10장 워드 클라우드

신봉균 20191624

2023-05-10

## 코드 10-1

Sys.setenv(JAVA\_HOME='C:/Program Files/Java/jre1.8.0\_211')  
library(wordcloud) #워드클라우드

## Loading required package: RColorBrewer

library(KoNLP) # 한국어 처리

## java.home option:

## JAVA\_HOME environment variable: C:/Program Files/Java/jre1.8.0\_211

## Warning in fun(libname, pkgname): Java home setting is INVALID, it will be ignored.  
## Please do NOT set it unless you want to override system settings.

## Checking user defined dictionary!

library(RColorBrewer) # 색상 선택   
text <- readLines("./coding\_study/R/10장/mis\_document.txt", encoding ="UTF-8" ) #파일 읽기   
buildDictionary(ext\_dic = "woorimalsam") #'우리 말씀' 한글사전 로딩

## 629897 words dictionary was built.

pal2 <- brewer.pal(8, "Dark2") #팔레트 생성  
noun <- sapply(text,extractNoun, USE.NAMES=F) #명사 추출  
noun #추출된 명사 출력

## [[1]]  
## [1] "<U+FEFF>1" "여성" "가족" "부" "불법" "촬영" "등" "디지털"  
## [9] "성범죄" "안전" "한" "사회" "국민" "들" "말"   
##   
## [[2]]  
## [1] ""  
##   
## [[3]]  
## [1] "존경" "하" "국민" "여러분"  
##   
## [[4]]  
## [1] "여성" "가족" "부" "장관" "정현" "백"   
## [7] "우리사회" "들불" "미" "투" "운동" "계기"   
## [13] "일상" "화" "폭력" "차별" "맞선" "여성"   
## [19] "들" "연대" "이" "이" "성평등" "한"   
## [25] "세상" "우리" "사회" "무엇" "강력" "추동"   
## [31] "일상" "성평등" "뿌리" "문제" "우리" "앞"   
## [37] "불법" "촬영" "촬영" "물" "유포" "포함"   
## [43] "하" "디지털" "성범죄"   
##   
## [[5]]  
## [1] ""  
##   
## [[6]]  
## [1] "디지털" "성범죄" "피해" "고통" "들" "언제"   
## [7] "나" "자신" "끔찍" "한" "범죄" "대상"   
## [13] "불안감" "호소" "하" "국민" "들" "앞"   
## [19] "저" "송구" "마음" "순간" "여성" "몸"   
## [25] "상품" "화한" "불법" "영상물" "광범위" "소비"   
## [31] "되" "이" "누군가" "경제" "적" "이득"   
## [37] "저" "자리" "정부" "대표" "해" "정부"   
## [43] "디지털" "성범죄" "피해" "심각" "성" "현안"   
## [49] "중대" "인식" "비장" "한" "각오" "정책추진"  
##   
## [[7]]  
## [1] ""  
##   
## [[8]]  
## [1] "지난해" "9" "월" "26" "디지털" "성범죄"   
## [7] "피해" "방지" "종합대책" "수립" "이래" "범정부"   
## [13] "차원" "종합적" "대응체계" "가동" "되" "신속"   
## [19] "엄정" "한" "수사" "처벌" "피해자" "보호"   
## [25] "지원" "디지털" "성범죄" "차단" "예방" "조치"   
## [31] "들이"   
##   
## [[9]]  
## [1] ""  
##   
## [[10]]  
## [1] "경찰" "청" "엄정" "신속" "한"   
## [6] "수사" "지방" "경찰" "청" "사이버"   
## [11] "성폭력" "수사" "팀" "신설" "방송통신"   
## [16] "심의위원회" "불법" "영상물" "신속" "한"   
## [21] "삭제" "차단" "디지털" "성범죄" "대응"   
## [26] "팀" "별도" "여성" "가족" "부"   
## [31] "상담" "삭제" "지원" "수사" "지원"   
## [36] "사후" "모니터링" "등" "피해자" "종합적"   
## [41] "서비스" "지원" "디지털" "성범죄" "피해자"   
## [46] "지원" "센터" "문" "정부" "우선"   
## [51] "적" "신속" "한" "삭제" "지원"   
## [56] "이" "소요" "되" "비용" "가해자"   
## [61] "구상권" "청구" "할" "계획"   
##   
## [[11]]  
## [1] ""  
##   
## [[12]]  
