막 확인

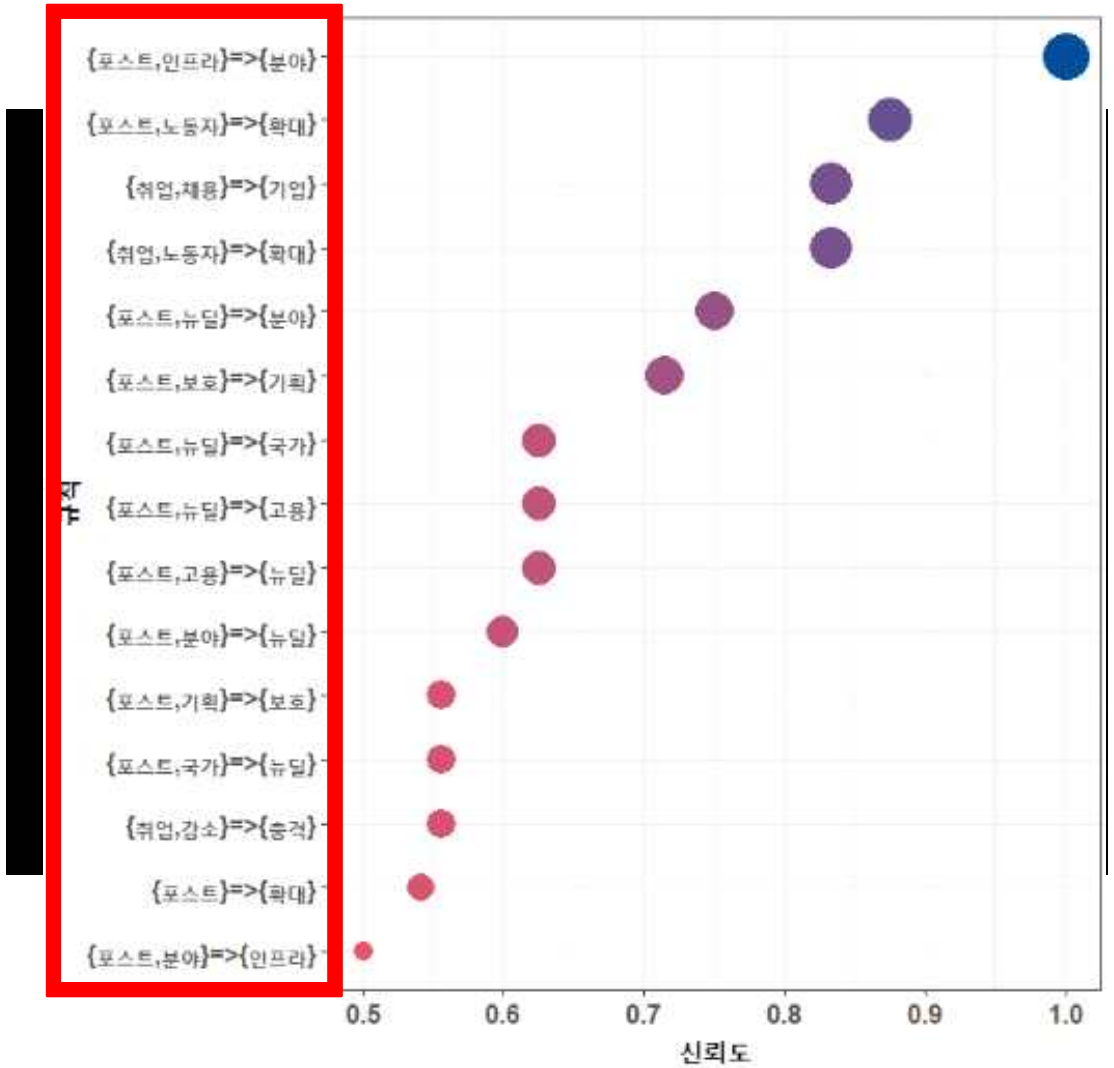
```
for texts in url_list:
    a = Article(texts, language='ko')
    a.download()
    a.parse()
    if ('코로나' in a.title  and '채용' in a.title)  or \
       ('코로나' in a.title and '구직' in a.title) or \
       ('코로나' in a.title and '취업' in a.title): # 그 단어가 포함된 것으로
        n=[]# 리스트 생성
        f.write(str(cnt))
        for i in okt.pos(a.text): # 형태소 단위로 리스트에 넣고
            if len(i[0])>=2 and (i[1]=='Noun' or i[1]=='Alpha'):
                n.append(i[0])
        f.write(str(cnt))
        for nm in n:
            if n.count(nm)>=2 and n.count(nm)<=5 :
                if nm not in sublist:
                    f.write(','+nm)
        f.write('\n') # 리스트 끝나면 줄바꿈
        cnt+=1 # 카운트 증가
```

