mini

## 크롤링 데이터 전처리

다음과 같은 원데이터에서

## New names:  
## \* `` -> `...1`

## # A tibble: 6 x 3  
## coname title career  
## <chr> <chr> <chr>   
## 1 "[<a href=\"/company/221\">아이지에이웍스</a>]" [<h1>머신러닝 엔지~ [<dd>~  
## 2 "[<a href=\"/company/588\">미식의시대</a>]" [<h1>신한은행 배달~ [<dd>~  
## 3 "[<a href=\"/company/566\">포디리플레이코리아</a>]" [<h1>Computer Visi~ [<dd>~  
## 4 "[<a href=\"/company/791\">카비</a>]" [<h1>영상인식 엔지~ [<dd>~  
## 5 "[<a href=\"/company/1620\">리플에이아이</a>]" [<h1>머신러닝 엔지~ [<dd>~  
## 6 "[<a href=\"/company/854\">포지큐브</a>]" [<h1>음성/자연어 ~ [<dd>~

CSS 태그가 함께 붙은 크롤링 데이터를 정제함.

## # A tibble: 6 x 3  
## coname title career   
## <chr> <chr> <chr>   
## 1 아이지에이웍스 머신러닝 엔지니어 (디파이너리) 신입 - ~  
## 2 미식의시대 신한은행 배달앱 '땡겨요' 개인화 큐레이션 개발 총괄 5 - 10년  
## 3 포디리플레이코리아 Computer Vision\_3D Vision 3 - 10년  
## 4 카비 영상인식 엔지니어 3 - 10년  
## 5 리플에이아이 머신러닝 엔지니어 신입 - ~  
## 6 포지큐브 음성/자연어 딥러닝 신입

파싱이 완료된 데이터를 통합한 후, 변수 전처리를 시행함.

(원데이터)

## # A tibble: 6 x 5  
## coname keyword qualif prefer career  
## <chr> <chr> <chr> <chr> <chr>   
## 1 포디리플레이코리아 opencv, c, c++, tensorflow, pytorch,~ "Open~ "비전~ 3 - 1~  
## 2 카비 deeplearning, machinelearning, c, c+~ "관련~ "영상~ 3 - 1~  
## 3 리플에이아이 python, pytorch, docker, machinelear~ "컴퓨~ "ML ~ 신입 ~  
## 4 브릭 React, spring boot, Vue, 소프트웨어~ <NA> <NA> 신입   
## 5 허리업코리아 데이터베이스, DBA, MySQL, .NET, ASP.~ <NA> <NA> 경력~   
## 6 콘텐츠웨이브(wavve) <NA> "- Ja~ "- MV~ <NA>

회사 위치 컬럼을 시/구 단위로 나눈 후, 서울을 서울특별시로 정규화하였음.

## # A tibble: 6 x 2  
## coname loc   
## <chr> <chr>   
## 1 포디리플레이코리아 서울특별시 강남구  
## 2 카비 서울특별시 금천구  
## 3 리플에이아이 서울특별시 관악구  
## 4 브릭 경기도 성남시   
## 5 허리업코리아 서울특별시 강남구  
## 6 콘텐츠웨이브(wavve) 서울특별시 강남구

경력 컬럼을 정제 후, 경력 무관은 0으로, 신입은 1로, 경력직은 2로 할당하여 새롭게 저장해 줌.

## # A tibble: 6 x 3  
## coname career career2  
## <chr> <chr> <dbl>  
## 1 포디리플레이코리아 3-10 2  
## 2 카비 3-10 2  
## 3 리플에이아이 0-20 1  
## 4 브릭 0 1  
## 5 허리업코리아 무관 0  
## 6 콘텐츠웨이브(wavve) <NA> NA

자격요건/우대조건 컬럼을 정제해줌.

## # A tibble: 6 x 1  
## qualif   
## <chr>   
## 1 "OpenCV, C/C++, Python\r\r\n유관업무 3년 이상 경력"   
## 2 "관련 업무 경력 3년 이상\r\r\n컴퓨터공학 또는 영상처리 관련학과 석사(대학원) ~  
## 3 "컴퓨터 공학, 머신러닝 관련 분야 석/박사 혹은 그에 준하는 경험이 있으신 분\r\~  
## 4 <NA>   
## 5 <NA>   
## 6 "- Java, Kotlin 기반 Android 앱 개발 경험 / 3년 이상/- Restful API 이용 서비~

먼저 소문자로 변환 후, 제어 문자를 삭제함.  
마크업 언어를 제거해주고, 특수문자 및 숫자를 삭제함.  
그리고 딥러닝/dl 과 같이 뜻이 동일한 단어들을 영어로 변환해 통일함.