## [1] "여기" "정부" "관련" "전문가" "들" "현장" "제언" "반영"   
## [9] "해" "강력" "세부적" "근절" "방안" "들" "추진"   
##   
## [[13]]  
## [1] ""  
##   
## [[14]]  
## [1] "첫째" "물통" "형" "카메라" "단추"   
## [6] "형" "카메라" "등" "누구" "구입"   
## [11] "해" "불법" "촬영" "악용" "할"   
## [16] "수" "변형" "카메라" "등록제" "도입"   
## [21] "변형" "카메라" "제조" "수입" "판매"   
## [26] "하고" "등록" "판매" "이력" "관리"   
## [31] "이력" "정보시스템" "구축" "운영" "할"   
## [36] "계획" "이" "관련" "법안" "들이"   
## [41] "국회" "제출" "상태" "과학" "기술"   
## [46] "정보통신부" "중심" "제도" "시행" "할"   
## [51] "수" "노력" "할" "것"   
##   
## [[15]]  
## [1] ""  
##   
## [[16]]  
## [1] "둘째" "불법" "영상물" "유포"   
## [5] "기술" "개발" "추진" "속도"   
## [9] "내" "과학" "기술" "정보통신부"   
## [13] "인공지능(AI)" "빅데이터를" "활용" "한"   
## [17] "불법" "영상" "실시간" "차단기"   
## [21] "술" "개발" "진행" "중"   
## [25] "내년" "유해" "동영상" "실시간"   
## [29] "차단" "시제품" "개발" "완료"   
## [33] "계획" "방송" "통신위원회" "불법"   
## [37] "촬영" "물" "편집" "변형"   
## [41] "유통" "하" "것" "원천적"   
## [45] "차단" "할" "수" "디앤에이(DNA)"  
## [49] "필터링" "기술" "개발" "올해"   
## [53] "안" "완료" "목표"   
##   
## [[17]]  
## [1] ""  
##   
## [[18]]  
## [1] "셋째" "법무부" "미국" "일본" "등" "양자"   
## [7] "사법공조" "회의" "개최" "경찰" "청" "아동"   
## [13] "음란물" "대부분" "국가" "불법" "점" "착안"   
## [19] "아동" "음란물" "공략" "과제" "해외" "수사기관"  
## [25] "적극" "적" "공조" "수사" "등" "가능"   
## [31] "한" "방법" "총동원" "해외" "사이트" "불법"   
## [37] "영상물" "유포" "하" "끝" "추적" "해서"   
## [43] "처벌"   
##   
## [[19]]  
## [1] ""  
##   
## [[20]]  
## [1] "넷째" "행정" "안전" "부" "지방자치단체"  
## [6] "협력" "해" "전국" "공중" "화장실"   
## [11] "불법" "촬영" "카메라" "설치" "여부"   
## [16] "등" "현장점검" "강화" "경찰" "청"   
## [21] "재유" "포" "차단" "불법" "촬영"   
## [26] "물" "주요" "공급" "망인" "음란"   
## [31] "사이트" "웹하드" "등" "강력" "한"   
## [36] "단속과" "수사" "진행" "할" "예정"   
##   
## [[21]]  
## [1] ""  
##   
## [[22]]  
## [1] "얼마" "전" "혜화" "역" "시위" "그동안" "장구" "한"   
## [9] "세월" "우리" "사회" "여성" "들" "상처" "깊이" "정도"   
## [17] "사건" "여성" "가족" "부" "포함" "해" "정부" "전체"   
## [25] "여성" "들이" "불안" "고통" "공감"   
##   
## [[23]]  
## [1] ""  
##   
## [[24]]  
## [1] "디지털" "성범죄" "지하철" "공중" "화장실" "평범" "한" "일상"   
## [9] "공간" "누구" "피해자" "수" "확산" "속" "인간" "영혼"   
## [17] "파괴" "할" "수" "범죄" "그것" "촬영" "하" "것"   
## [25] "유포" "하" "것" "것" "명백" "한" "범죄" "정부"   
## [33] "가능" "한" "수단" "자원" "총동원" "근절" "되" "날"   
## [41] "혼신" "노력" "것" "약속" "드립" "니"   
##   
## [[25]]  
## [1] ""  
##   
## [[26]]  
## [1] "그동안" "국민" "들" "앞" "발표" "한" "관련" "대책"   
## [9] "이행" "국민" "일상" "속" "효과" "발휘" "여성" "가족"   
## [17] "부" "점검" "미진" "한" "부분" "관계" "부처" "지속적"  
## [25] "협의" "해" "끝" "관철"   
##   
## [[27]]  
## [1] ""  
##   
## [[28]]  
## [1] "막중" "한" "불법" "촬영" "물"   
## [6] "유포자" "징역형" "으로만" "처벌" "등"   
## [11] "가해자" "처벌" "강화" "하" "성폭력범죄"  
