An Introduction to the WEKA Data Mining System

Zdravko Markov
Central Connecticut State University
markovz@ccsu.edu

Ingrid Russell
University of Hartford
irussell@hartford.edu

Data Mining

- **"Drowning in Data yet Starving for Knowledge"**
  ???

- **"Computers have promised us a fountain of wisdom but delivered a flood of data"**
  William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus

- ***Data Mining: "The non trivial extraction of implicit, previously unknown, and potentially useful information from data"***
  William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus

- **Data mining finds valuable information hidden in large volumes of data.**

- **Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.**

- **Data Mining is an interdisciplinary field involving:**
  - Databases
  - Statistics
  - Machine Learning
  - High Performance Computing
  - Visualization
  - Mathematics

Data Mining Software

KDnuggets : Polls : Data Mining Tools You Used in 2005 (May 2005) PollData mining/Analytic tools you used in 2005 [376 voters, 860 votes total]

- **Enterprise-level: (US $10,000 and more)** Fair Isaac, IBM, Insightful, KXEN, Oracle, SAS, and SPSS
- **Department-level: (from $1,000 to $9,999)** Angoss, CART/MARS/TreeNet/Random Forests, Equbits, GhostMiner, Gornik, Mineset, MATLAB, Megaputer, Microsoft SQL Server, Statsoft Statistica, ThinkAnalytics
- **Personal-level: (from $1 to $999)**: Excel, See5
- **Free**: C4.5, R, Weka, Xelopes

Poll: Data Mining Tools You Used in 2005 - Microsoft Internet Explorer

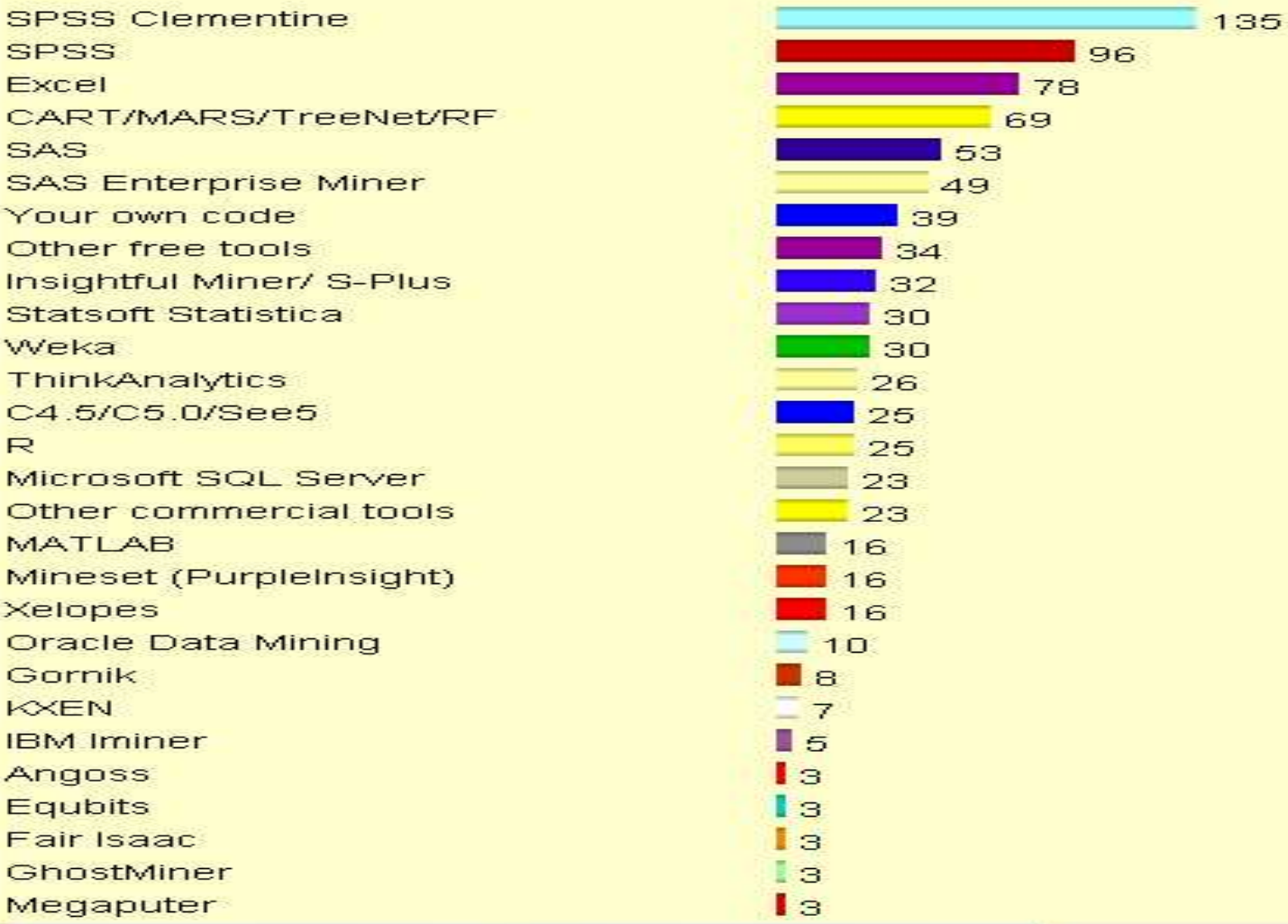
File Edit View Favorites Tools Help

Address http://www.kdnuggets.com/polls/2005/data_mining_tools.htm Go Links

KDnuggets : Polls : Data Mining Tools You Used in 2005 (May 2005)

Poll

Data mining/Analytic tools you used in 2005 [376 voters, 860 votes total]



Weka Data Mining Software

KDnuggets : News : 2005 : n13 : item2

SIGKDD Service Award is the highest service award in the field of data mining and knowledge discovery. It is is given to one individual or one group who has performed significant service to the data mining and knowledge discovery field, including professional volunteer services in disseminating technical information to the field, education, and research funding.

The **2005 ACM SIGKDD Service Award** is presented to **the Weka team** for their development of the freely-available Weka Data Mining Software, including the accompanying book Data Mining: Practical Machine Learning Tools and Techniques (now in second edition) and much other documentation.

The Weka team includes **Ian H. Witten** and **Eibe Frank**, and the following major contributors (in alphabetical order of last names): Remco R. Bouckaert, John G. Cleary, Sally Jo Cunningham, Andrew Donkin, Dale Fletcher, Steve Garner, Mark A. Hall, Geoffrey Holmes, Matt Humphrey, Lyn Hunt, Stuart Inglis, Ashraf M. Kibriya, Richard Kirkby, Brent Martin, Bob McQueen, Craig G. Nevill-Manning, Bernhard Pfahringer, Peter Reutemann, Gabi Schmidberger, Lloyd A. Smith, Tony C. Smith, Kai Ming Ting, Leonard E. Trigg, Yong Wang, Malcolm Ware, and Xin Xu.

The Weka team has put a tremendous amount of effort into continuously developing and maintaining the system **since 1994**. The development of Weka was funded by a grant from the New Zealand Government's Foundation for Research, Science and Technology.

The **key features** responsible for Weka's success are:

- it provides many different algorithms for data mining and machine learning
- is is open source and freely available
- it is platform-independent
- it is easily useable by people who are not data mining specialists
- it provides flexible facilities for scripting experiments
- it has kept up-to-date, with new algorithms being added as they appear in the research literature.

Weka Data Mining Software

KDnuggets : News : 2005 : n13 : item2 (cont.)

The Weka Data Mining Software has been downloaded **200,000 times** since it was put on SourceForge in April 2000, and is currently downloaded at a rate of 10,000/month. The Weka mailing list has over **1100 subscribers in 50 countries**, including subscribers from many major companies.

There are **15 well-documented substantial projects** that incorporate, wrap or extend Weka, and no doubt many more that have not been reported on Sourceforge.

Ian H. Witten and Eibe Frank also wrote a **very popular book "Data Mining: Practical Machine Learning Tools and Techniques"** (now in the second edition), that seamlessly integrates Weka system into teaching of data mining and machine learning. In addition, they provided **excellent teaching material** on the book website.

This book became one of the most popular textbooks for data mining and machine learning, and is **very frequently cited in scientific publications**.

Weka is a **landmark system in the history of the data mining and machine learning** research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time (the first version of Weka was released 11 years ago). Other data mining and machine learning systems that have achieved this are individual systems, such as C4.5, not toolkits.

Since Weka is freely available for download and offers many powerful features (sometimes not found in commercial data mining software), it has become one of the most widely used data mining systems. Weka also became one of the favorite vehicles for data mining research and helped to advance it by making many powerful features available to all.

In sum, the Weka team has made an outstanding contribution to the data mining field.

Machine Learning Project - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.cs.waikato.ac.nz/~ml/index.html Go Links

WEKA
The University of Waikato

Project

project · software · book · publications · people · related

Interested in doing an MSc or PhD in Machine Learning here at Waikato and spending some time overseas while working on your project? Then check this out.

Weka Machine Learning Project

An exciting and potentially far-reaching development in computer science is the invention and application of methods of machine learning. These enable a computer program to automatically analyse a large body of data and decide what information is most relevant. This crystallised information can then be used to automatically make predictions or to help people make decisions faster and more accurately.

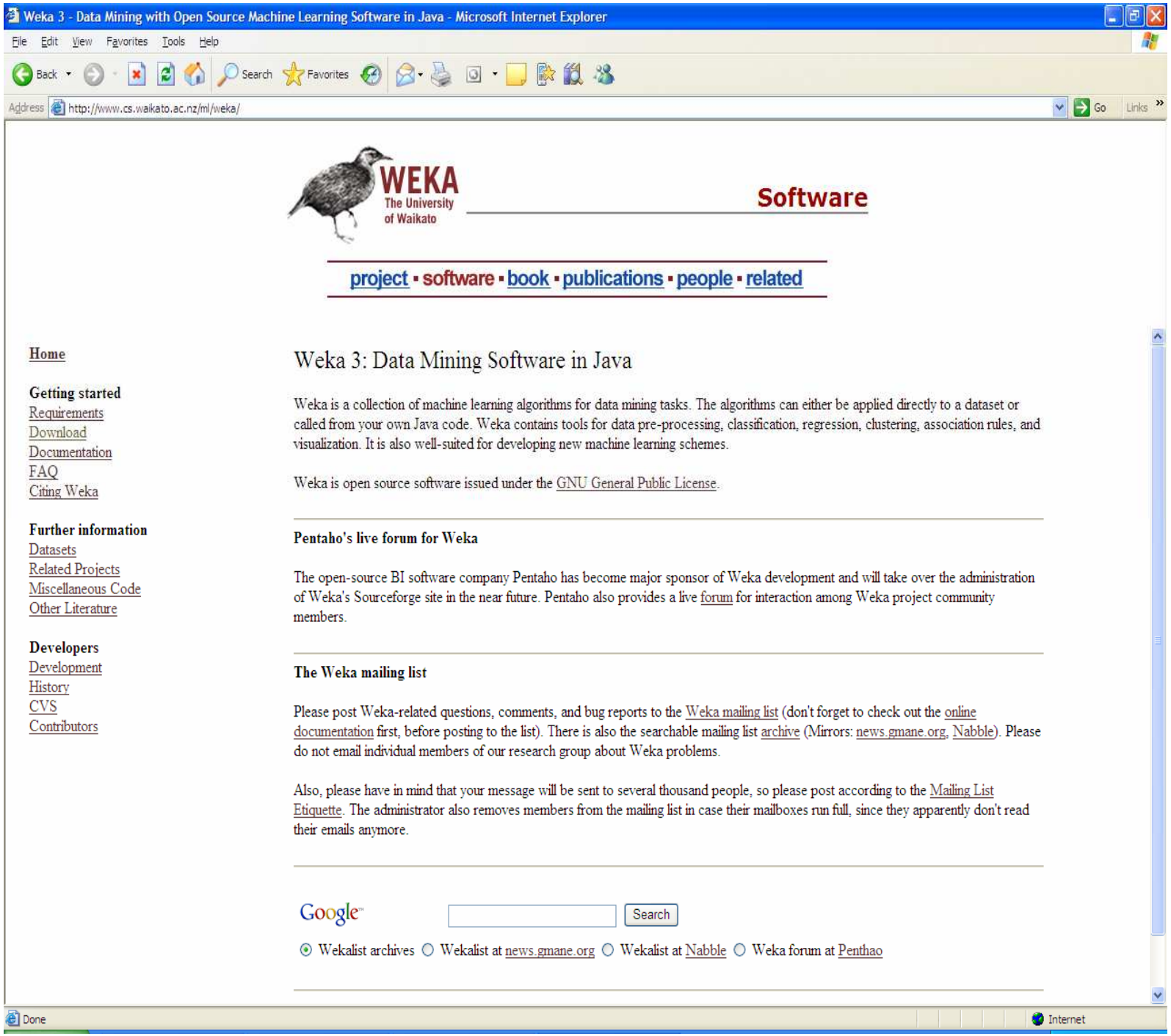
The overall goal of our project is to build a state-of-the-art facility for developing machine learning (ML) techniques and to apply them to real-world data mining problems. Our team has incorporated several standard ML techniques into a software "workbench" called WEKA, for Waikato Environment for Knowledge Analysis. With it, a specialist in a particular field is able to use ML to derive useful knowledge from databases that are far too large to be analysed by hand. WEKA's users are ML researchers and industrial scientists, but it is also widely used for teaching.

Our objectives are to

- make ML techniques generally available;
- apply them to practical problems that matter to New Zealand industry;
- develop new machine learning algorithms and give them to the world;
- contribute to a theoretical framework for the field.

Our machine learning package is publically available and presents a collection of algorithms for solving real-world data mining problems. The software is written entirely in Java and includes a uniform interface to a number of standard ML techniques. Please feel free to browse around.

Found only on the islands of New Zealand, the weka is a flightless bird with an inquisitive nature. (How should you pronounce WEKA? What does the weka sound like.)


Weka 3 - Data Mining with Open Source Machine Learning Software in Java - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.cs.waikato.ac.nz/ml/weka/
Go Links
WEKA
The University of Waikato
Software
project · software · book · publications · people · related
Home
Getting started
Requirements
Download
Documentation
FAQ
Citing Weka
Further information
Datasets
Related Projects
Miscellaneous Code
Other Literature
Developers
Development
History
CVS
Contributors
Weka 3: Data Mining Software in Java
Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.
Weka is open source software issued under the GNU General Public License.
Pentaho's live forum for Weka
The open-source BI software company Pentaho has become major sponsor of Weka development and will take over the administration of Weka's Sourceforge site in the near future. Pentaho also provides a live forum for interaction among Weka project community members.
The Weka mailing list
Please post Weka-related questions, comments, and bug reports to the Weka mailing list (don't forget to check out the online documentation first, before posting to the list). There is also the searchable mailing list archive (Mirrors: news.gmane.org, Nabble). Please do not email individual members of our research group about Weka problems.
Also, please have in mind that your message will be sent to several thousand people, so please post according to the Mailing List Etiquette. The administrator also removes members from the mailing list in case their mailboxes run full, since they apparently don't read their emails anymore.
Google
Search
Wekalist archives Wekalist at news.gmane.org Wekalist at Nabble Weka forum at Penthao
Done
Internet

Data Mining: Practical Machine Learning Tools and Techniques - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.cs.waikato.ac.nz/~ml/weka/book.html

WEKA The University of Waikato

Book

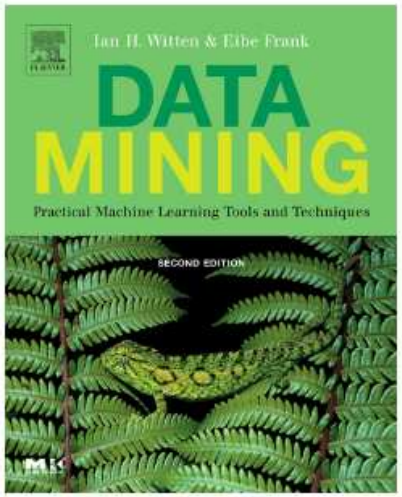
project ▪ software ▪ book ▪ publications ▪ people ▪ related

Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)

Ian H. Witten, Eibe Frank

Morgan
Kaufmann
June 2005
525 pages
Paper
ISBN 0-12-
088407-0

Eibe Frank and Ian Witten



Click here to order from Amazon.com

Comments

"If you have data that you want to analyze and understand, this book and the associated Weka toolkit are an excellent way to start."

-Jim Gray, Microsoft Research

Using Weka to teach Machine Learning, Data and Web Mining

http://uhaweb.hartford.edu/compsci/ccli/

Machine Learning, Data and Web Mining by Example ("learning by doing" approach)

- Data preprocessing and visualization
- Attribute selection
- Classification (OneR, Decision trees)
- Prediction (Nearest neighbor)
- Model evaluation
- Clustering (K-means, Cobweb)
- Association rules

Data preprocessing and visualization

Initial Data Preparation
(Weka data input)

- Raw data (Japanese loan data)
- Web/Text documents (Department data)

Data preprocessing and visualization

Japanese loan data (a sample from a loan history database of a Japanese bank)

Clients: s1,..., s20

- Approved loan: s1, s2, s4, s5, s6, s7, s8, s9, s14, s15, s17, s18, s19
- Rejected loan: s3, s10, s11, s12, s13, s16, s20

Clients data:

- unemployed clients: s3, s10, s12
- loan is to buy a personal computer: s1, s2, s3, s4, s5, s6, s7, s8, s9, s10
- loan is to buy a car: s11, s12, s13, s14, s15, s16, s17, s18, s19, s20
- male clients: s6, s7, s8, s9, s10, s16, s17, s18, s19, s20
- not married: s1, s2, s5, s6, s7, s11, s13, s14, s16, s18
- live in problematic area: s3, s5
- age: s1=18, s2=20, s3=25, s4=40, s5=50, s6=18, s7=22, s8=28, s9=40, s10=50, s11=18, s12=20, s13=25, s14=38, s15=50, s16=19, s17=21, s18=25, s19=38, s20=50
- money in a bank (x10000 yen): s1=20, s2=10, s3=5, s4=5, s5=5, s6=10, s7=10, s8=15, s9=20, s10=5, s11=50, s12=50, s13=50, s14=150, s15=50, s16=50, s17=150, s18=150, s19=100, s20=50
- monthly pay (x10000 yen): s1=2, s2=2, s3=4, s4=7, s5=4, s6=5, s7=3, s8=4, s9=2, s10=4, s11=8, s12=10, s13=5, s14=10, s15=15, s16=7, s17=3, s18=10, s19=10, s20=10
- months for the loan: s1=15, s2=20, s3=12, s4=12, s5=12, s6=8, s7=8, s8=10, s9=20, s10=12, s11=20, s12=20, s13=20, s14=20, s15=20, s16=20, s17=20, s18=20, s19=20, s20=30
- years with the last employer: s1=1, s2=2, s3=0, s4=2, s5=25, s6=1, s7=4, s8=5, s9=15, s10=0, s11=1, s12=2, s13=5, s14=15, s15=8, s16=2, s17=3, s18=2, s19=15, s20=2

Data preprocessing and visualization

Relations, attributes, tuples (instances)

LoanData - Notepad

File Edit Format View Help

```
ID,sex,married,age,money,pay,months,buy,emp,lastemp,area,approved
1,f,n,18,20,2,15,pc,y,1,good,y
2,f,n,20,10,2,20,pc,y,2,good,y
3,f,y,25,5,4,12,pc,n,0,bad,n
4,f,y,40,5,7,12,pc,y,2,good,y
5,f,n,50,5,4,12,pc,y,25,bad,y
6,m,n,18,10,5,8,pc,y,1,good,y
7,m,n,22,10,3,8,pc,y,4,good,y
8,m,y,28,15,4,10,pc,y,5,good,y
9,m,y,40,20,2,20,pc,y,15,good,
10,m,y,50,5,4,12,pc,n,0,good,n
11,f,n,18,50,8,20,car,y,1,good
12,f,y,20,50,10,20,car,n,2,goo
13,f,n,25,50,5,20,car,y,5,good
14,f,n,38,150,10,20,car,y,15,g
15,f,y,50,50,15,20,car,y,8,goo
16,m,n,19,50,7,20,car,y,2,good
17,m,y,21,150,3,20,car,y,3,goo
18,m,n,25,150,10,20,car,y,2,go
19,m,y,38,100,10,20,car,y,15,g
20,m,y,50,50,10,30,car,y,2,goo
```

Loan data – CVS format (LoanData.cvs)

Microsoft Excel - LoanData

File Edit View Insert Format Tools Data Window Help

Type a question for help

N23

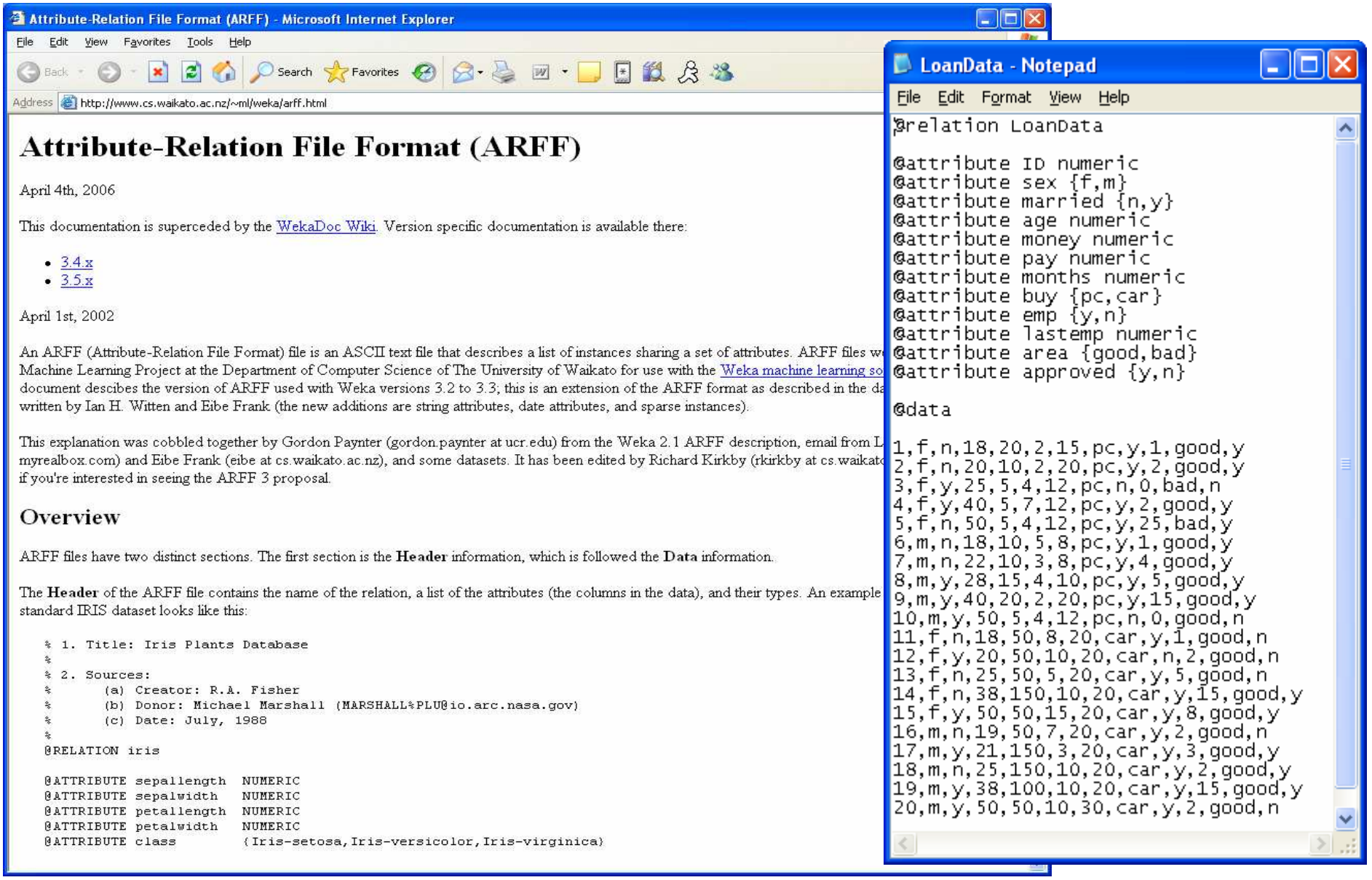
	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	sex	married	age	money	pay	months	buy	emp	lastemp	area	approved
2	1	f	n	18	20	2	15	pc	y	1	good	y
3	2	f	n	20	10	2	20	pc	y	2	good	y
4	3	f	y	25	5	4	12	pc	n	0	bad	n
5	4	f	y	40	5	7	12	pc	y	2	good	y
6	5	f	n	50	5	4	12	pc	y	25	bad	y
7	6	m	n	18	10	5	8	pc	y	1	good	y
8	7	m	n	22	10	3	8	pc	y	4	good	y
9	8	m	y	28	15	4	10	pc	y	5	good	y
10	9	m	y	40	20	2	20	pc	y	15	good	y
11	10	m	y	50	5	4	12	pc	n	0	good	n
12	11	f	n	18	50	8	20	car	y	1	good	n
13	12	f	y	20	50	10	20	car	n	2	good	n
14	13	f	n	25	50	5	20	car	y	5	good	n
15	14	f	n	38	150	10	20	car	y	15	good	y
16	15	f	y	50	50	15	20	car	y	8	good	y
17	16	m	n	19	50	7	20	car	y	2	good	n
18	17	m	y	21	150	3	20	car	y	3	good	y
19	18	m	n	25	150	10	20	car	y	2	good	y
20	19	m	y	38	100	10	20	car	y	15	good	y
21	20	m	y	50	50	10	30	car	y	2	good	n
22												

LoanData

Ready

Data preprocessing and visualization

Attribute-Relation File Format (ARFF) - http://www.cs.waikato.ac.nz/~ml/weka/arff.html



Attribute-Relation File Format (ARFF) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Attribute-Relation File Format (ARFF)

April 4th, 2006

This documentation is superceded by the WekaDoc Wiki. Version specific documentation is available there:

- 3.4.x
- 3.5.x

April 1st, 2002

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files w
Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning so
document descibes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the da
written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date attributes, and sparse instances).

