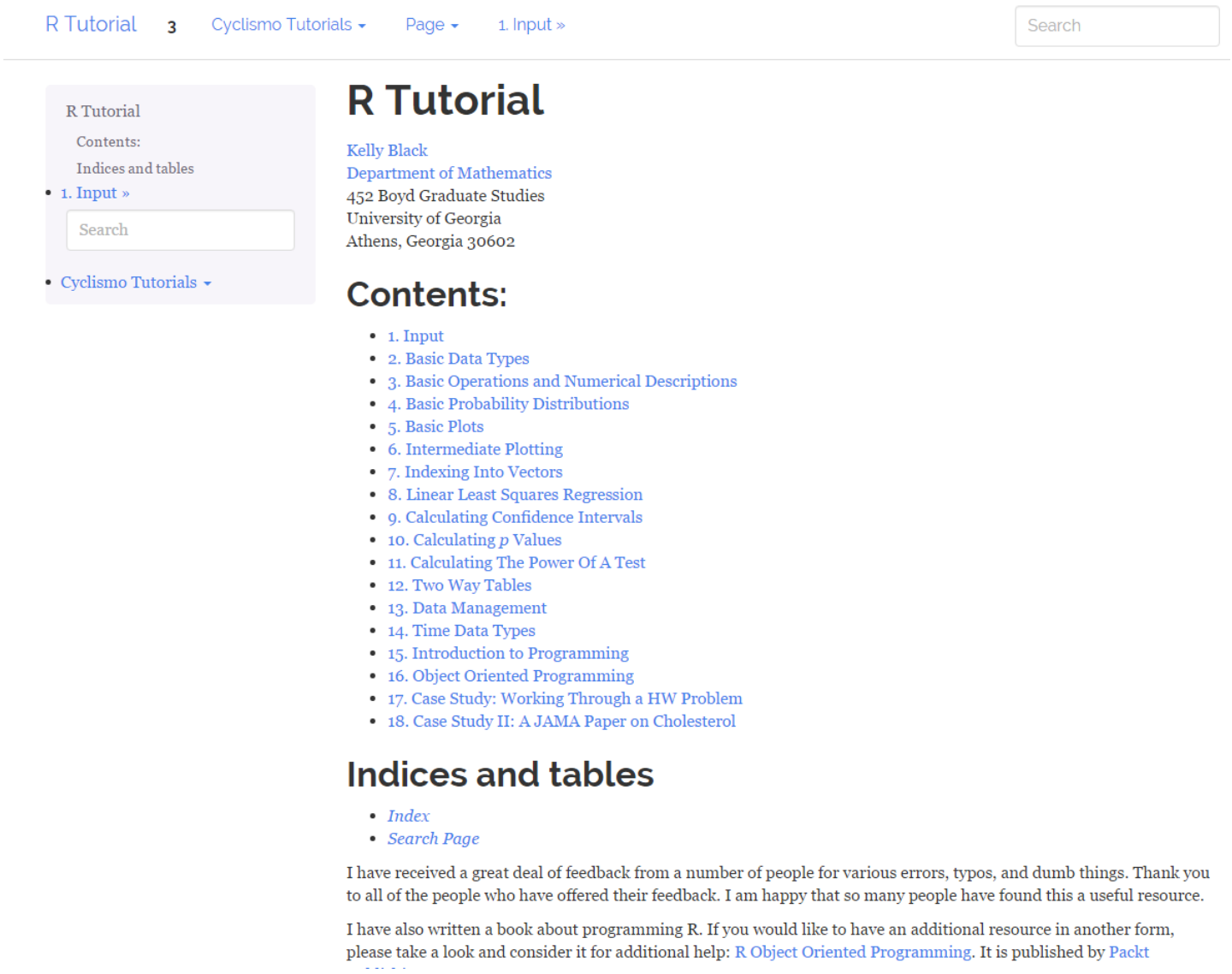
R 추가 자료 및 Weka 소개

October 7th, 2015

SNU Data Mining Center
Han Kyul Kim

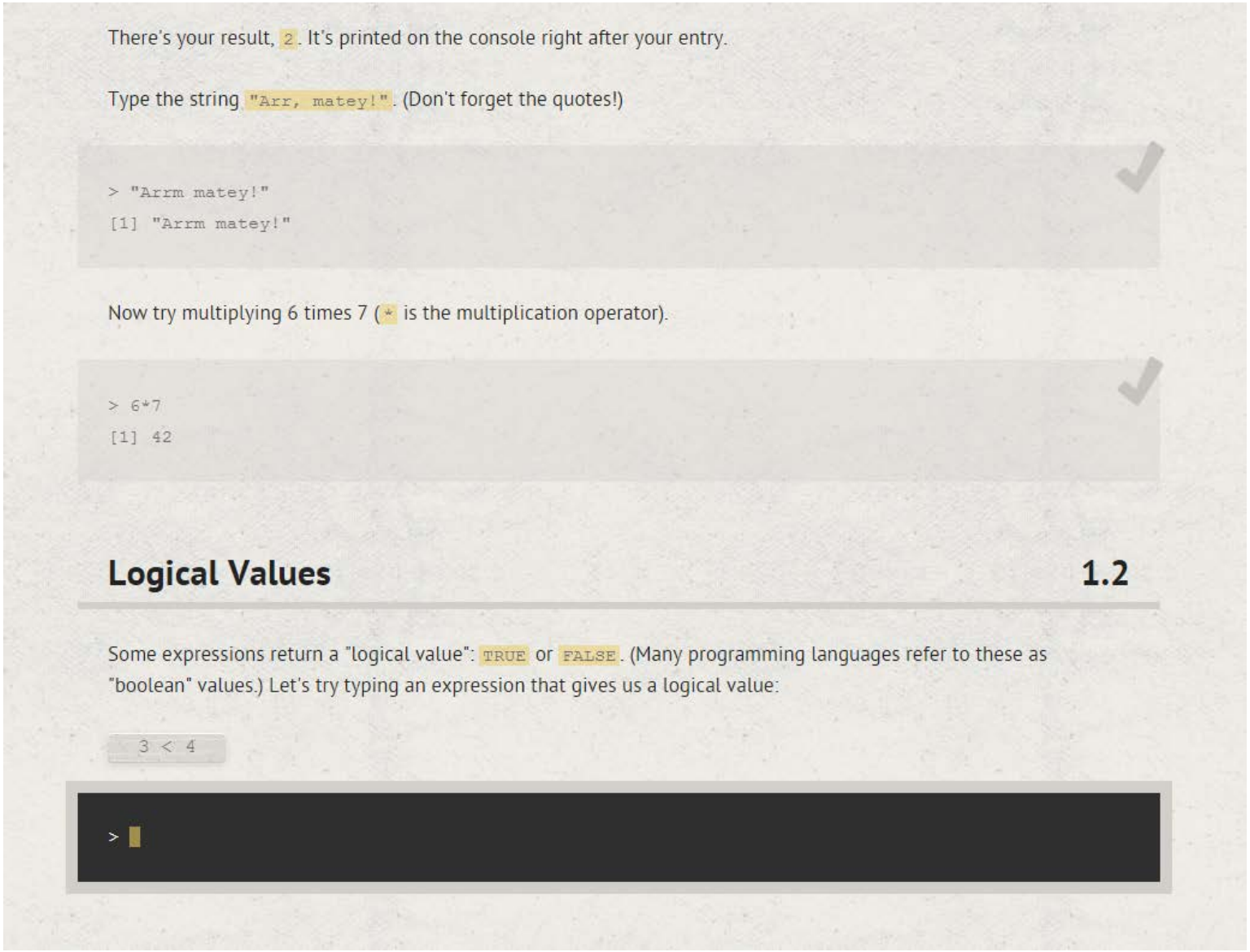
기본적인 사용법 메뉴얼

http://www.cyclismo.org/tutorial/R/



기본적인 사용법 실습

http://tryr.codeschool.com/



There's your result, `2`. It's printed on the console right after your entry.

Type the string `"Arr, matey!"`. (Don't forget the quotes!)

```
> "Arrm matey!"
[1] "Arrm matey!"
```

Now try multiplying 6 times 7 (`*` is the multiplication operator).

```
> 6*7
[1] 42
```

