

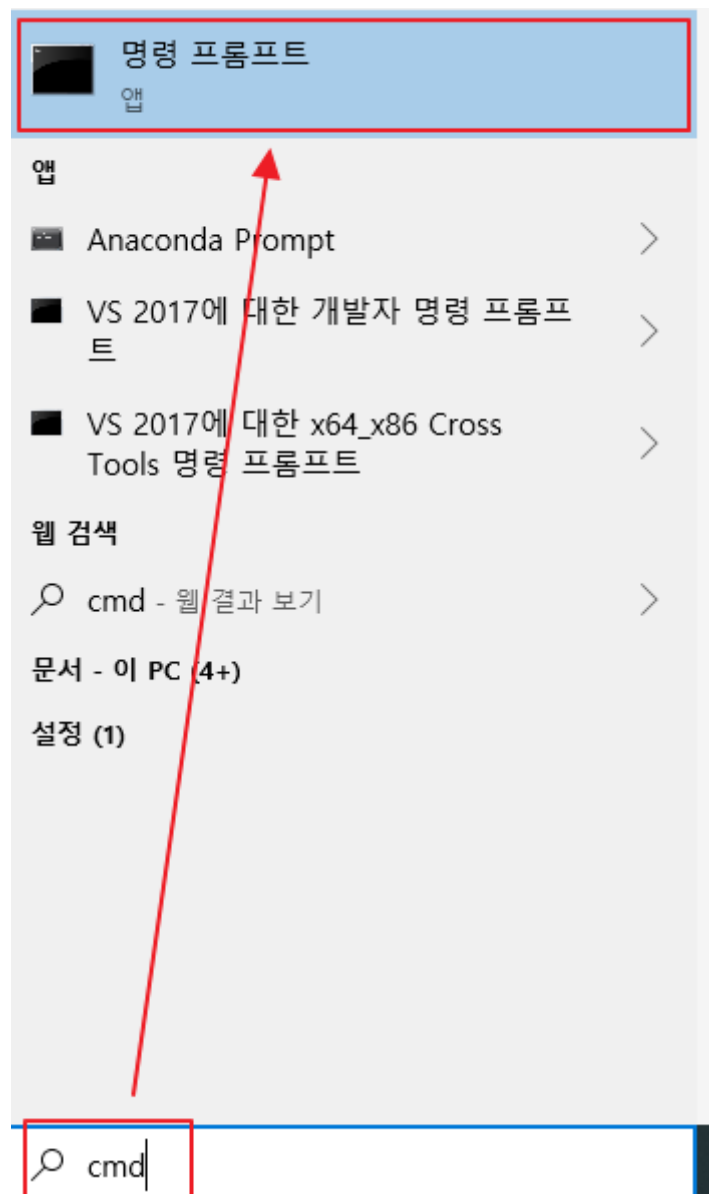
L-UP_02 텍스트 마이닝

학습 내용

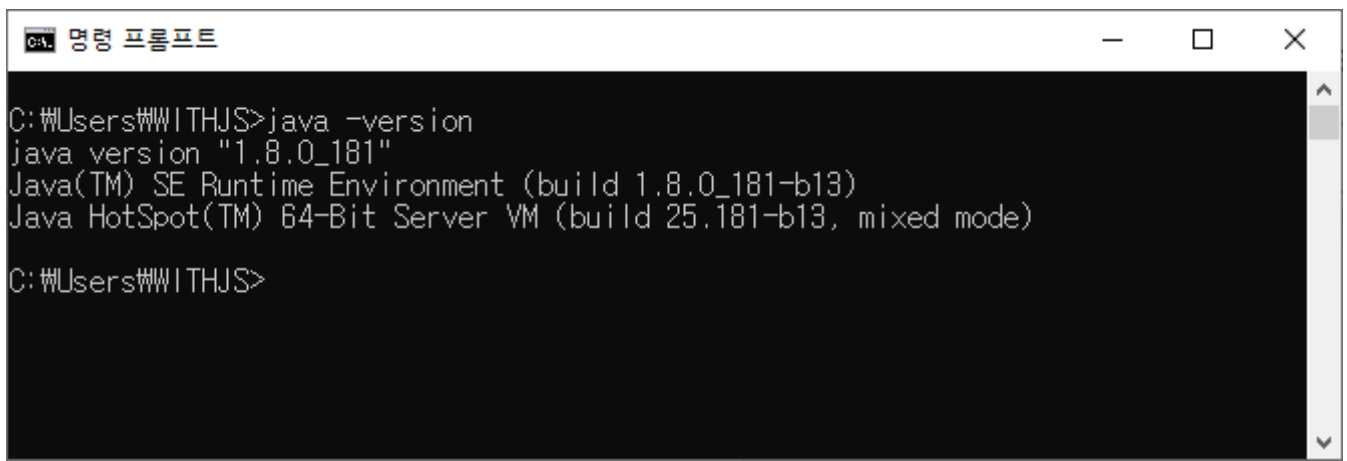
- 텍스트 마이닝을 위한 패키지 설치 및 불러오기
- 텍스트 데이터 불러오기
- 문자열 전처리