## [16] "처벌" "등" "특례법" "개정안" "국회"   
## [21] "계류" "중" "개정안" "국회" "통과"   
## [26] "수" "역량" "총동원"   
##   
## [[29]]  
## [1] ""  
##   
## [[30]]  
## [1] "앞" "입법"   
## [3] "제도" "보완"   
## [5] "등" "진행"   
## [7] "상황" "계속"   
## [9] "해서" "국민여러분께"   
## [11] "소통" "마지막"   
## [13] "피해" "분"   
## [15] "여성" "가족"   
## [17] "부" "디지털"   
## [19] "성범죄" "피해자"   
## [21] "지원센터(02-735-8994)'로" "도움"   
## [23] "요청" "것"   
## [25] "당부" "드립"   
## [27] "니" "감사"   
##   
## [[31]]  
## [1] ""  
##   
## [[32]]  
## [1] ""  
##   
## [[33]]  
## [1] "2" "행정" "안전" "부"   
## [5] "여성" "범죄인" "불법촬영(몰카)을" "근절하겠습니"   
##   
## [[34]]  
## [1] ""  
##   
## [[35]]  
## [1] "여성" "들이" "거리" "나" "여성" "대상"   
## [7] "불법" "촬영" "몰카" "각종" "성차별" "극심"   
## [13] "해지" "때문" "성차별" "일상" "화" "오늘"   
## [19] "날" "우리사회" "현실" "절규"   
##   
## [[36]]  
## [1] ""  
##   
## [[37]]  
## [1] "인간" "쾌적" "편안" "공간" "화장실" "불법"   
## [7] "촬영" "카메라" "때문" "여성" "들" "공포의"   
## [13] "공간" "몰카" "문명사회" "수" "짓" "중대"   
## [19] "한" "범죄행위"  
##   
## [[38]]  
## [1] ""  
##   
## [[39]]  
## [1] "정부" "이상" "좌시" "하지" "사안" "시급성" "때문" "공중"   
## [9] "화장실" "시작" "특별" "재원" "몰카" "탐지기" "대량" "확보"   
## [17] "탐지기" "동해" "여성" "이용" "하" "공중" "화장실" "일상"   
## [25] "생활" "공간" "화장실" "점검"   
##   
## [[40]]  
## [1] ""  
##   
## [[41]]  
## [1] "지자체" "경찰" "공공기관" "등" "동원" "시민"   
## [7] "단체" "점검" "앞" "화장실" "불법" "촬영"   
## [13] "적발" "되" "용납" "하지"   
##   
## [[42]]  
## [1] ""  
##   
## [[43]]  
## [1] "돈" "뭐" "비열한" "배금주의" "고통" "여성"   
## [7] "들" "공포" "분노" "정부" "공감"   
##   
## [[44]]  
## [1] ""  
##   
## [[45]]  
## [1] "여성" "들이" "길" "때" "화장실" "때" "생활" "할"   
## [9] "때" "불안" "두려움" "달라" "이상"   
##   
## [[46]]  
## [1] ""  
##   
## [[47]]  
## [1] "법무부" "경찰" "청" "등" "수사기관" "불법"   
## [7] "촬영" "유포" "불법행위" "신속" "수사" "피해자"   
## [13] "고통" "최소화" "범죄자" "단호" "처리"   
##   
## [[48]]  
## [1] ""  
##   
## [[49]]  
## [1] "경찰" "청" "여성" "청소년" "수사" "인력" "확충" "법"   
## [9] "미진" "한" "것" "사실" "법" "보완" "해" "그때"   
## [17] "법" "허용" "하" "범위" "안" "수단" "방법" "동원"   
##   
## [[50]]  
## [1] ""  
##   
## [[51]]  
## [1] "화장실" "시작" "하지" "여성" "대상" "범죄"   
## [7] "전면" "전" "선포" "반문명적" "범죄행위" "끝"   
## [13] "추적" "단속" "감시하"   
##   
## [[52]]  
## [1] ""  
##   
## [[53]]  
## [1] "세상" "절반" "여성" "안심" "할"   
## [6] "수" "편안" "하지" "우리" "사회"   
## [11] "야만(野蠻)" "여성" "문제" "여성" "만"   
## [16] "문제" "우리" "오늘" "오후" "저"   
## [21] "현장" "나" "불법" "촬영" "카메라"   
## [26] "점검"   
##   
## [[54]]  
## [1] ""  
##   
## [[55]]  
## [1] "감사"  
##   
## [[56]]  
## [1] ""  
##   
## [[57]]  
## [1] "2018." "6." "15"   
##   
## [[58]]  
## [1] "행정" "안전" "부장관" "김" "부" "겸"   