Logical Values 1.2

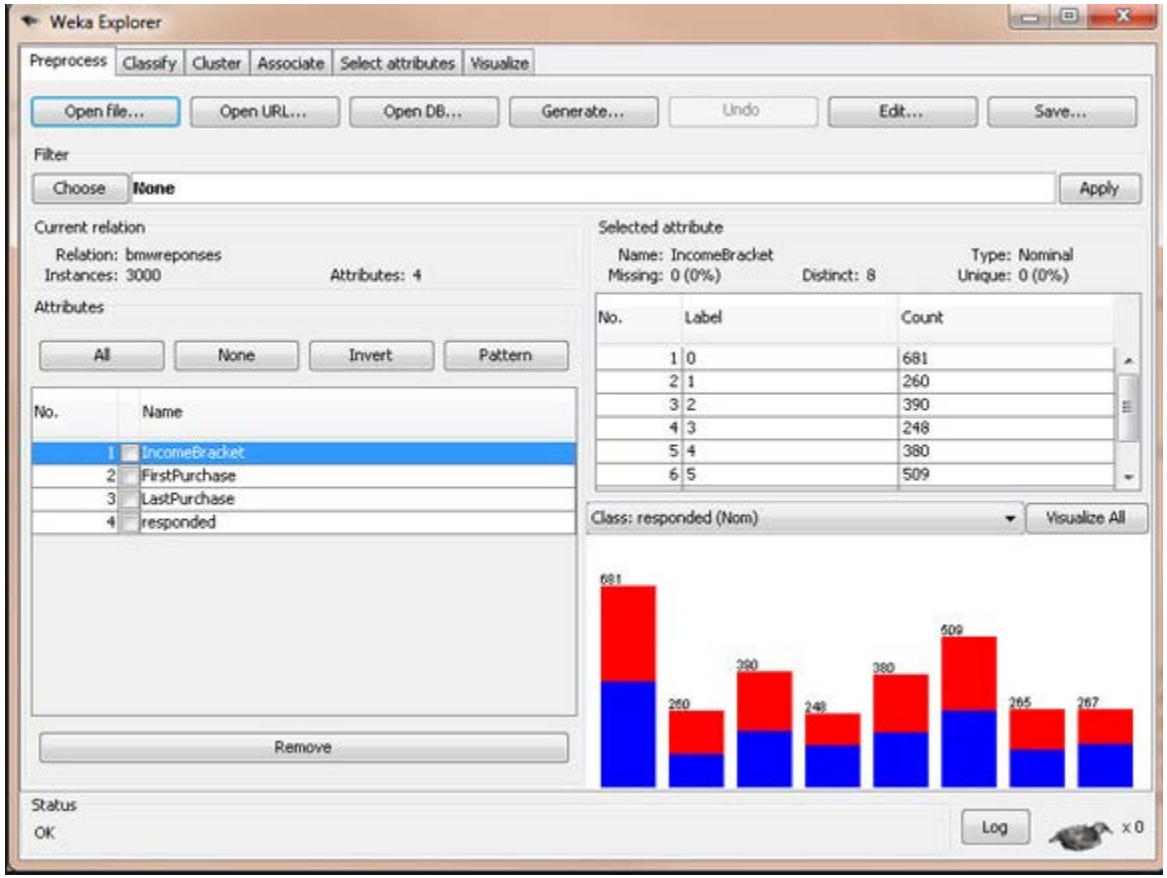
Some expressions return a "logical value": `TRUE` or `FALSE`. (Many programming languages refer to these as "boolean" values.) Let's try typing an expression that gives us a logical value:

```
3 < 4
```

```
>
```

Weka (Waikato Environment for Knowledge Analysis)