01 준비하기

- 한글 텍스트 마이닝을 위해서 KoNLP(Korean Natural Language Processing)을 이용한다.
- KoNLP는 자바(Java)가 설치되어 있어야 한다.
- 확인 사항(내 컴퓨터가 32bit, 64bit) 확인 후, 자바 사이트에 가서 다운로드



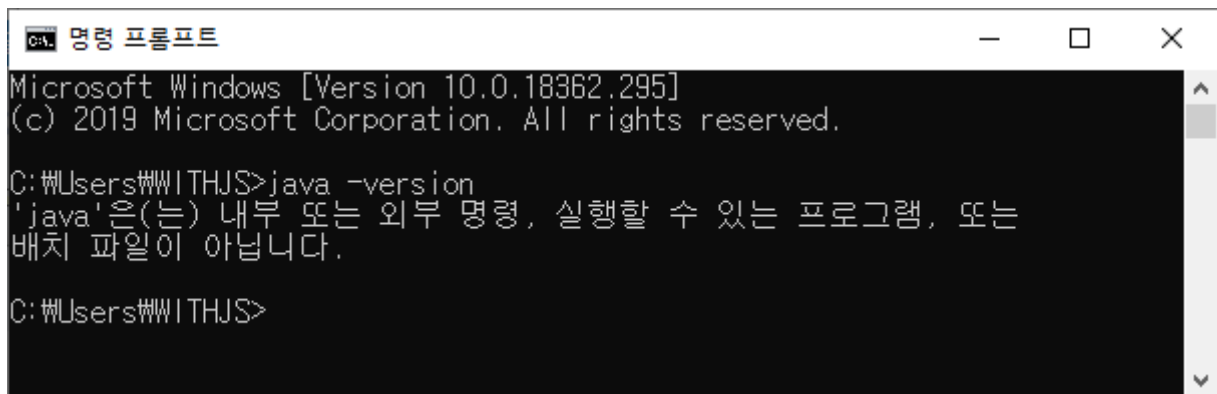
설치되어 있을 경우,



```
CA. 명령 프롬프트
C:\Users\WITHJS>java -version
java version "1.8.0_181"
Java(TM) SE Runtime Environment (build 1.8.0_181-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.181-b13, mixed mode)

C:\Users\WITHJS>
```

설치가 안되어 있을 경우



```
CA. 명령 프롬프트
Microsoft Windows [Version 10.0.18362.295]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\WITHJS>java -version
'java'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.

C:\Users\WITHJS>
```

- 검색 - cmd 입력
- 명령 프롬프트 실행
- java 설치 확인
 - java -version 으로 확인되면 설치되어 있음.

JAVA 설치하기

- 사전에 설치가 되어 있다면 제어판-프로그램 제거
- JAVA 홈페이지에서 다운로드 후, 설치(필요하다면 오라클 계정 가입이 필요)
- 사전에 동의가 필요
- 다운로드 후, 설치 진행

Java SE Development Kit 8u221

You must accept the [Oracle Technology Network License Agreement](#) for Oracle Java SE to download this software.

Thank you for accepting the Oracle Technology Network License Agreement for Oracle Java SE; you may now download this software.