##   
## [[59]]  
## [1] ""  
##   
## [[60]]  
## [1] ""  
##   
## [[61]]  
## [1] "3" "경찰" "청" "불법" "촬영" "예방" "관련" "경찰" "청" "메시"  
## [11] "지"   
##   
## [[62]]  
## [1] ""  
##   
## [[63]]  
## [1] "경찰" "청" "차장"  
##   
## [[64]]  
## [1] ""  
##   
## [[65]]  
## [1] "경찰" "청" "여성" "분" "안전" "확보" "불안감" "해소"   
## [9] "불법" "촬영" "유포" "수사" "차단" "최선"   
##   
## [[66]]  
## [1] ""  
##   
## [[67]]  
## [1] "불법" "촬영" "예방" "수사"  
##   
## [[68]]  
## [1] ""  
##   
## [[69]]  
## [1] "지자체" "등" "유관기관" "다중" "이용" "시설"   
## [7] "내" "공중" "화장실" "탈의실" "기숙" "학교"   
## [13] "등" "물론" "휴가" "시즌" "대비" "피서지"   
## [19] "불법" "카메라" "설치" "여부" "세밀" "점검"   
## [25] "구멍" "등" "초소형" "카메라" "설치" "흔적"   
## [31] "선정" "적" "낙" "등" "불안" "요소"   
## [37] "경우" "시설주" "개선" "권고" "앞" "안부"   
## [43] "교육" "부" "지자체" "등" "긴밀" "협력"   
## [49] "상시" "점검" "여성" "불안감" "해소" "노력"   
##   
## [[70]]  
## [1] ""  
##   
## [[71]]  
## [1] "취약" "시간" "장소" "단속활동" "강화" "퇴근"   
## [7] "시간" "지하철역" "등" "불법" "촬영" "다발"   
## [13] "시간" "장소" "예방" "단속활동" "강화" "7"   
## [19] "월" "피서지" "여름" "경찰관서" "성범죄" "전담"   
## [25] "팀" "운영(78" "개소" "할" "예정"   
##   
## [[72]]  
## [1] ""  
##   
## [[73]]  
## [1] "불법" "촬영" "행위" "적발"   
## [5] "한" "경우" "신속" "강력"   
## [9] "수사" "현행범" "체포" "등"   
## [13] "강제수사" "압수수색" "디지털포렌식을" "증거"   
## [17] "확보" "추가" "범죄" "확인"   
## [21] "구속영장" "청구" "할" "방침"   
##   
## [[74]]  
## [1] ""  
##   
## [[75]]  
## [1] "중앙" "전파" "관리" "소" "세관" "등" "합동" "장형"   
## [9] "카메라" "유통" "등" "단속" "병행"   
##   
## [[76]]  
## [1] ""  
##   
## [[77]]  
## [1] "불법" "촬영" "물" "음란물" "유포" "수사" "차단" "ㆍ"   
## [9] "에"   
##   
## [[78]]  
## [1] ""  
##   
## [[79]]  
## [1] "사이버" "여청" "수사" "인력" "활용" "불법" "촬영" "물"   
## [9] "공급자" "수사" "강화"   
##   
## [[80]]  
## [1] "이" "시민단체(한사성" "300" "건"   
## [5] "등" "누리캅스" "등" "협력단체"   
## [9] "긴밀" "협력" "해" "첩보"   
## [13] "수집" "음란" "사이트" "운영자"   
## [17] "웹하드" "헤비업로더" "SNS" "상습"   
## [21] "유포자" "중심" "단속"   
##   
## [[81]]  
## [1] "사설" "불법"   
## [3] "정보" "삭제업자(디지털장의사)가"  
## [5] "사이트" "운영자"   
## [7] "등" "공모"   
## [9] "한" "경우"   
## [11] "음란물" "유포"   
## [13] "방조범(정보통신망법" "제44조의7"   
## [15] "위반" "강력"   
## [17] "수사"   
##   
## [[82]]  
## [1] "외국" "서버" "경우"   
## [4] "수사" "서버" "폐쇄"   
## [7] "점" "아동" "음란물"   
## [10] "유무" "자금흐름" "추적"   
## [13] "연계" "사이트" "광고주"   
## [16] "등" "수사" "사법처리"   
## [19] "서버" "폐쇄" "추진"   
## [22] "하" "한" "편"   
## [25] "美" "국토안보수사청(HSI)" "등"   
## [28] "해외" "수사기관" "공조"   
## [31] "강화"   
##   
## [[83]]  
## [1] ""  
##   
## [[84]]  
## [1] "피해" "영상물" "신속" "삭제" "차단" "되"   
## [7] "피해" "회복" "수" "수사" "초기단계" "방심"   
## [13] "위" "여가" "부" "등" "긴밀" "협력"

한국어 자연어 처리를 지원하는 **KoNLP** 패키지를 이용하려면 java가 설치되어 있어야 하는데, 위 명령문은 자바 실행 환경(JRE)을 이용할 수 있도록 설정하는 내용이다. 컴퓨터에 따라 설치된 폴더와 **JRE** 버전이 다를 수 있으니 자신의 컴퓨터에 맞게 경로를 설정해 주어야 한다.

워드클라우드 작성에 필요한 패키지들을 불러온다. 해당 패키지들은 미리 설치되어 있어야 한다. **KoNLP** 패키지의 경우 설치가 되지 않으면 출판사에서 제공하는 별도 자료를 참조하도록 한다.

**buildDictionary** 함수는 워드클라우드 작성 시 필요한 한글 사전을 불러오는 역할을 한다. 워드클라우드 작성 대상 파일은 문장 형태로 되어있는데 문장으로부터 단어를 추출하러면 사전이 필요하다. 즉, 어떤 문자열이 사전에 있으면 단어라고 인식한다.

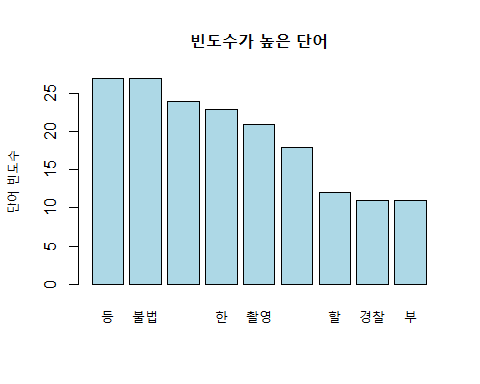
대국민 담화문 내용으로부터 명사를 추출하여 noun에 저장하고 noun의 내용을 출력한다. 내용을 보면 담화문 의 문장하나가 리스트 noun에 하나의 원소로 저장된 것을 알 수 있다. 하나의 문장은 단어 단위로 분리되어 있다.

## 코드 10-2

noun2 <- unlist(noun) #추출된 명사 통합   
wordcount <- table(noun2) #단어 빈도수 계산  
temp <- sort(wordcount, decreasing=T)[1:10] #빈도수 높은 단어 10개의 추출  
temp

## noun2  
## 등 불법 여성 한 촬영 수사 할 경찰 부   
## 40 27 27 24 23 21 18 12 11 11

temp <- temp[-1] #공백 단어 제거  
barplot(temp, #막대그래프 작성  
 names.arg = names(temp), #막대 이름을 단어로 표시  
 col ="lightblue", #막대의 색상 지정  
 main ="빈도수가 높은 단어", ylab = "단어 빈도수")



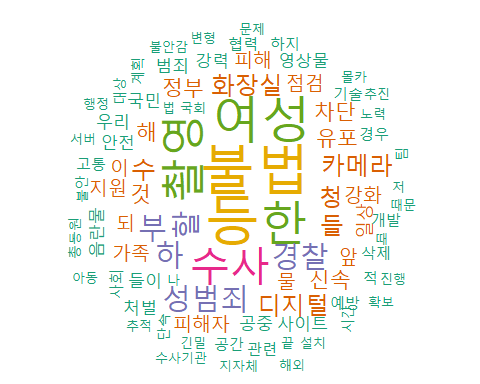
**unlist()**함수는 리스트의 형태의 자료를 벡터 형태로 바꾼다. 이로서 **noun**에서 각 원소별로 나누어져 있던 단어들이 하나의 벡터로 통합된다. **table()** 함수를 통해 단어별로 빈도수를 계산하여 **wordcount**에 저장한다.