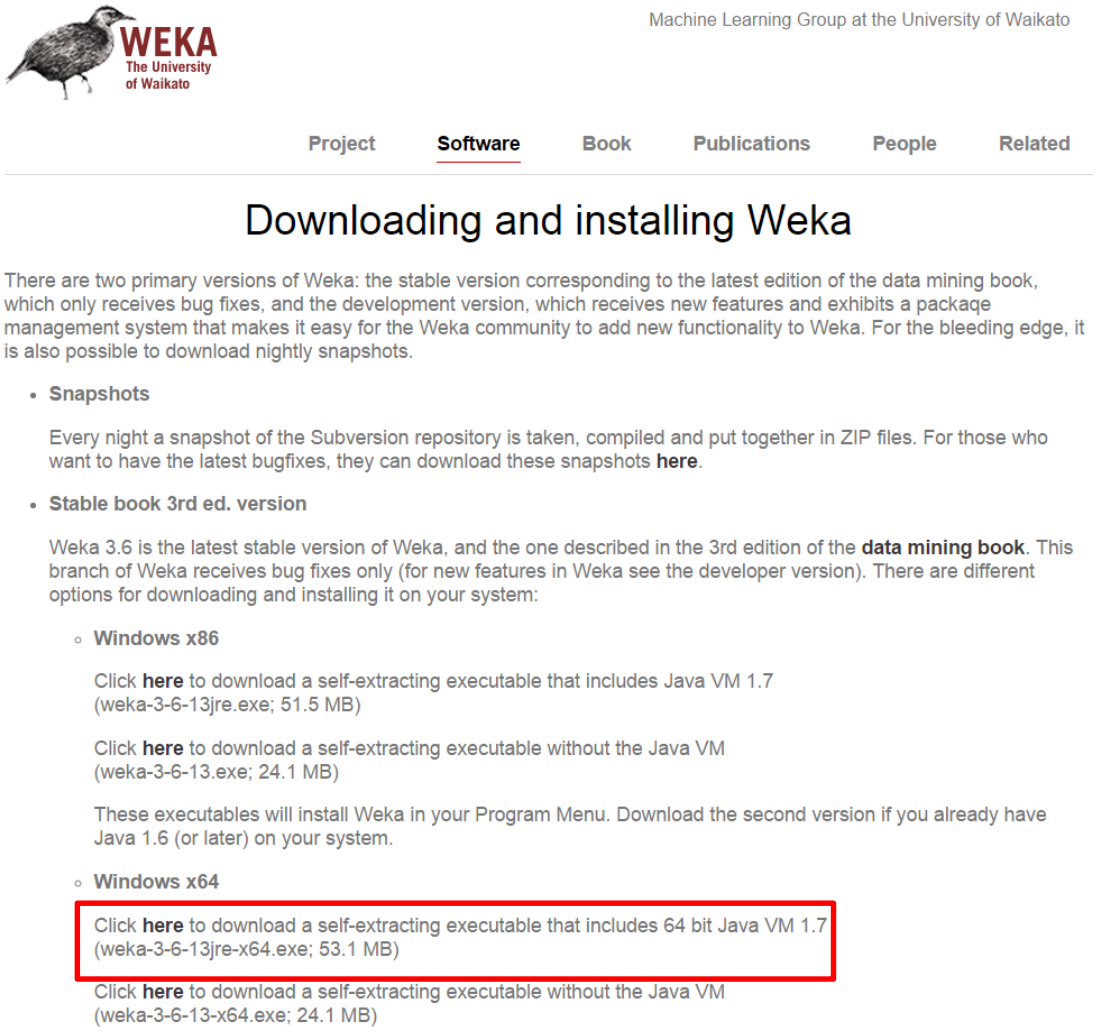
- 뉴질랜드 University of Waikato에서 개발했으며 GNU 라이선스로 공개된 데이터 분석 / 머신 러닝 소프트웨어
- Java 기반으로 만들어졌으나 사용하기 쉬운 GUI 버전 또한 있으니 코딩 없이도 간단한 분석이 가능



Weka 사용법

1. http://www.cs.waikato.ac.nz/ml/weka/downloading.html 접속

- Java 1.6 이상을 설치하지 않았다면 " self-extracting executable that includes 64 bit Java VM 1.7" 다운



WEKA
The University of Waikato

Machine Learning Group at the University of Waikato

Project | Software | Book | Publications | People | Related

Downloading and installing Weka

There are two primary versions of Weka: the stable version corresponding to the latest edition of the data mining book, which only receives bug fixes, and the development version, which receives new features and exhibits a package management system that makes it easy for the Weka community to add new functionality to Weka. For the bleeding edge, it is also possible to download nightly snapshots.

- **Snapshots**

  Every night a snapshot of the Subversion repository is taken, compiled and put together in ZIP files. For those who want to have the latest bugfixes, they can download these snapshots **here**.

- **Stable book 3rd ed. version**

  Weka 3.6 is the latest stable version of Weka, and the one described in the 3rd edition of the **data mining book**. This branch of Weka receives bug fixes only (for new features in Weka see the developer version). There are different options for downloading and installing it on your system:

  - **Windows x86**

    Click **here** to download a self-extracting executable that includes Java VM 1.7
    (weka-3-6-13jre.exe; 51.5 MB)

    Click **here** to download a self-extracting executable without the Java VM
    (weka-3-6-13.exe; 24.1 MB)

    These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.6 (or later) on your system.