This explanation was cobbled together by Gordon Paynter (gordon.paynter at ucr.edu) from the Weka 2.1 ARFF description, email from L
myrealbox.com) and Eibe Frank (eibe at cs.waikato.ac.nz), and some datasets. It has been edited by Richard Kirkby (rkirkby at cs.waikat
if you're interested in seeing the ARFF 3 proposal.

Overview

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types. An example
standard IRIS dataset looks like this:

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

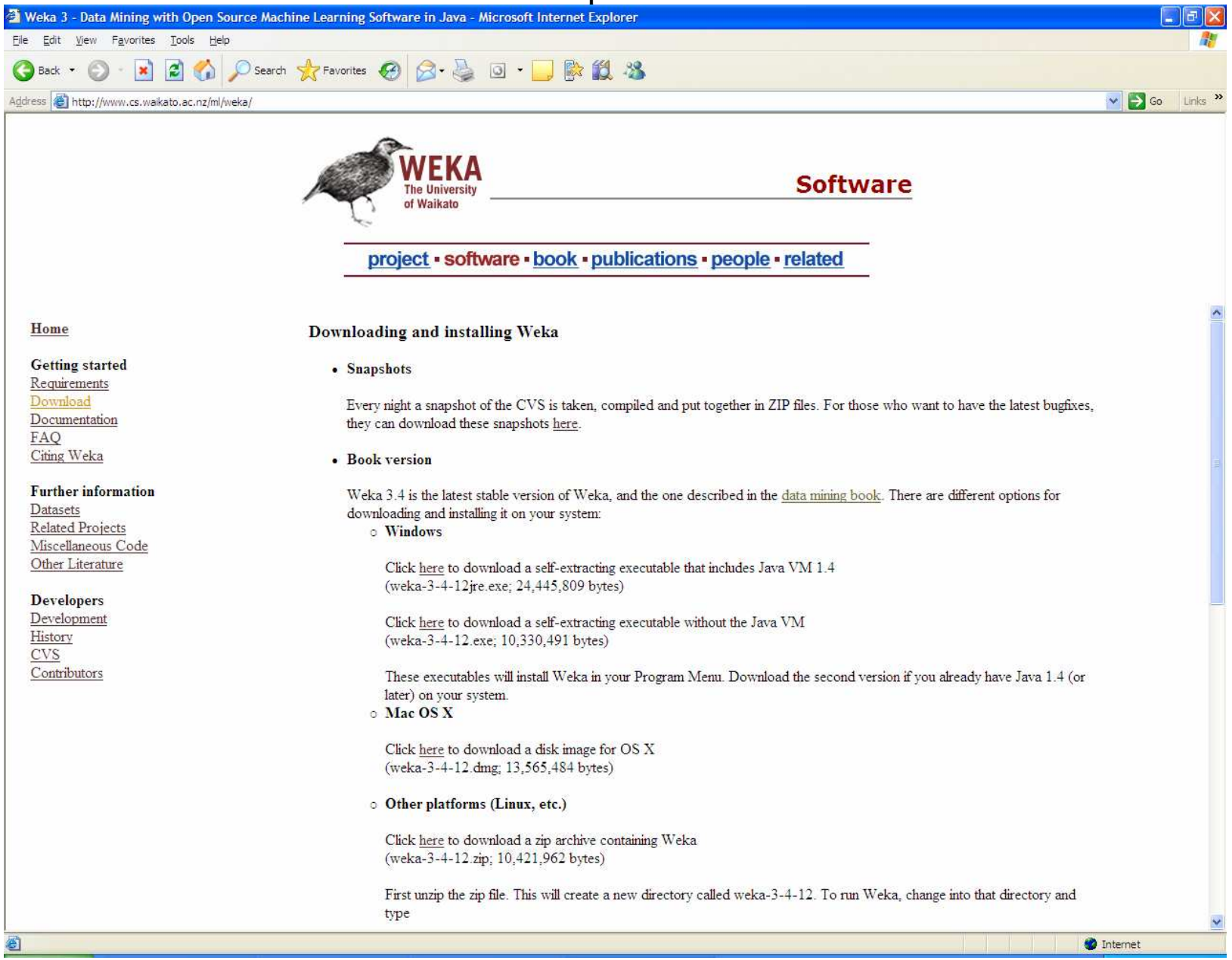
LoanData - Notepad

File Edit Format View Help

```
@relation LoanData

@attribute ID numeric
@attribute sex {f,m}
@attribute married {n,y}
@attribute age numeric
@attribute money numeric
@attribute pay numeric
@attribute months numeric
@attribute buy {pc,car}
@attribute emp {y,n}
@attribute lastemp numeric
@attribute area {good,bad}
@attribute approved {y,n}

@data

1,f,n,18,20,2,15,pc,y,1,good,y
2,f,n,20,10,2,20,pc,y,2,good,y
3,f,y,25,5,4,12,pc,n,0,bad,n
4,f,y,40,5,7,12,pc,y,2,good,y
5,f,n,50,5,4,12,pc,y,25,bad,y
6,m,n,18,10,5,8,pc,y,1,good,y
7,m,n,22,10,3,8,pc,y,4,good,y
8,m,y,28,15,4,10,pc,y,5,good,y
9,m,y,40,20,2,20,pc,y,15,good,y
10,m,y,50,5,4,12,pc,n,0,good,n
11,f,n,18,50,8,20,car,y,1,good,n
12,f,y,20,50,10,20,car,n,2,good,n
13,f,n,25,50,5,20,car,y,5,good,n
14,f,n,38,150,10,20,car,y,15,good,y
15,f,y,50,50,15,20,car,y,8,good,y
16,m,n,19,50,7,20,car,y,2,good,n
17,m,y,21,150,3,20,car,y,3,good,y
18,m,n,25,150,10,20,car,y,2,good,y
19,m,y,38,100,10,20,car,y,15,good,y
20,m,y,50,50,10,30,car,y,2,good,n
```

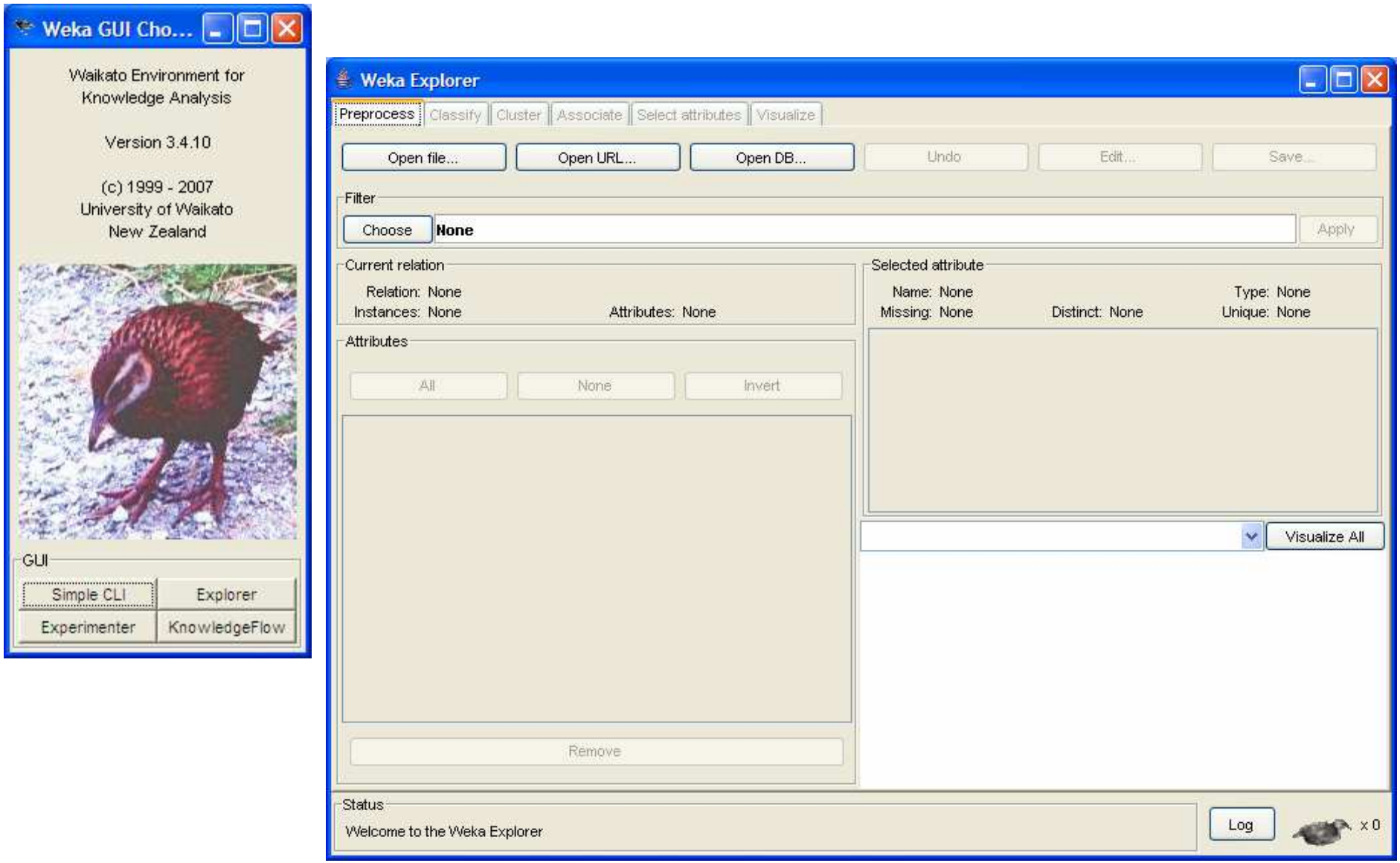
Data preprocessing and visualization

Download and install Weka - http://www.cs.waikato.ac.nz/~ml/weka/



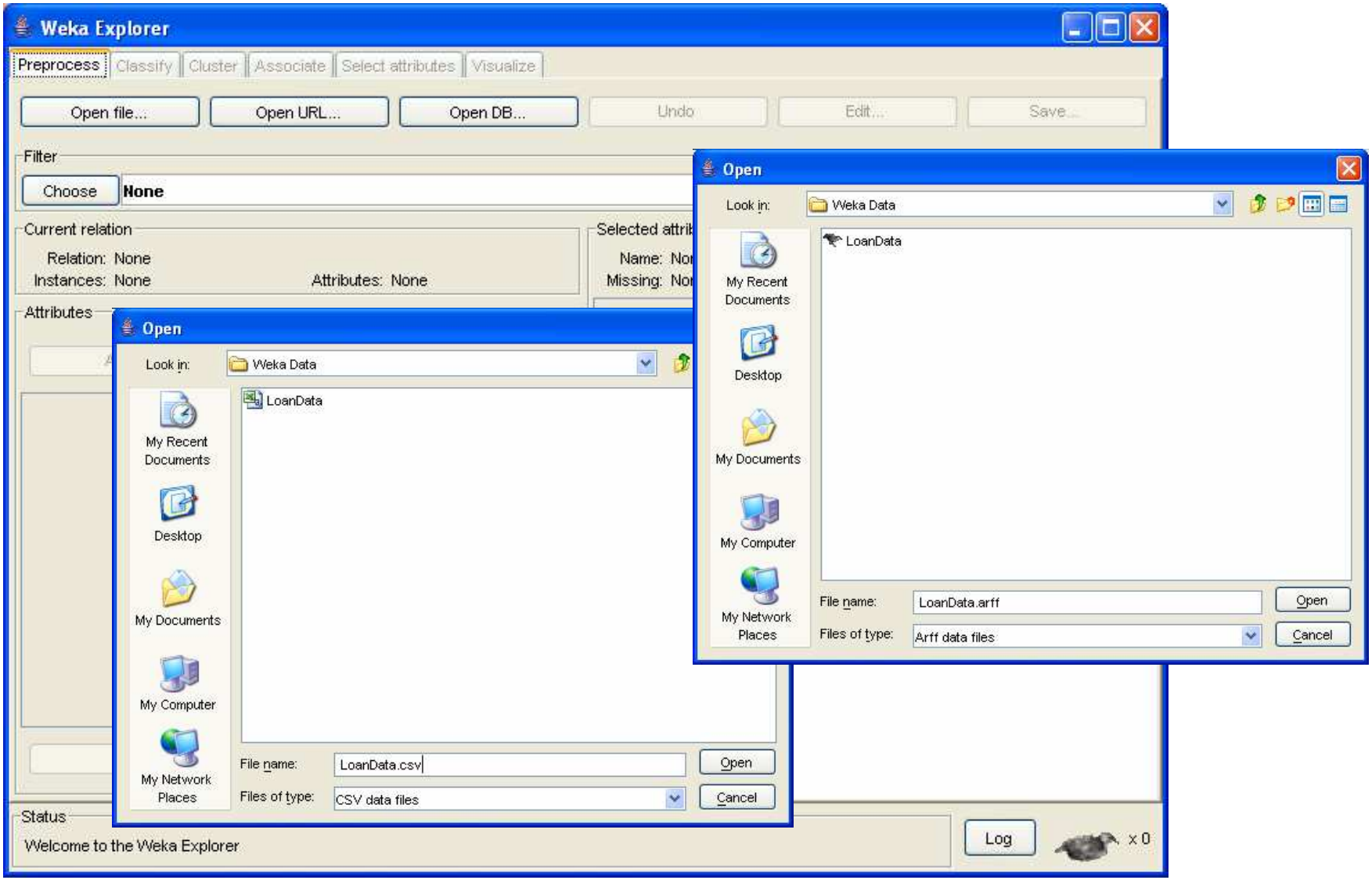
Data preprocessing and visualization

Run Weka and select the Explorer



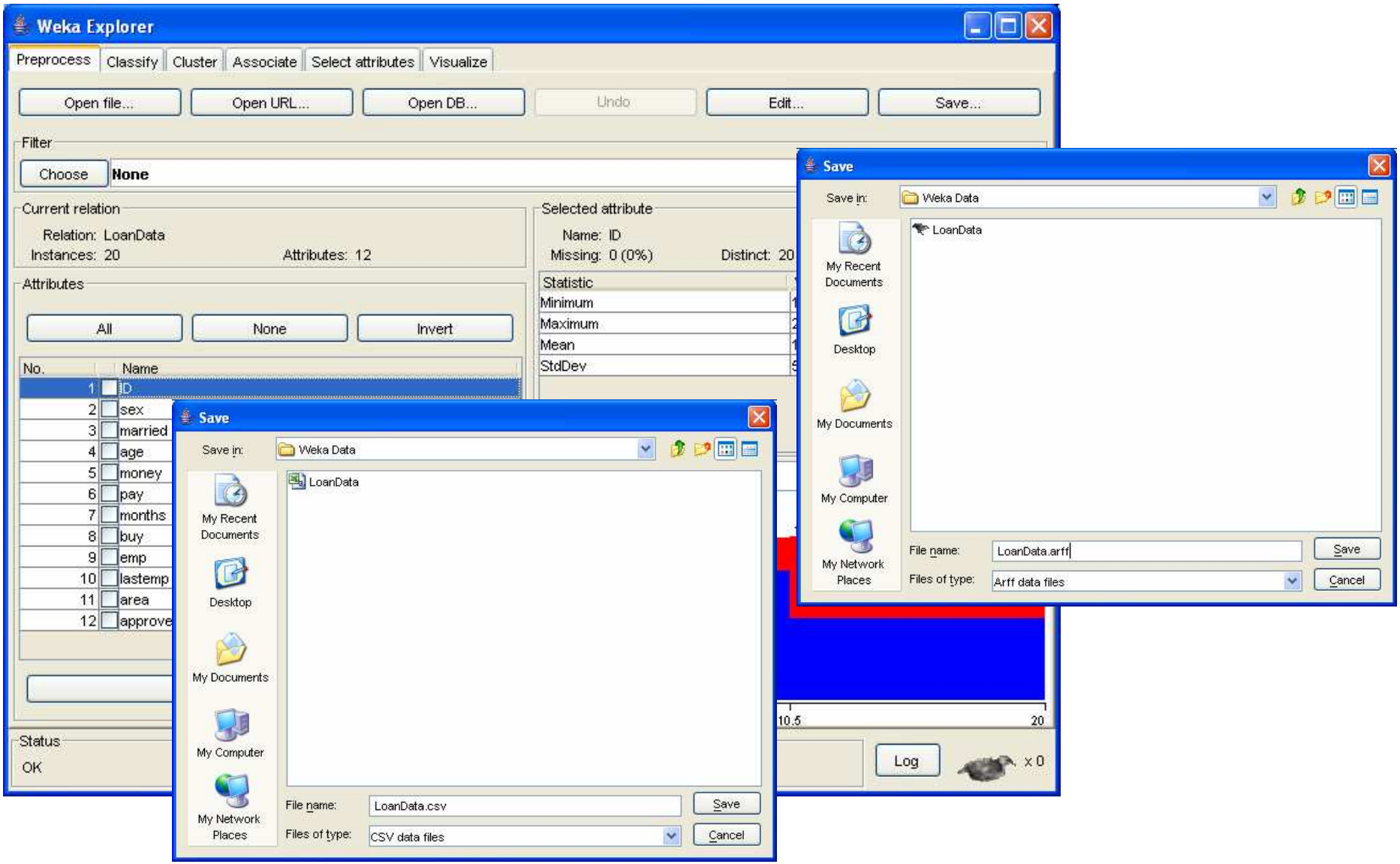
Data preprocessing and visualization

Load data into Weka – ARFF format or CVS format (click on “Open file…”)



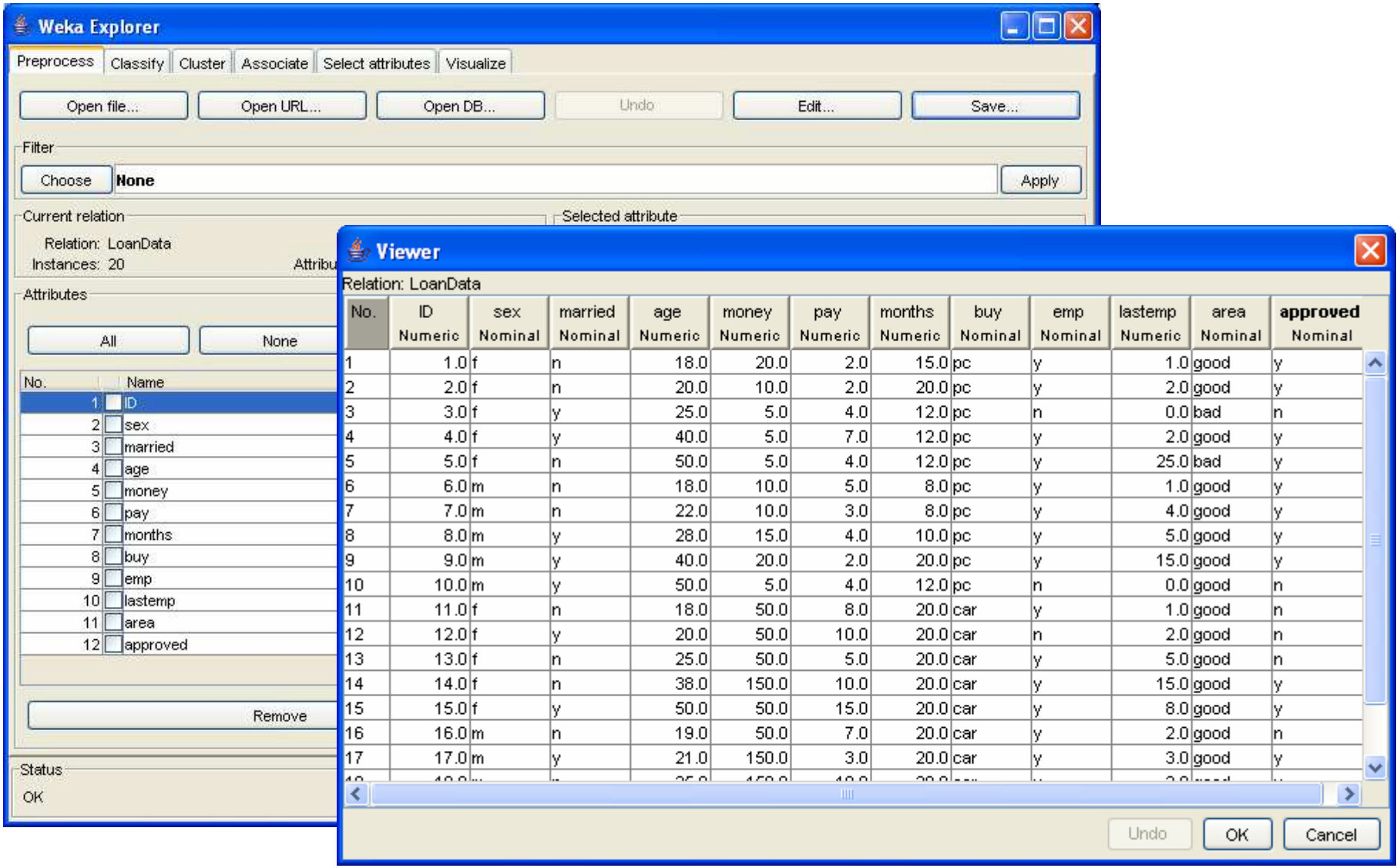
Data preprocessing and visualization

Converting data formats through Weka (click on “Save…”)