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.9 MB	jdk-8u221-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.81 MB	jdk-8u221-linux-arm64-vfp-hflt.tar.gz
Linux x86	174.18 MB	jdk-8u221-linux-i586.rpm
Linux x86	189.03 MB	jdk-8u221-linux-i586.tar.gz
Linux x64	171.19 MB	jdk-8u221-linux-x64.rpm
Linux x64	186.06 MB	jdk-8u221-linux-x64.tar.gz
Mac OS X x64	252.52 MB	jdk-8u221-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	132.99 MB	jdk-8u221-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	94.23 MB	jdk-8u221-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.66 MB	jdk-8u221-solaris-x64.tar.Z
Solaris x64	91.95 MB	jdk-8u221-solaris-x64.tar.gz
Windows x86	202.73 MB	jdk-8u221-windows-i586.exe
Windows x64	215.35 MB	jdk-8u221-windows-x64.exe

ORACLE

오라클 계정 로그인

사용자 이름



.....@gmail.com



암호



.....



! 사용자 이름 및 암호를 입력하십시오

로그인

도움이 필요하십니까?

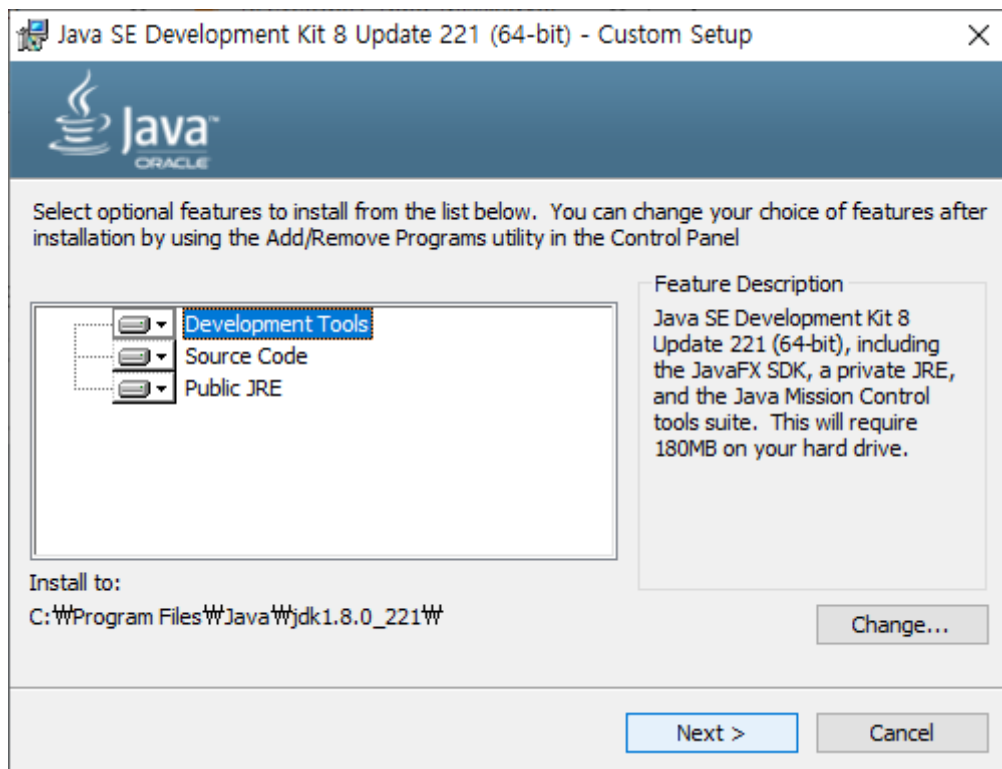
계정 만들기



jdk-8u221-windo....exe
32.6/215MB, 2분 남음

전체 보기





설치 진행 후, 환경변수 설정 필요

- 제어판 - 시스템 - 고급 시스템 설정
- 환경 변수 선택
- 시스템 변수 중 path를 선택 후, 편집
- 새로 만들기를 누르고 java 의 폴더 경로를 추가한다.
- cmd를 실행시키고, java -version으로 확인

rJava 설치를 위한 rtool설치

- url : <https://cran.rstudio.com/bin/windows/Rtools/> (<https://cran.rstudio.com/bin/windows/Rtools/>)
- (19.08.23 추천버전 3.5)
- 다운로드 후, 설치 진행
- 설치 위치 : C:\Rtools
- Select Components와 Select Additional Tasks는 아래 이미지 참조

Building R for Windows

This document is a collection of resources for building packages for R under Microsoft Windows, or for building R itself (version 1.9.0 or later). The original collection was put together by Prof. Brian Ripley and Duncan Murdoch; it is currently maintained by Jeroen Ooms.