**wordcount**를 내림차순으로 정렬한 후 빈도수가 높은 단어 10개를 추출하여 temp에 저장한다. temp의 내용 보면 공백 문자가 가장 빈도수가 높게 나오는데, 의미 없는 단어이므로 temp[-1]을 통해 temp에서 제거한다.

**temp**로 막대그래프를 작성한다. ‘불법’,’여성’과 같은 단어가 상위에 랭크된 것을 알 수 있는데 불법촬영이 범죄임을 강조하다 보니 ’불법’이라는 단어가 많이 등장하고 불법촬영의 피해자 중 여성이 차지하는 비중이 높아 ’여성’이라는 단어가 자주 등장한다고 판단된다.

## 코드 10-3

wordcloud(names(wordcount), #단어들   
 freq=wordcount, # 단어들의 빈도  
 scale=c(6,0.7), #단어의 폰트 크기  
 min.freq=3, #단어의 최소 빈도  
 random.order=F, #단어의 출력위치  
 rot.per=.1, #90도 회전 단어 비율  
 colors=pal2) #단어의 색



**names(wordcount)** 워드칼라우드 상에 표시할 단어를 지정한다.

**freq=wordcount** 워드클라우드 상에 표시할 단어의 빈도수를 지정한다.

**scale=c(6,0.7)** 표시할 단어의 폰트 크기를 지정한다. 여기서 6은 폰트의 최대 크기, 0.7은 폰트의 최소 크기를 의미한다.

**min.freq=3** 빈도수가 3 이상인 단어들만 표시한다.

**random.order=F** 단어가 표시될 위치를 지정한다 T는 단어의 표시 위치를 무작위로 지정할 수 있고, F는 빈도수가 높은 다어일수록 중앙쪽에 배치된다.

**rot.per=.1**단어를 표시할 때 세로 방향으로 표시할 단어의 비율을 지정한ㄷ. 여기서 -1은 10%를 의미한다.

**colors=pal2** 빈도수에 따라 pal2에 있는 색으로 단어의 색을 지정한다.

## 코드 10-4

# 빈도수 높은데 워드클라우드에 없으면 사용자 사전에 추가  
buildDictionary(ext\_dic = "woorimalsam",   
 user\_dic=data.frame("정치","ncn"),  
 replace\_usr\_dic = T)

## 629898 words dictionary was built.

noun <- sapply(text,extractNoun, USE.NAMES=F)  
noun2 <- unlist(noun) #추출된 명사 통합  
  
# 무의미한 단어 제거  
noun2 <- noun2[nchar(noun2)>1] #1글자 단어 제거   
noun2 <- gsub("하지","", noun2) #'하지' 제거   
noun2 <- gsub("때문","", noun2) #'떄문' 제거   
  
wordcount <- table(noun2)   
wordcloud(names(wordcount), #단어 빈도수 계산  
 freq=wordcount,   
 scale=c(6,0.7),   
 min.freq=3,  
 random.order=F,  
 rot.per=.1,  
 colors=pal2)

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 불안감 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 설치 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 수사기관 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 지자체 could not be fit on page. It will not be plotted.

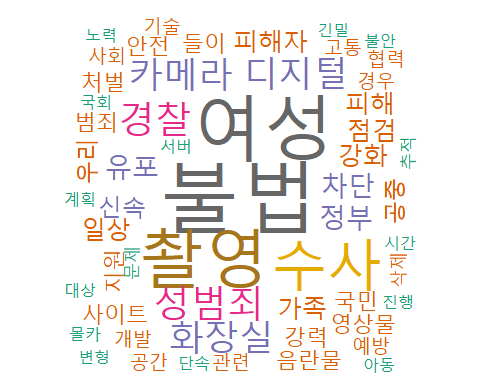
## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 총동원 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 추진 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 해외 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 행정 could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(wordcount), freq = wordcount, scale = c(6, 0.7), :  
## 확보 could not be fit on page. It will not be plotted.



위 명령문 ’**정치**’라는 단어가 사전에 없다고 가정하고 사전에 추가하는 과정을 보여준다.

단어의 길이가 한글자인 단어는 대체로 별 의미가 없으므로 제거한다. 또한 **‘하지’**,**‘때문’**과 같이 의미 없는 단어도 제거한다. **gsub()**함수는 문자열을 바꾸는 함수로, **gsub(‘하지’,’‘,noun2)는 noun2에 있는 단어 중 ’하지’**는 **’’**으로 바꾸라는 의미이다.