☞WordCloud / 그래프 비교

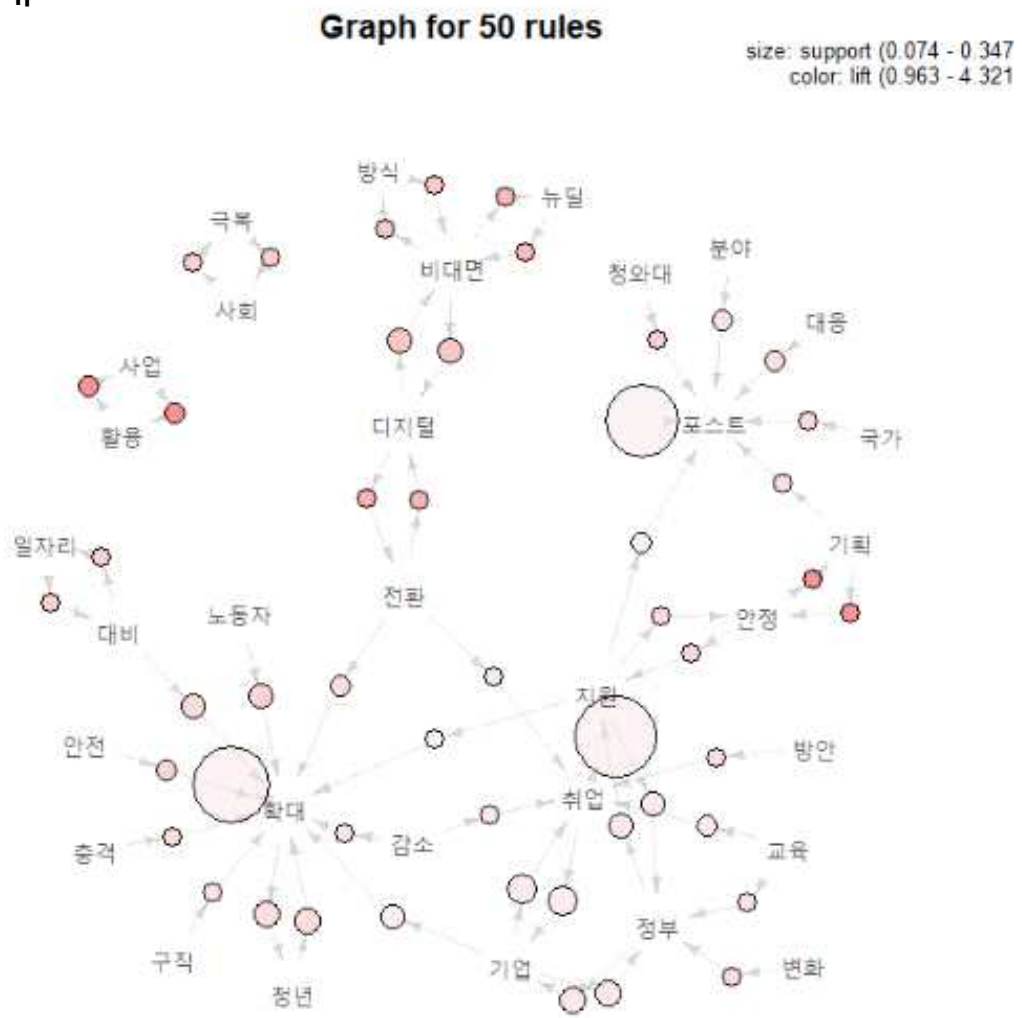
불용어 감소

☞WordCloud / 그래프 비교





네트워크 형태 그래프


Graph for 50 rules
size: support (0.074 - 0.347
color: lift (0.963 - 4.321
방식
뉴딜
극복
비대면
사회
청와대
분야
대응
사업
활용
디지털
포스트
국가
일자리
기획
전환
대비
노동자
안정
지원
안전
방안
확대
취업
충격
감소
교육
구직
정부
청년
기업
변화