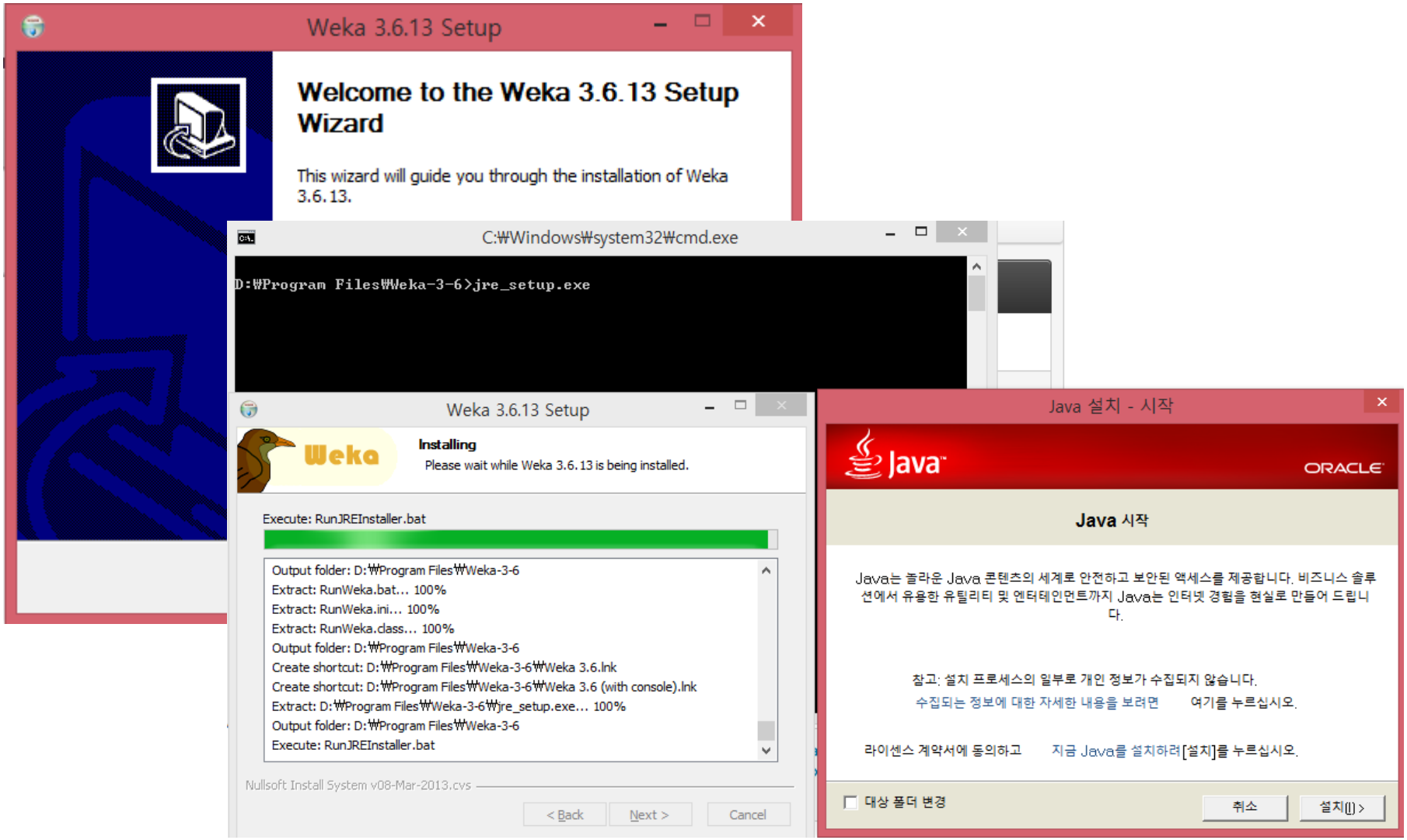
  - **Windows x64**

    Click **here** to download a self-extracting executable that includes 64 bit Java VM 1.7
    (weka-3-6-13jre-x64.exe; 53.1 MB)

    Click **here** to download a self-extracting executable without the Java VM
    (weka-3-6-13-x64.exe; 24.1 MB)

2. 파일 설치

- 자바가 없다면 자바 또한 설치

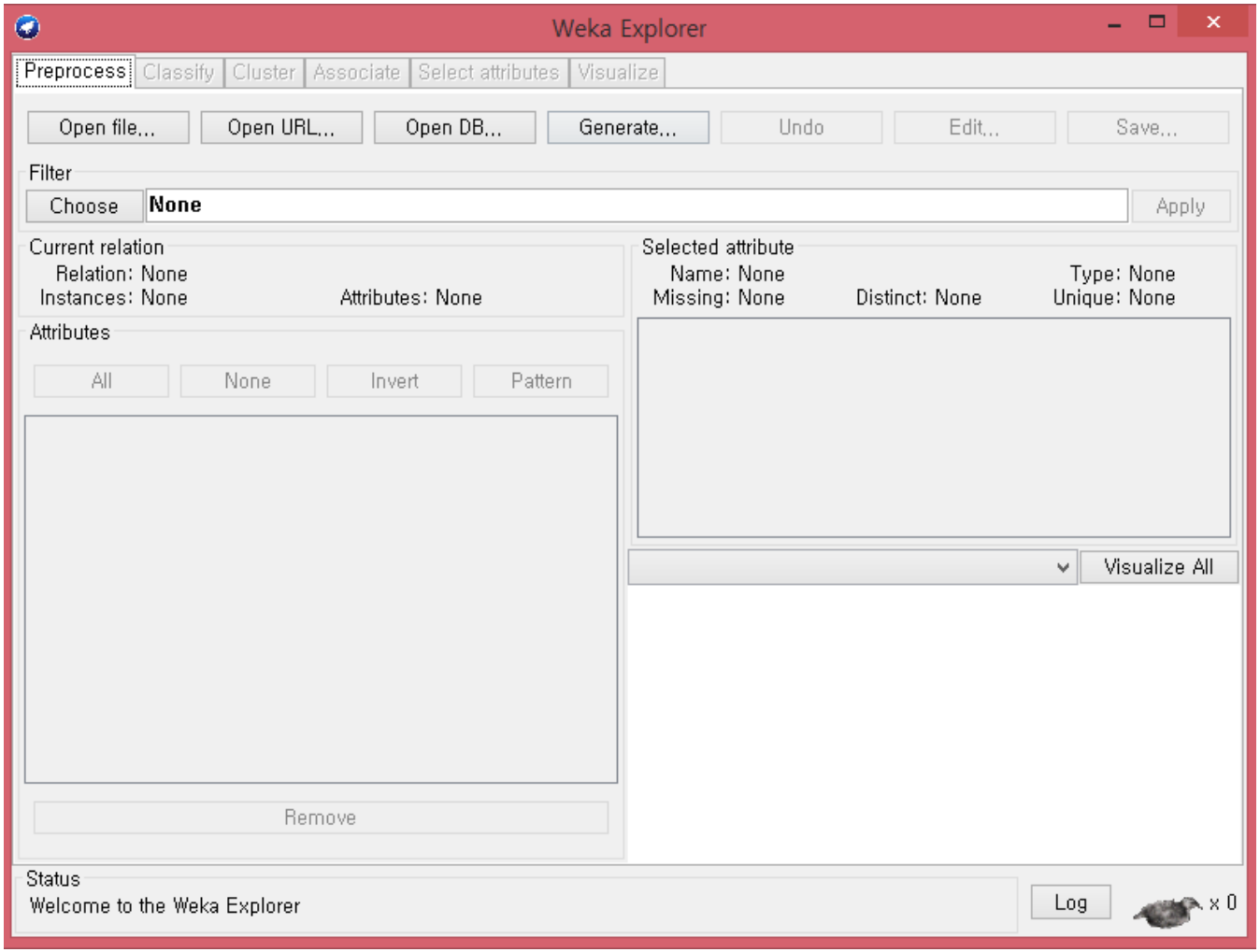


3. 파일 실행 후 Explorer 선택

Weka 사용법

4. 데이터를 불러온 후 원하는 분석 도구를 선택해서 사용

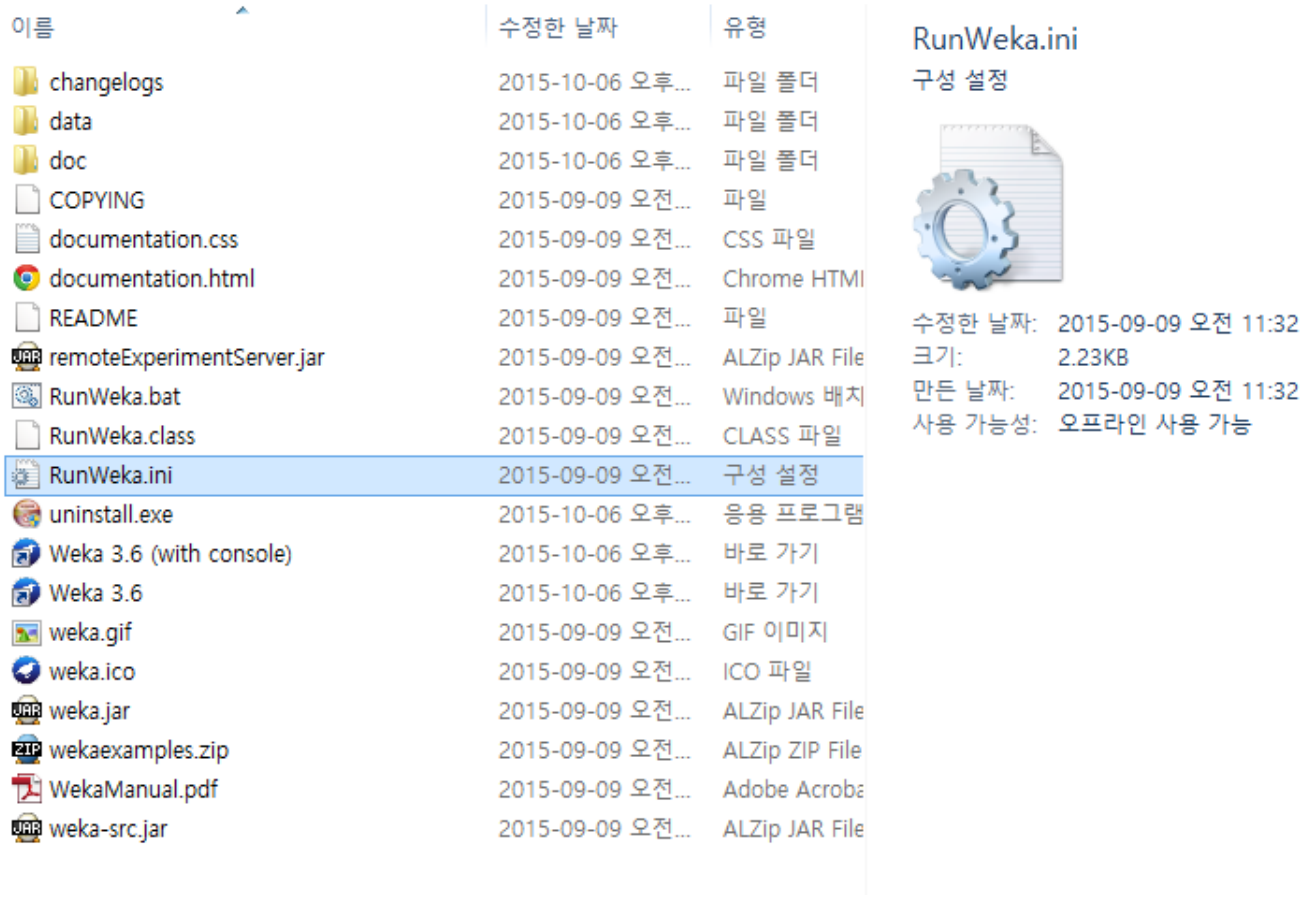
- 분석 가능 task: classification, clustering, association, variable selection, visualization, tree models



Weka 사용법

5. Java의 메모리 제한 설정 해제

- Weka는 Java 기반으로 만들어졌기 때문에 기본적인 setting에서는 한 번에 컴퓨터의 메모리에 올릴 수 있는 데이터의 양이 제한되어 있음(512 MB ~ 1 GB)
- 설정을 바꿔서 메모리 제한을 풀 수 있음 (더 큰 데이터를 다룰 수 있게 됨)
- **Weka가 설치되어 있는 폴더로 이동 후 메모장을 사용해서 "RunWeka.ini" 파일을 연다**



5. Java의 메모리 제한 설정 해제

- “maxheap”에 들어있는 값을 할당하고 싶은 메모리 양으로 바꿔줌

```
#   key can only be listed ONCE.
#
# Author  FracPete (fracpete at waikato dot ac dot nz)
# Version $Revision: 1.3 $

# setups (prefixed with "cmd_")
cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxheap# #javaOpts# -classpath "#wekaja
cmd_console=cmd.exe /K start cmd.exe /K "java -Dfile.encoding=#fileEncoding# -Xmx#maxheap# #j
cmd_explorer=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap# #javaOpts# -classpath "#wekaja
cmd_knowledgeFlow=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap# #javaOpts# -classpath "#w

# placeholders ("#bla#" in command gets replaced with content of key "bla")
# Note: "#wekajar#" gets replaced by the launcher class, since that jar gets
#       provided as parameter
maxheap=1024M
# The MDI GUI
#mainclass=weka.gui.Main
# The GUIChooser
mainclass=weka.gui.GUIChooser
# The file encoding; use "utf-8" instead of "Cp1252" to display UTF-8 characters in the
# GUI, e.g., the Explorer
fileEncoding=Cp1252
# The JAVA_OPTS environment variable (if set). Can be used as an alternative way to set
# the heap size (or any other JVM option)
javaOpts=%JAVA_OPTS%
# The classpath placeholder. Add any environment variables or jars to it that
# you need for your Weka environment.
# Example with an enviroment variable (e.g., THIRD_PARTY_LIBS):
#   cp=%CLASSPATH%;%THIRD_PARTY_LIBS%
# Example with an extra jar (located at D:\libraries\libsvm.jar):
#   cp=%CLASSPATH%;D:\\\\libraries\\\\libsvm.jar
# Or in order to avoid quadrupled backslashes, you can also use slashes "/":
#   cp=%CLASSPATH%;D:/libraries/libsvm.jar
cp=%CLASSPATH%
```