Data preprocessing and visualization

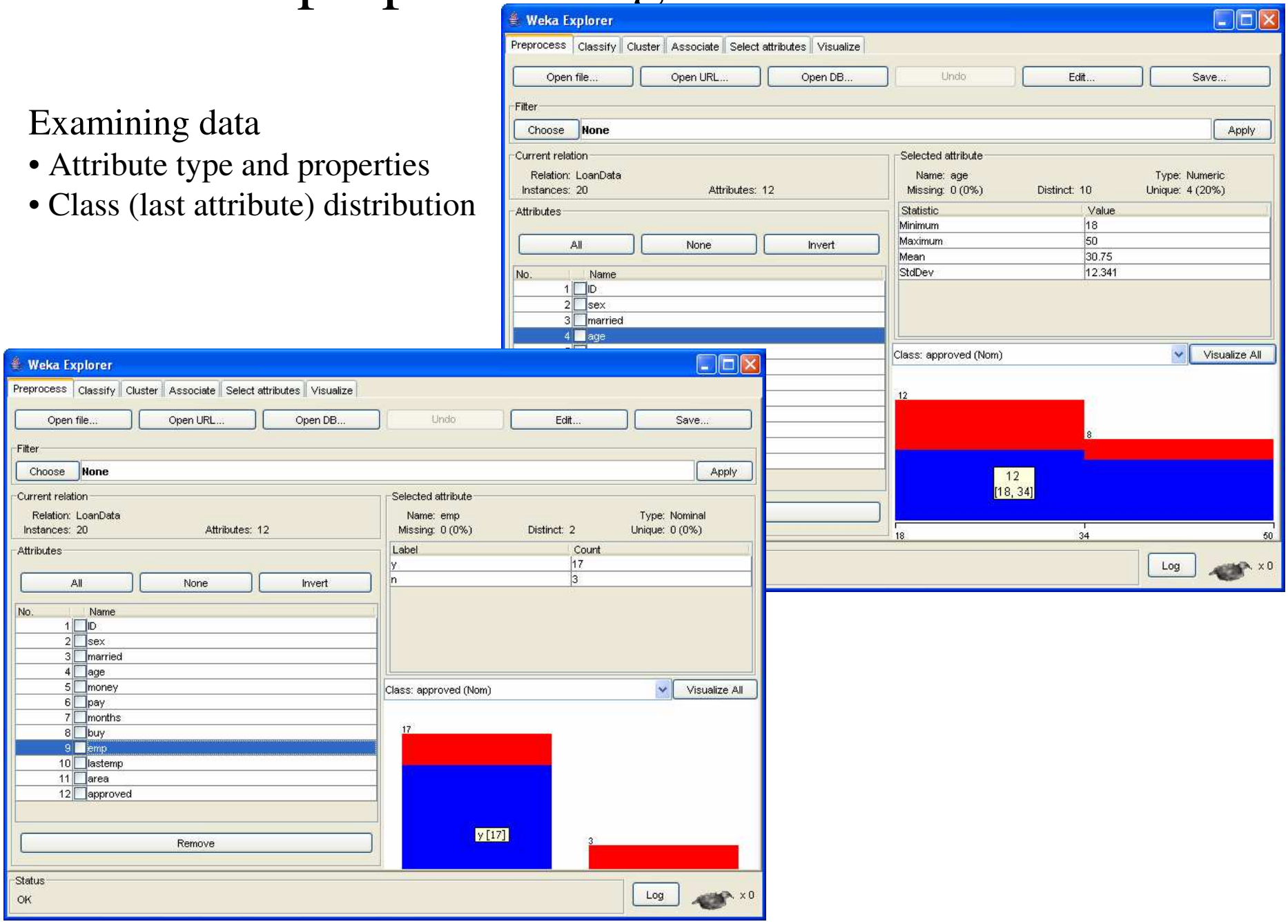
Editing data in Weka (click on ”Edit…”)



Data preprocessing and visualization

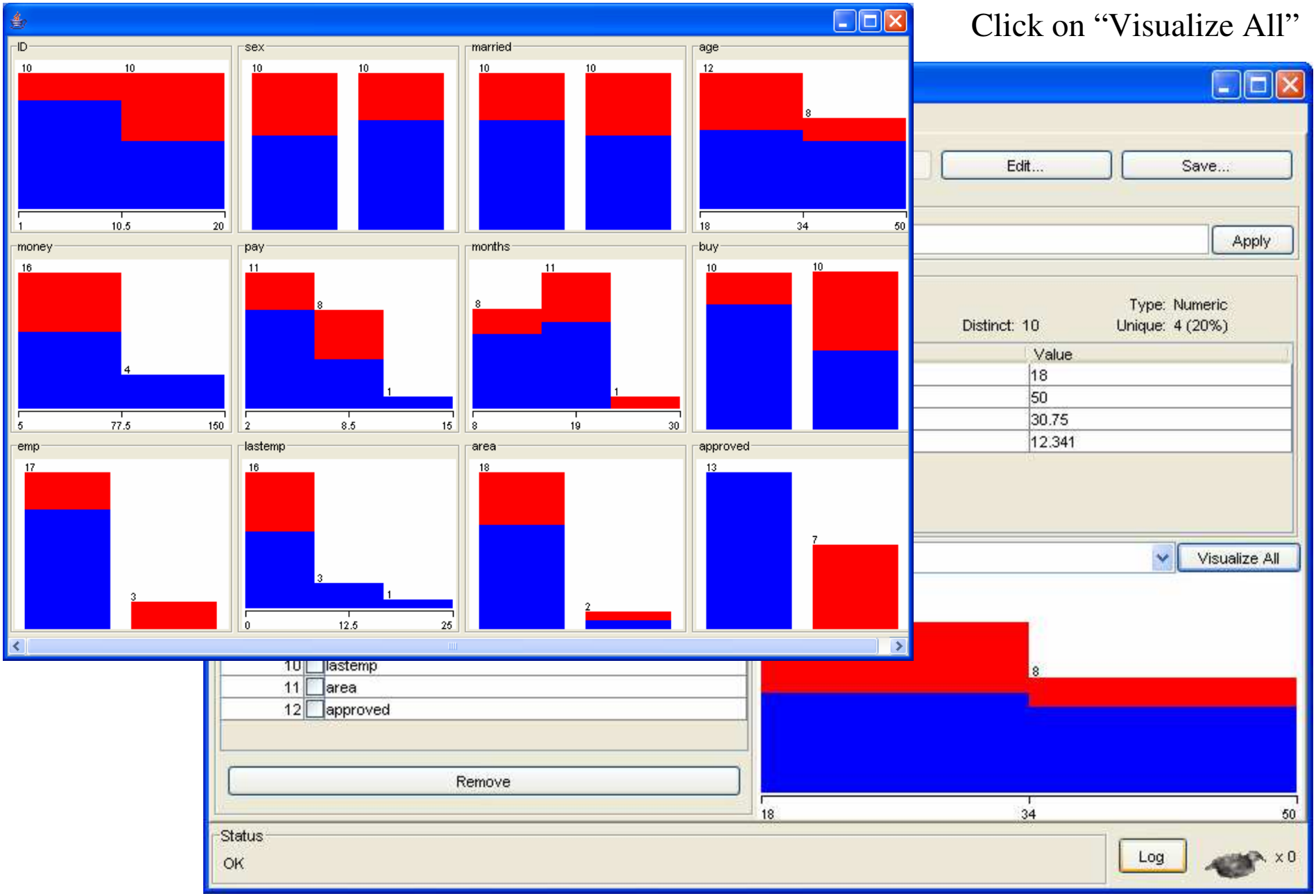
Examining data

- Attribute type and properties
- Class (last attribute) distribution



Data preprocessing and visualization

Click on “Visualize All”

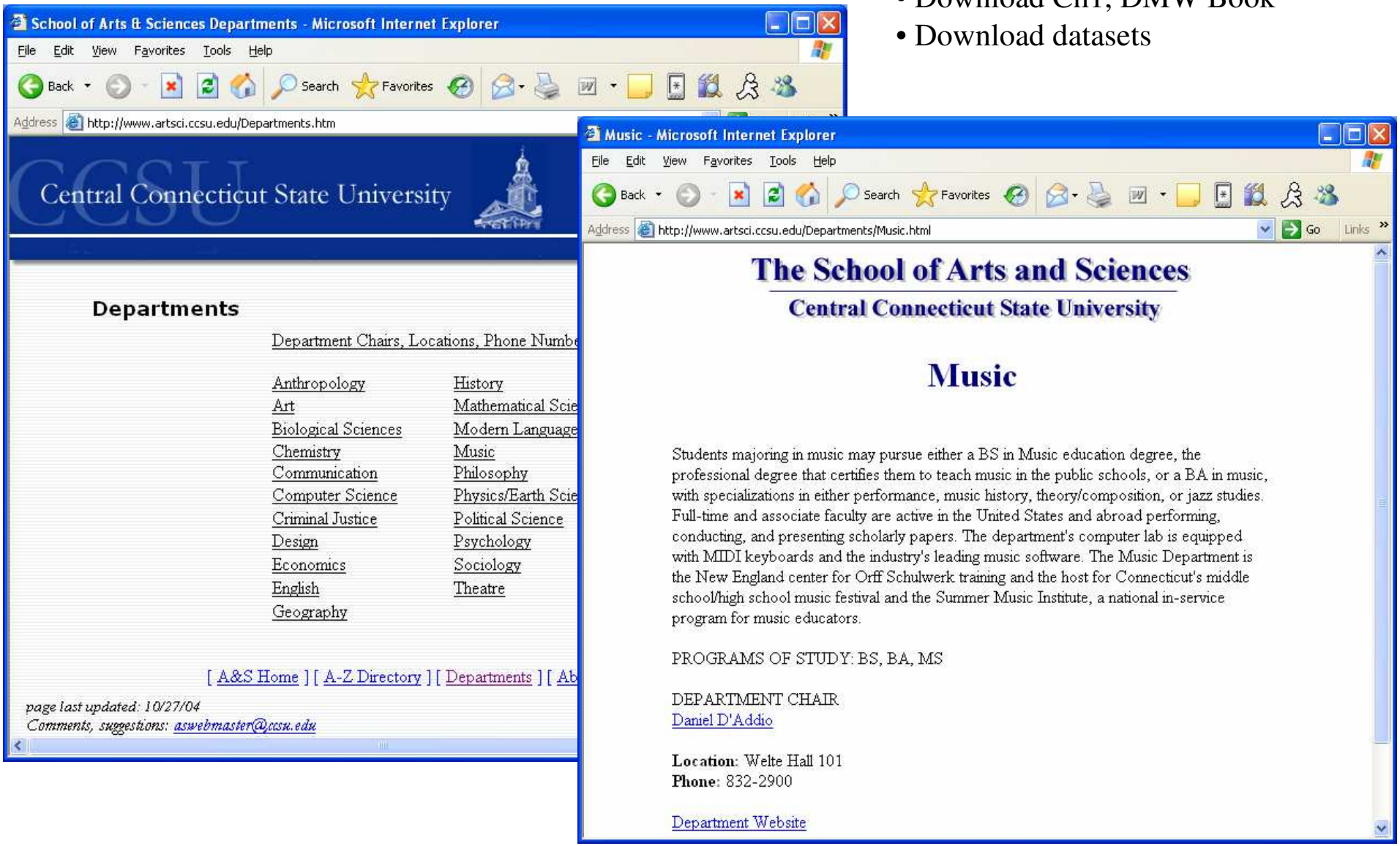


Data preprocessing and visualization

Web/Text documents - Department data

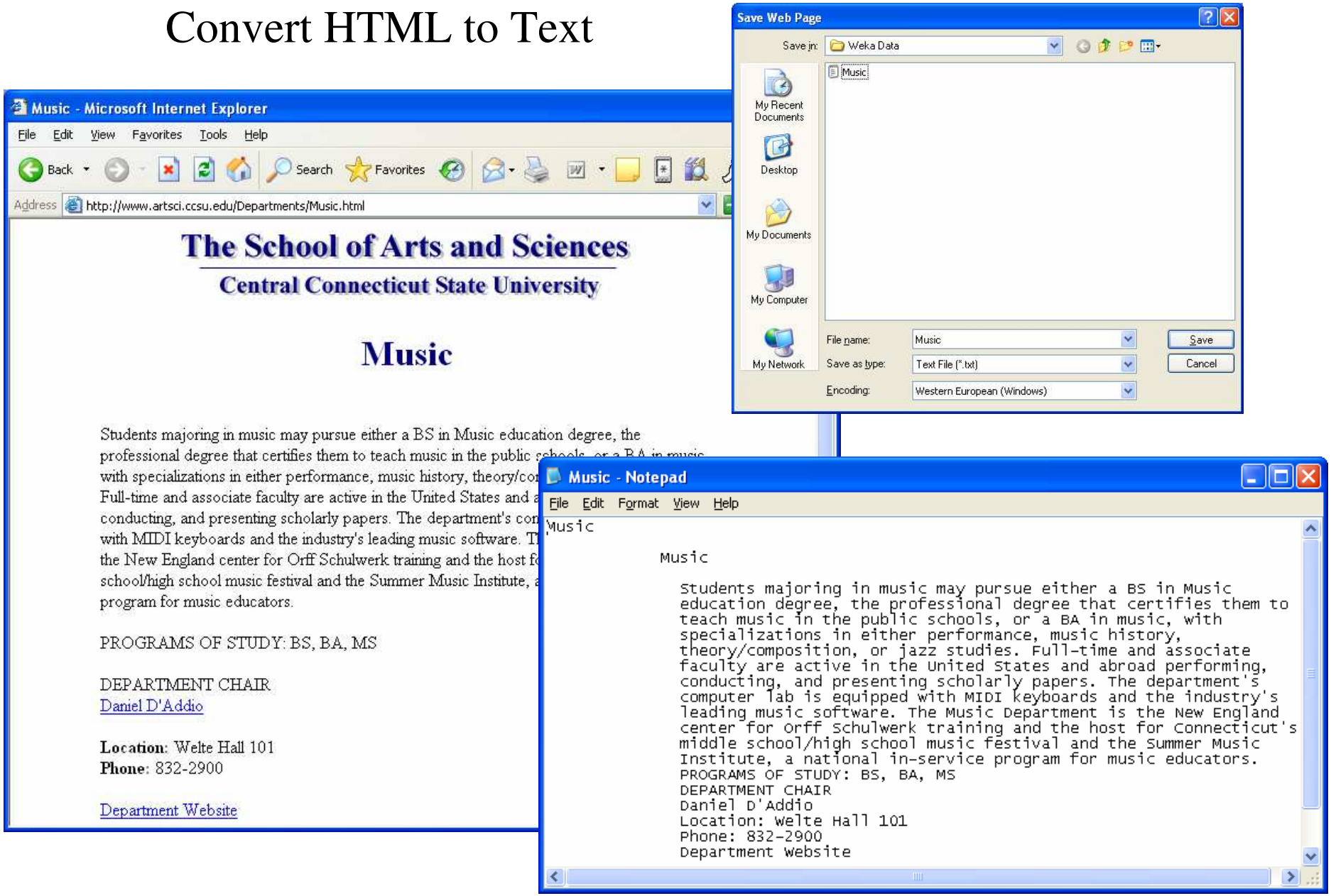
http://www.cs.ccsu.edu/~markov/

- Download Ch1, DMW Book
- Download datasets

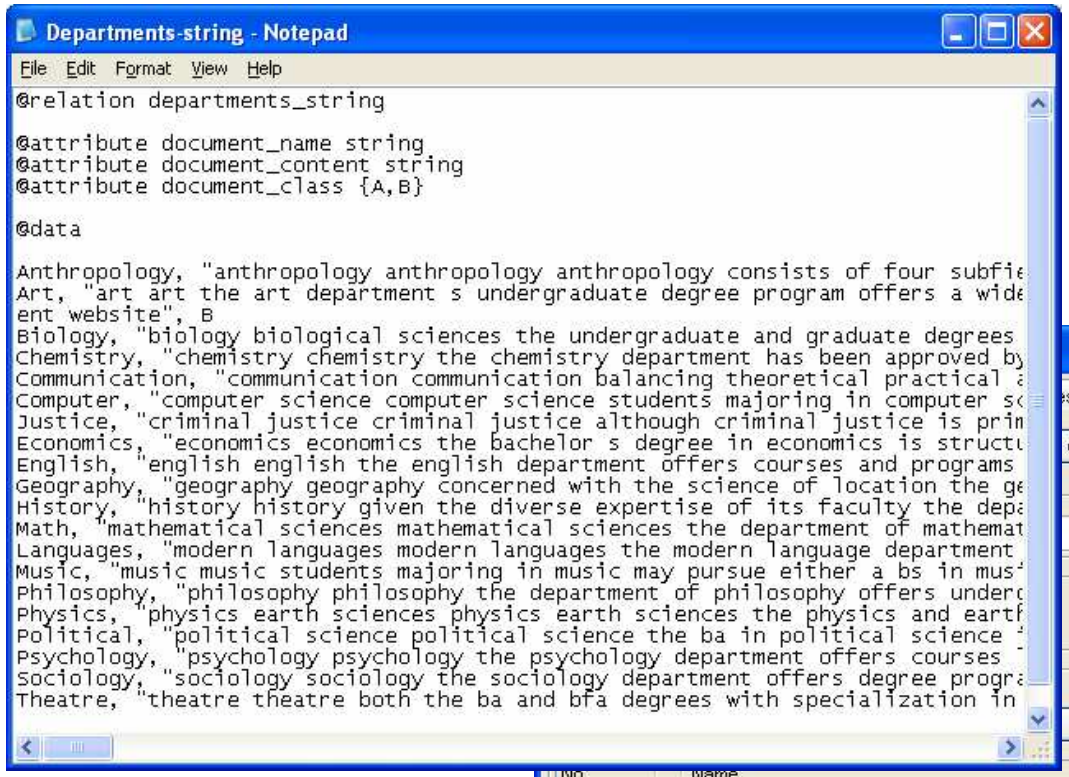


Data preprocessing and visualization

Convert HTML to Text

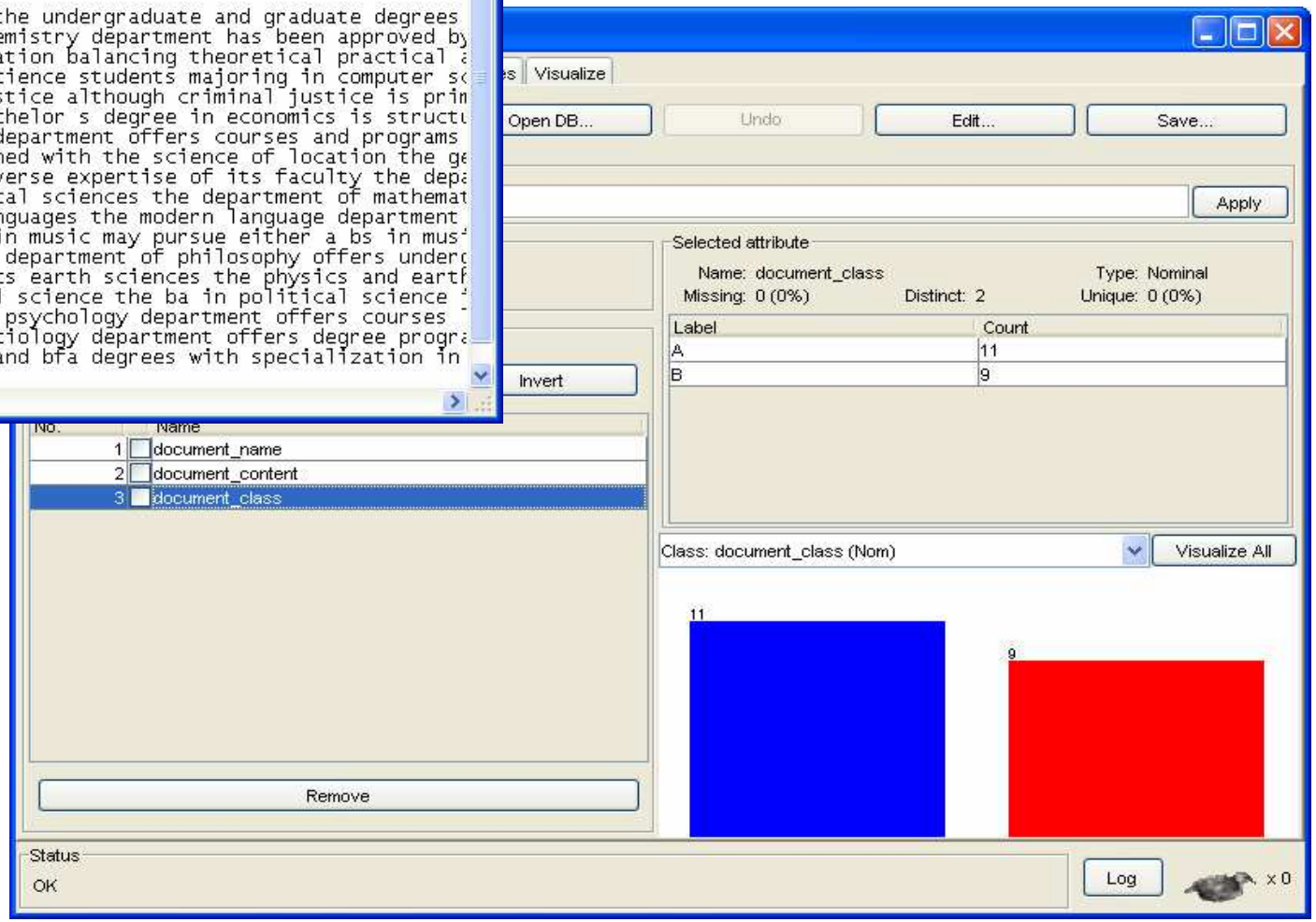


Data preprocessing and visualization



Loading text data in Weka

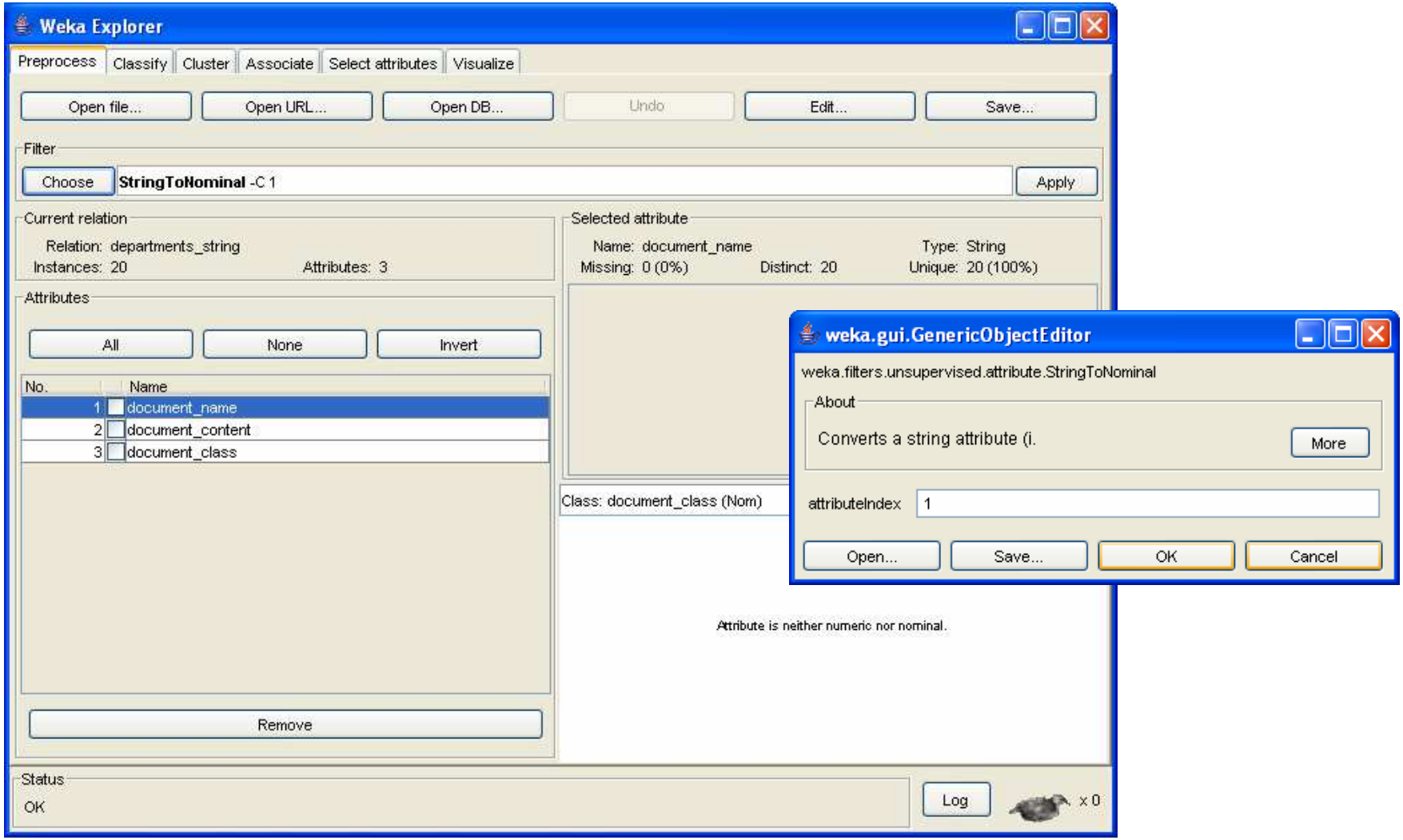
- String format for ID and content
- One document per line
- Add class (nominal) if needed