분석 결과

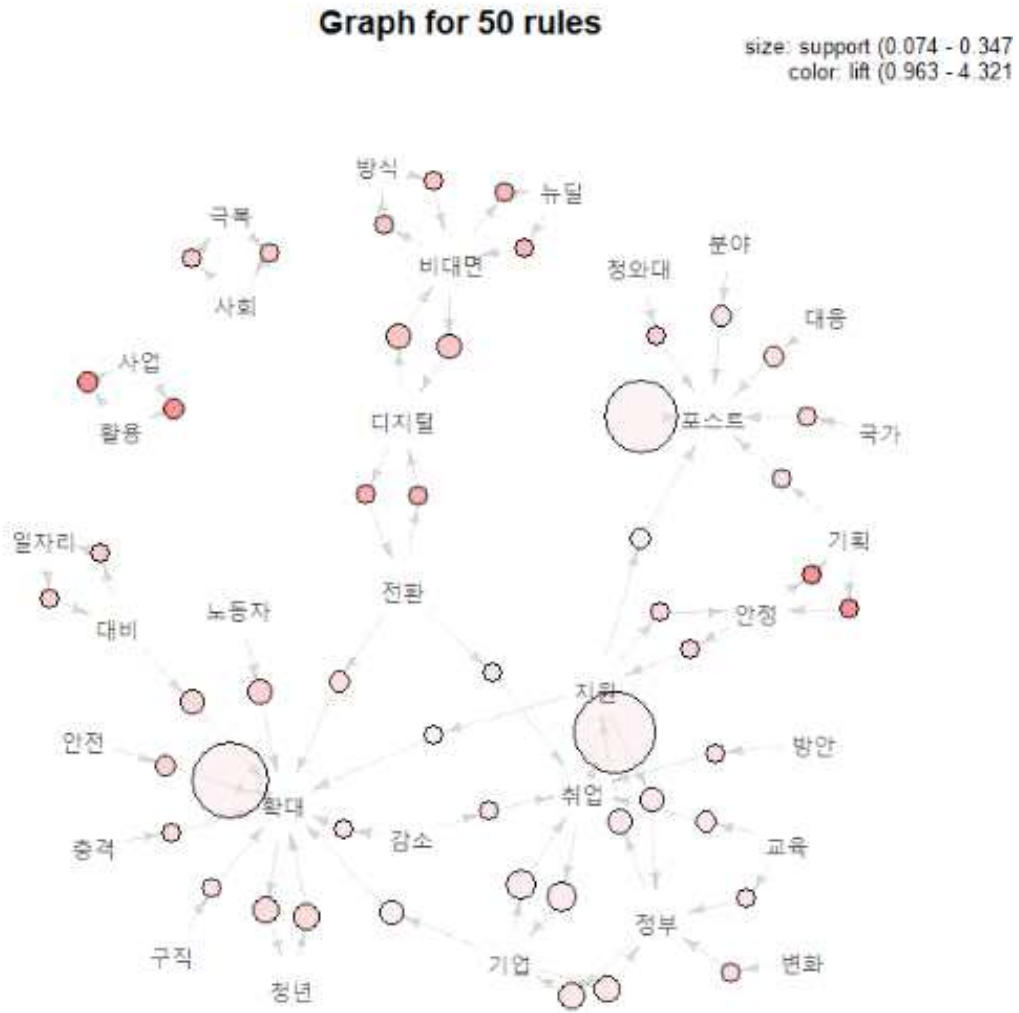
기사 제목 내 필수 조건
-> 코로나 + (채용 or 구직 or 취업)
-> 해당 단어는 모든 기사에 함께 들어있다고 가정

가장 원(지지도)의 크기가 큰 3가지 단어
-> 포스트, 취업, 확대

1. **포스트**
   청와대, 분야, 국가, 대응, 기획, 안정등의 단어와 연관
   코로나 이후 채용에는 국가적인 노력이 수반된다고 예측됨

2. **취업**
   지원, 정부, 교육, 기업, 감소, 방안, 정부 등의 단어와 연관
   기업 및 정부에서 교육을 실시해주는 방안들이 예측됨

3. **확대**
   가까운 연관 규칙의 단어와는 크게 의미를 찾을수 없음
   2차 연관으로 디지털, 전환, 비대면, 방식, 뉴딜, 지원 ,기업 등의 단어가 연관 규칙을 나타낸다.
   취업이나 채용의 방식이 디지털이나 비대면 방식으로 전환,
   한국의 뉴딜 프로젝트 비대면 방식으로 진행될 것으로 예측됨



프로젝트를 진행하며

프로젝트 진행 간 다양한 기사들을 통해 최근 동향 및 가까운 미래를 구경할 수 있었다.

연관 분석은 분석가, 개개인에 따라 해석이 달라질 수 있다고 느꼈다.

각자의 해석의 근거가 될 수 있는 지지도, 신뢰도, 향상도에 대한 정확한 숙지와 더 나아가서 IS측도, 교차지지도 등
이론적으로 단단해지면 더욱 더 좋은 분석과 해석을 할 수 있다고 생각했다.

머리속이나 기타 간단한 예제들로만 알고리즘을 접했을 때 보다 연관 규칙에 관한 깊은 이해를 할 수 있었다.

다른 알고리즘들 역시 실생활 및 방대한 양의 실무 데이터를 다룰 수 있는 능력을 갖춰야 한다는 생각을 하게 되었다.

프로젝트 분석 결과를 이해하며 취업난에 부담을 갖거나 겁먹지 말아야 겠다고 느꼈다.

나라 및 기업들도 구직자들을 위해 노력해줄 것을 믿으며 노력에 부응할 수 있도록 나만의 준비를 열심히 해야 할 것 같다.

비대면으로 전환될 수 있는 분야(면접, 고객상담, 업무) 등에 준비하는 것이 좋을 것 같다.

우리반 화이팅!!