Weka Tutorial

http://www.cs.waikato.ac.nz/ml/weka/documentation.html
https://www.youtube.com/watch?v=m7kpIBGEdkI

Documentation

For an overview of the techniques implemented in Weka, and the software itself, you may want to consider taking a look at the **data mining book**. However, there is a large amount of freely available information as well. Weka 3.6.x and 3.7.x have extensive help facilities built in and come with a comprehensive manual.

- General documentation
  - The **Weka Wiki**, including **frequently asked questions (FAQ)**, help on **trouble-shooting** Weka, and information on **using the package manager** and **how to structure packages** in Weka 3.7.
  - The Weka **mailing list** (**archive**).
  - **Community documentation** for Weka at **Pentaho**.
  - The Weka manual (**Weka 3.6.13**, **Weka 3.7.13**), as included in the distribution.
  - The Weka API, extracted from the Javadoc (**Weka 3.6**,**Weka 3.7**), as included in the distribution.
  - **A list of packages** for Weka 3.7 that can be installed via its package manager. Corresponding Javadoc for the packages is available at http://weka.sourceforge.net/doc.packages/[name of package].
- Miscellaneous information
  - **Mark Hall's Weka-related blog**
  - A **presentation** entitled "WEKA in the Ecosystem for Scientific Computing". Covers how to access the WEKA data mining software from Octave/Matlab, R, and Python. Also discusses how some R functionality can be applied from within Weka and facilities for distributed computation in Weka.

    A **tutorial on connecting Weka to MongoDB using a JDBC driver.**
  - **A somewhat outdated, but still useful introduction**, written by Alex K. Seewald, to using Weka 3.4.6 from the command line.
  - A **presentation** demonstrating all graphical user interfaces (GUI) in Weka. (Warning: large file.)

HW: 목욕비누 구매자 세분화 (18. 4)

광고/프로모션 전략을 세우기 위해 고객을 군집화하려고 한다. K-means clustering을 이용해 군집화하라.

1. **클러스터링을 수행할 때는, 사용하는 변수의 종류에 따라 결과가 크게 영향을 받는다. 자동으로 변수를 선택하지 못하고 모든 변수를 활용해서 거리를 계산하기 때문이다. 그래서 분석가가 판단하여 관련 없는 변수를 제거하거나 의미있는 변수를 생성하는 것은 중요하다.**
   1) Member ID와 Demographics 에 해당하는 변수들은 클러스터링 수행 시에는 제외하자. 추후 각 클러스터의 특성을 들여다 볼 때는 함께 볼 수도 있다.
   2) brand wise volume에 해당하는 9개의 변수를 대신하여, '브랜드 충성도'라는 유도 변수(derived variable) 한 개를 생성해보자. 여러 브랜드가 아닌 특정 브랜드에 몰려 있을수록 큰 숫자가 나올 수 있도록 자유롭게 만들어 보자.

2. **적절한 K값을 정해보자. 3~5개의 광고/프로모션 전략이 가능하다는 제약조건이 있다. 즉 K는 3~5 사이의 숫자 중에 결정하자.**
   1) 데이터를 normalize하라. K-means clustering에는 거리 계산하는 과정이 포함되므로 normalize과정이 꼭 필요하다.
   2) k=3로 설정해서 k-means clustering을 수행해라.
   3) silhouette 결과 값을 확인하라.
   4) k를 4,5로 바꿔가며 silhouette 결과 값을 확인하라. 그리고 최적의 k값을 선정하자.

3. **선정된 K로 클러스터링 된 결과를 이용해서, 각 군집의 특징을 파악해보자. 특징을 기반으로 군집별로 광고/프로모션 전략을 세우려고 한다. 각 군집은 어떤 특징을 가지고 있는지 기술하라. (Spotfire 활용 가능)**

Clustering with R

```
library(cluster)
kc <- kmeans(pb, centers = 3);

str(kc)
kc$centers
kc$size
kc$cluster

dissE<-daisy(pb)
sk<-silhouette(kc$cl,dissE)

pdf('D:/DM_TA/my_nice_plot.pdf')
plot(sk)
dev.off()
```