Data preprocessing and visualization

Converting a string attribute into nominal

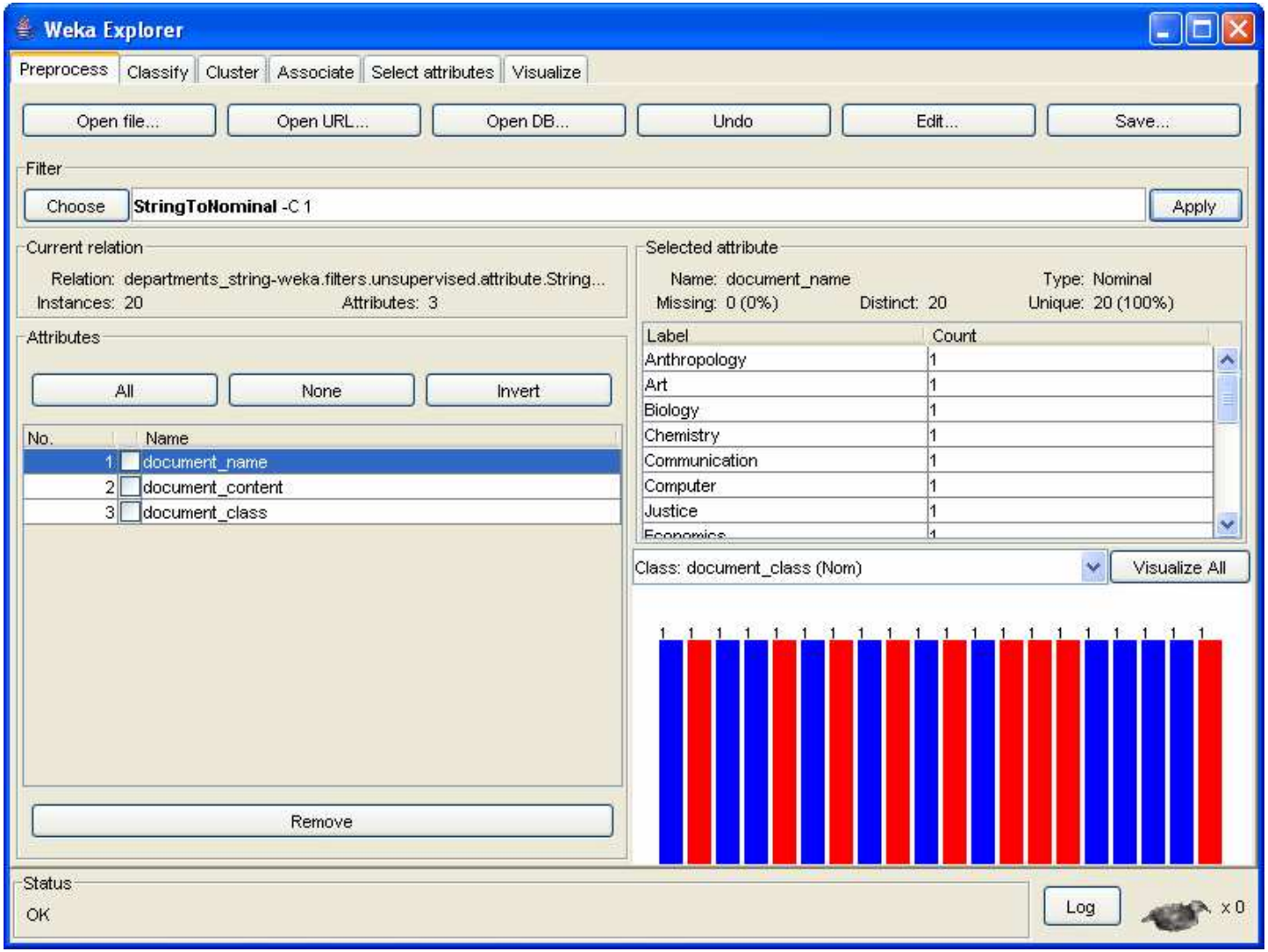
Choose filters/unsupervised/attribute/StringToNominaland and set the index to 1



Data preprocessing and visualization

Converting a string attribute into nominal

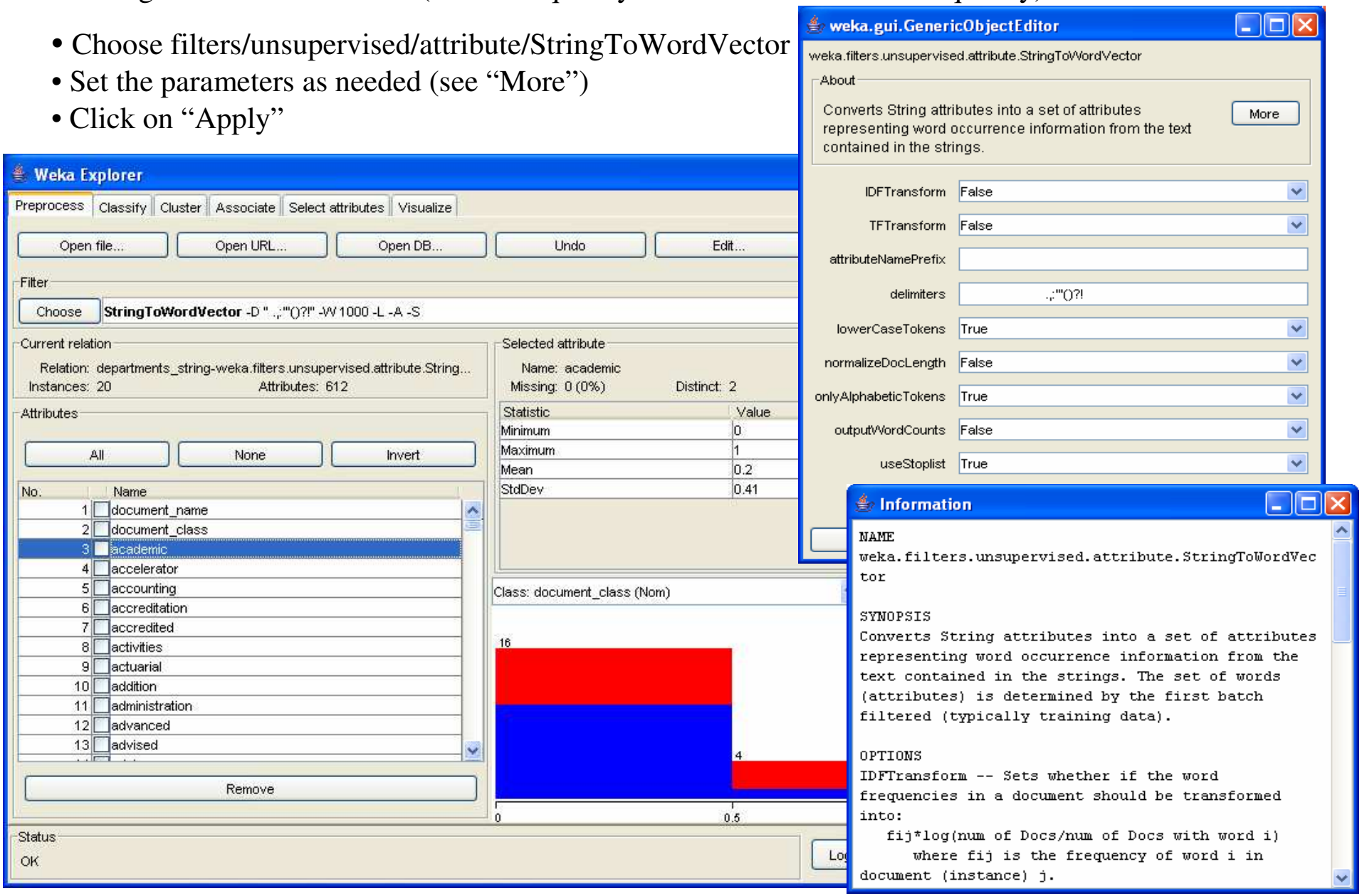
Click on Apply – document_name is now nominal



Data preprocessing and visualization

Converting text data into TFIDF (Term Frequency – Inverted Document Frequency) attribute format

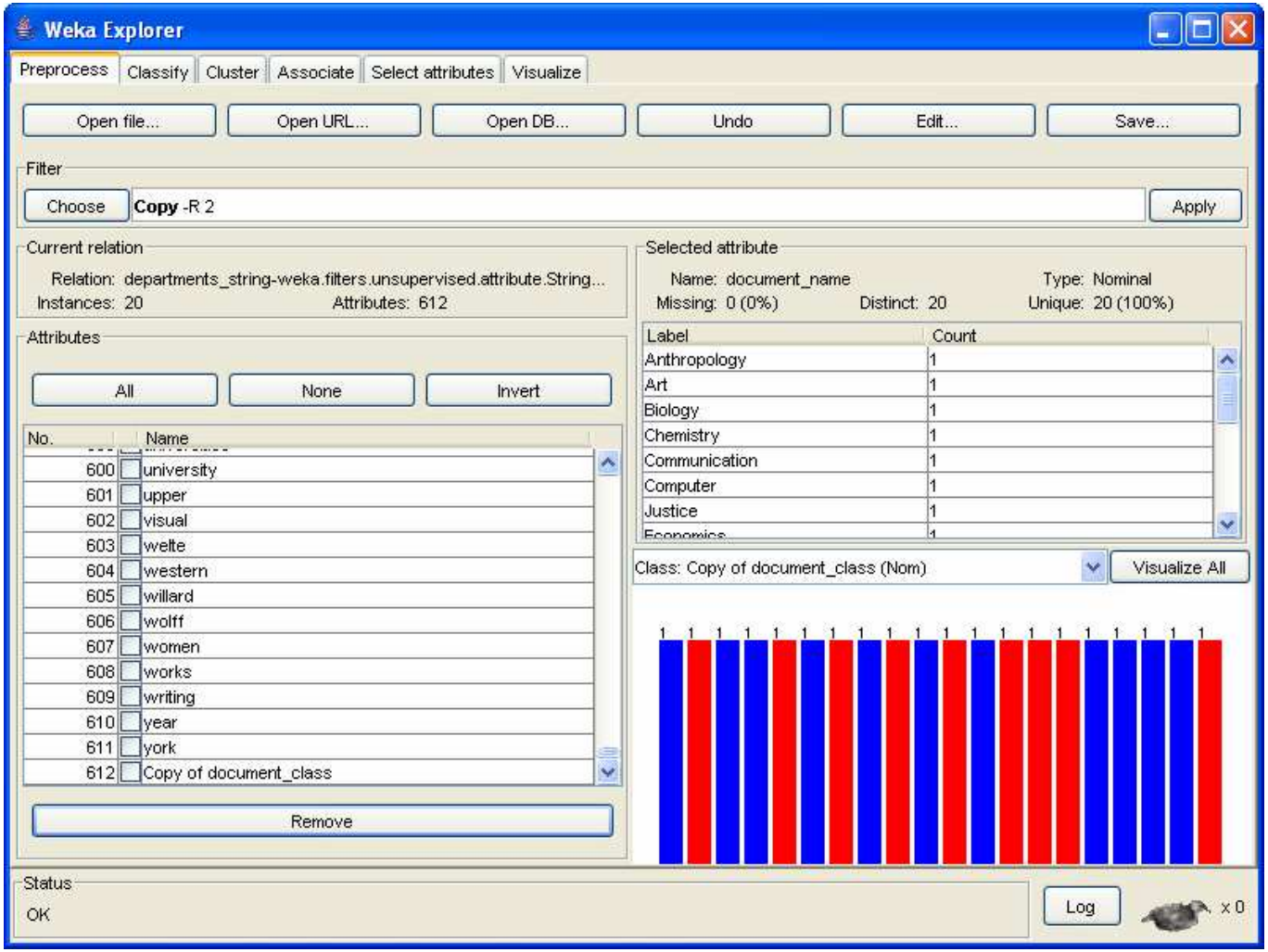
- Choose filters/unsupervised/attribute/StringToWordVector
- Set the parameters as needed (see “More”)
- Click on “Apply”



Data preprocessing and visualization

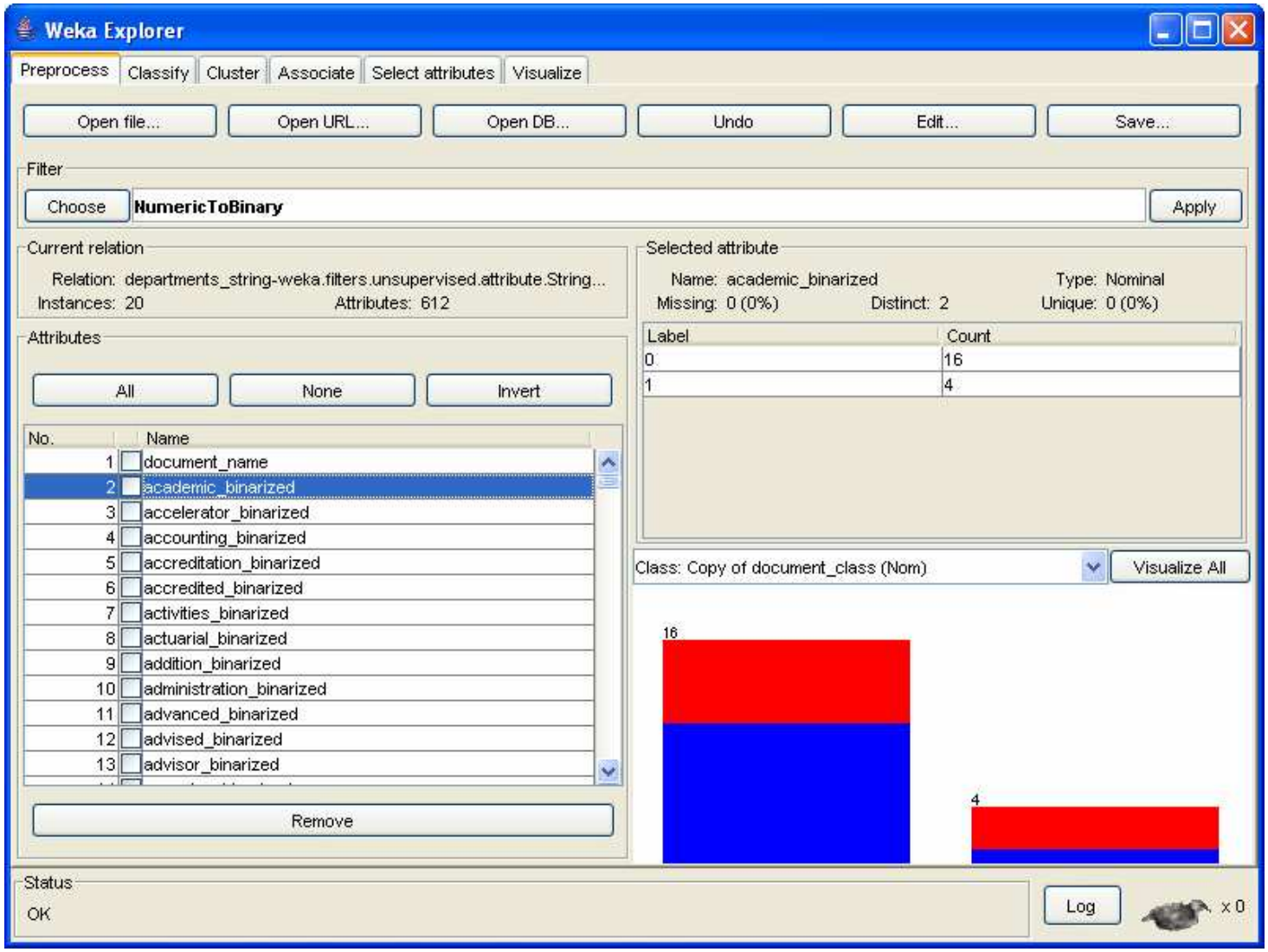
Make the class attribute last

- Choose filters/unsupervised/attribute/Copy
- Set the index to 2 and click on Apply
- Remove attribute 2



Data preprocessing and visualization

- Change the attributes to nominal (use NumericToBinary filter)
- Save data on a file for further use



Data preprocessing and visualization

ARFF file representing the department data in binary format (NonSparse)

Note the format (see SparseToNonSparse instance filter)

Departments-binary - WordPad

File Edit View Insert Format Help

```
@relation 'departments_string-weka.filters.unsupervised.attribute.StringToNominal-C1-weka.filters.unsupervised.attr

@attribute document_name {Anthropology,Art,Biology,Chemistry,Communication,Computer,Justice,Economics,English,Geogr
@attribute academic_binarized {0,1}
@attribute accelerator_binarized {0,1}
@attribute accounting_binarized {0,1}
@attribute accreditation_binarized {0,1}
@attribute accredited_binarized {0,1}
@attribute activities_binarized {0,1}
@attribute actuarial_binarized {0,1}
@attribute addition_binarized {0,1}
@attribute administration_binarized {0,1}
@attribute advanced_binarized {0,1}
@attribute advised_binarized {0,1}
@attribute advisor_binarized {0,1}
@attribute agencies_binarized {0,1}
@attribute air_binarized {0,1}
@attribute ali_binarized {0,1}
@attribute alternative_binarized {0,1}
@attribute american_binarized {0,1}
@attribute anesthesia_binarized {0,1}
@attribute annually_binarized {0,1}
@attribute antar_binarized {0,1}
@attribute anthropology_binarized {0,1}
@attribute applied_binarized {0,1}
@attribute approach_binarized {0,1}
@attribute approved_binarized {0,1}
@attribute archaeology_binarized {0,1}
@attribute area_binarized {0,1}
@attribute areas_binarized {0,1}
@attribute artificial_binarized {0,1}
@attribute assist_binarized {0,1}
@attribute associate_binarized {0,1}
@attribute astronomy_binarized {0,1}
@attribute astrophysics_binarized {0,1}
@attribute atomic_binarized {0,1}
@attribute attend_binarized {0,1}
```

For Help, press F1

Departments-binary.arff - WordPad

File Edit View Insert Format Help

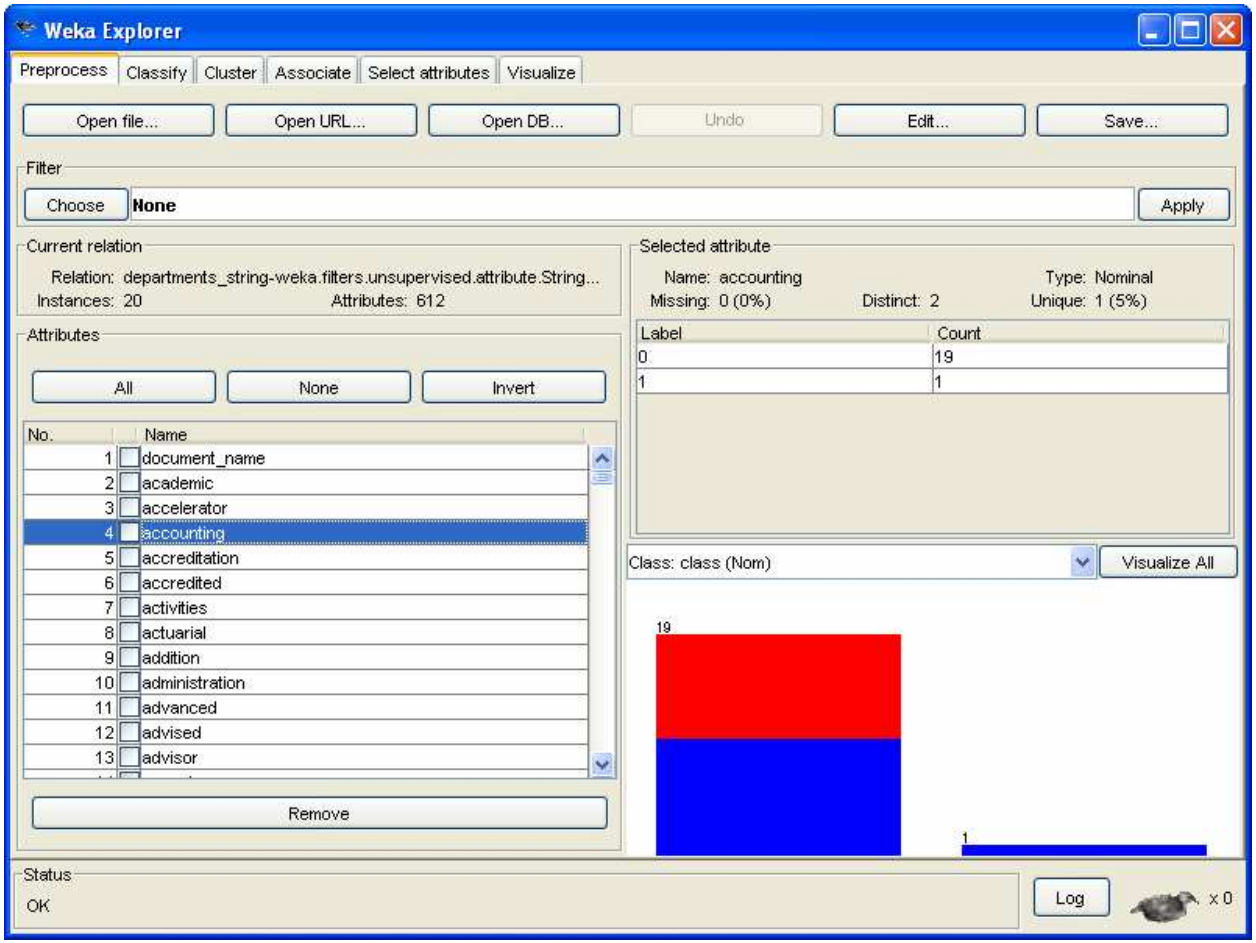
```
@attribute western_binarized {0,1}
@attribute willard_binarized {0,1}
@attribute wolff_binarized {0,1}
@attribute women_binarized {0,1}
@attribute works_binarized {0,1}
@attribute writing_binarized {0,1}
@attribute year_binarized {0,1}
@attribute york_binarized {0,1}
@attribute class {A,B}

@data

Anthropology,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,0,1,0,1,0,1,0,1,1,0,0,0,1,0,0,0,
Art,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Biology,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,1,0,1,0,0,0
Chemistry,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,1,0
Communication,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0
Computer,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,
Justice,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
Economics,1,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,1,0,1,0
English,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0
Geography,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0
History,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
Math,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
Languages,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
Music,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0
Philosophy,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,
Physics,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,1,0,1,0,0,0,0,0,1,0,1,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
Political,0,0,0,0,0,0,1,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Psychology,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,
Sociology,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0
Theatre,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0
```

For Help, press F1

Attribute Selection

Finding a minimal set of attributes that preserve the class distribution

Attribute relevance with respect to the class – not relevant attribute (*accounting*)

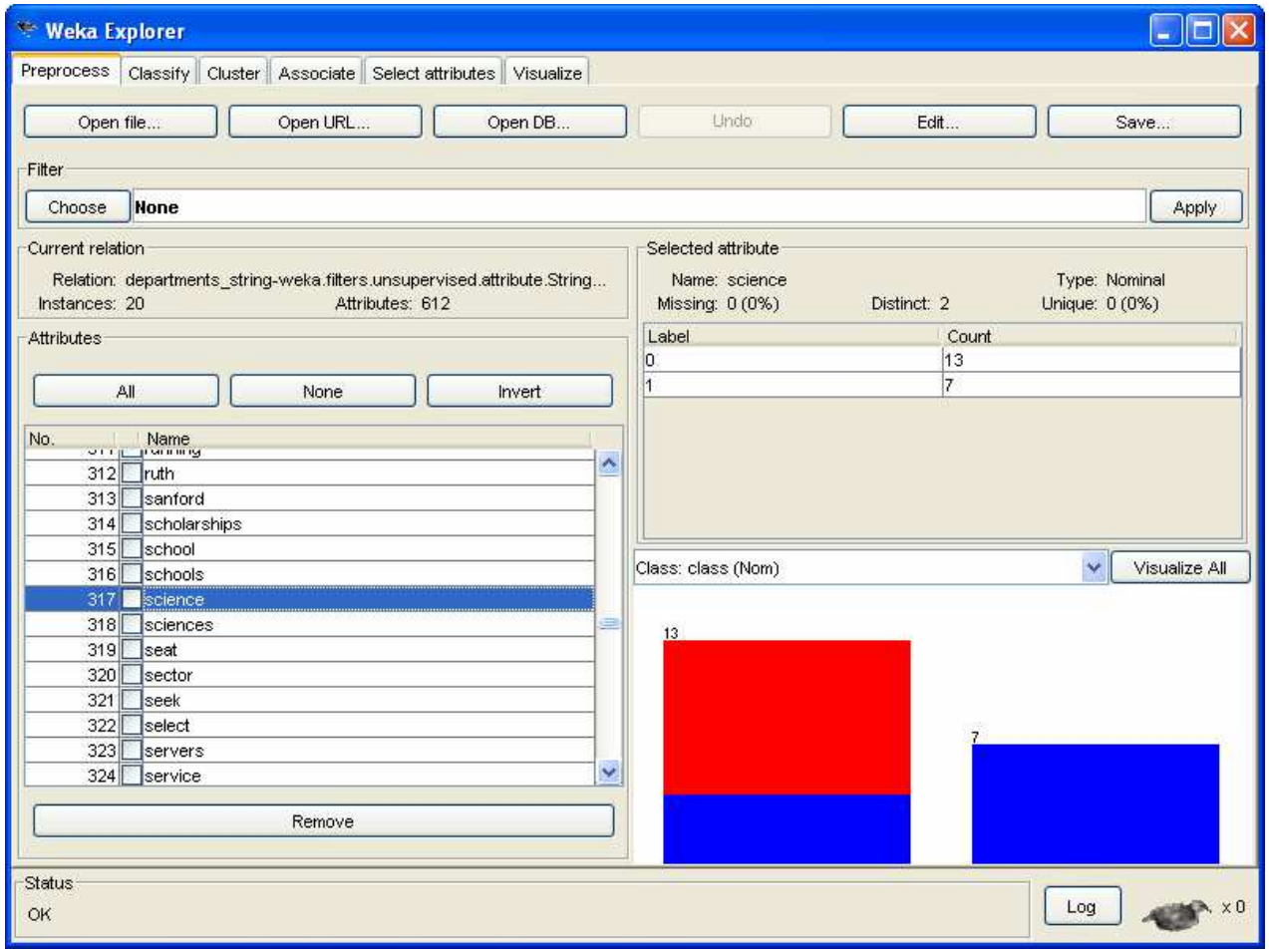


IF accounting=1 THEN class=A (Error=0, Coverage = 1 instance → **overfitting**)

IF accounting=0 THEN class=B (Error=10/19, Coverage = 19 instances → **low accuracy**)

Attribute Selection

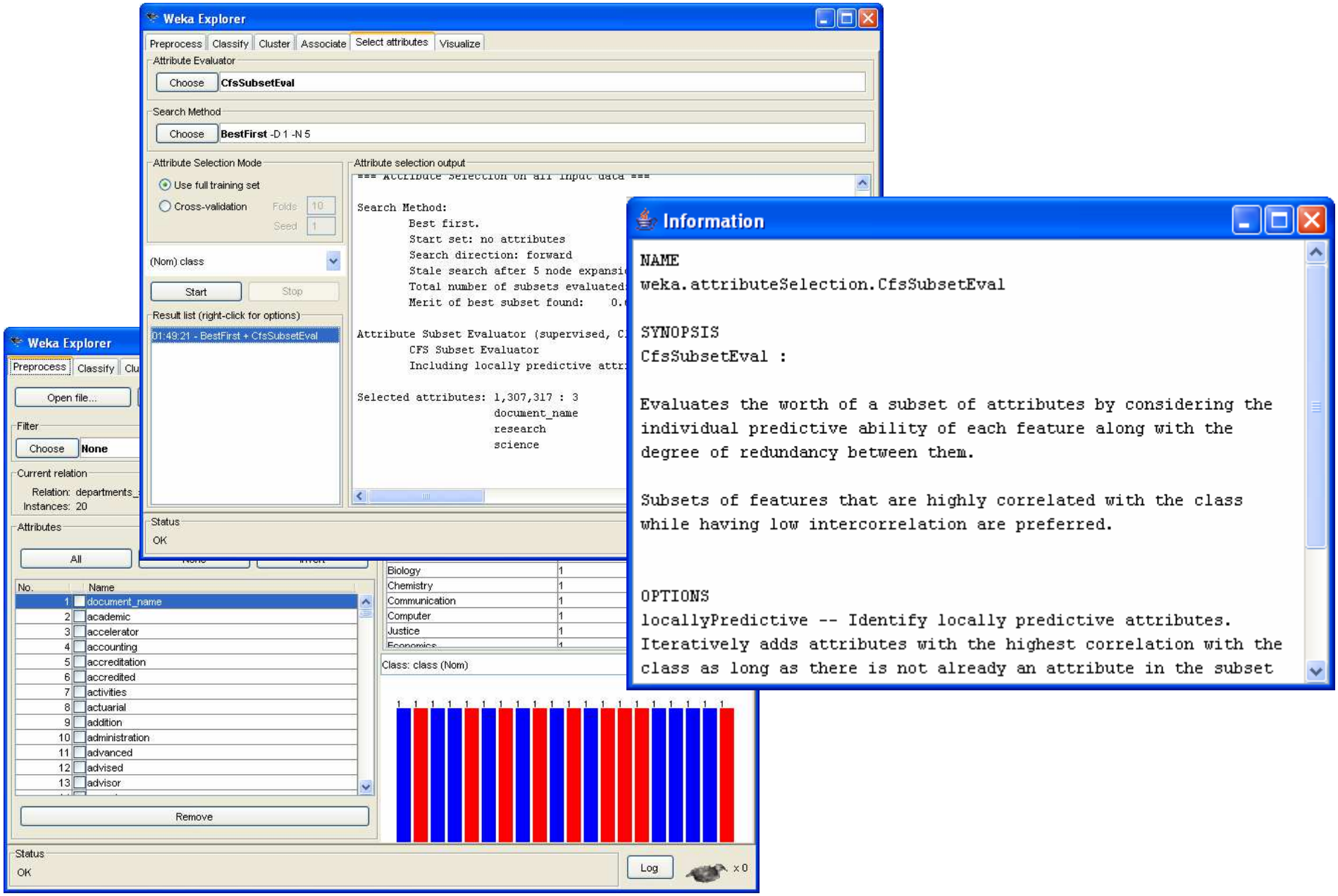
Attribute relevance with respect to the class – relevant attribute (*science*)



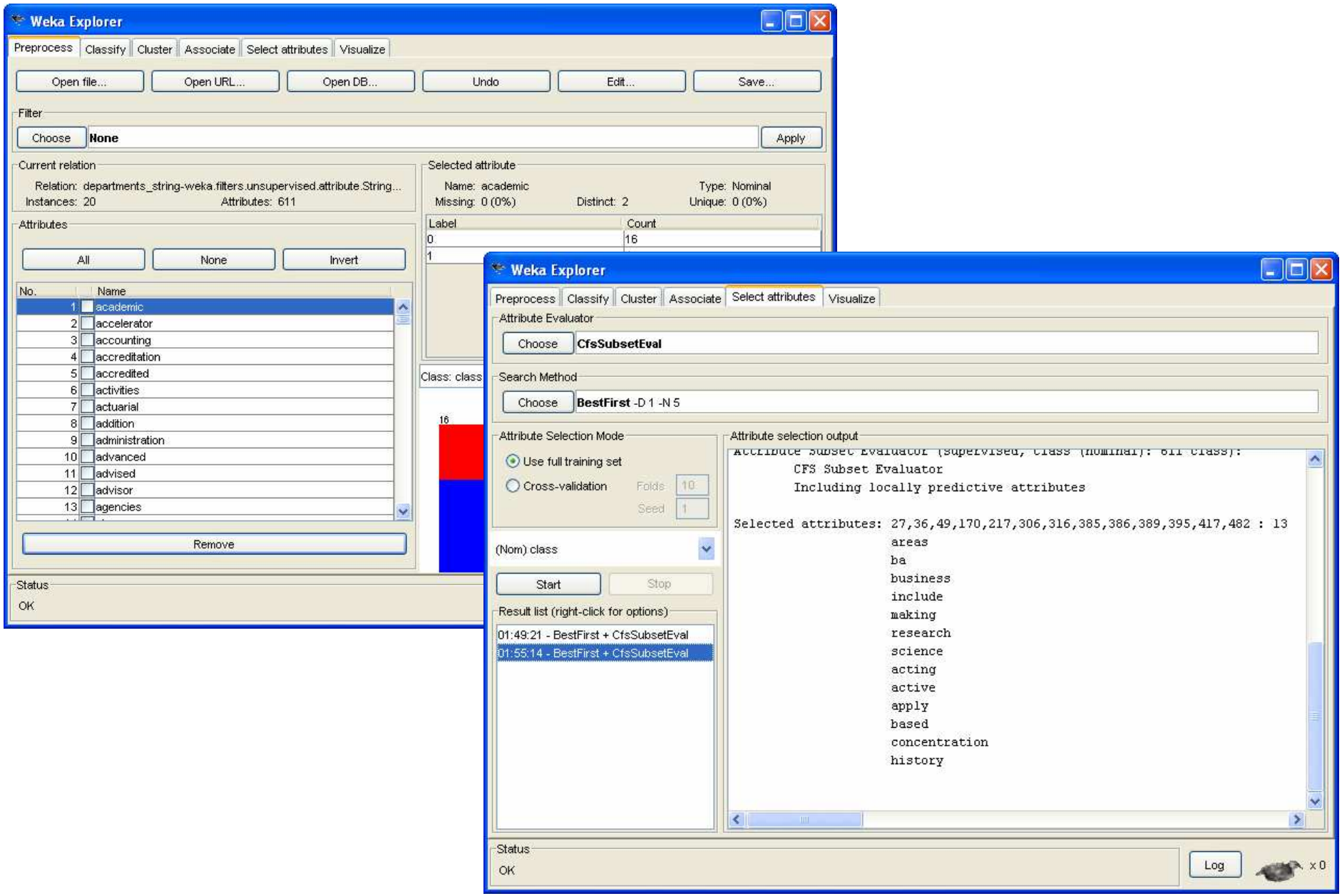
IF accounting=1 THEN class=A (Error=0, Coverage = 7 instance)

IF accounting=0 THEN class=B (Error=4/13, Coverage = 13 instances)

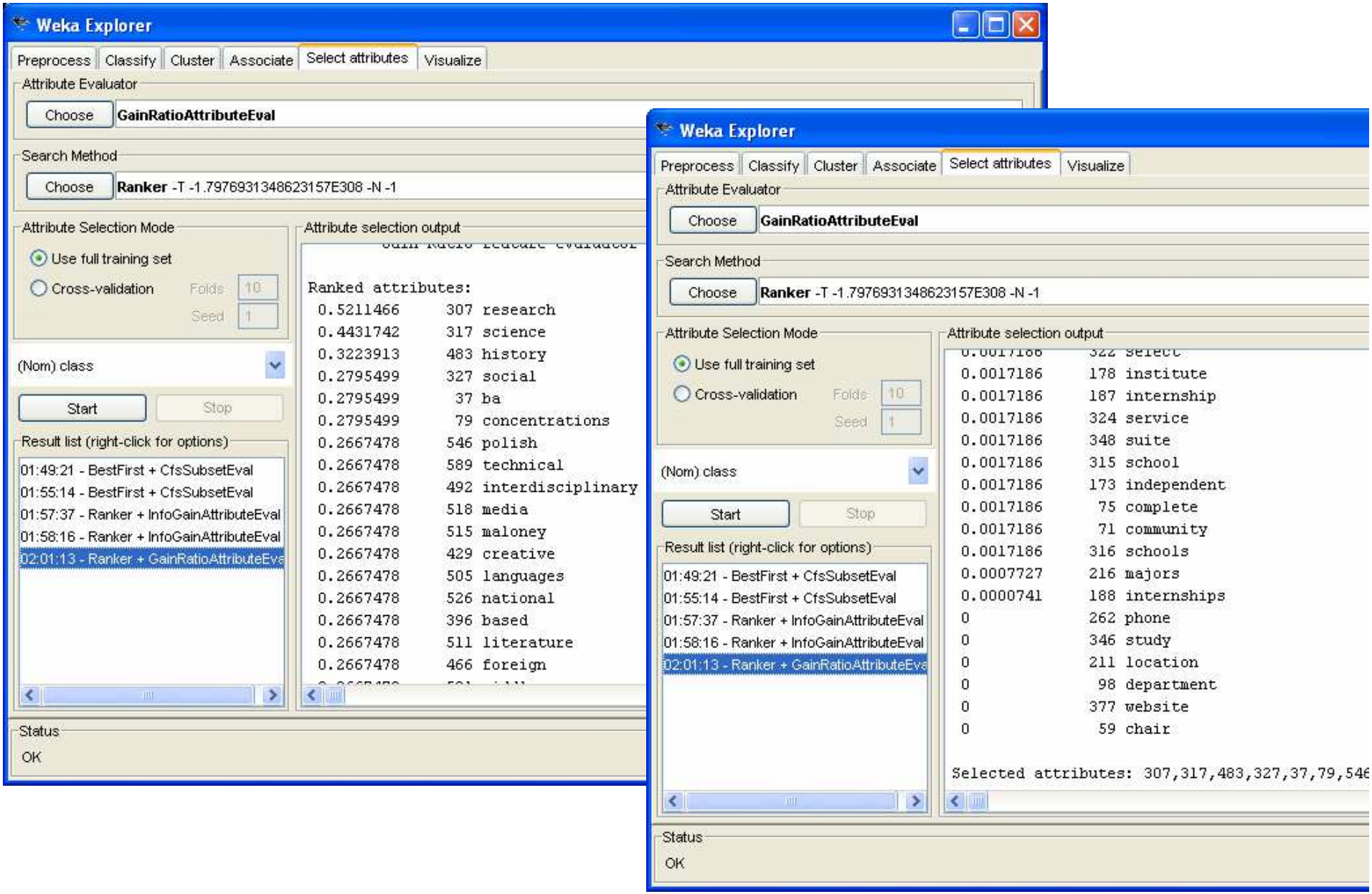
Attribute Selection (with document_name)



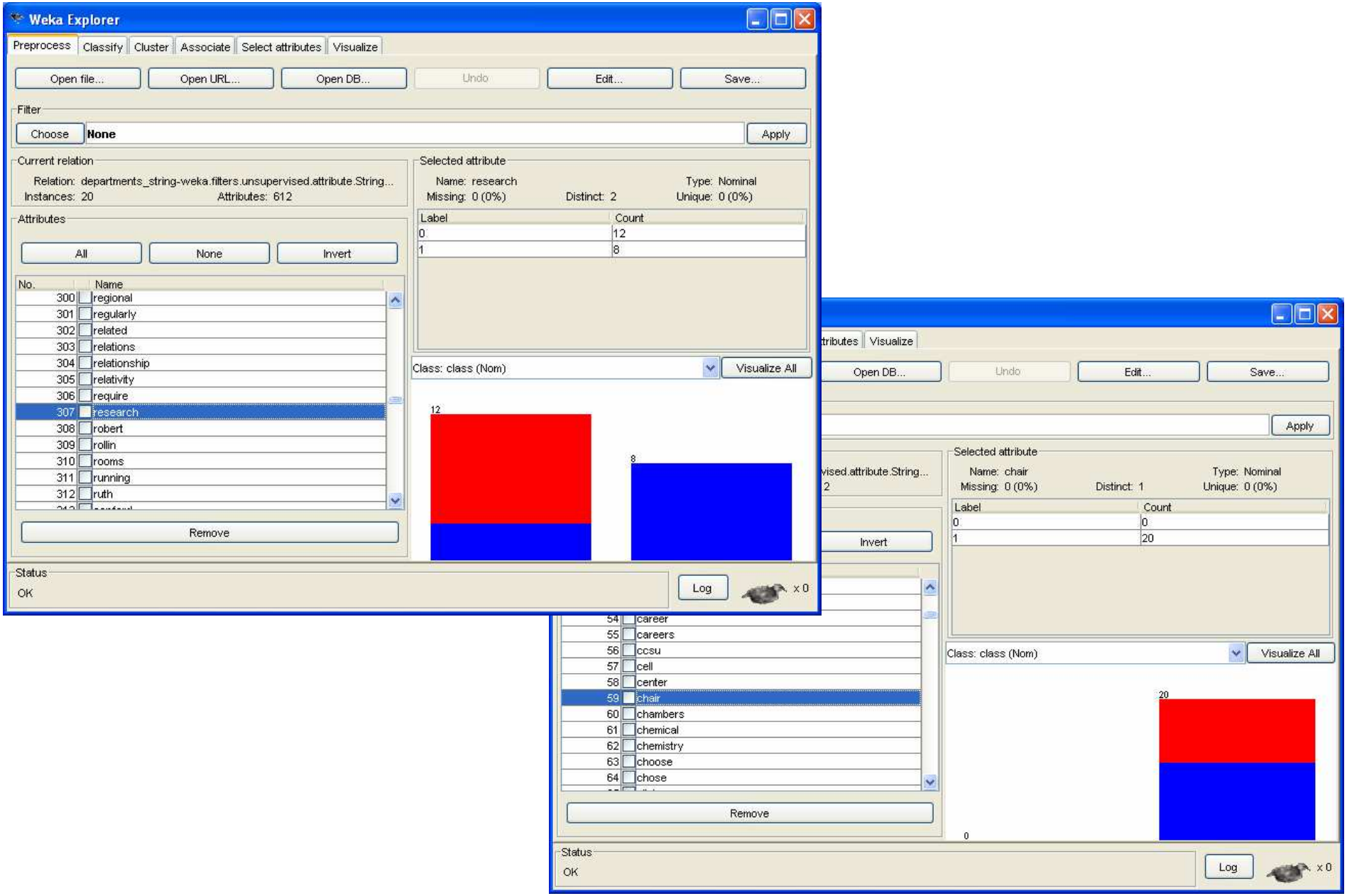
Attribute Selection (without document_name)



Attribute Selection (ranking)

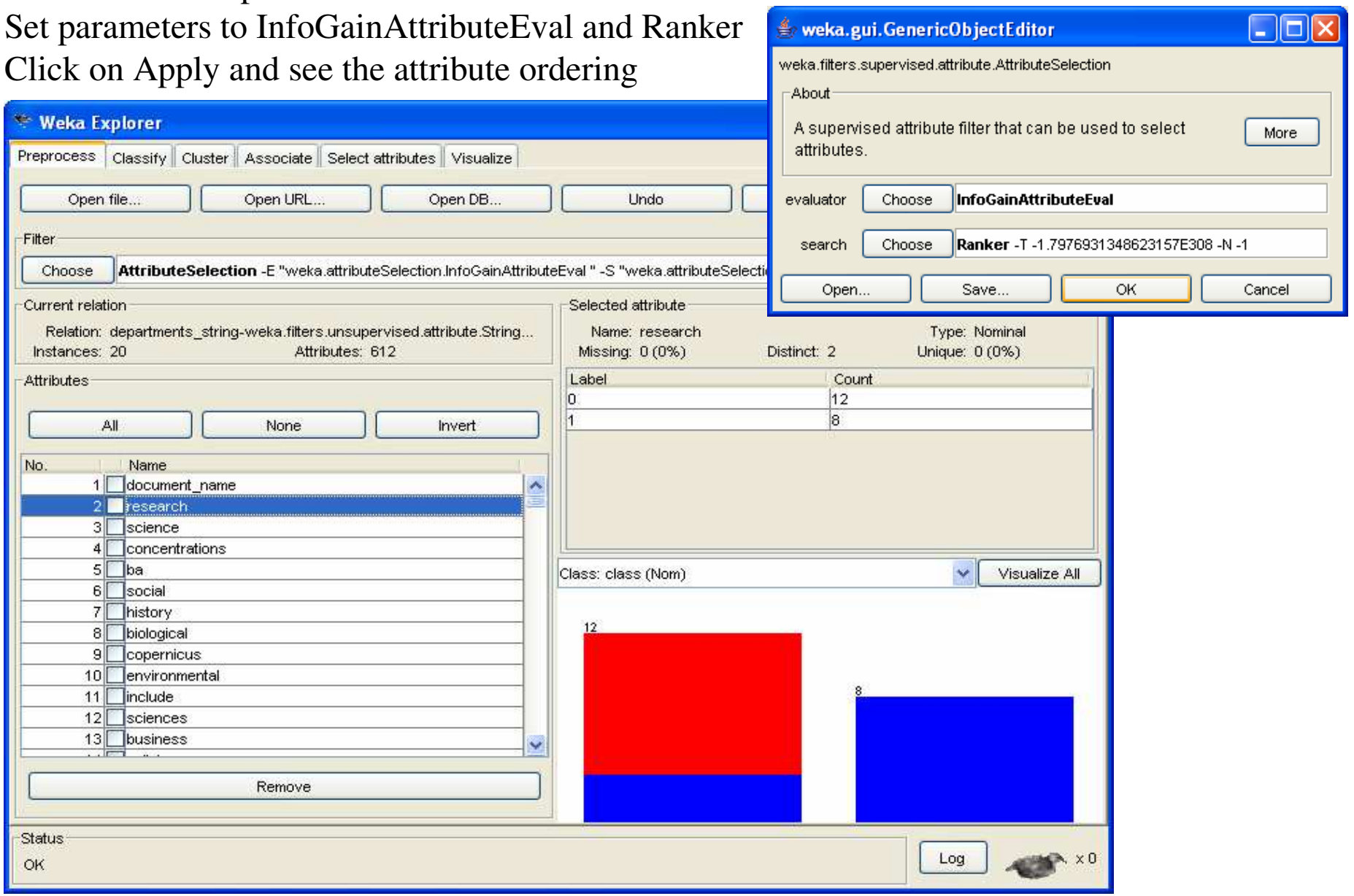


Attribute Selection (explanation of ranking)