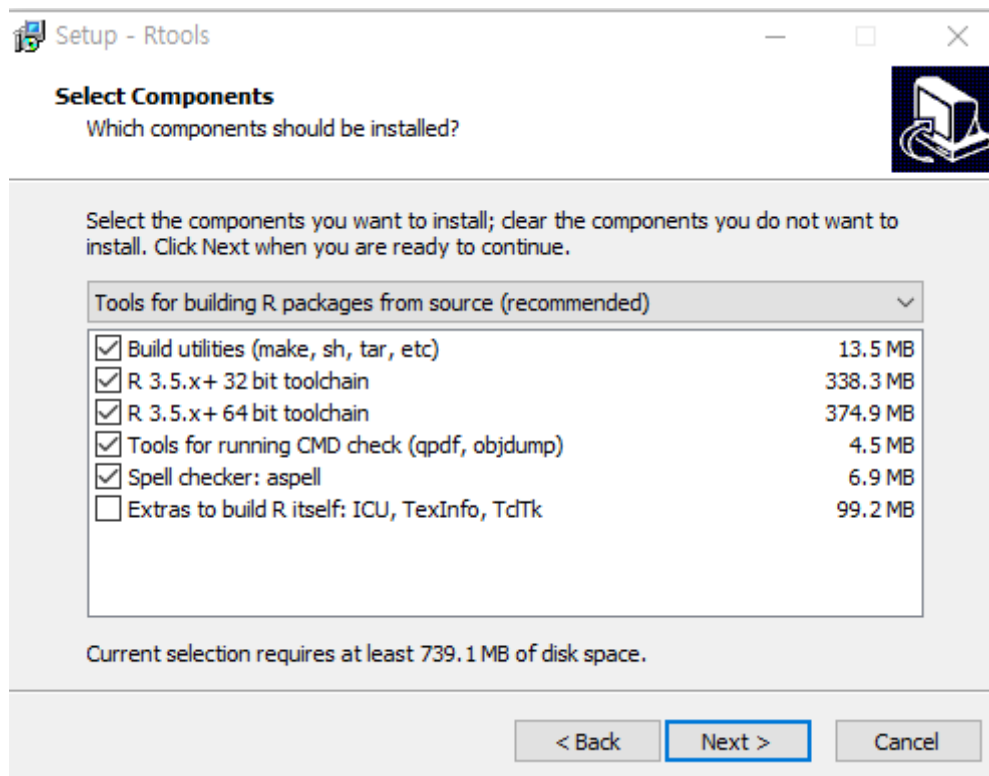
The authoritative source of information for tools to work with the current release of R is the "R Administration and Installation" manual. In particular, please read the ["Windows Toolset" appendix](#).