## # A tibble: 6 x 1  
## qualif   
## <chr>   
## 1 opencv, c c++, python 유관업무 년 이상 경력   
## 2 관련 업무 경력 년 이상 컴퓨터공학 또는 영상처리 관련학과 석사, 대학원 이상~  
## 3 컴퓨터공학, ml 관련 분야 석 박사 혹은 그에 준하는 경험이 있으신 분 pytorch~  
## 4 <NA>   
## 5 <NA>   
## 6 - java, kotlin 기반 android app 개발 경험 년 이상 - restful-api 이용 서비~

정제된 자격요건/우대조건 컬럼을 토큰화 해줌.

## Warning in readLines("C:/R/R/stack.txt", encoding = "UTF-8"): 'C:/R/R/  
## stack.txt'에서 불완전한 마지막 행이 발견되었습니다

## 375826 words dictionary was built.

## # A tibble: 6 x 1  
## qualif2   
## <chr>   
## 1 opencv, c, c++, python, 유관, 업무, 년, 이상, 경, 력   
## 2 관련, 업무, 경력, 년, 이상, 컴퓨터공학, 영상처리, 관련학과, 석사, 대학원, 학~   
## 3 컴퓨터공학, ml, 관련, 분야, 박사, 그, 경험, 분, pytorch, 이용, 한, 프로젝트, ~  
## 4 NA   
## 5 NA   
## 6 java, kotlin, 기반, android, app, 개발, 경험, 년, 이상, restful, api, 이용, ~

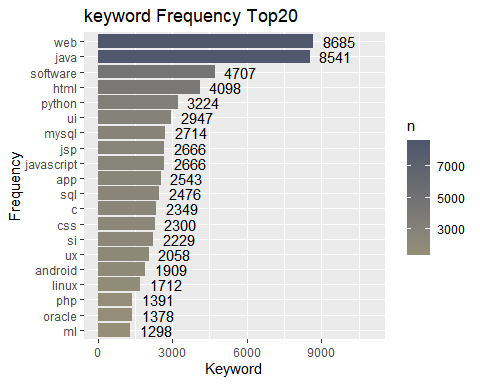
토큰화된 단어 중 핵심어들을 keyword2에 저장함.

## Warning in ex$keyword[is.na(ex$keyword)] <- gsub("[가-힣]", "", ex$qualif2):  
## number of items to replace is not a multiple of replacement length

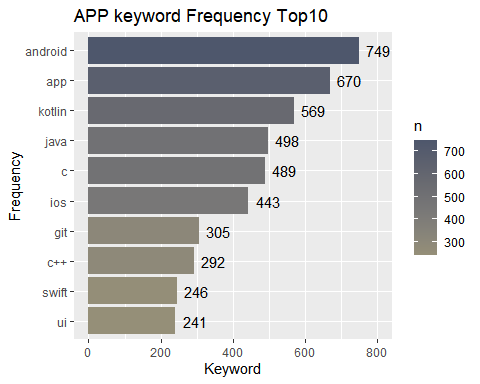
## # A tibble: 6 x 1  
## keyword2   
## <chr>   
## 1 "opencv, c, c++, tensorflow, pytorch, python, ar, deeplearning"   
## 2 "deeplearning, machinelearning, c, c++, python"   
## 3 "python, pytorch, docker, machinelearning"   
## 4 "react, spring boot, vue, bigdata, si, java, c#, java, python, ui·ux, "   
## 5 "dba, mysql, .net, asp.net, java, jsp, mysql, asp.net, c#, java, oracle, sql,~  
## 6 "opencv, c, c++, python, "

## 키워드 시각화

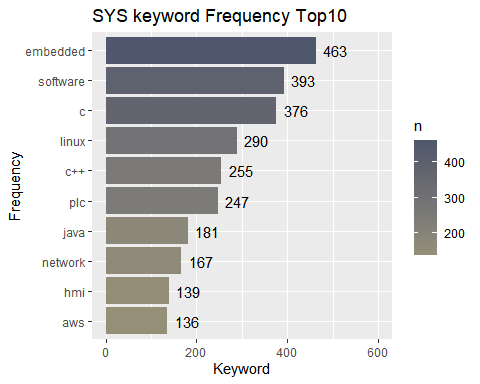
토큰화한 키워드를 이용하여 어떤 키워드가 어떤 직군의 문서에서 자주 나왔는지를 언어 및 직무 스택을 중심으로 살펴봄. Count 함수를 이용하여 단어의 등장 빈도를 세주었고, 이를 막대 그래프를 이용해 시각화하였음.



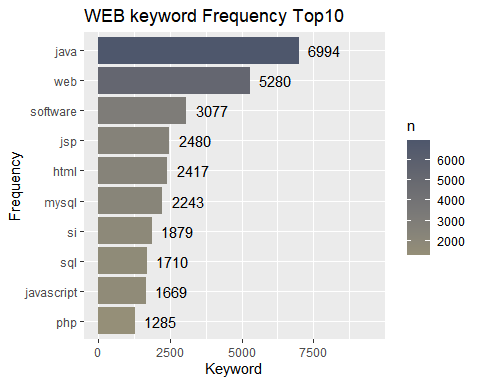
전체 키워드 시각화 결과에서는 web과 java가 약 9000번으로 다른 키워드에 비해 월등히 자주 등장한 것으로 나타남. 이는 웹프로그래머 공고가 다른 직군에 비해 많기 때문으로 웹프로그래머와 관련이 있는 HTML, JSP, JavaScript, PHP 등에 대한 키워드 역시 자주 나타남. 이어서 Python, MySQL, SQL, Oracle, 머신러닝 등 데이터 분석 및 데이터베이스에 대한 키워드가 자주 나타났고, C, App, Android가 이어서 자주 등장하였음.



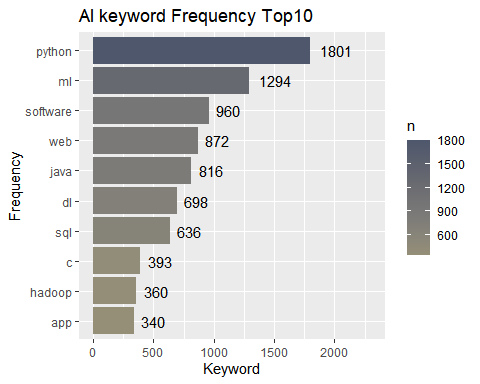
어플리케이션 개발 직군에서는 Android, Kotlin 등 안드로이드 개발 관련 키워드가 가장 많이 나타났고, 이어서 Java, C 등 개발 언어가 뒤를 이었음. Ios, Swift 등 Ios 어플리케이션 개발에 대한 키워드는 각각 6번째와, 9번째로 자주 등장한 것으로 나타남. 한국 어플리케이션 개발은 안드로이드 개발 직군에 대한 수요가 더 많은 것으로 나타남.



시스템, DBA 관련 직군에서는 임베디드 키워드가 가장 자주 나타났고, 그 뒤로 C, C++, Java 등 개발 언어에 대한 키워드가 자주 나타났음. Linux가 다른 직군에 비해 높은 순위를 차지한 것을 보아 Linux 개발 환경을 많이 사용하는 것으로 추측해볼 수 있음. 이어서 PLC, HMI 등 직무 관련 키워드가 많이 등장하였음.



웹 개발 직군에서는 Java가 가장 자주 등장하였고, JSP, JavaScript 등 관련 키워드 역시 자주 등장함. Web, HTML, PHP 등 웹 개발 키워드 역시 자주 등장하였음. 이밖에 MySQL, SI, SQL 등 백엔드 개발 관련 키워드가 각각 6번째, 7번째 8번째를 차지하였음.



AI 및 데이터 관련 직군에서는 Python이 가장 자주 나타났고, 머신러닝과 딥러닝 키워드가 상당히 자주 나타난 것으로 보아 기계학습 관련 직무에 수요가 큰 것으로 보임. 이어서 Java 및 C언어가 자주 나타났고, SQL, Hadoop 등 데이터베이스 관련 키워드도 자주 나왔음.