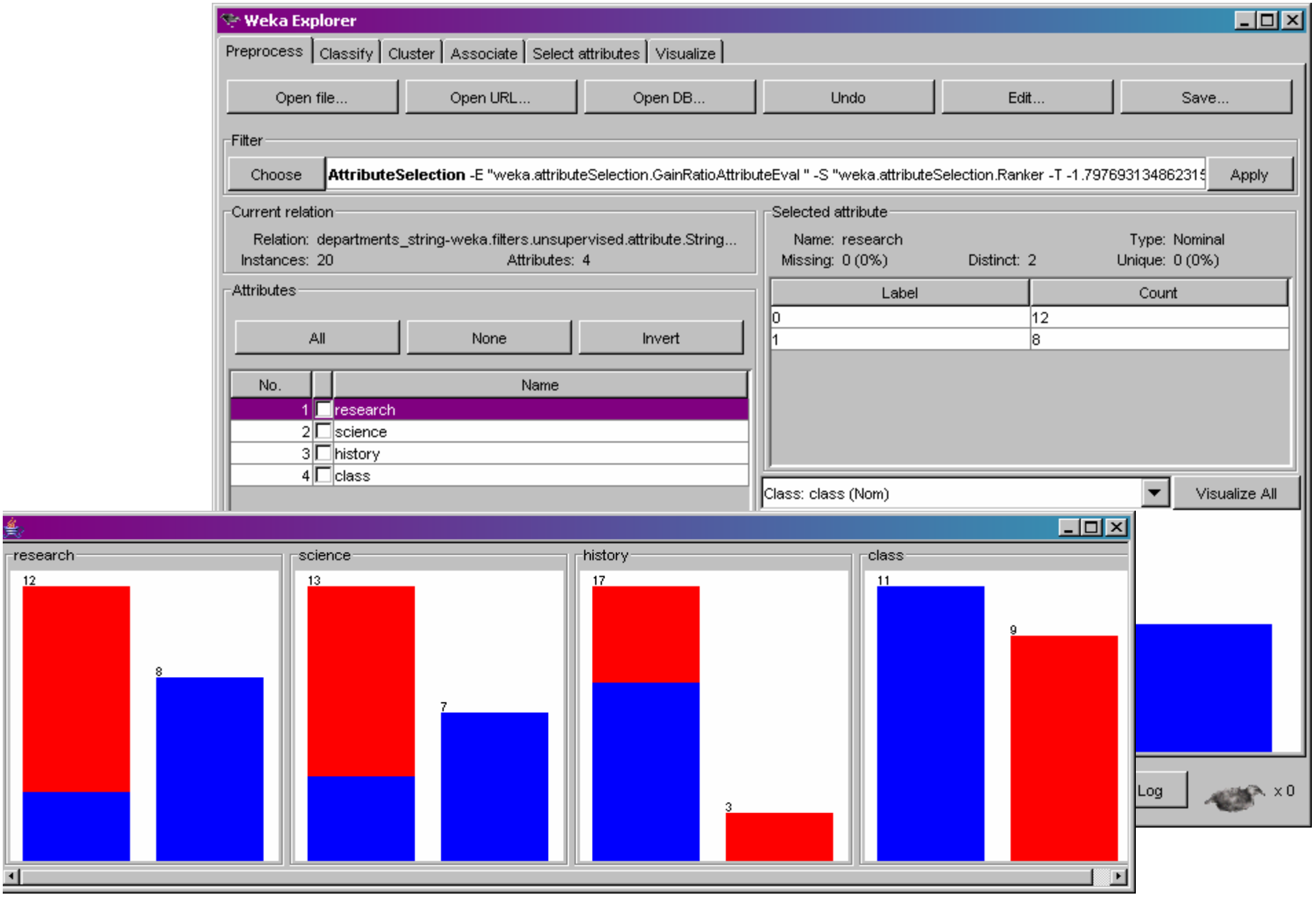


Attribute Selection (using filters)

- Choose filters/supervised/attribute/AttributeSelection
- Set parameters to InfoGainAttributeEval and Ranker
- Click on Apply and see the attribute ordering



Attribute Selection (using filters)



Classification – creating models (hypotheses)

Mapping (independent attributes -> class)

Inferring rudimentary rules - OneR

Weather data (weather.nominal.arff)

Viewer

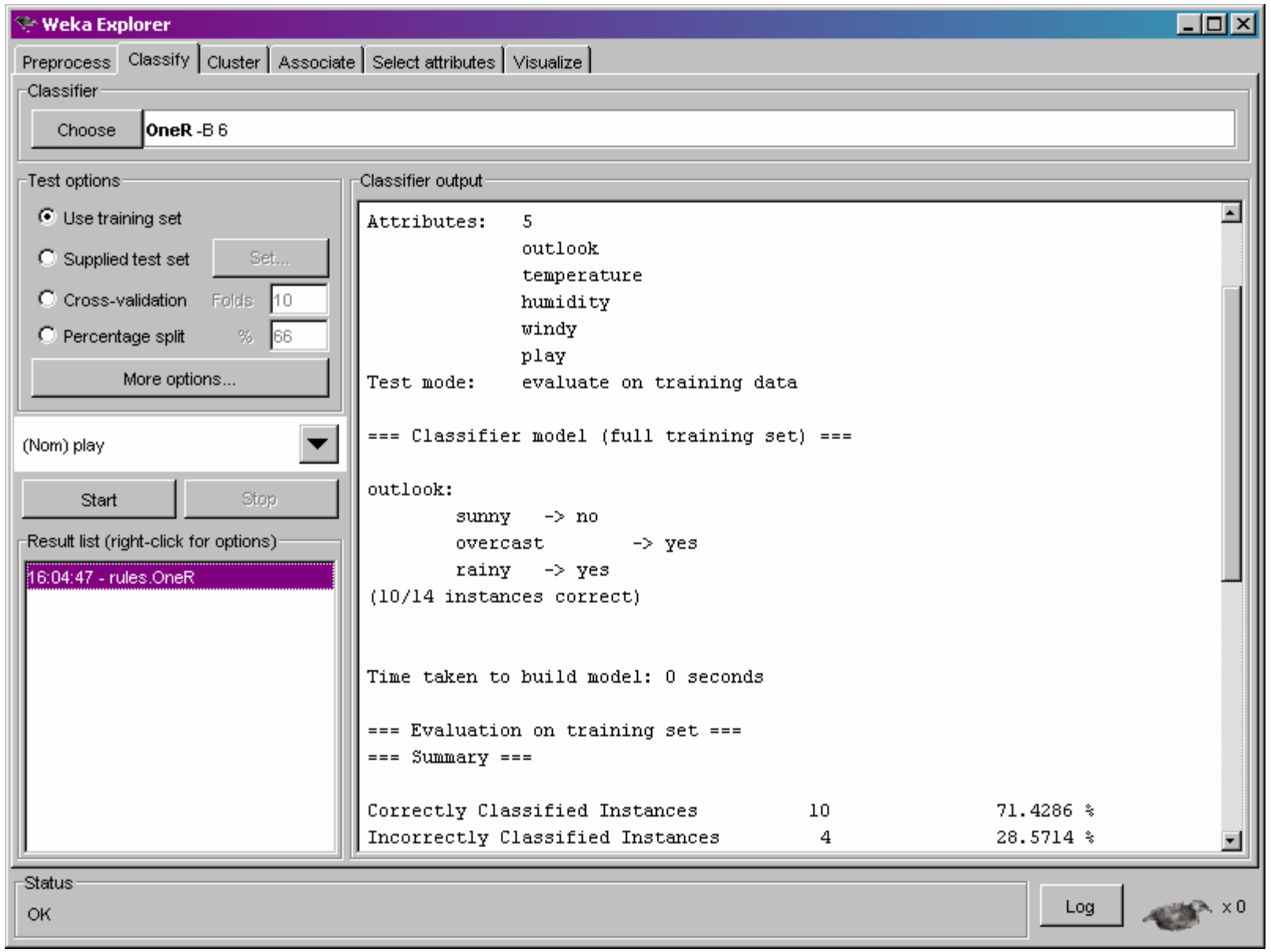
Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Undo OK Cancel

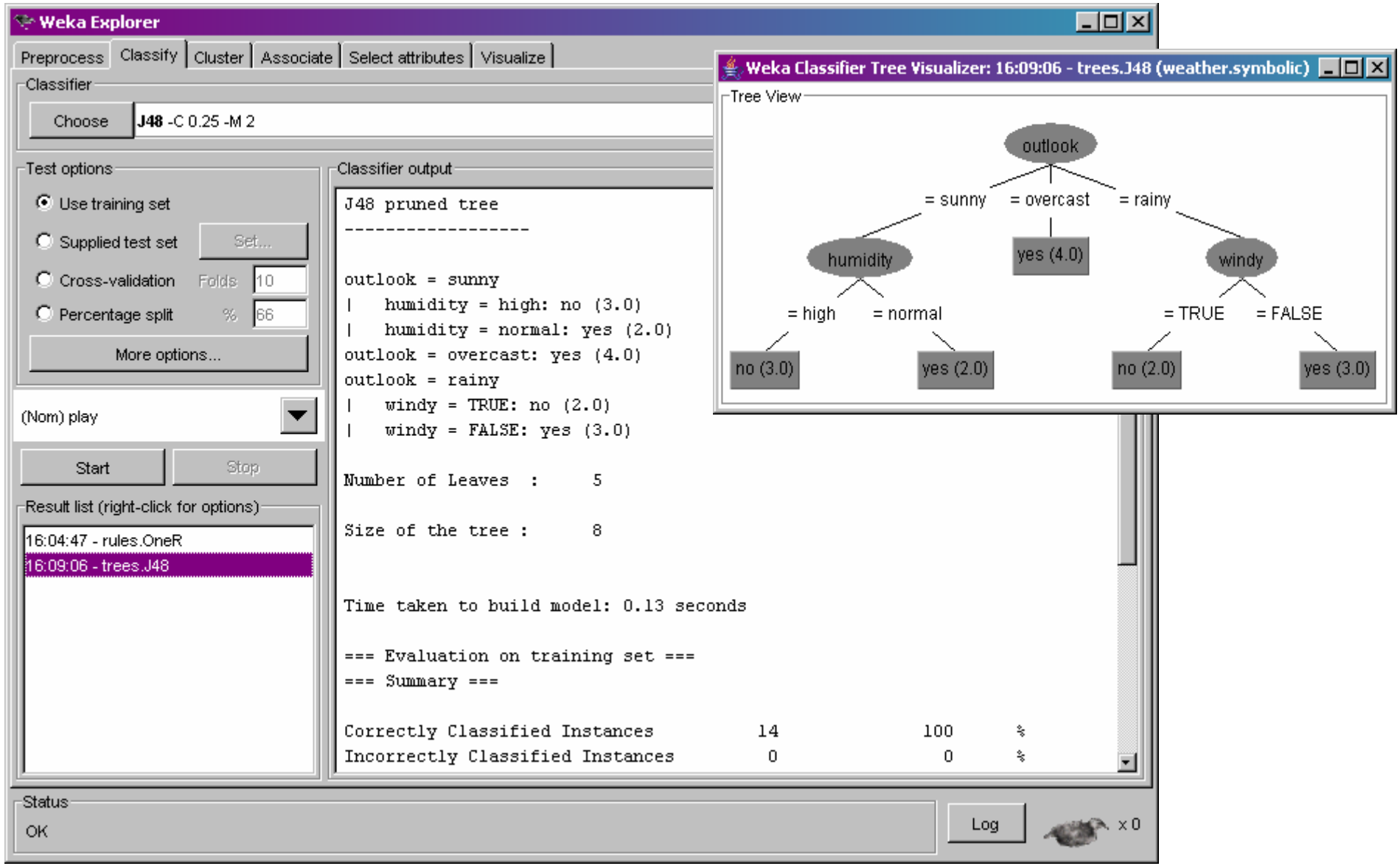
Attribute	Rules	Errors	Total error
outlook	sunny -> no overcast -> yes rainy -> yes	2/5 0/4 2/5	4/14
temperature	hot -> no mild -> yes cool -> yes	2/4 2/6 1/4	5/14
humidity	high -> no normal -> yes	3/7 1/7	4/14
windy	false -> yes true -> no	2/8 3/5	5/14

Classification – OneR



Classification – decision tree

Right click on the highlighted line in Result list and choose Visualize tree



Classification – decision tree

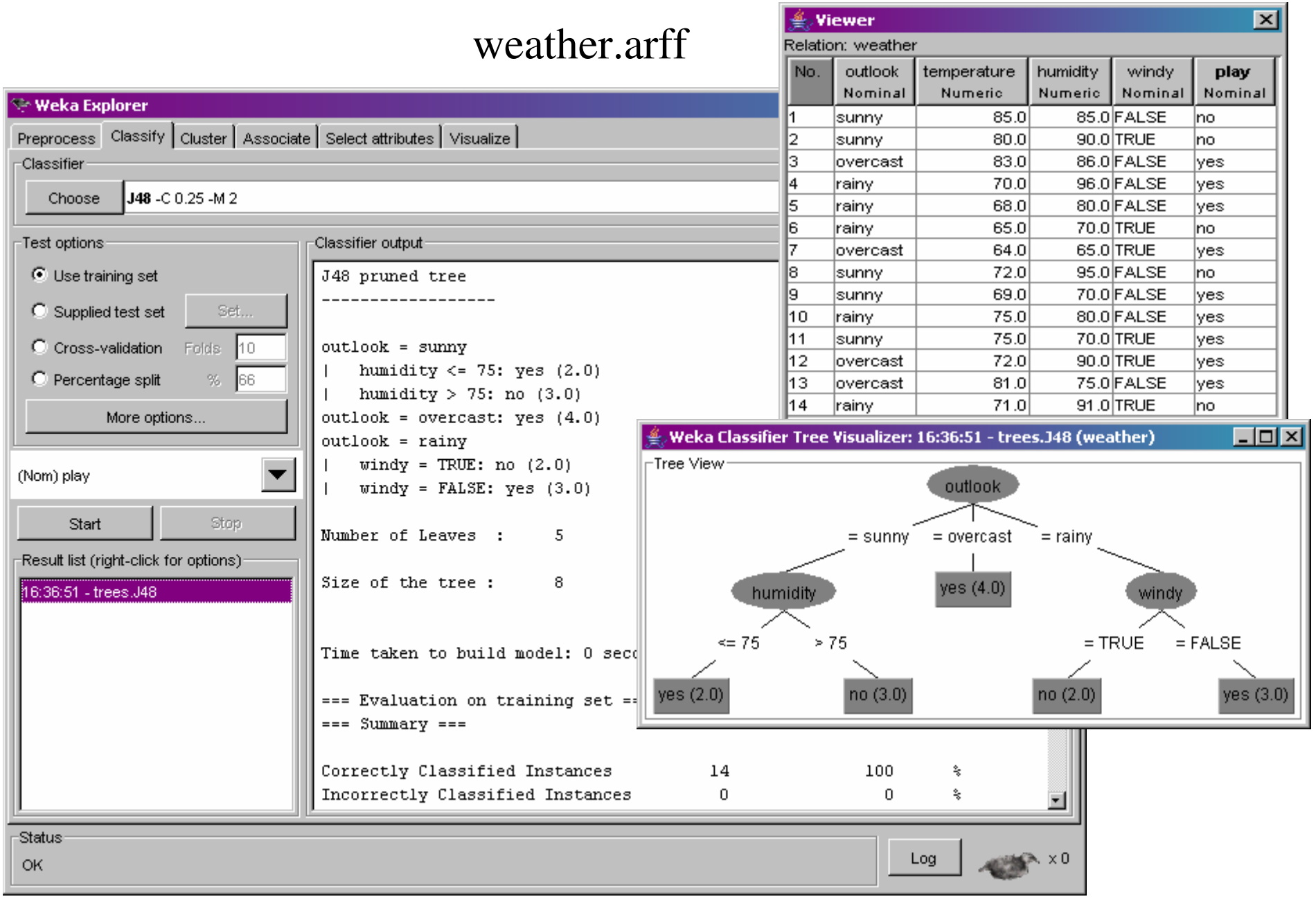
Top-down induction of decision trees (TDIDT, old approach know from pattern recognition):

- Select an attribute for root node and create a branch for each possible attribute value.
- Split the instances into subsets (one for each branch extending from the node).
- Repeat the procedure recursively for each branch, using only instances that reach the branch (those that satisfy the conditions along the path from the root to the branch).
- Stop if all instances have the same class.

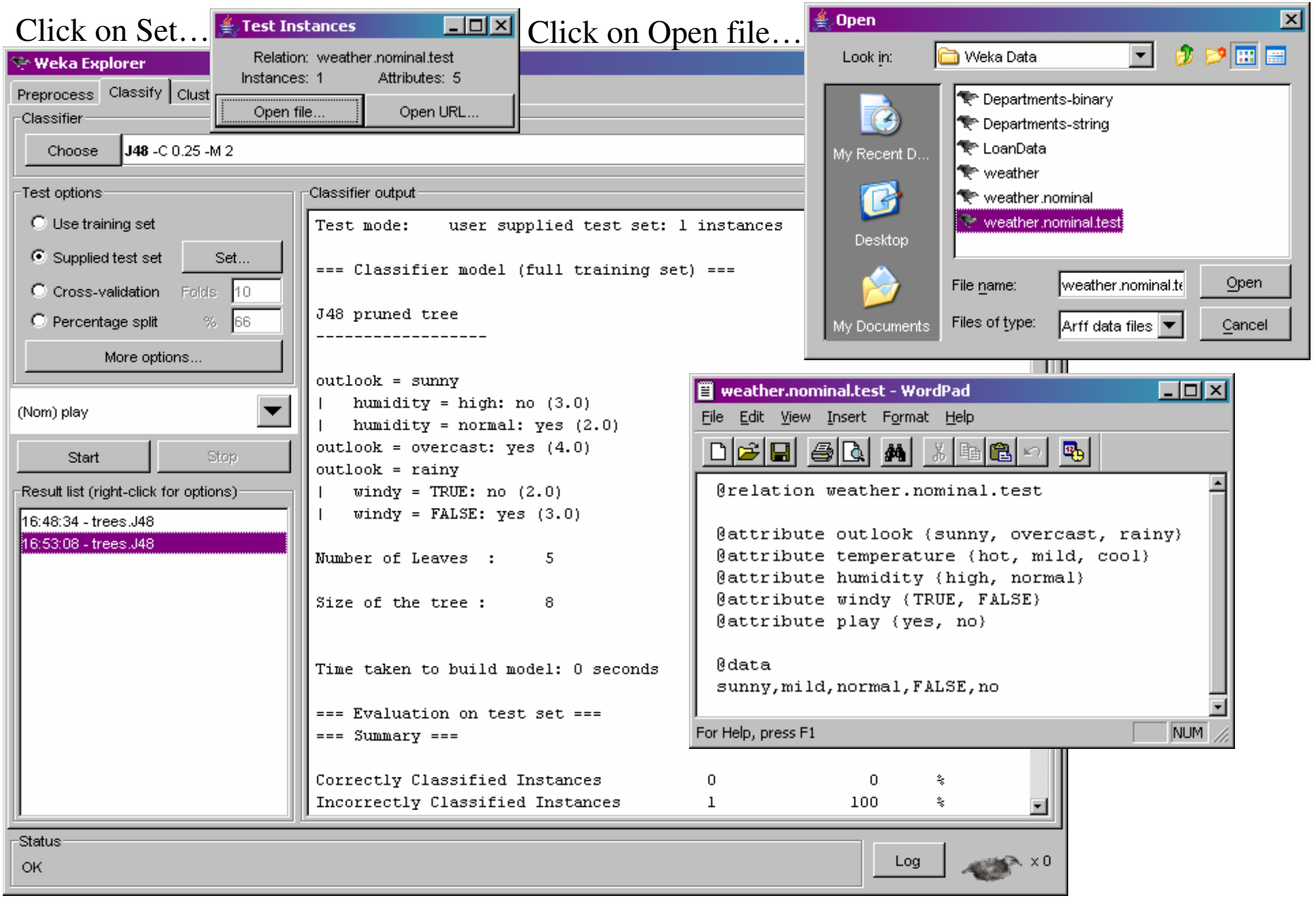
ID3, C4.5, J48 (Weka): Select the attribute that minimizes the class entropy in the split.

Classification – numeric attributes

weather.arff



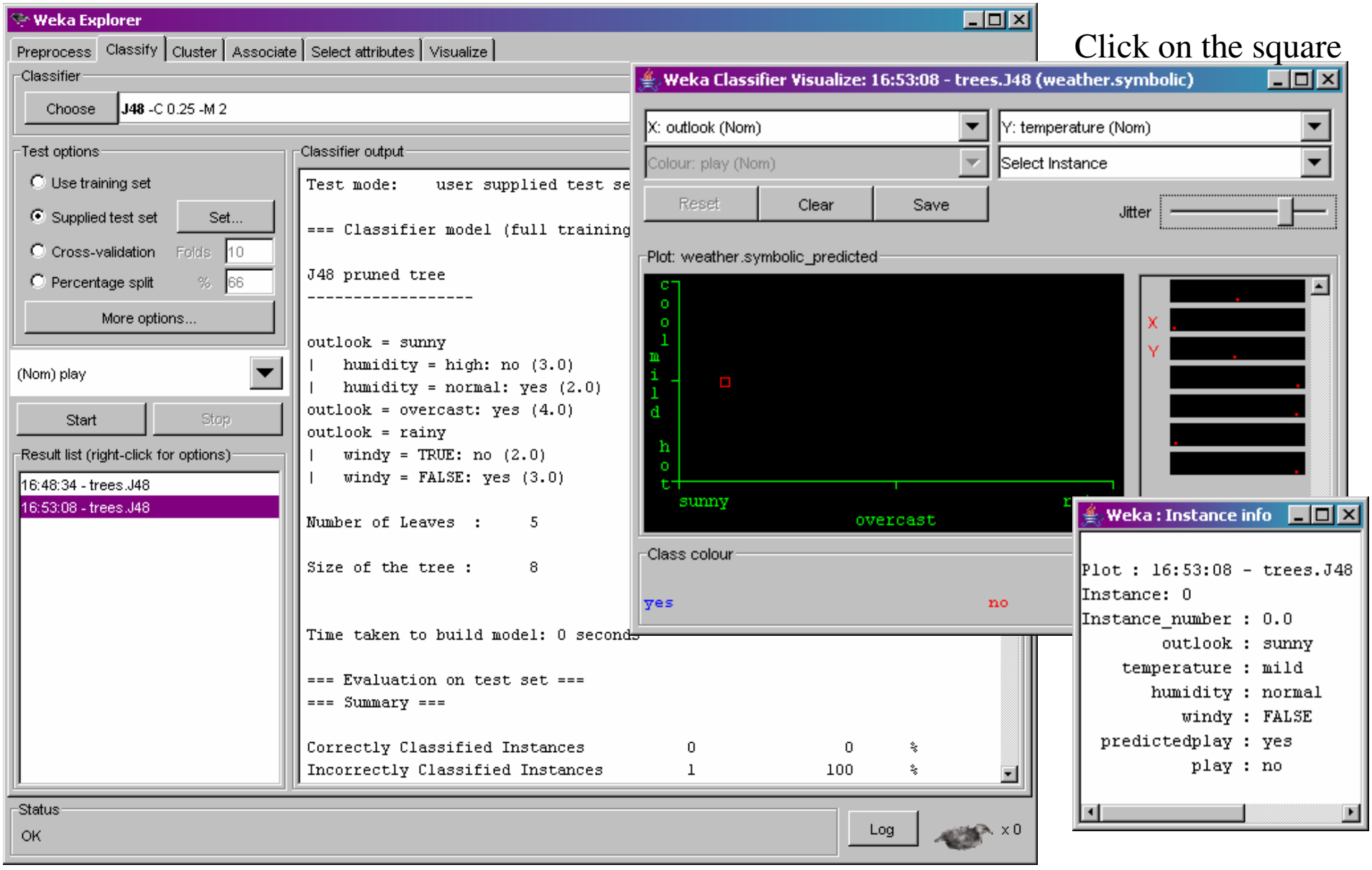
Classification – predicting class



Classification – predicting class

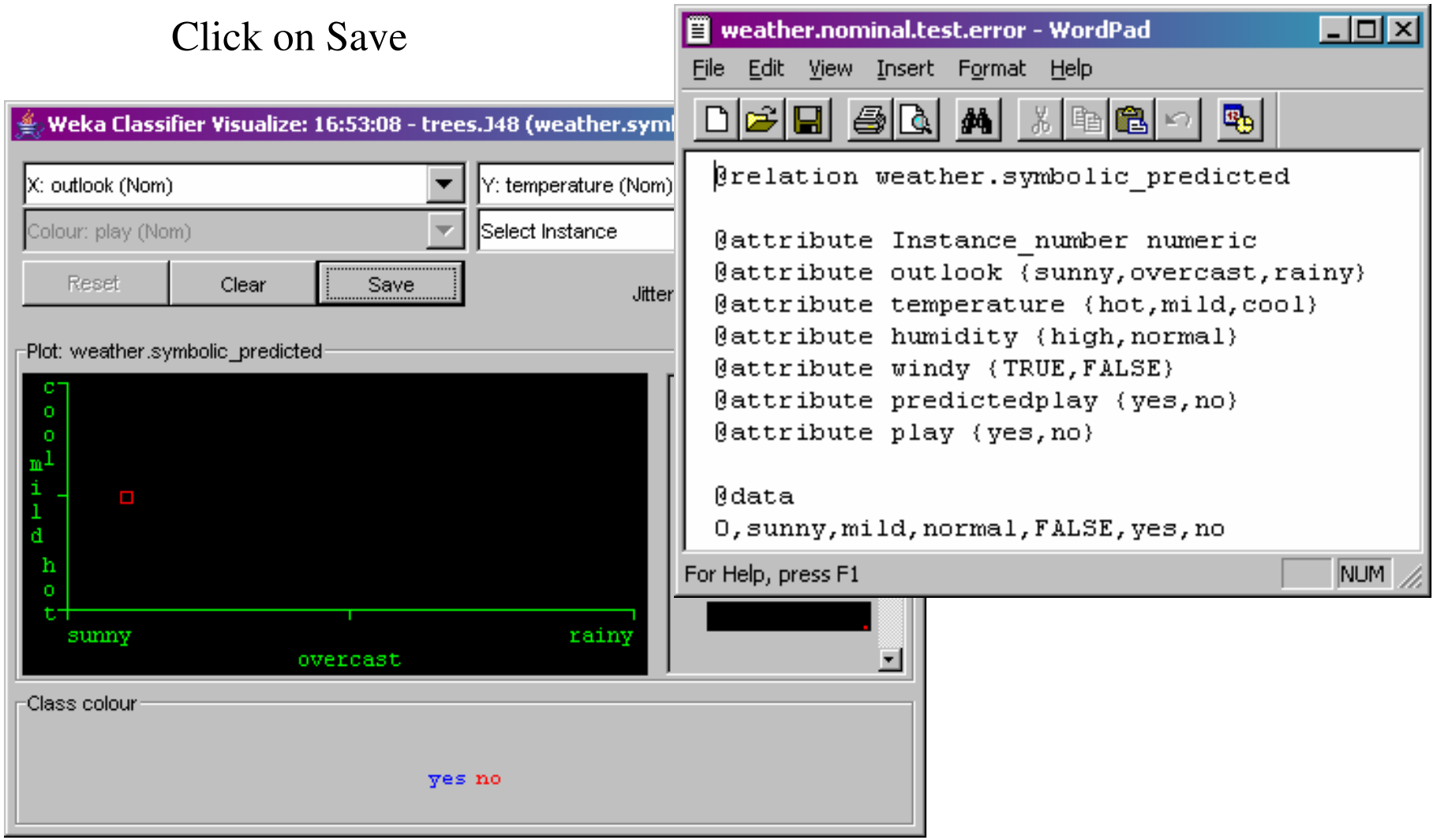
Right click on the highlighted line in Result list and choose Visualize classifier errors

Click on the square

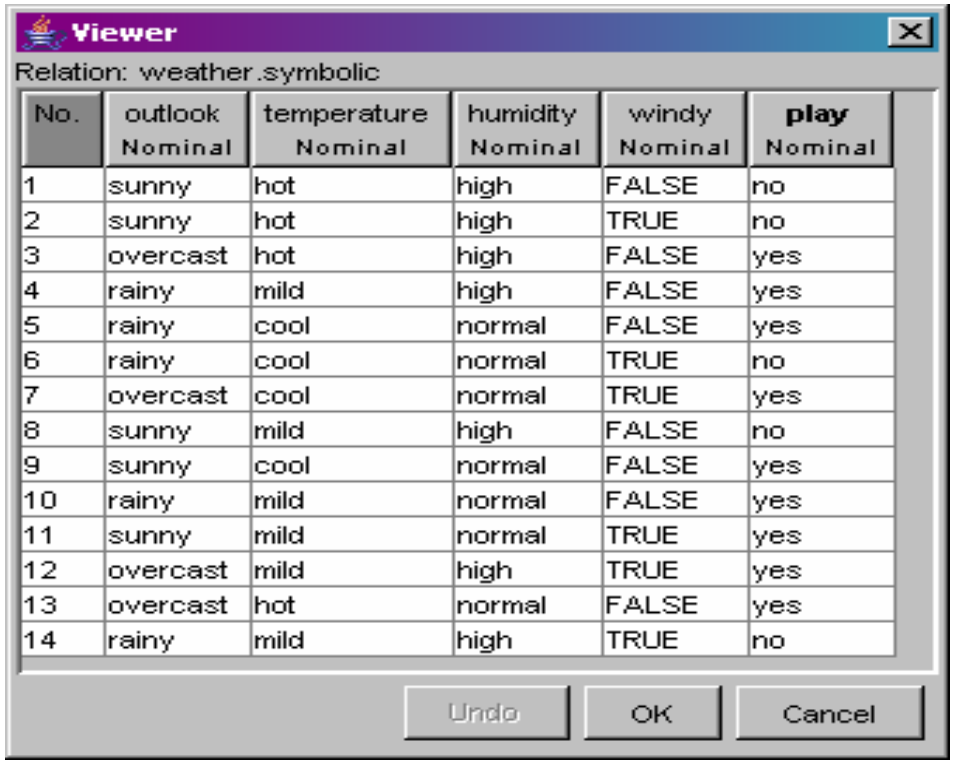


Classification – predicting class

Click on Save



Prediction (no model, lazy learning)



test: (sunny, cool, high, TRUE, ?)

- K-nearest neighbor (KNN, IBk)

  Take the class of the nearest neighbor or the majority class among K neighbors

  K=1 -> no

  K=3 -> no

  K=5 -> yes

  K=14 -> yes (Majority predictor, ZeroR)

- Weighted K-nearest neighbor

  K=5 -> undecided

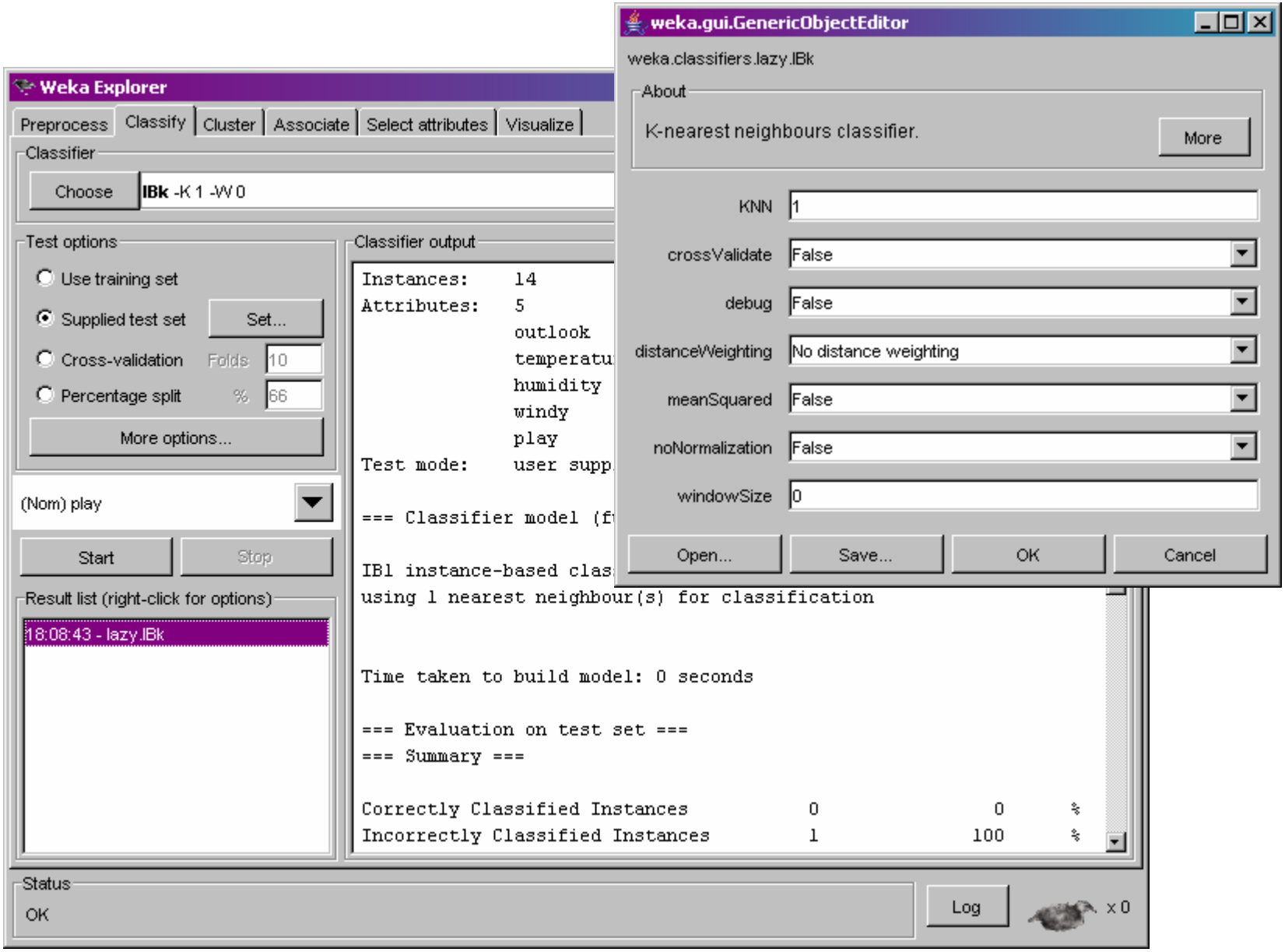
  no=1/1+1/2=1.5

  yes=1/2+1/2+1/2=1.5

X	2	8	9	11	12	…	10
Distance(test,X)	1	2	2	2	2	…	4
play	no	no	yes	yes	yes	…	yes

- Distance is calculated as the number of different attribute values
- Euclidean distance for numeric attributes

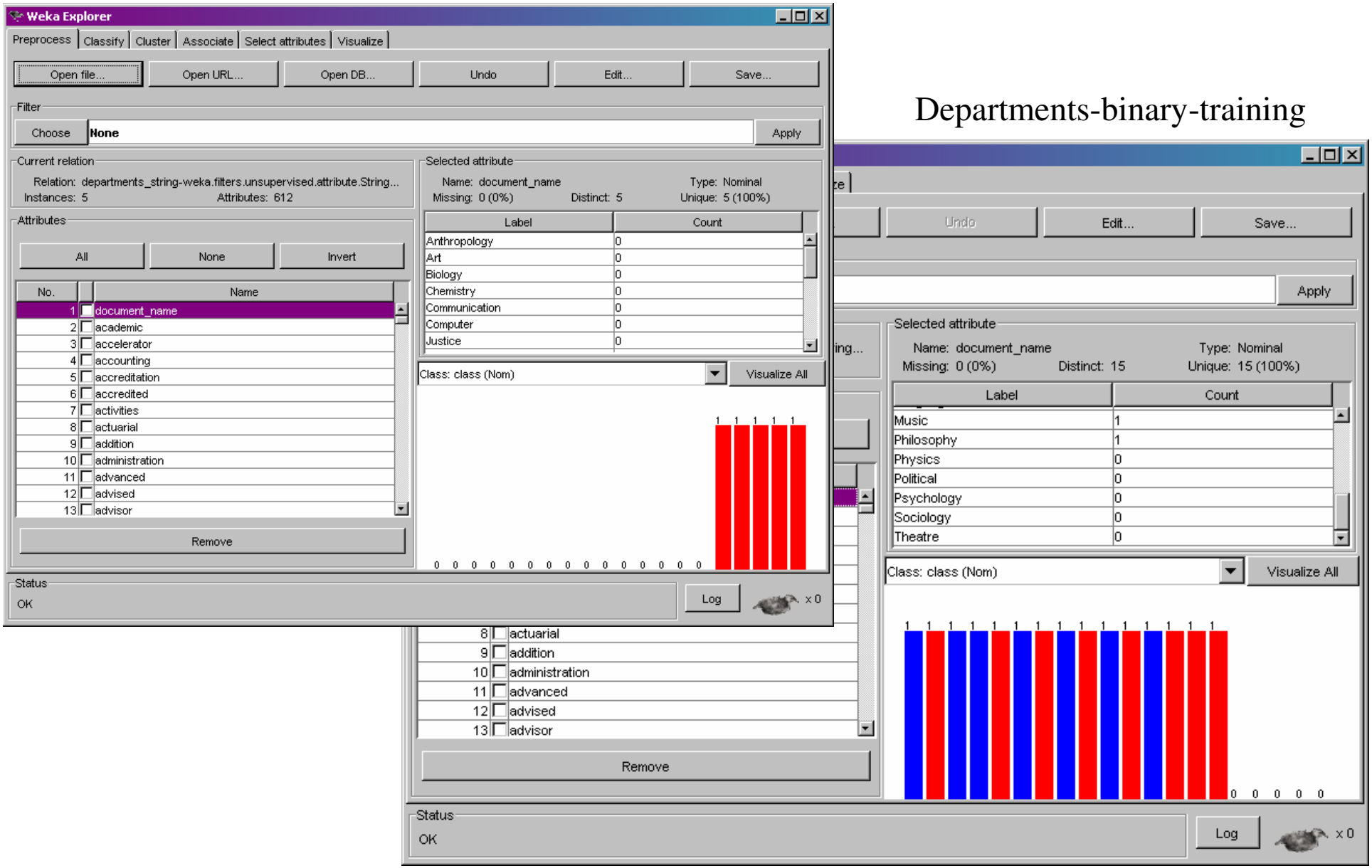
Prediction (no model, lazy learning)



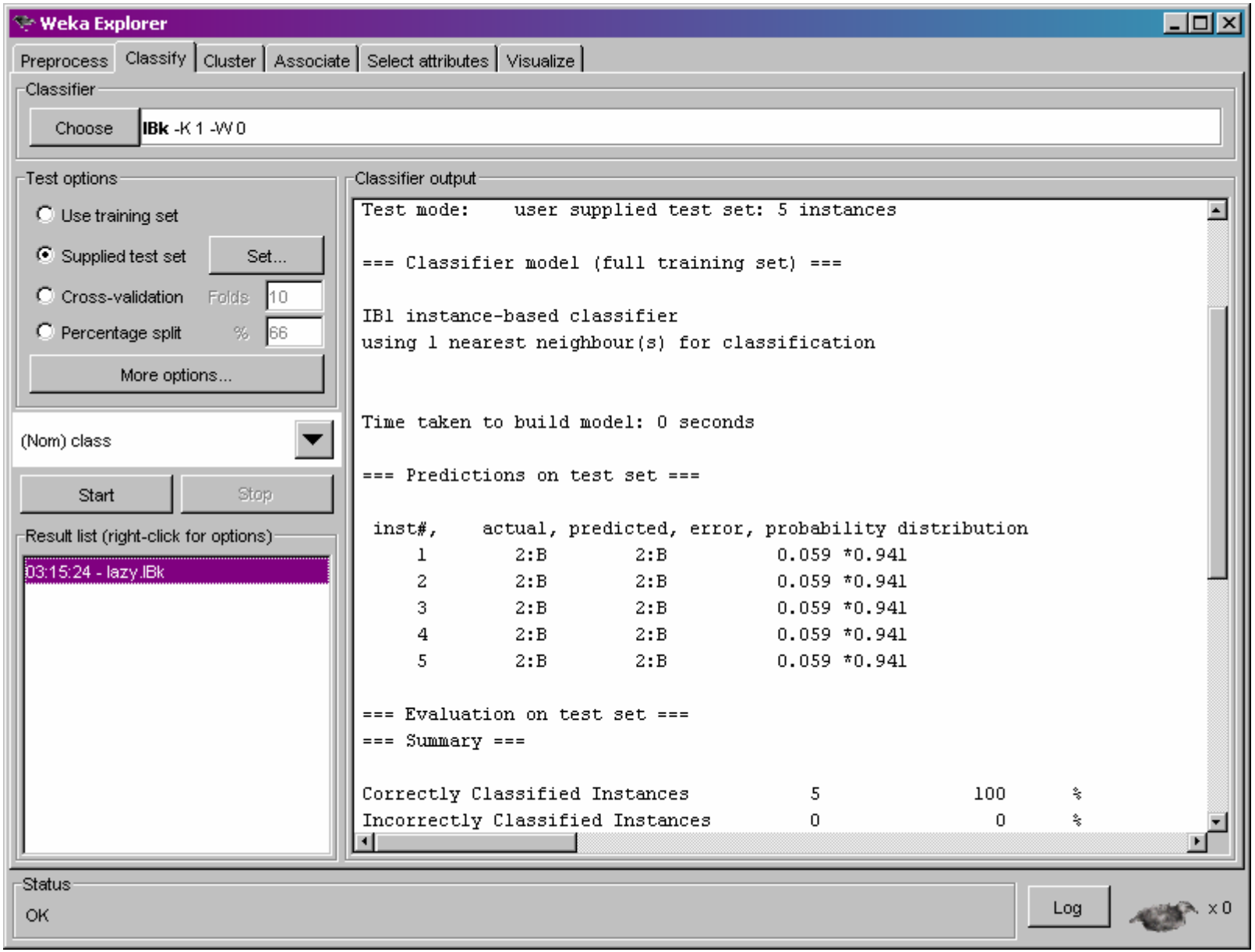
Prediction (no model, lazy learning)

Departments-binary-test.arff

Departments-binary-training

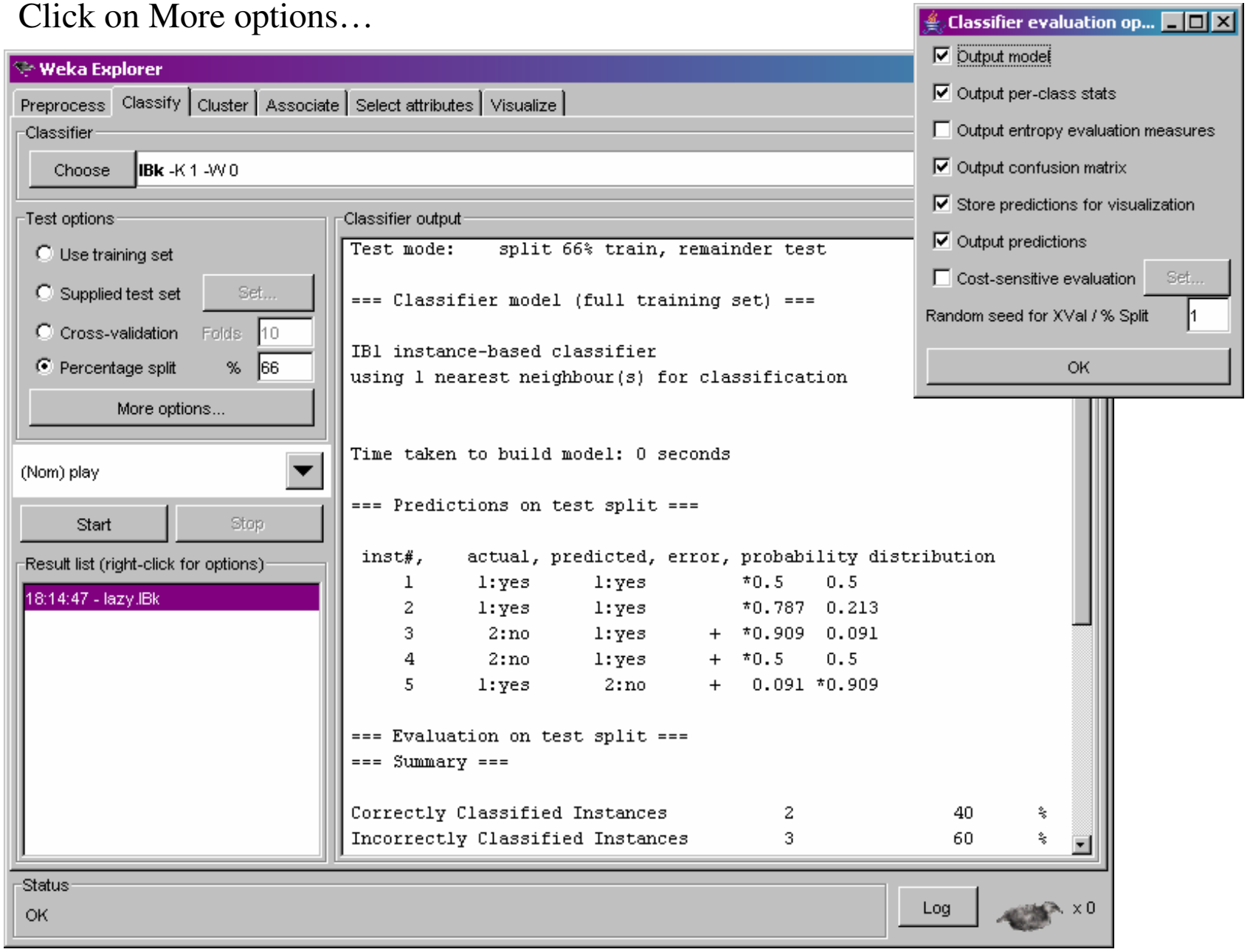


Prediction (no model, lazy learning)

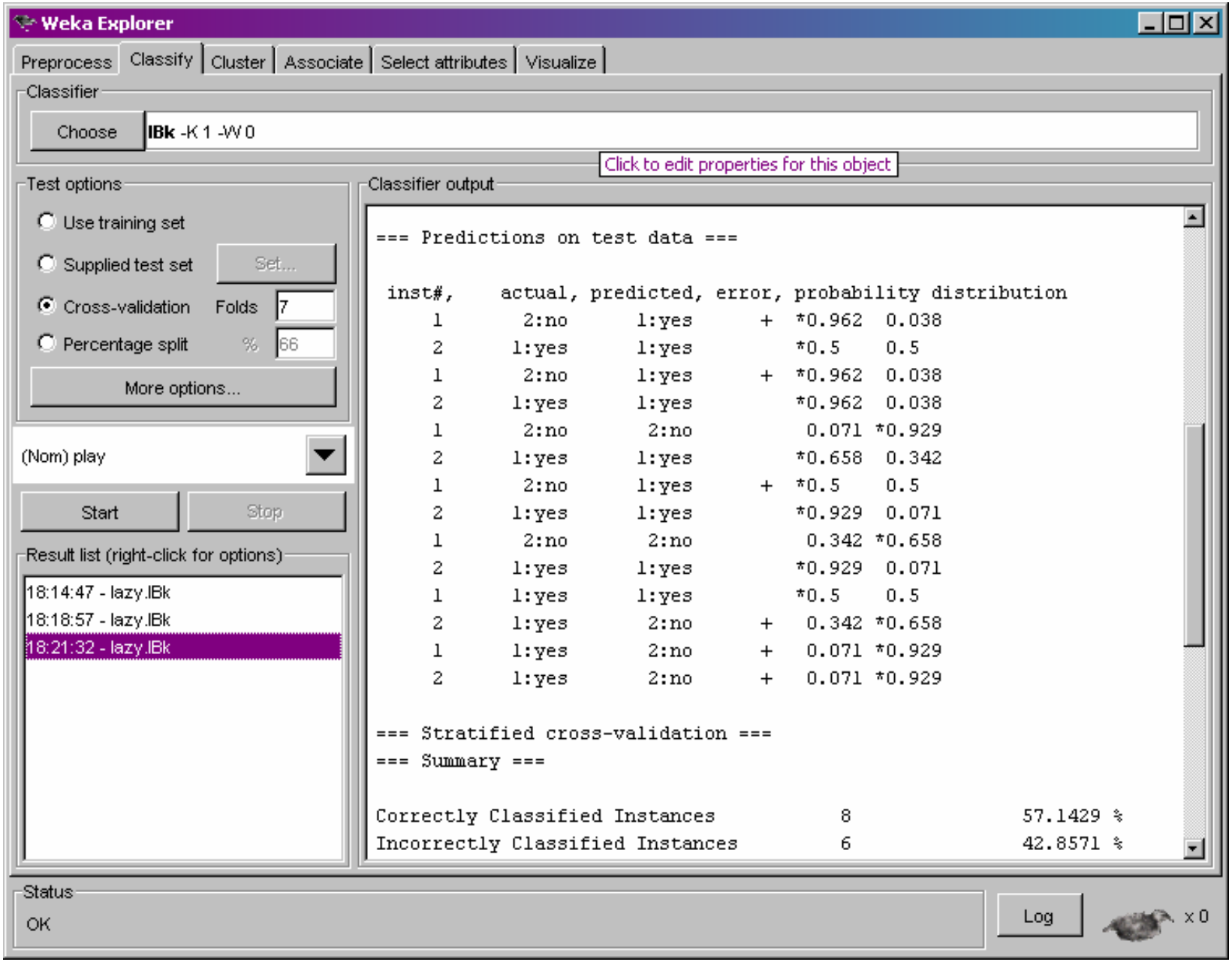


Model evaluation – holdout (percentage split)

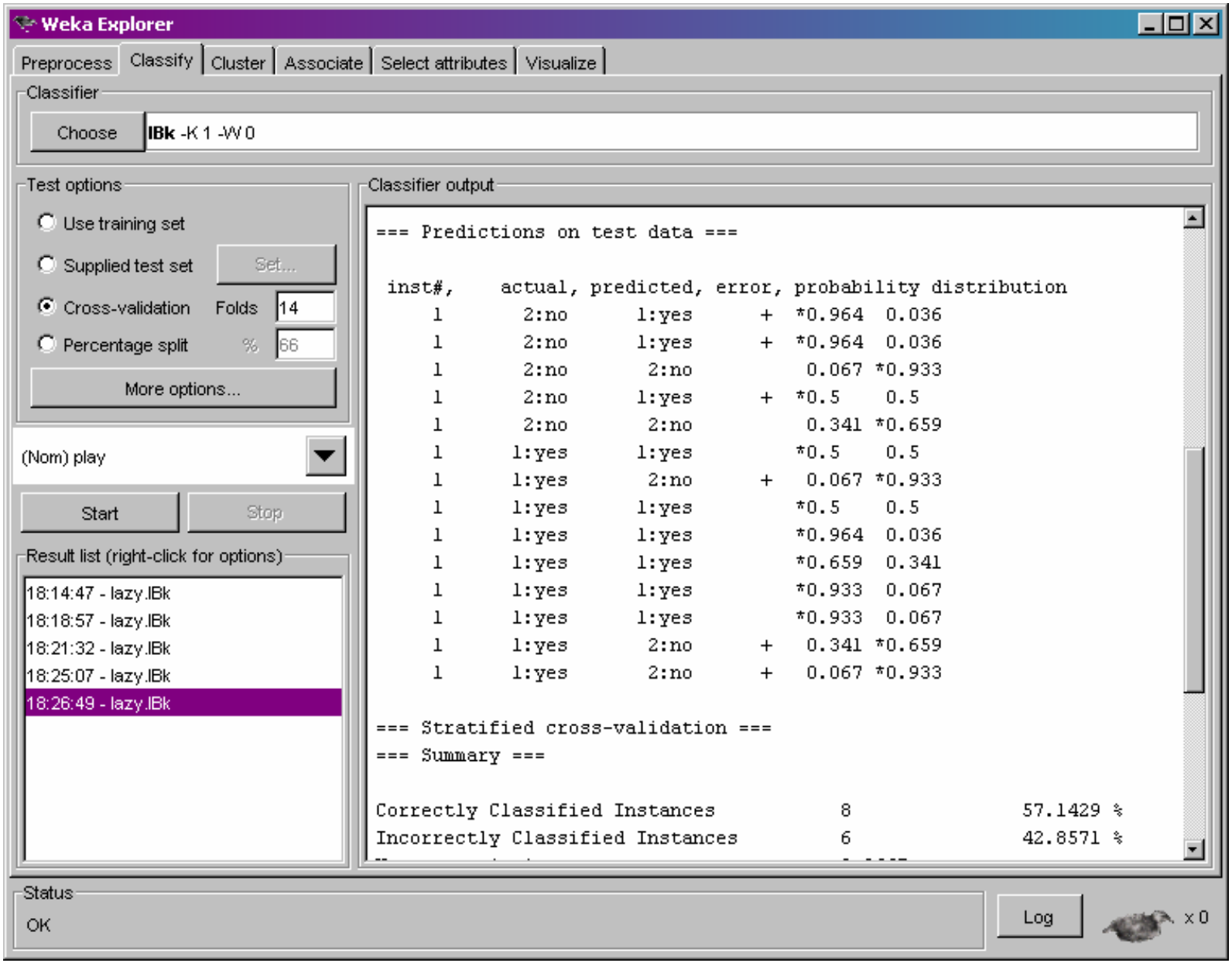
Click on More options…



Model evaluation – cross validation



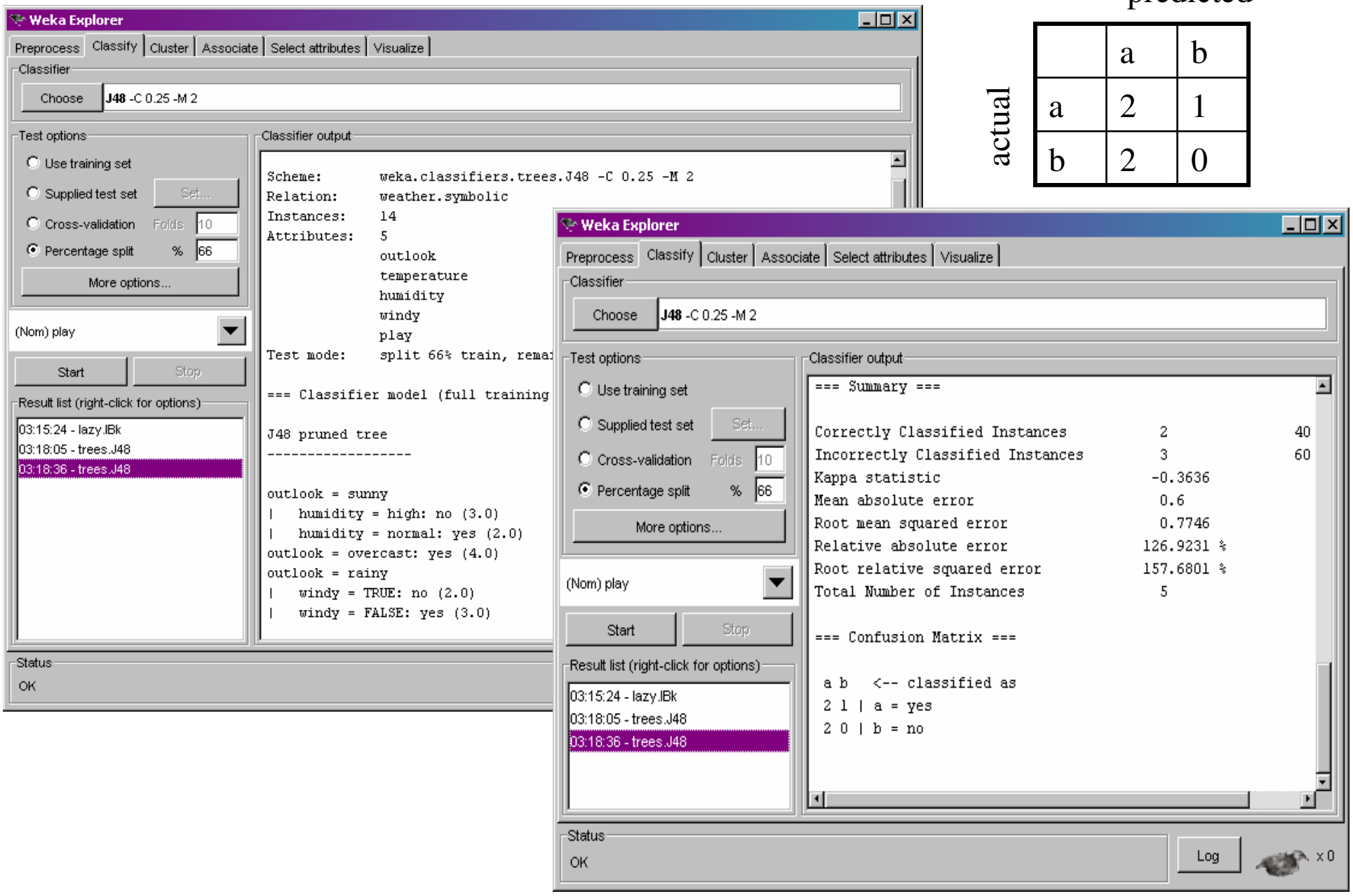
Model evaluation – leave one out cross validation



Model evaluation – confusion (contingency) matrix

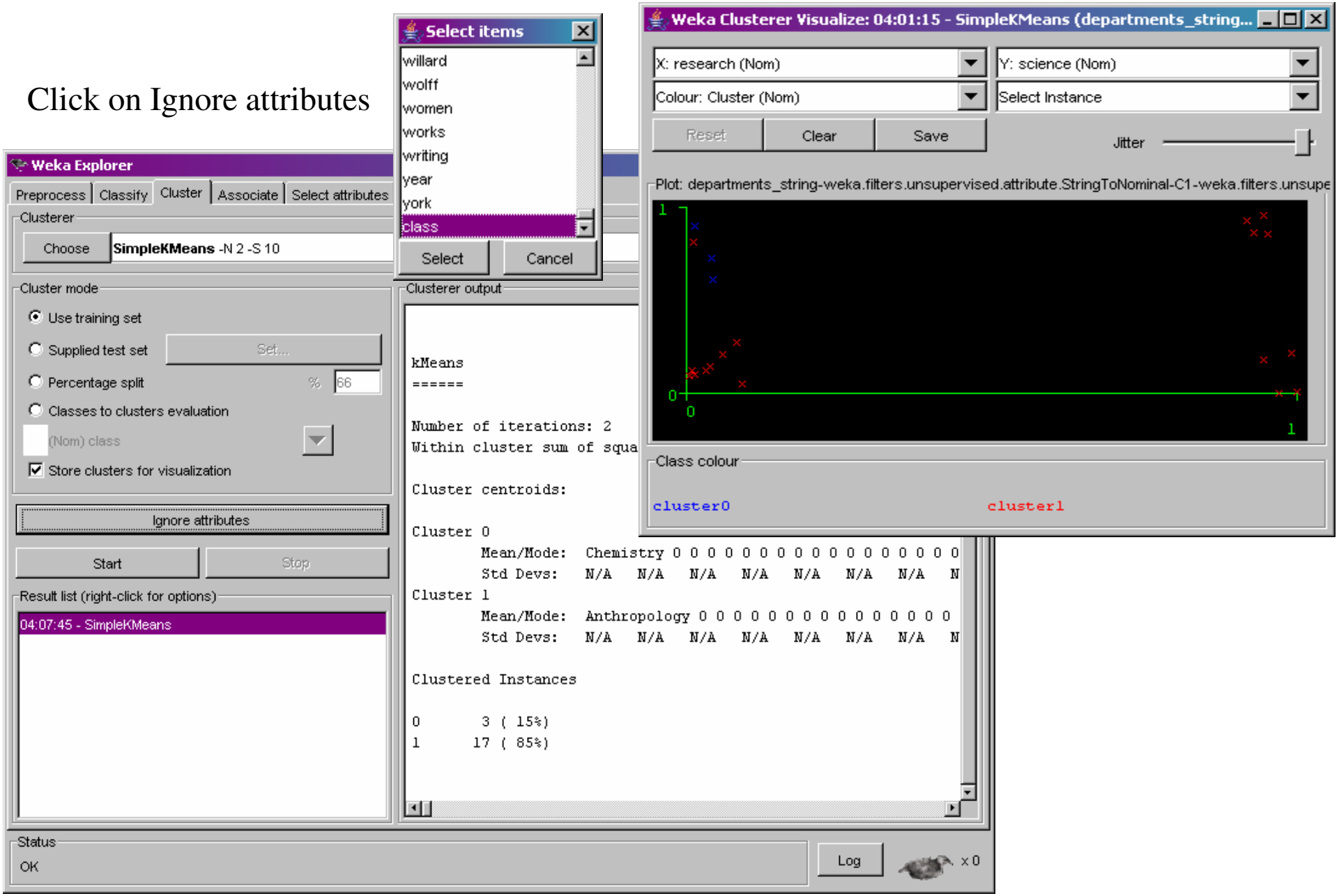
predicted

actual		a	b
	a	2	1
	b	2	0

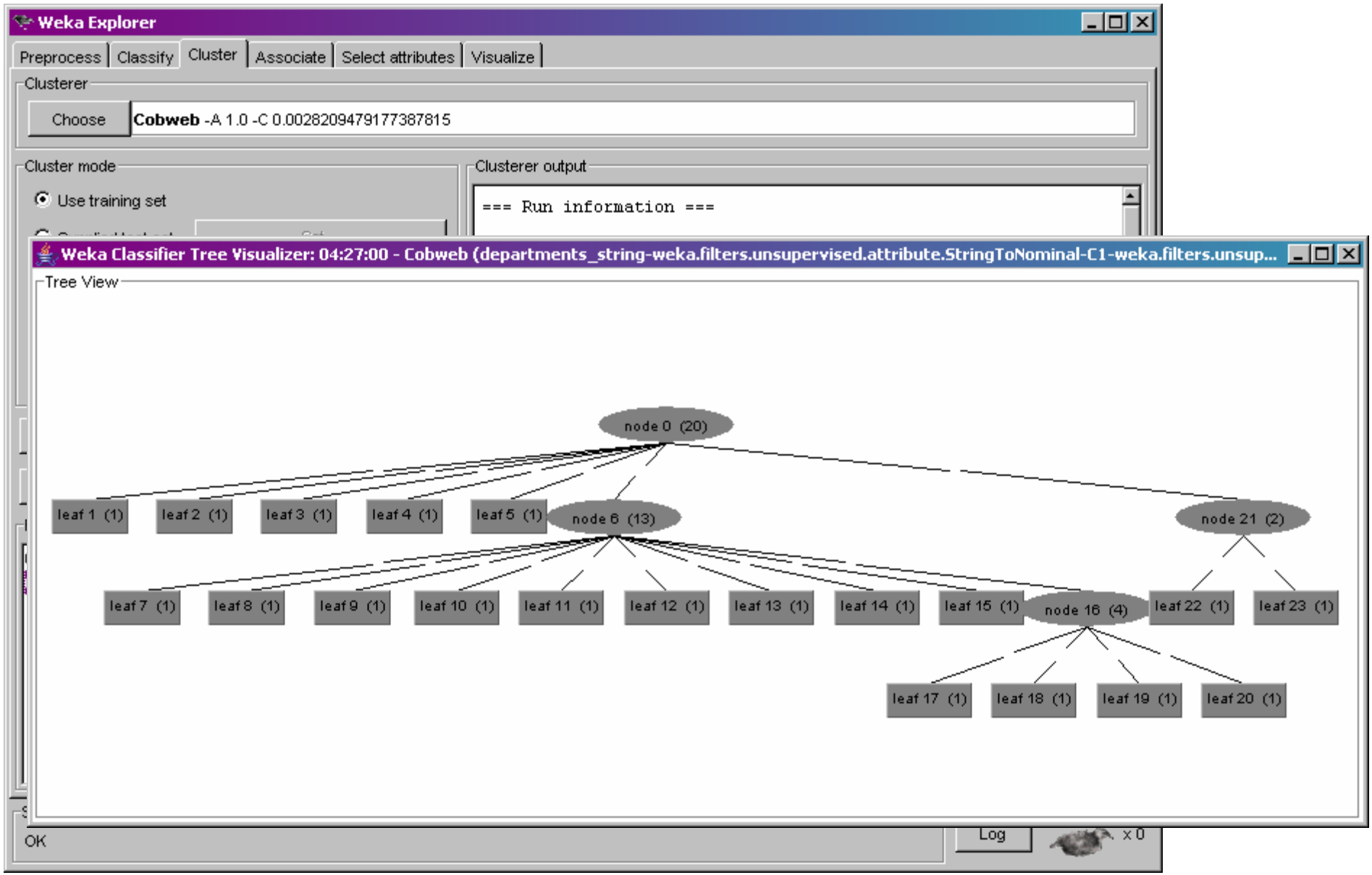


Clustering – k-means

Click on Ignore attributes

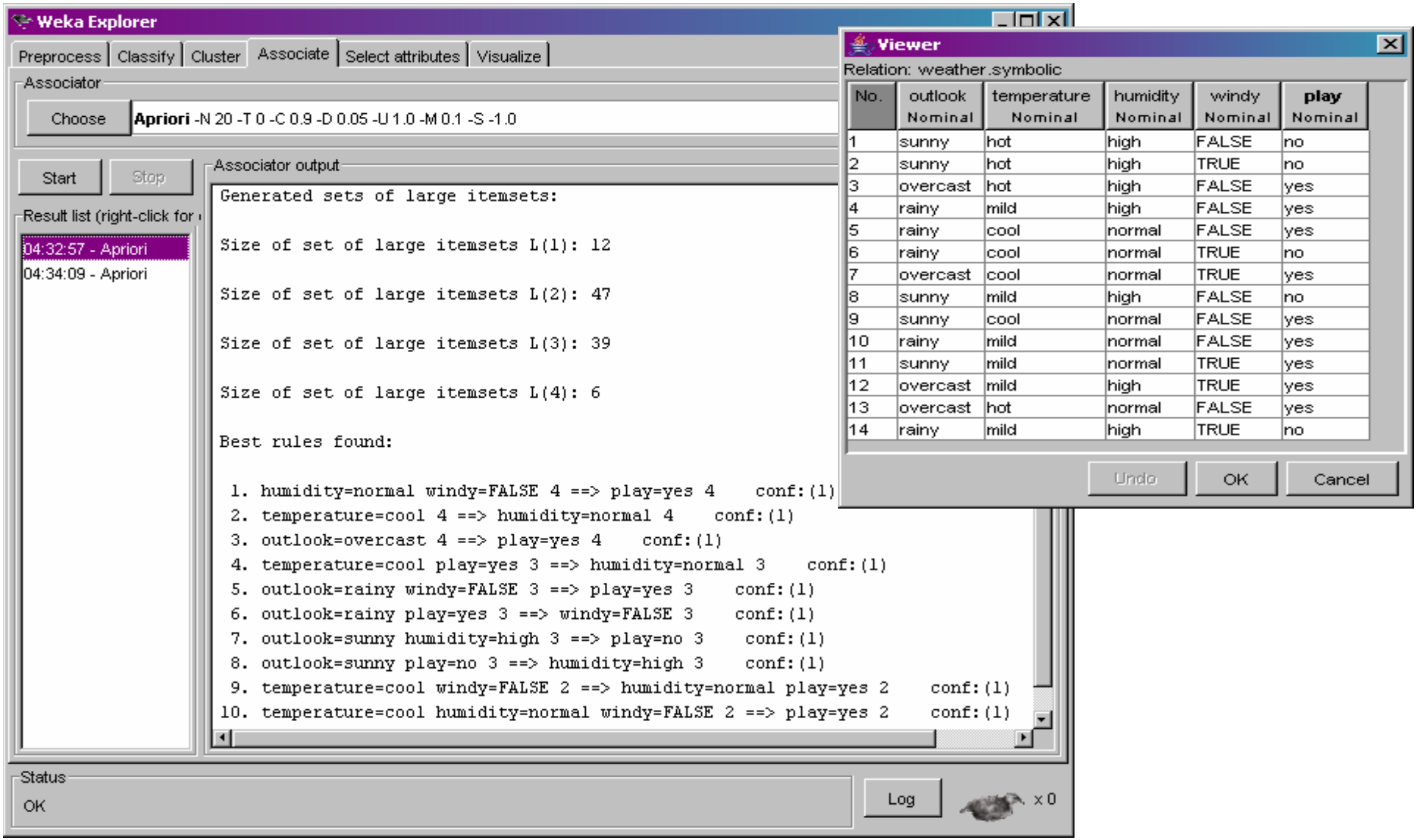


Hierarchical Clustering – Cobweb

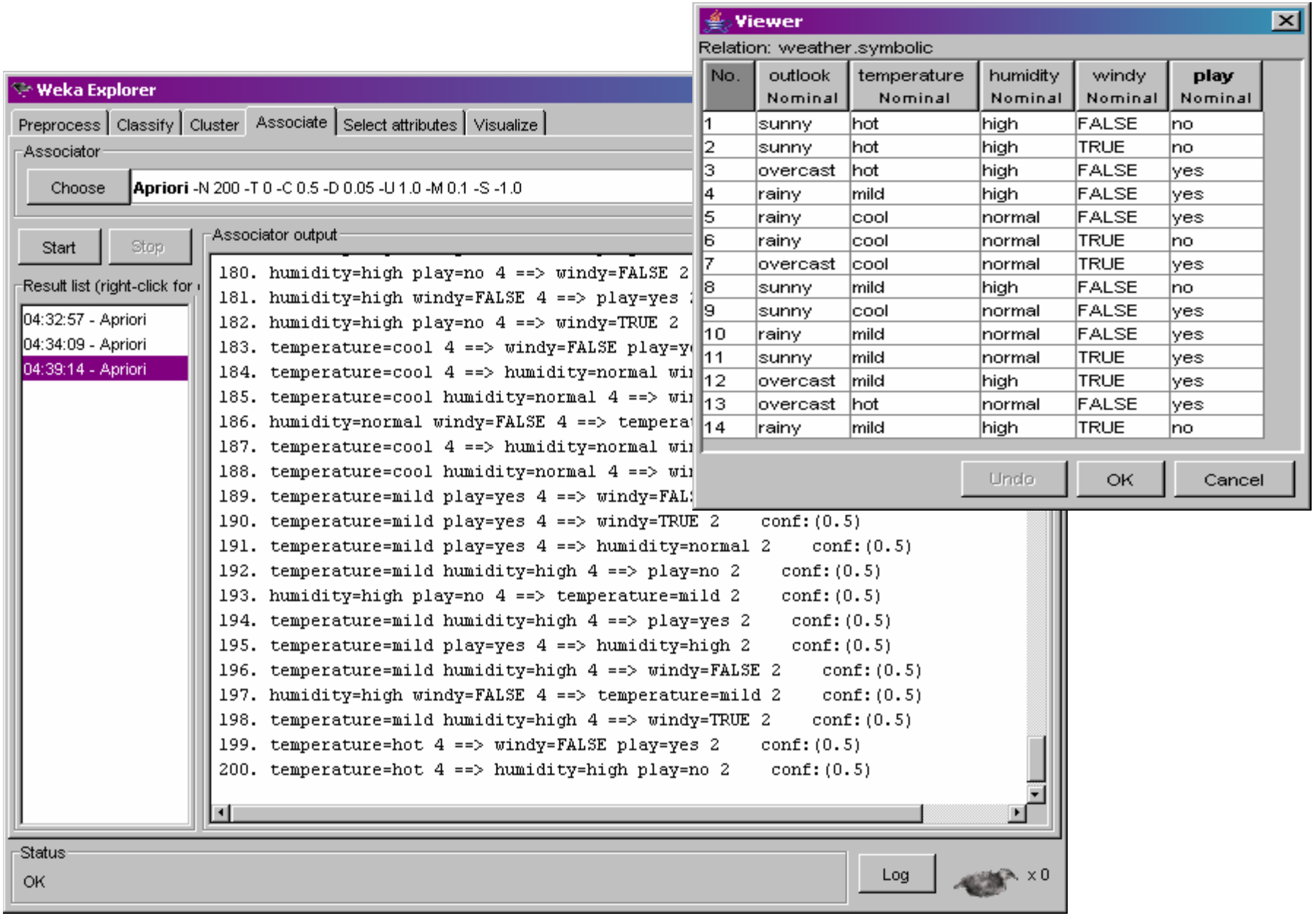


Association Rules (A => B)

- *Confidence* (accuracy): P(B|A) = (# of tuples containing both A and B) / (# of tuples containing A).
- *Support* (coverage): P(A,B) = (# of tuples containing both A and B) / (total # of tuples)



Association Rules



And many more …

Thank you!