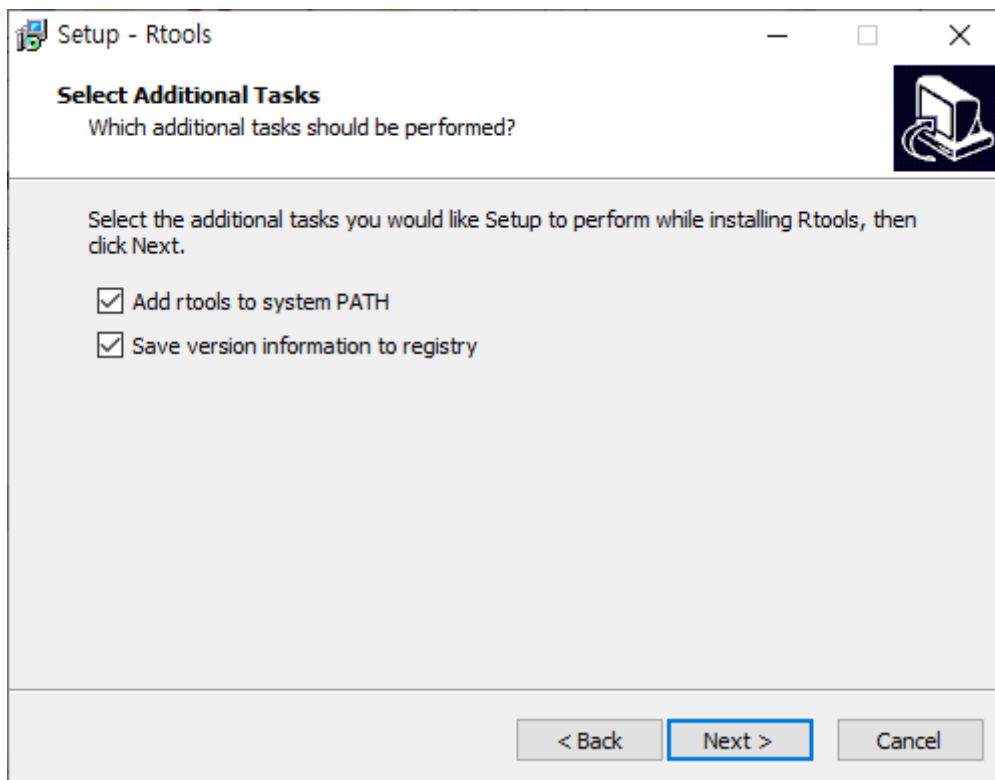
Rtools Downloads

Some of the tools are incompatible with obsolete versions of R. We maintain one actively updated version of the tools, and other "frozen" snapshots of them. We recommend that users use the latest release of Rtools with the latest release of R.

The current version of this file is recorded here: [VERSION.txt](#).

Download	R compatibility	Frozen?
Rtools40 (experimental)	Special R-testing build only, see documentation	-
Rtools35.exe (recommended)	R 3.3.x and later	No
Rtools34.exe	R 3.3.x and later	Yes
Rtools33.exe	R 3.2.x to 3.3.x	Yes
Rtools32.exe	R 3.1.x to 3.2.x	Yes
Rtools31.exe	R 3.0.x to 3.1.x	Yes





패키지 설치

- 패키지명 대소문자 구분하기
- KoNLP를 사용하기 위해서 rJava설치가 정상적으로 되는 것이 필요.
- rJava가 설치할때, Rtools가 설치가 되어야 정상적으로 동작함.
- 설치 후, 정상동작을 위해 r을 종료 후, 재실행

```
install.packages("rJava")  
install.packages("memoise")  
install.packages("KoNLP")
```

```
library(rJava)  
library(memoise)  
library(KoNLP)
```

```
### 사전 설정하기  
useNLADic()
```

```
### 데이터 준비하기(스파이더맨 데이터)  
txt <- readLines("D:\\dataset\\movieText\\SpiderMan.txt")  
head(txt)
```

```
### 문자열 처리하기  
install.packages("stringr")  
library(stringr)
```

```
### 특수문자 제거  
txt <- str_replace_all(txt, "www", " ")
```

```
txt
```

```
### 명사 추출
### KoNLP의 extractNoun()를 이용
extractNoun("오늘은 즐거운 날이다. 당신은 소중한 사람입니다.")
nouns <- extractNoun(txt)
wordCount <- table(unlist(nouns))

### 데이터 프레임 전환
df_word <- as.data.frame(wordCount, stringsAsFactors = F)

### 변수명 수정
library(dplyr)
df_word <- rename(df_word, word=Var1, freq=Freq)
df_word

### 두글자 이상 단어 추출
df_word <- filter(df_word, nchar(word) >= 2)
df_word
```

워드 클라우드(WordCloud)

```
####워드 클라우드
install.packages("wordcloud")

## 패키지 로드
library(wordcloud)
library(RColorBrewer)

## 색상 추출
pal <- brewer.pal(8, "Dark2")
set.seed(1004)

## 워드 클라우드 제작
wordcloud(word = df_word$word, # 단어
          freq = df_word$freq, # 빈도
          min.freq = 2,        # 최소 단어 빈도
          max.words = 100,     # 최대 표현 단어 수
          random.order = F,    # 고빈도 단어 중앙 배치(F: 하지 않음.)
          rot.per = 0.1,      # 회전 단어 비율 지정
          scale = c(5, 0.2),  # 단어 크기 범위
          colors = pal)       # 색상 목록
```



- 많이 사용된 단어일수록 글자가 크고 가운데 배치된다.
- 덜 사용된 단어일수록 글자가 작다.

REF

- KoNLP 깃허브 : <https://github.com/haven-jeon/KoNLP> (<https://github.com/haven-jeon/KoNLP>)

In []: