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

Agenda

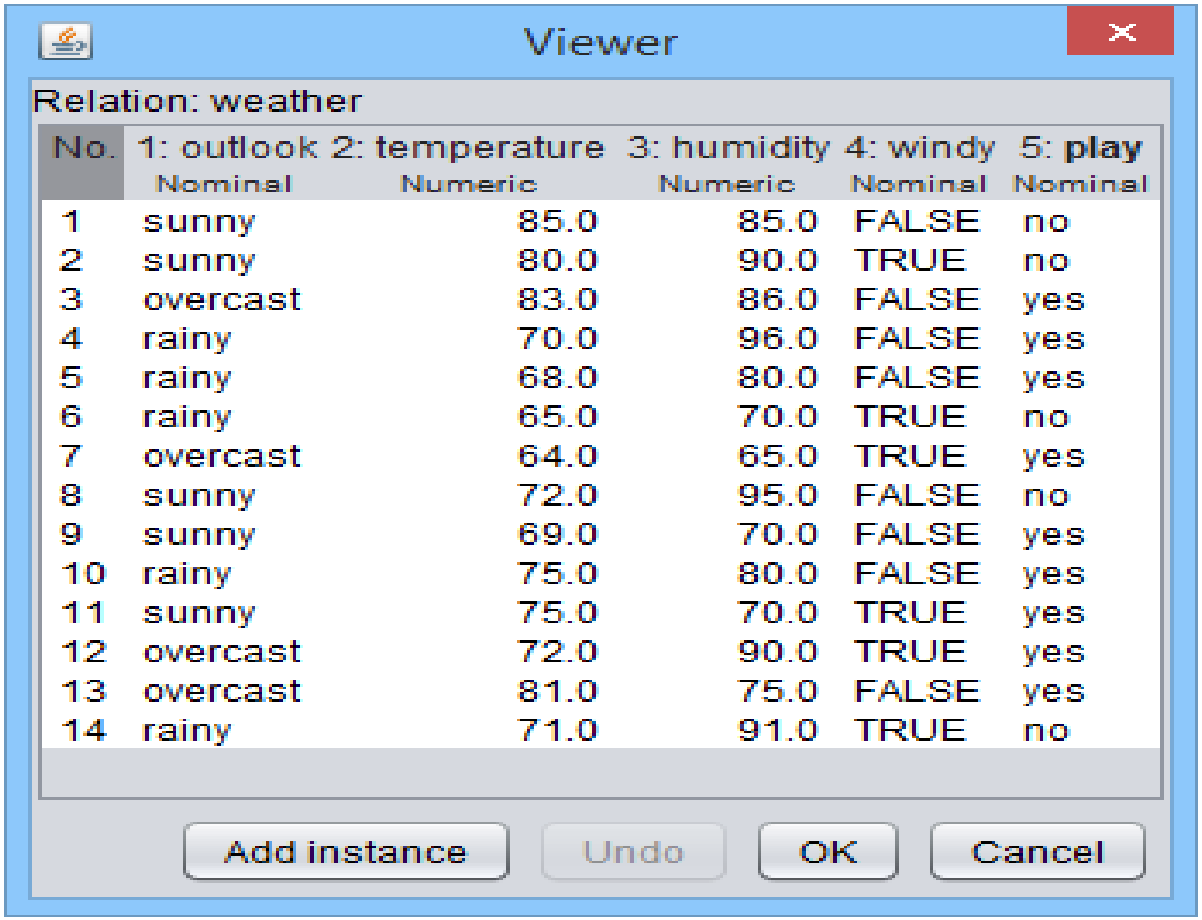
- Data Mining
- Weka Project
- Basic functionality of Weka by example
- Weka for document classification and clustering

Database management systems (DBMS), Online Analytical Processing (OLAP) and Data Mining

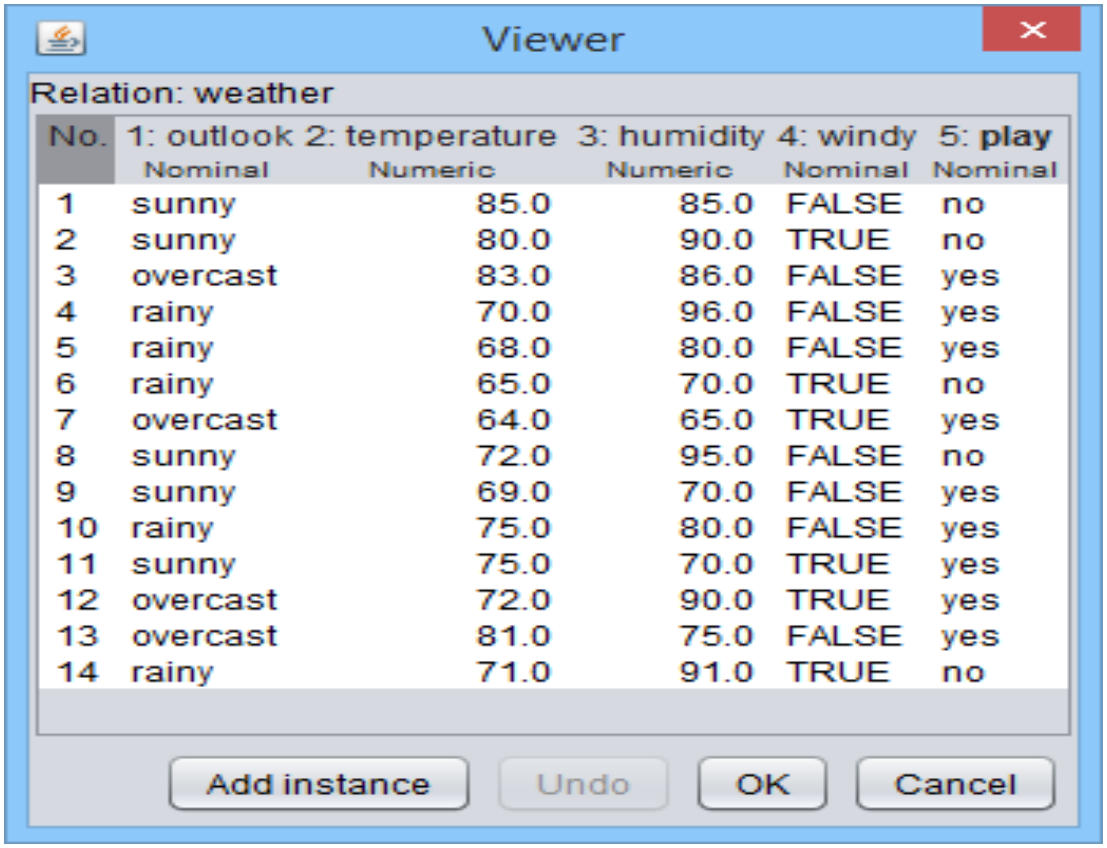
Area	DBMS	OLAP	Data Mining
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

Example of DBMS, OLAP and Data Mining: Weather data

Assume we have made a record of the weather conditions during a two-week period, along with the decisions of a tennis player whether or not to play tennis on each particular day. Thus we have generated **tuples** (or examples, instances) consisting of values of four **independent variables** (outlook, temperature, humidity, windy) and one **dependent variable** (play).



Data-Base Management System



- What was the temperature in the sunny days? {85, 80, 72, 69, 75}
- Which days the humidity was less than 75? {6, 7, 9, 11}
- Which days the temperature was greater than 70? {1, 2, 3, 8, 10, 11, 12, 13, 14}
- Which days the temperature was greater than 70 and the humidity was less than 75? The intersection of the above two: {11}

OLAP: Multidimensional Model (Data Cube)

Dimensions:

- Time: Week 1={1, 2, 3, 4, 5, 6, 7}, Week 2={8, 9, 10, 11, 12, 13, 14}
- Outlook: {sunny, rainy, overcast}

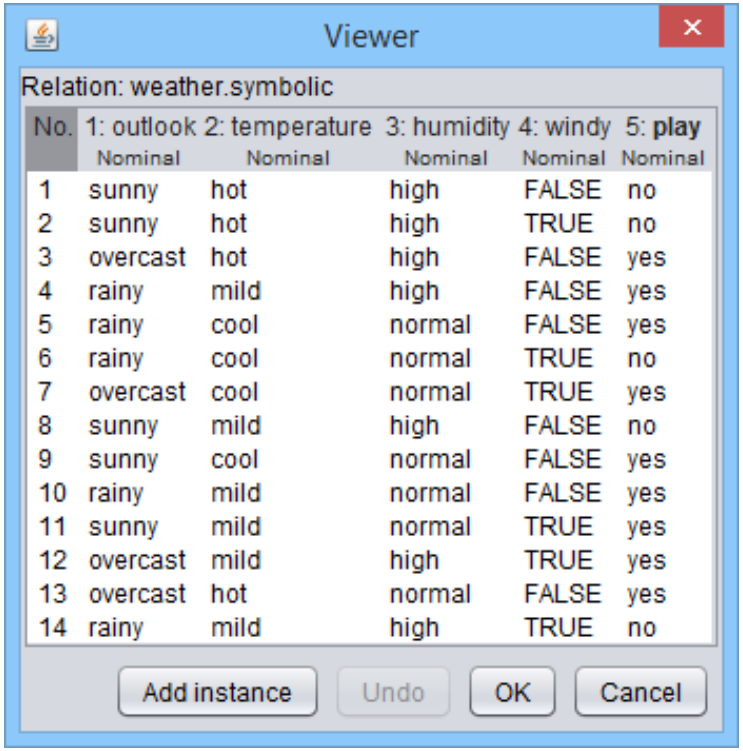
Unit: play (yes/no)

9 / 5	sunny	rainy	overcast
Week 1	0 / 2	2 / 1	2 / 0
Week 2	2 / 1	1 / 1	2 / 0

➔ if outlook = overcast then play = yes

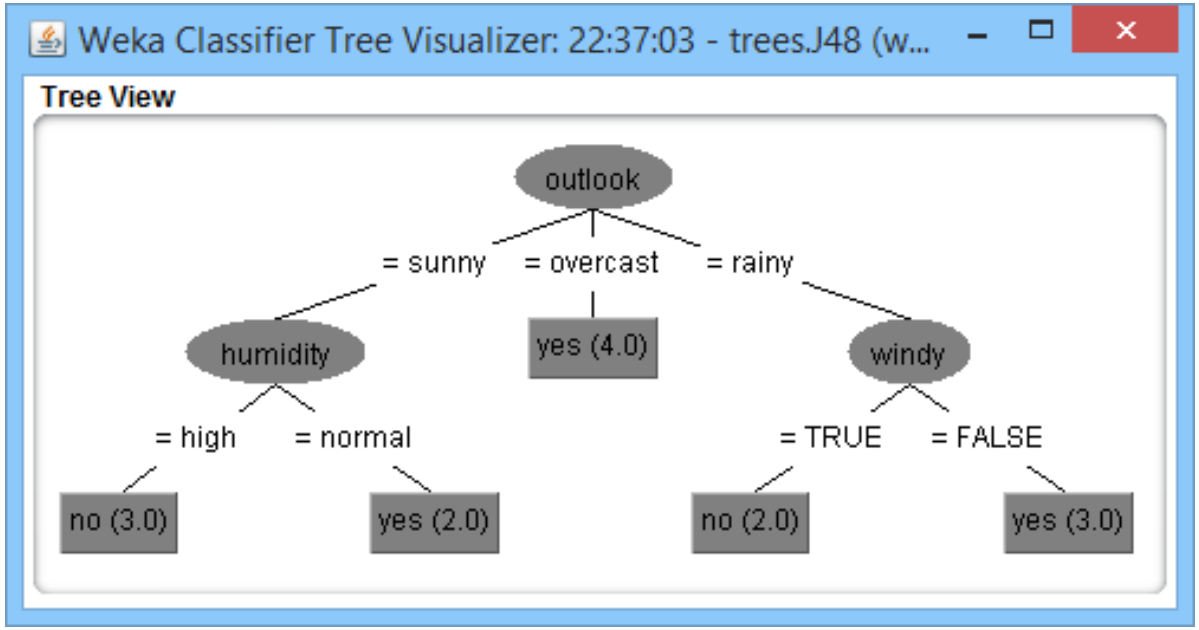
Data Mining: Association Rules

Discretize numeric attributes (data **pre-processing stage** in data mining). Group the temperature values in three intervals (hot, mild, cool) and humidity values in two (high, normal).



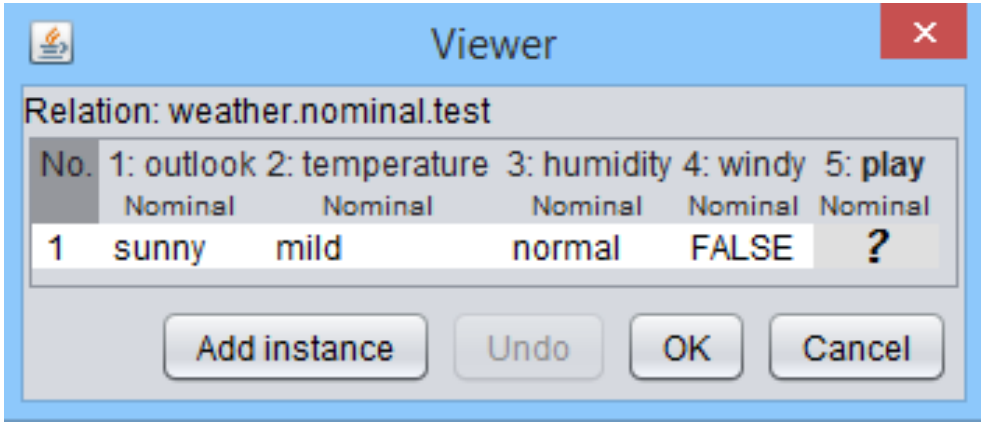
1. humidity=normal windy=false 4 ==> play=yes (4, 1)
2. temperature=cool 4 ==> humidity=normal (4, 1)
3. outlook=overcast 4 ==> play=yes (4, 1)
4. temperature=cool play=yes 3 ==> humidity=normal (3, 1)
5. outlook=rainy windy=false 3 ==> play=yes (3, 1)
6. outlook=rainy play=yes 3 ==> windy=false (3, 1)
7. outlook=sunny humidity=high 3 ==> play=no (3, 1)
8. outlook=sunny play=no 3 ==> humidity=high (3, 1)
9. temperature=cool windy=false 2 ==> humidity=normal play=yes (2, 1)
10. temperature=cool humidity=normal windy=false 2 ==> play=yes (2, 1)

Data Mining: Decision Tree and Rules



If outlook = overcast then yes
If humidity = normal and windy = false then yes
If temperature = mild and humidity = normal then yes
If outlook = rainy and windy = false then yes
If outlook = sunny and humidity = high then no
If outlook = rainy and windy = true then no

Data Mining: Prediction



P(play=yes | outlook=sunny, temperature=mild, humidity=normal, windy=false) = 0.8

P(play=no | outlook=sunny, temperature=mild, humidity=normal, windy=false) = 0.2

Weka Project

12 Years ago …

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 usable 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.

12 Years ago …

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.

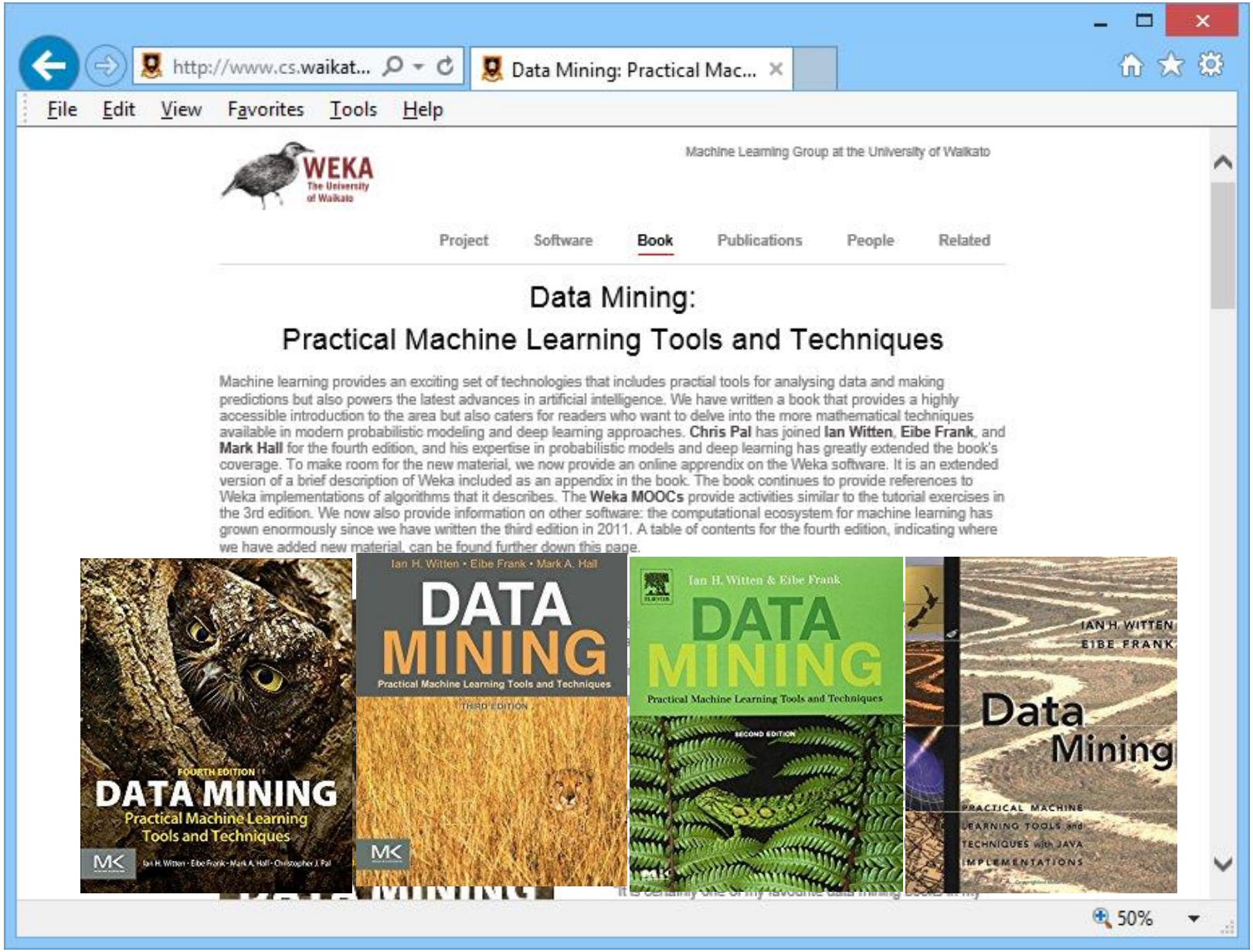
This book became one of the most popular textbooks for data mining and machine learning, and is **very frequently cited in scientific publications**.

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.

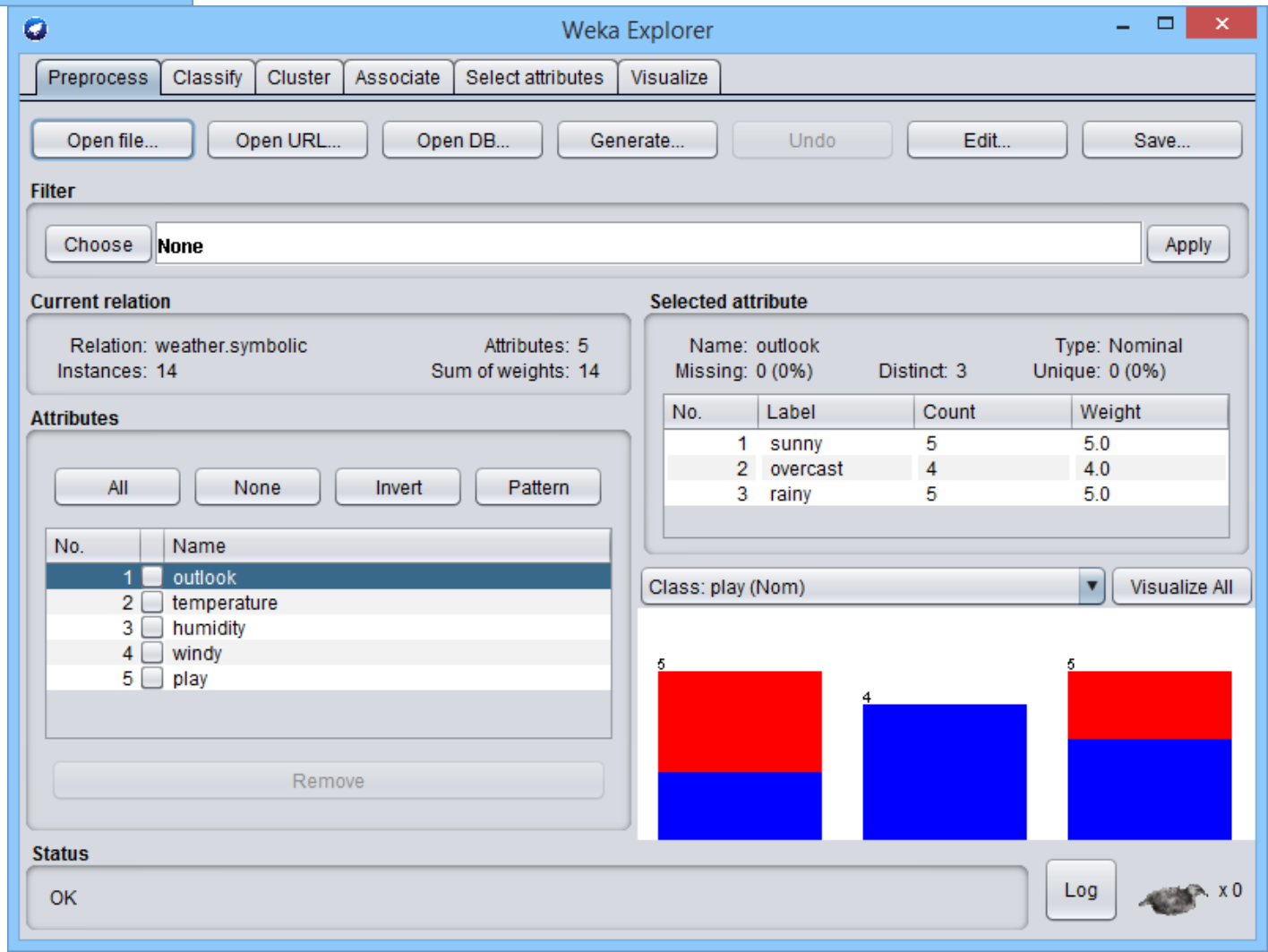
Since Weka is freely available for download and offers many powerful features (sometimes not found in commercial data mining software), it has become one of the most widely used data mining systems. Weka also became one of the favorite vehicles for data mining research and helped to advance it by making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

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

Now …


http://www.cs.waikat...
Data Mining: Practical Mac...
File Edit View Favorites Tools Help
WEKA
The University of Waikato
Machine Learning Group at the University of Waikato
Project Software Book Publications People Related
Data Mining:
Practical Machine Learning Tools and Techniques
Machine learning provides an exciting set of technologies that includes practial tools for analysing data and making predictions but also powers the latest advances in artificial intelligence. We have written a book that provides a highly accessible introduction to the area but also caters for readers who want to delve into the more mathematical techniques available in modern probabilistic modeling and deep learning approaches. Chris Pal has joined Ian Witten, Eibe Frank, and Mark Hall for the fourth edition, and his expertise in probabilistic models and deep learning has greatly extended the book's coverage. To make room for the new material, we now provide an online apprendix on the Weka software. It is an extended version of a brief description of Weka included as an appendix in the book. The book continues to provide references to Weka implementations of algorithms that it describes. The Weka MOOCs provide activities similar to the tutorial exercises in the 3rd edition. We now also provide information on other software: the computational ecosystem for machine learning has grown enormously since we have written the third edition in 2011. A table of contents for the fourth edition, indicating where we have added new material, can be found further down this page.
FOURTH EDITION
DATA MINING
Practical Machine Learning Tools and Techniques
Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
Ian H. Witten & Eibe Frank
DATA MINING
Practical Machine Learning Tools and Techniques
SECOND EDITION
IAN H. WITTEN
EIBE FRANK
Data Mining
PRACTICAL MACHINE LEARNING TOOLS and TECHNIQUES with JAVA IMPLEMENTATIONS
50%


Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No. Label Count Weight
1 sunny 5 5.0
2 overcast 4 4.0
3 rainy 5 5.0
Attributes
All
None
Invert
Pattern
No. Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Class: play (Nom)
Visualize All
5
4
5
Status
OK
Log
x 0

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)
- 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 – CSV format (LoanData.csv)

Microsoft Excel - LoanData

File Edit View Insert Format Tools Data Window Help

Type a question for help

N23

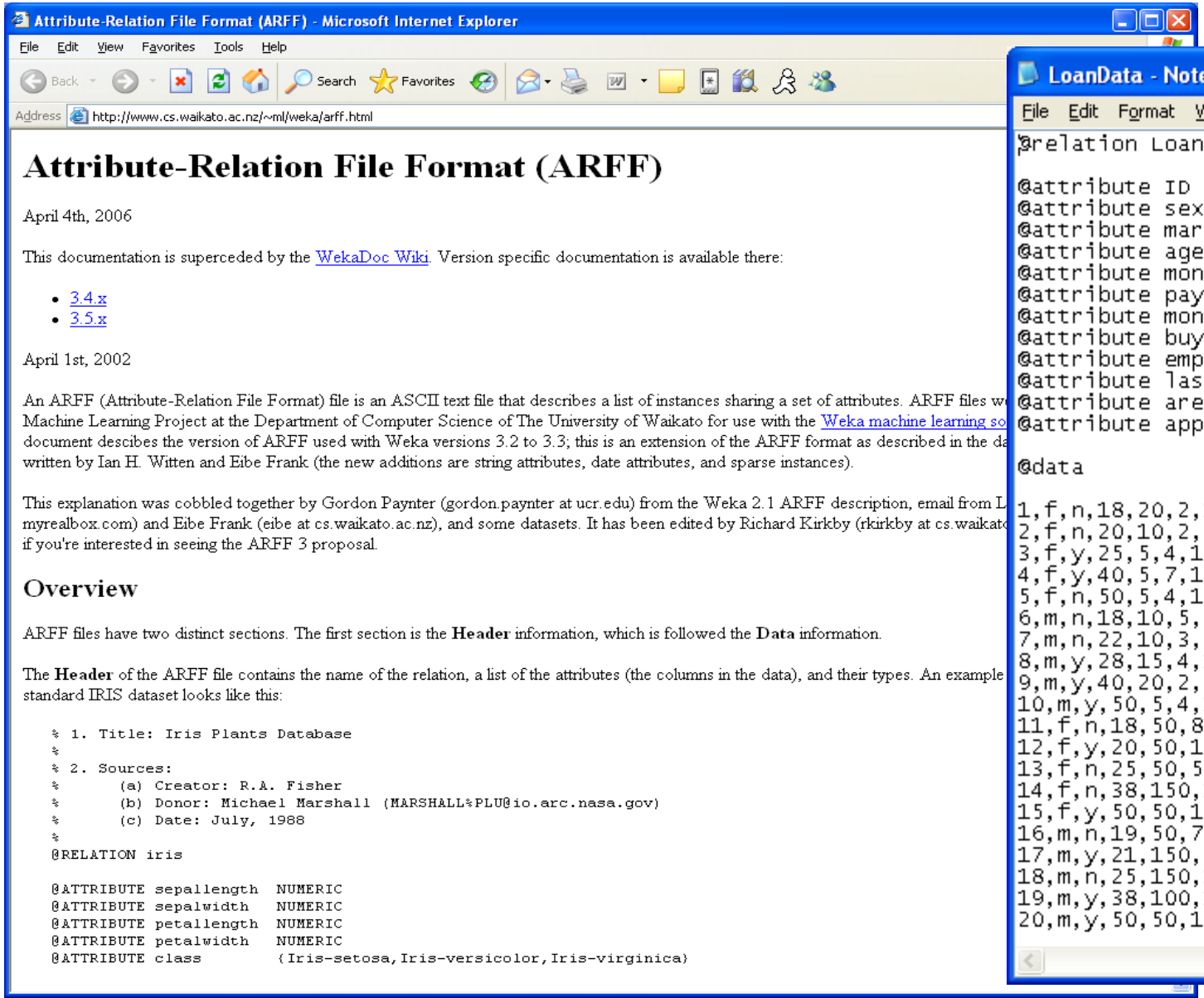
	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	sex	married	age	money	pay	months	buy	emp	lastemp	area	approved
2	1	f	n	18	20	2	15	pc	y	1	good	y
3	2	f	n	20	10	2	20	pc	y	2	good	y
4	3	f	y	25	5	4	12	pc	n	0	bad	n
5	4	f	y	40	5	7	12	pc	y	2	good	y
6	5	f	n	50	5	4	12	pc	y	25	bad	y
7	6	m	n	18	10	5	8	pc	y	1	good	y
8	7	m	n	22	10	3	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												

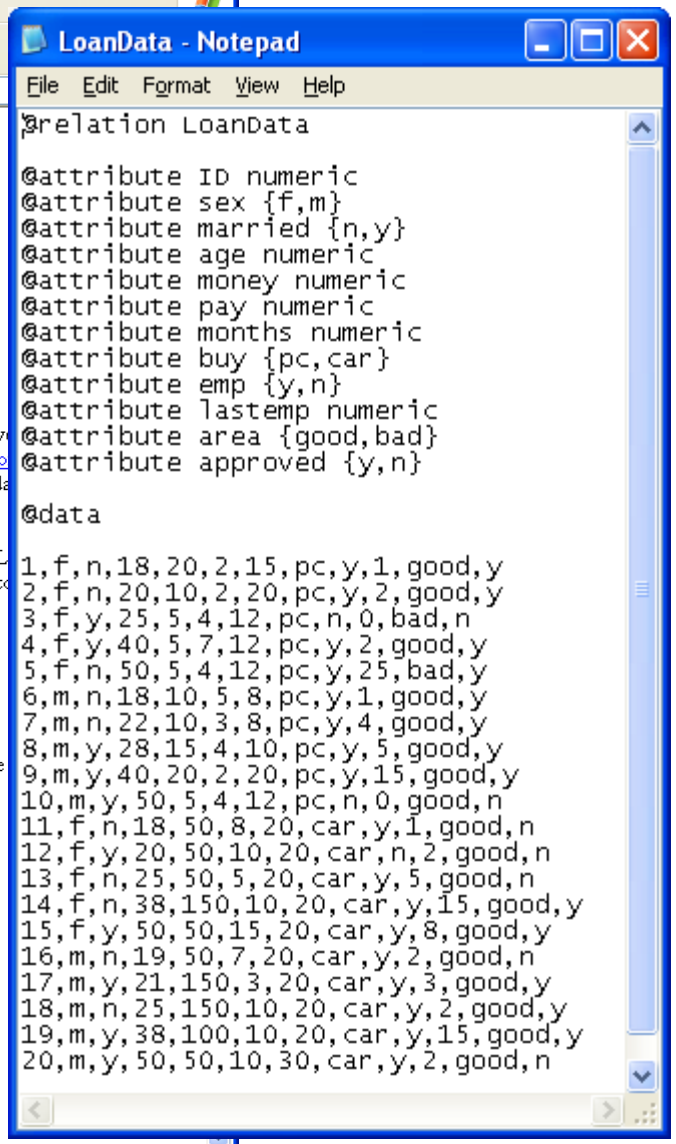
LoanData

Ready

Data preprocessing and visualization

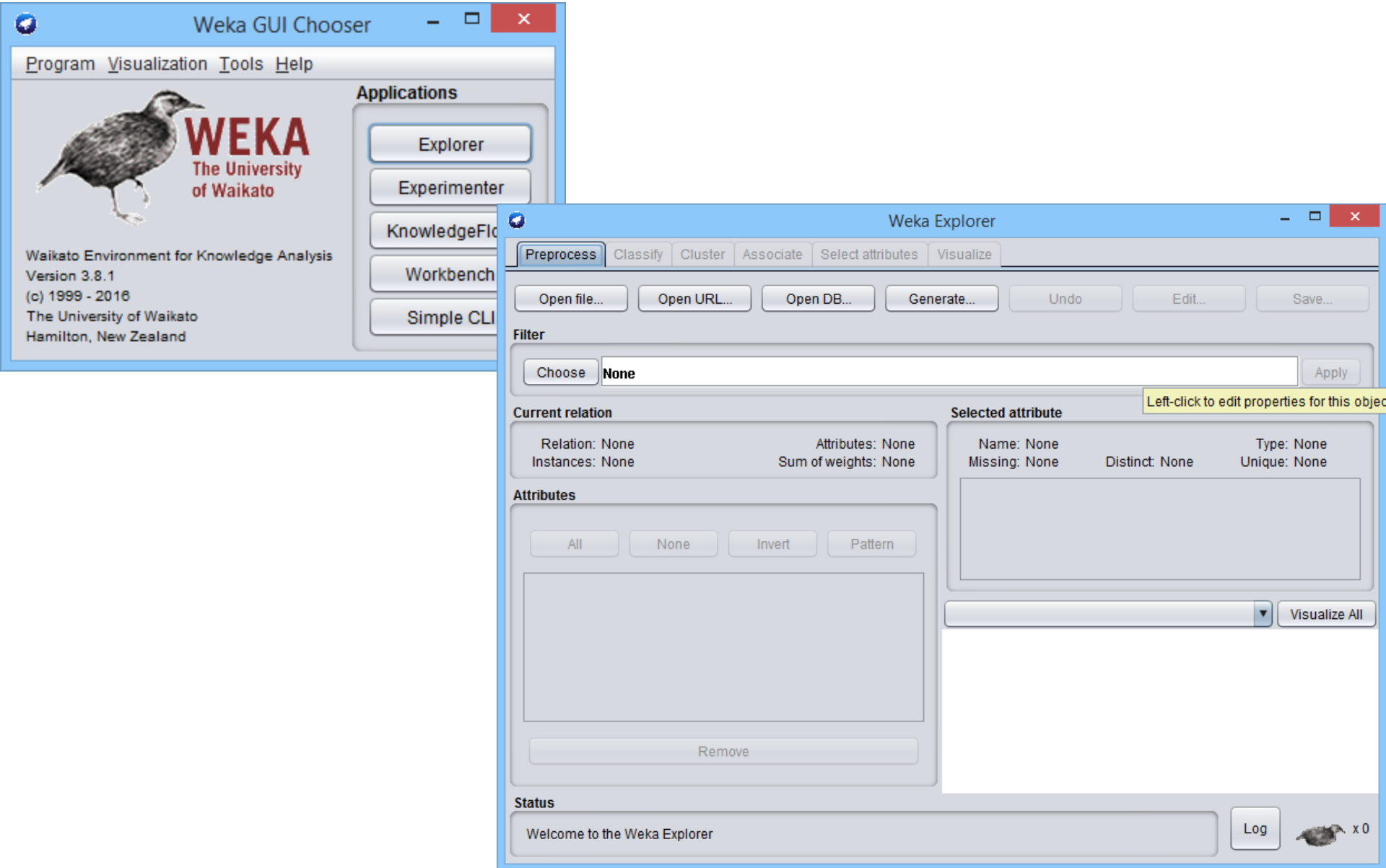
Attribute-Relation File Format (ARFF) - http://www.cs.waikato.ac.nz/~ml/weka/arff.html





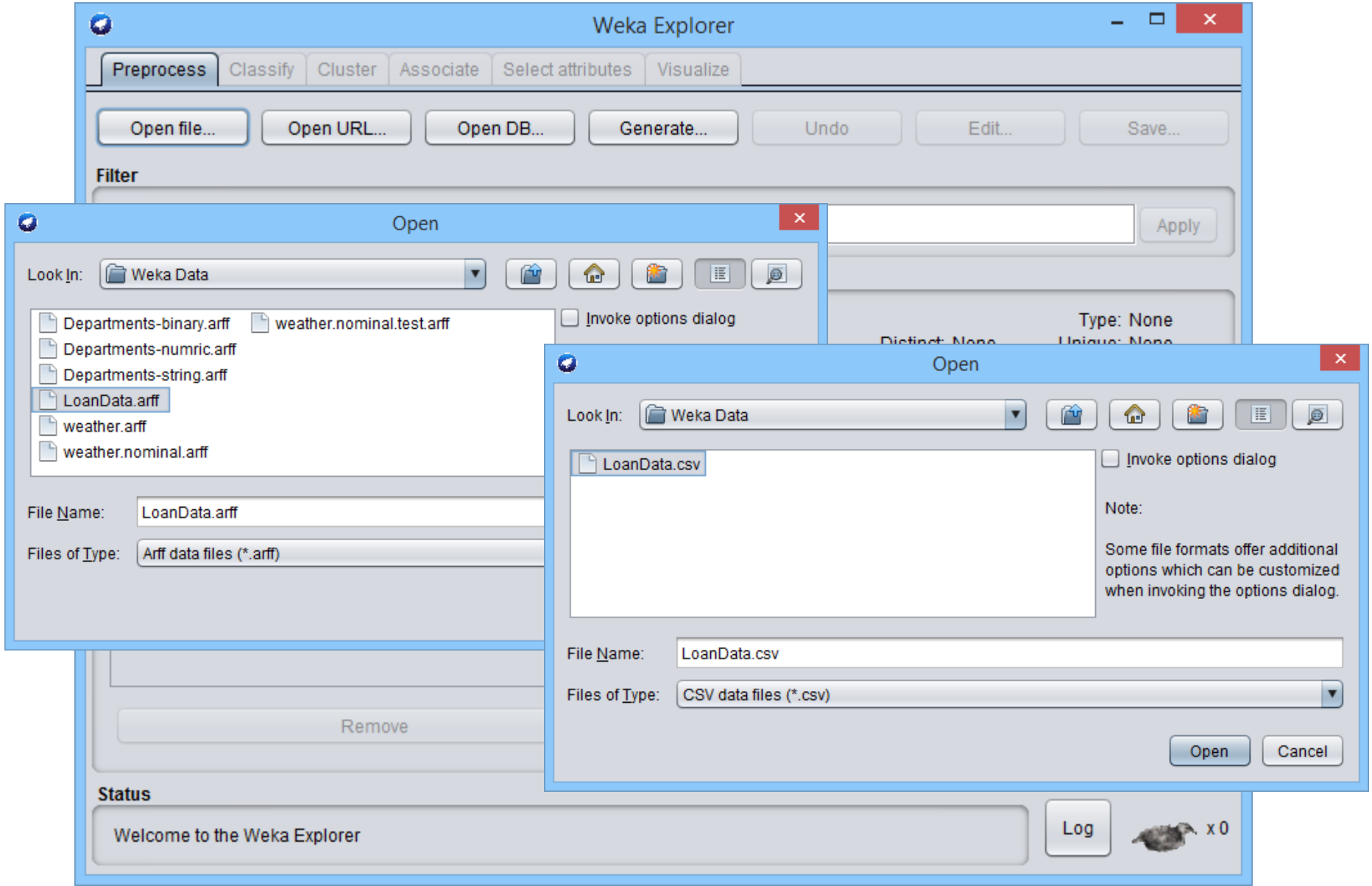
Data preprocessing and visualization

Run Weka and select the Explorer



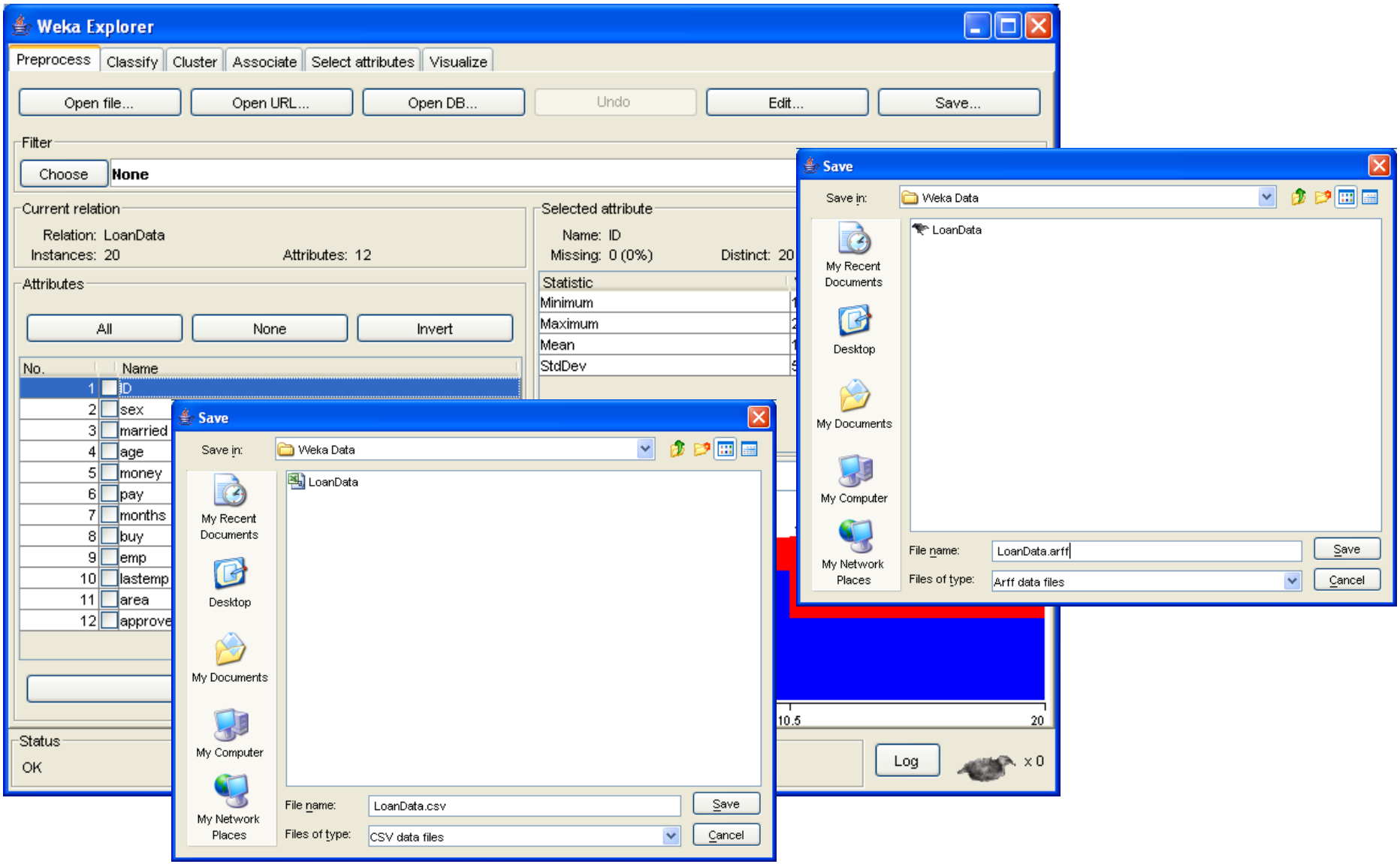
Data preprocessing and visualization

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



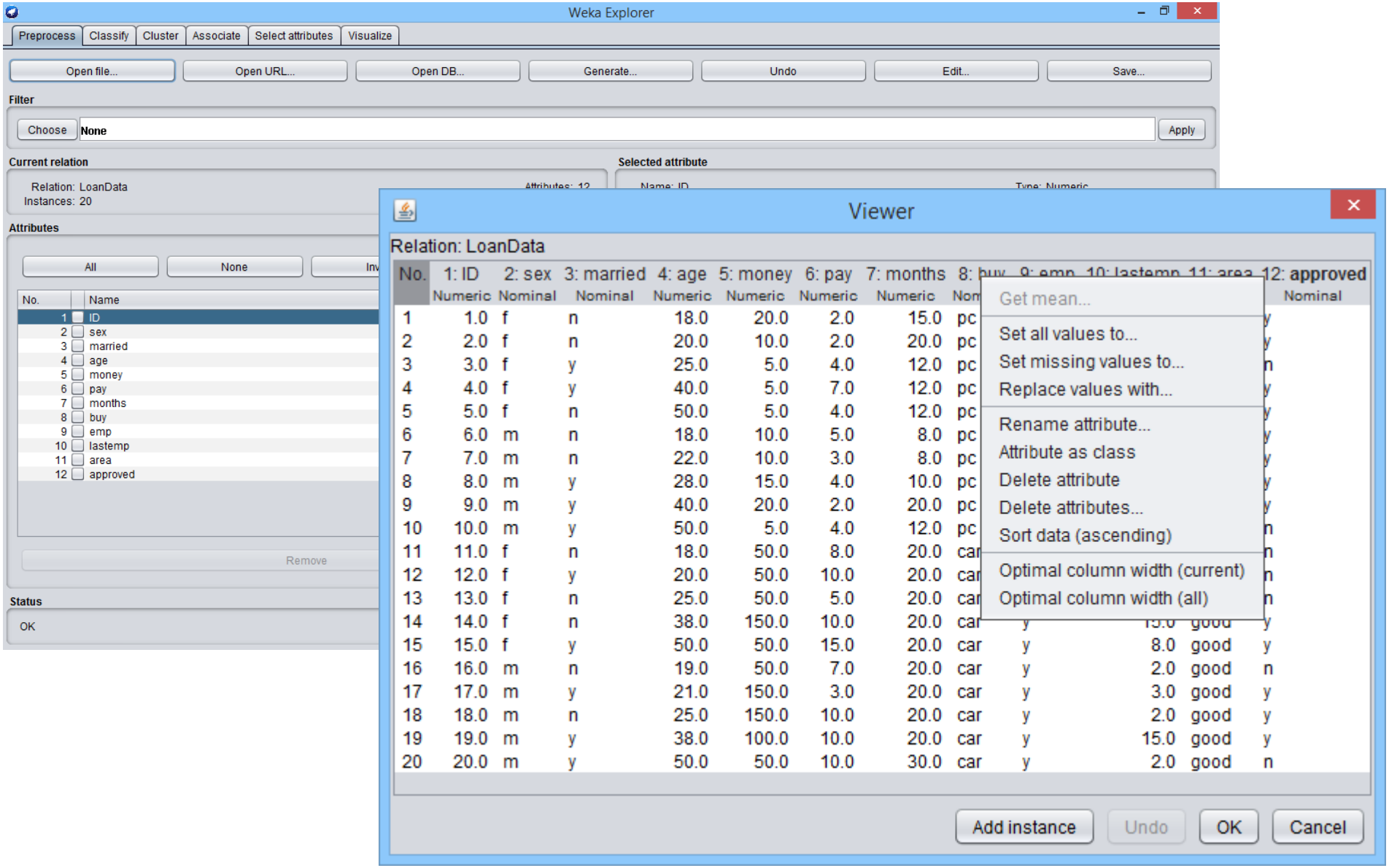
Data preprocessing and visualization

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



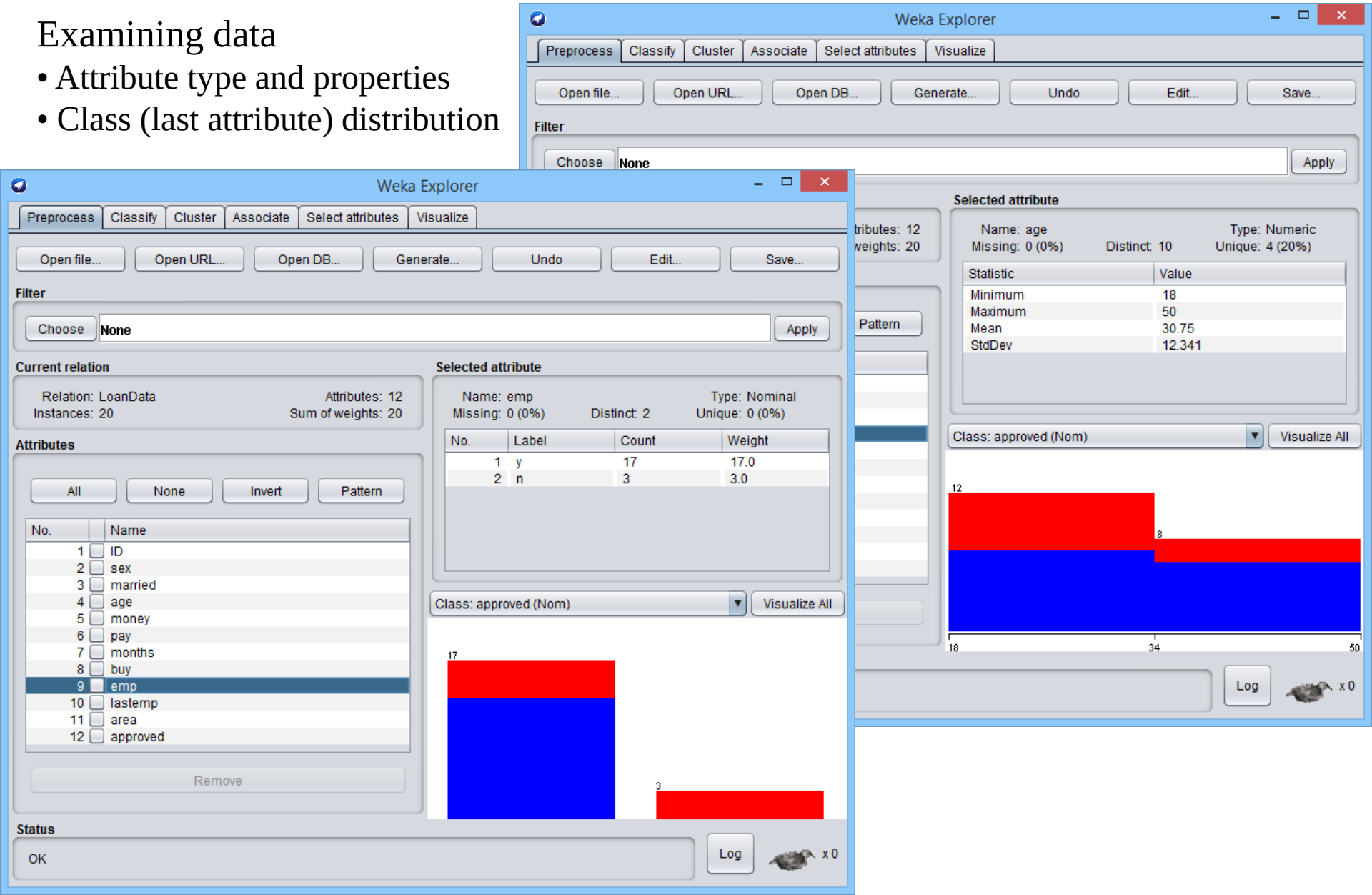
Data preprocessing and visualization

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



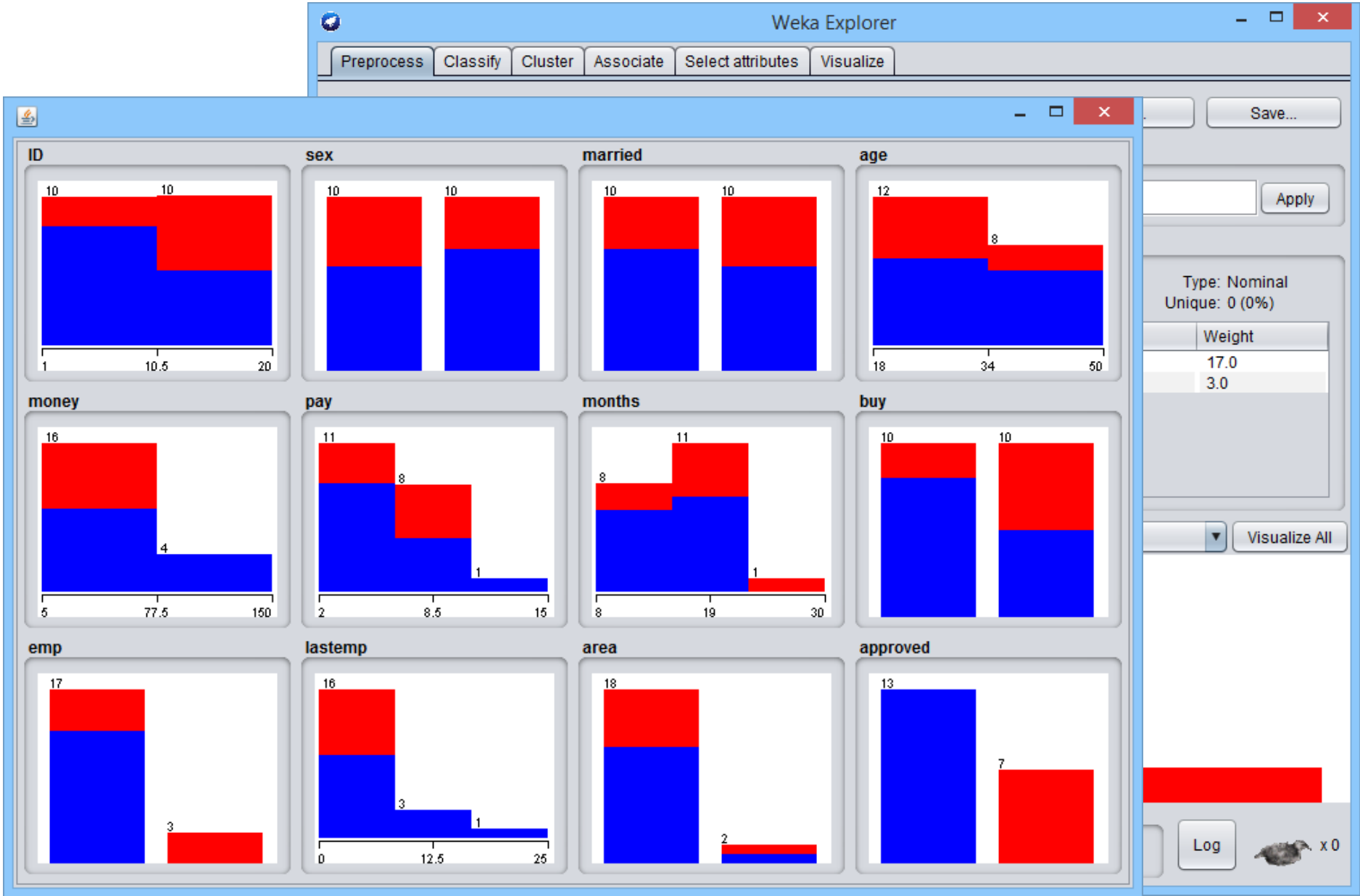
Data preprocessing and visualization

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



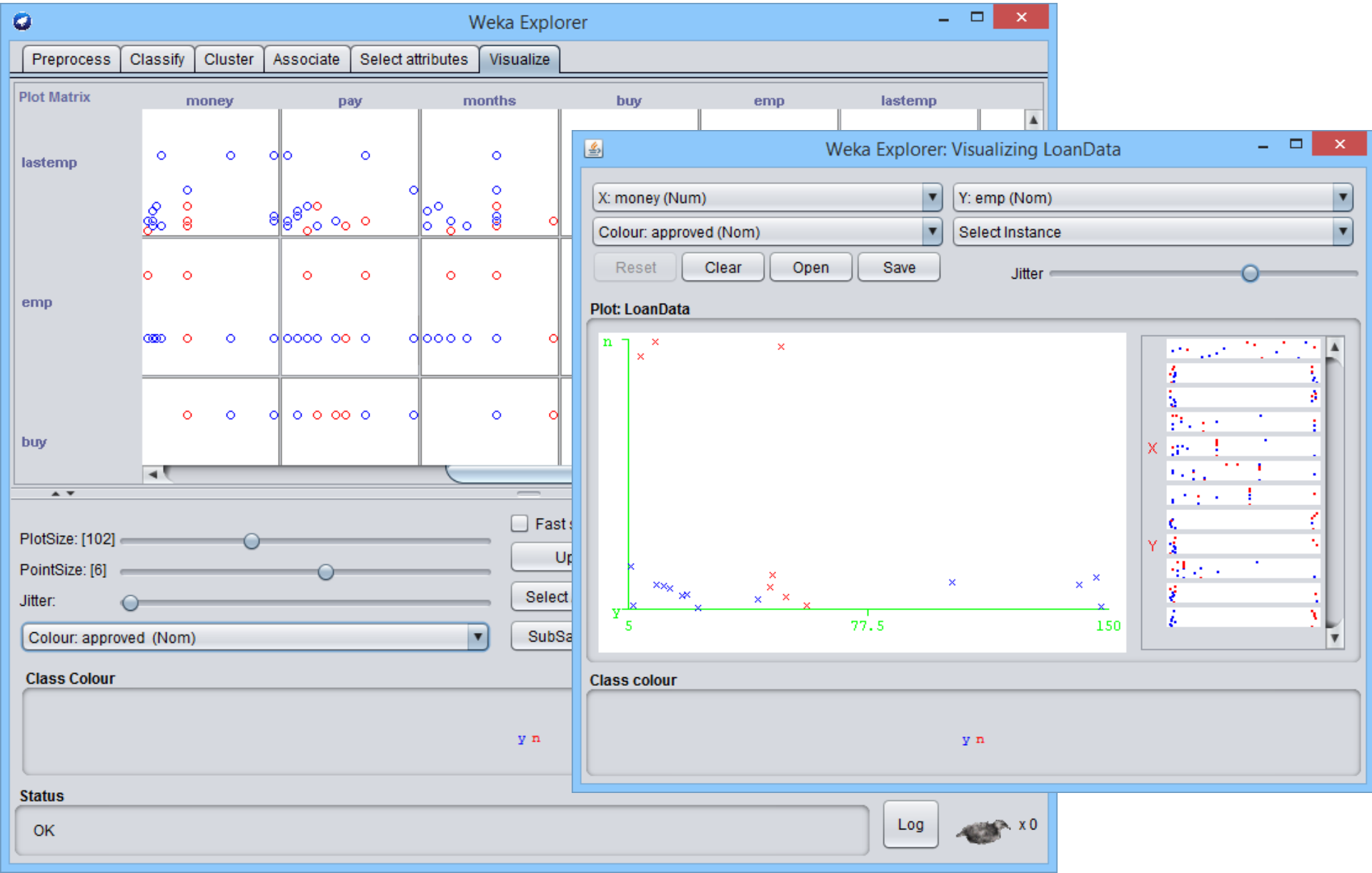
Data preprocessing and visualization

Click on “Visualize All”



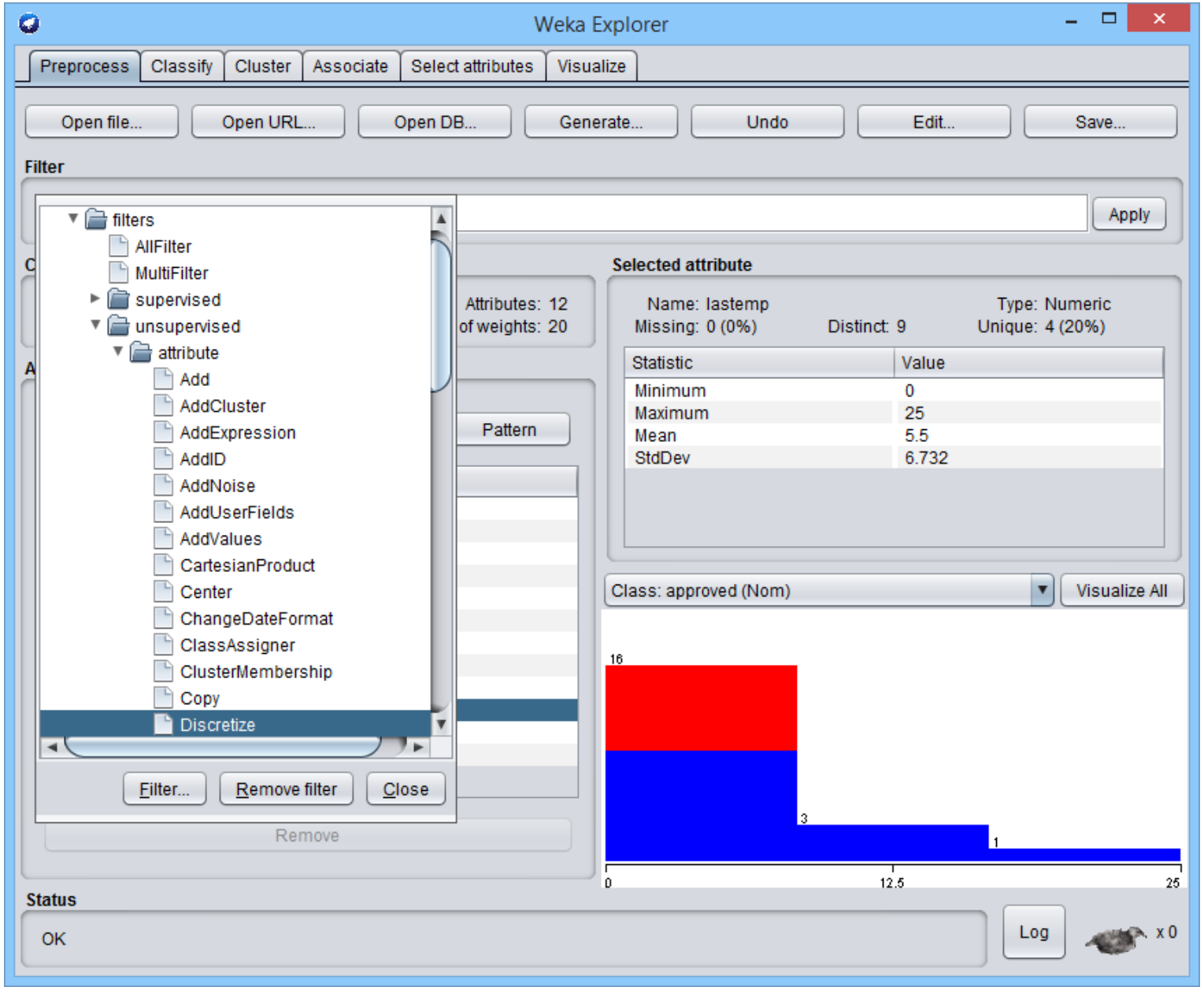
Data preprocessing and visualization

Click on “Visualize” tab, double-click on a plot to see the 2D projection of the instance space



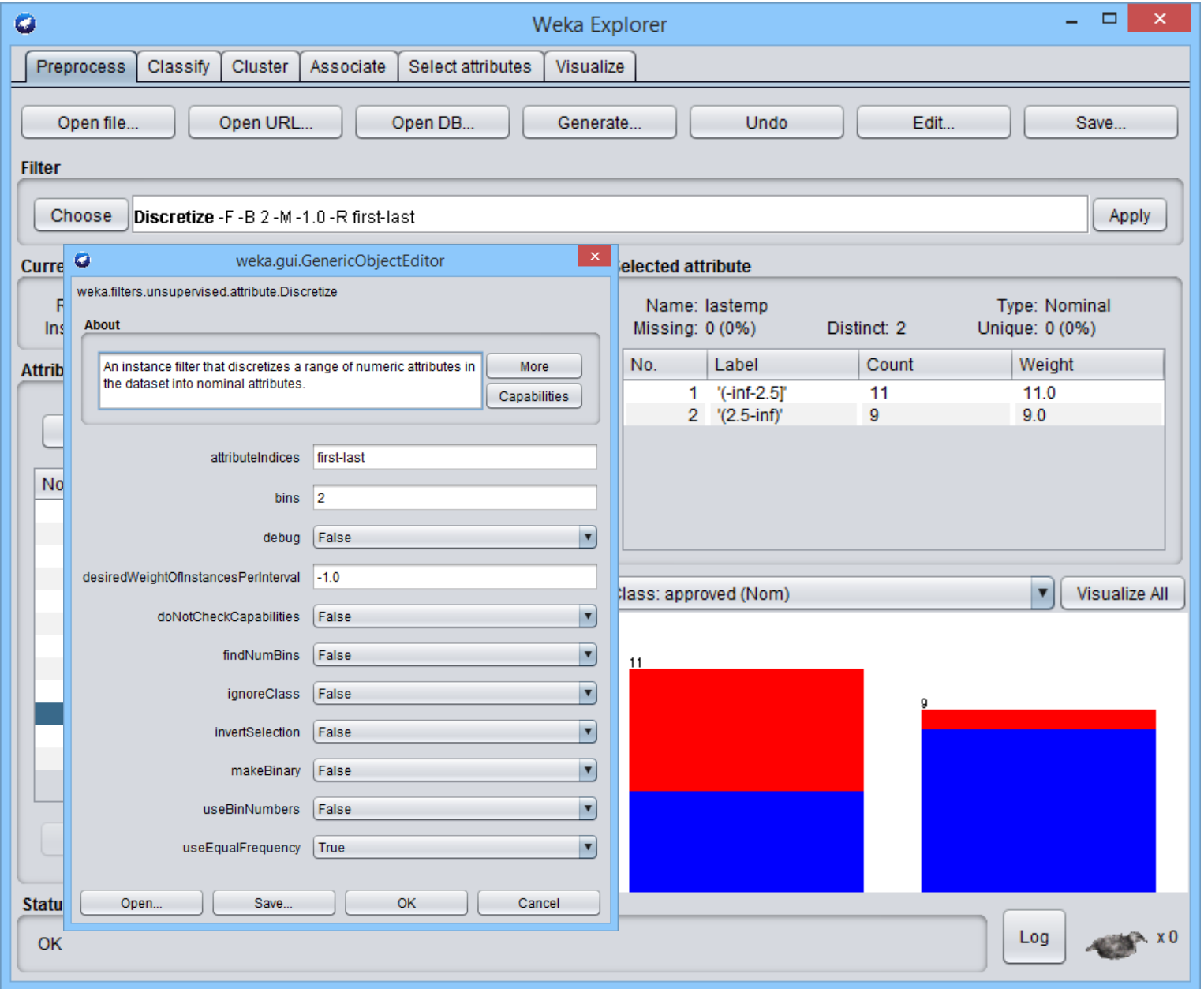
Data preprocessing and visualization

Using filters: click on “Choose” in the “Filter” window, select “Discretize”



Data preprocessing and visualization

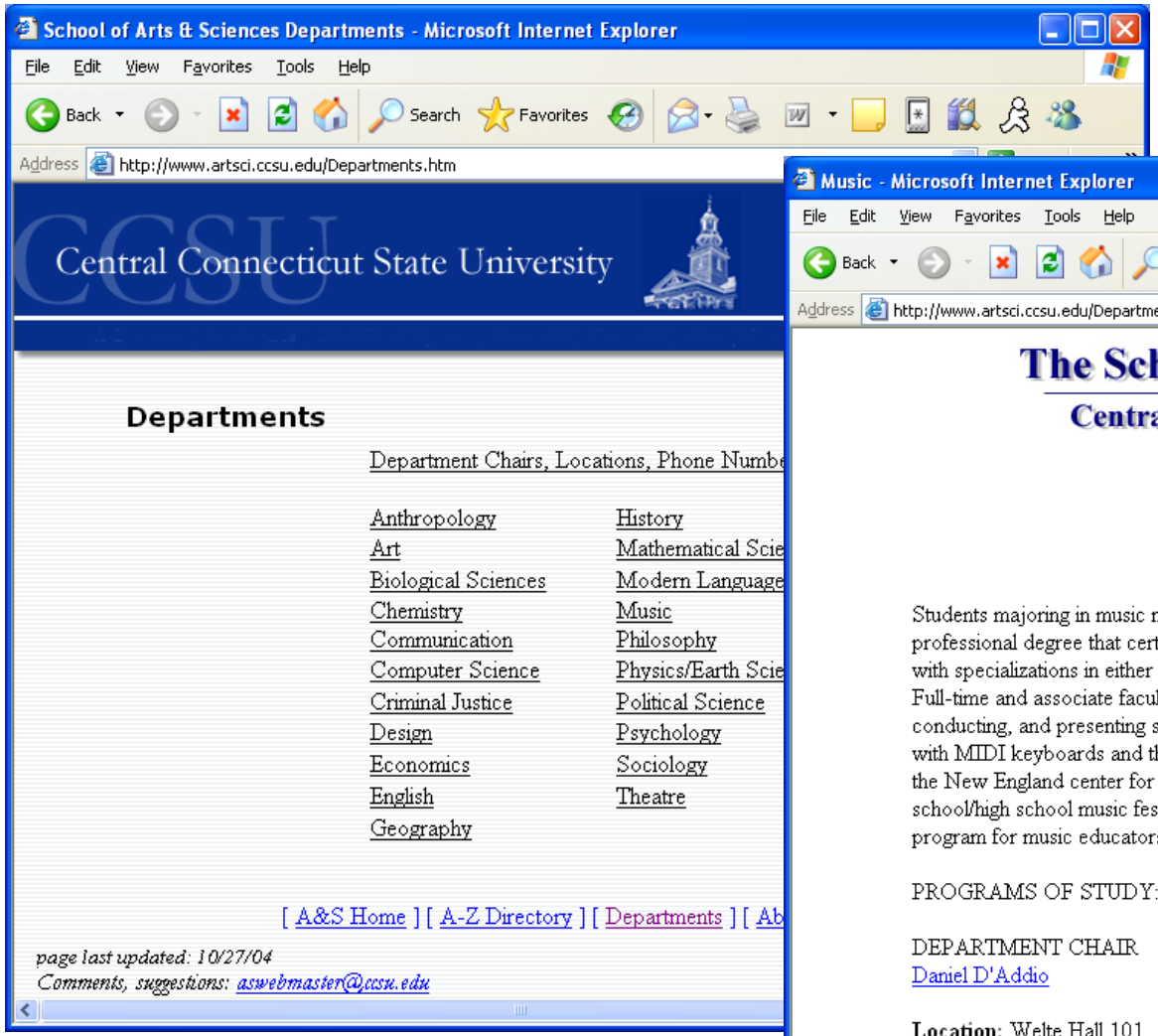
Click in the “Discretize” in the Filter window and choose parameters, then click on “Apply”

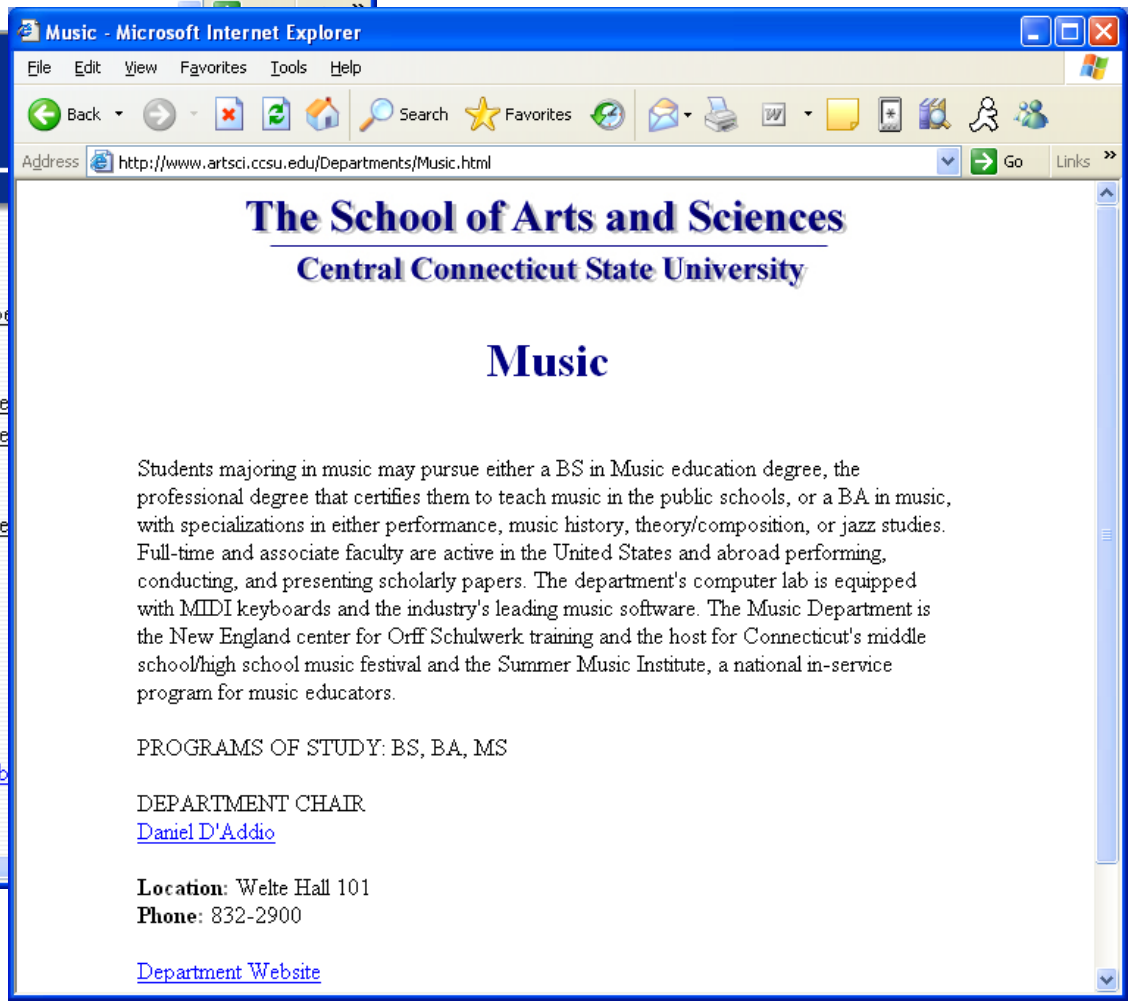


Note how the plot of “lastemp” changed.

Data preprocessing and visualization

Web/Text documents - Department data



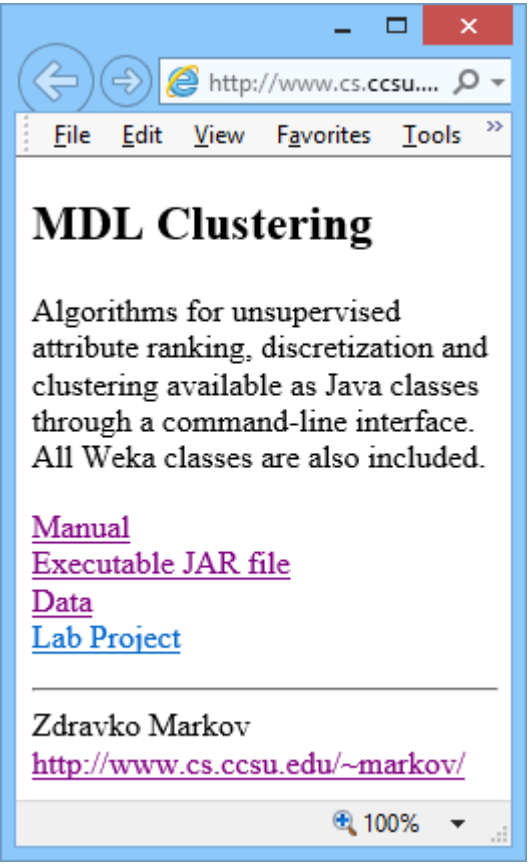


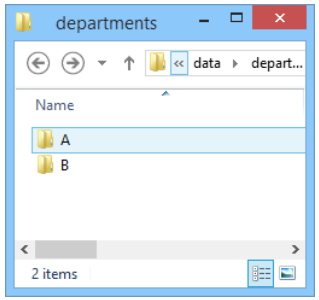
Data preprocessing and visualization

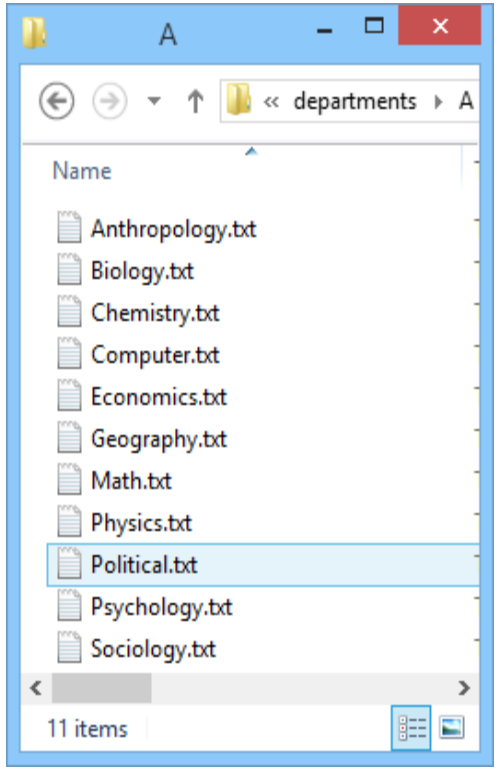
Department data document collection

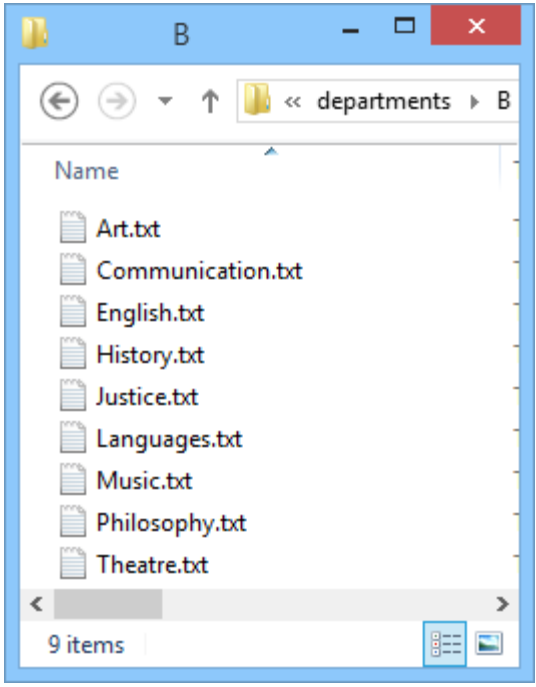
http://www.cs.ccsu.edu/~markov/MDLclustering

http://www.cs.ccsu.edu/~markov/MDLclustering/data.zip



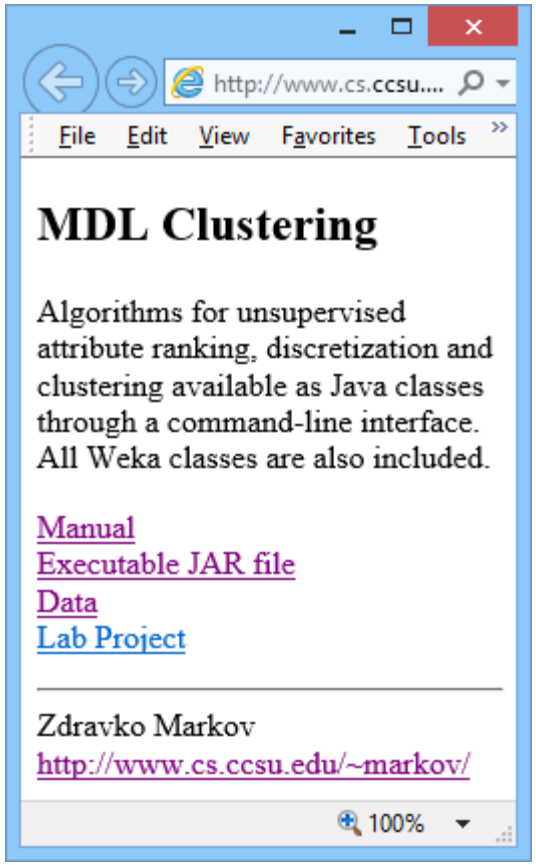




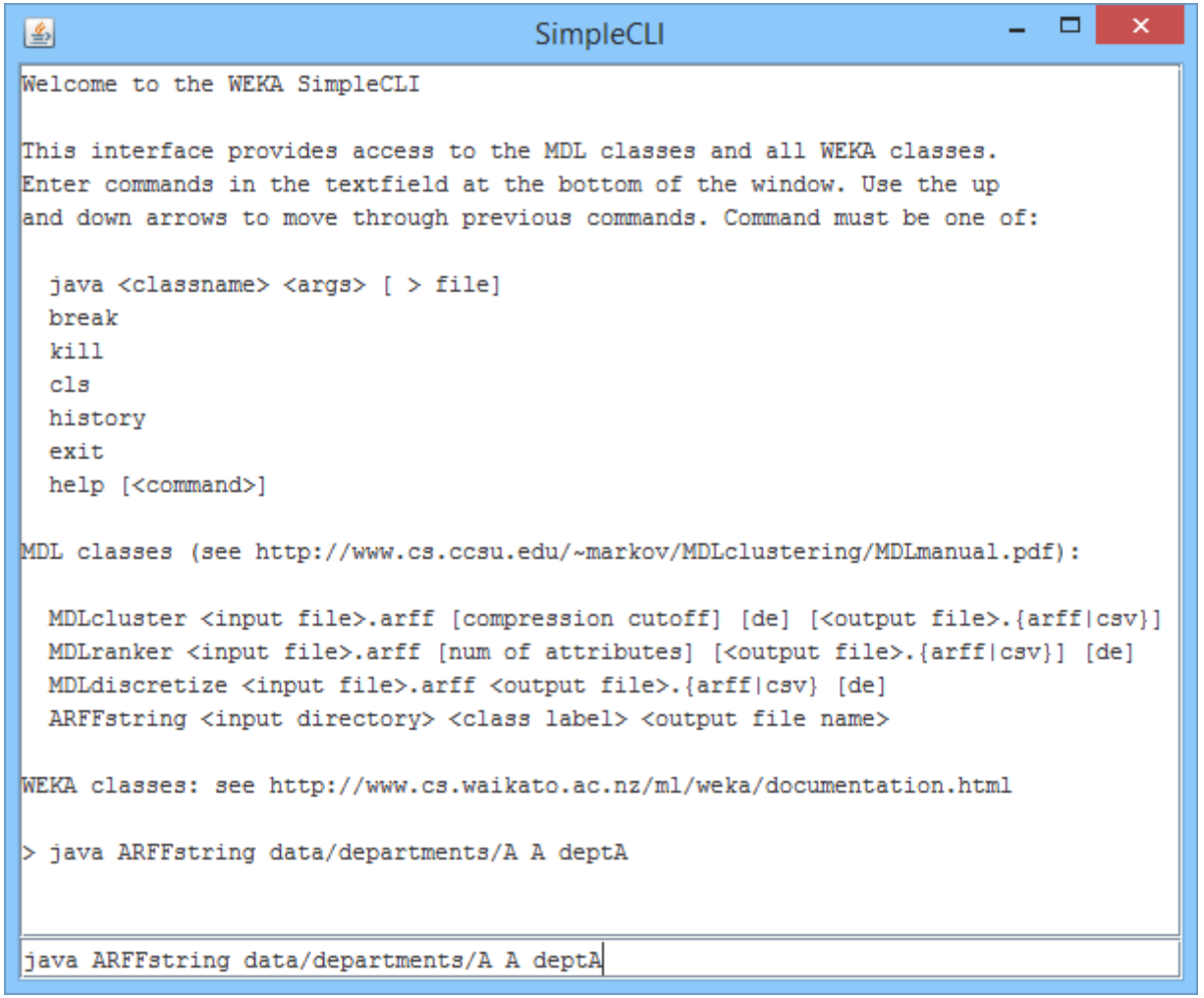


Data preprocessing and visualization

Department data: Create ARFF file



http://www.cs.ccsu.edu/~markov/MDLclustering/MDL.jar



Data preprocessing and visualization

Department data: Create ARFF file in string format (using SimpleCLI)

1. Create file deptA with the files in folder data/departments/A with class label A:

```
java ARFFstring data/departments/A A deptA
```

2. Create file deptB with the files in folder data/departments/B with class label B:

```
java ARFFstring data/departments/B B deptB
```

3. Merge deptA and deptB into one file departments-string.arff

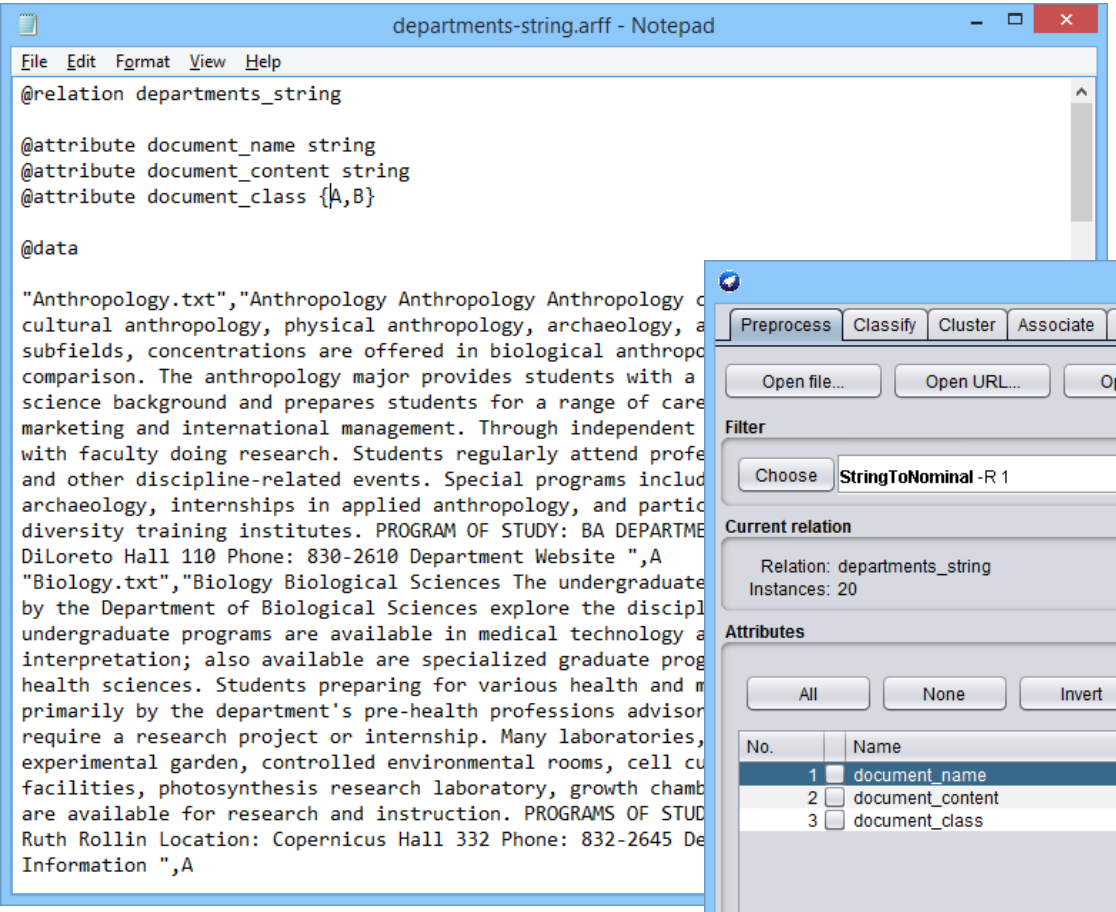
4. Add the following ARFF file header in the beginning of departments-string.arff:

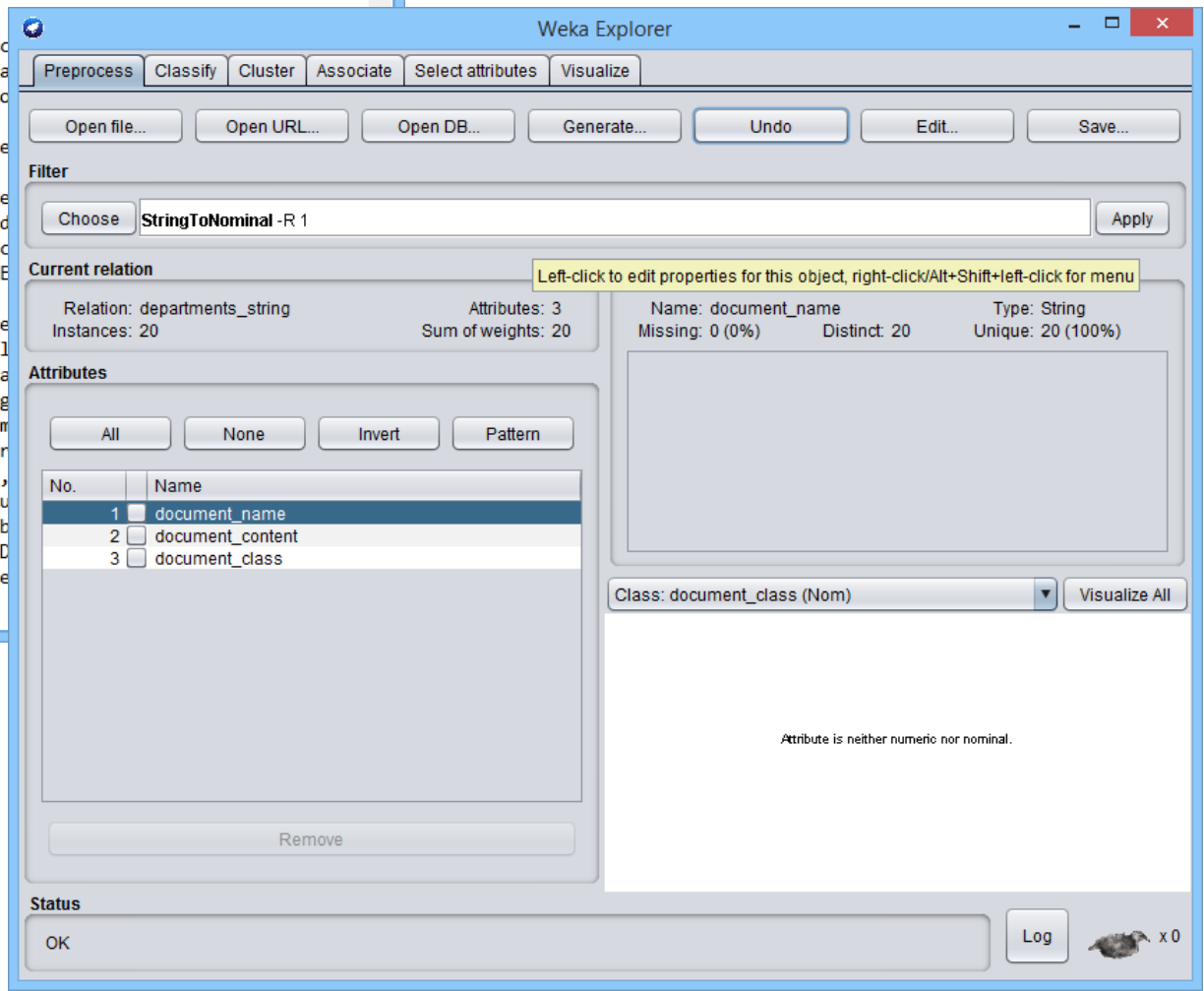
```
@relation departments_string
@attribute document_name string
@attribute document_content string
@attribute document_class {A,B}
@data
```

Data preprocessing and visualization

Loading text data in Weka

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

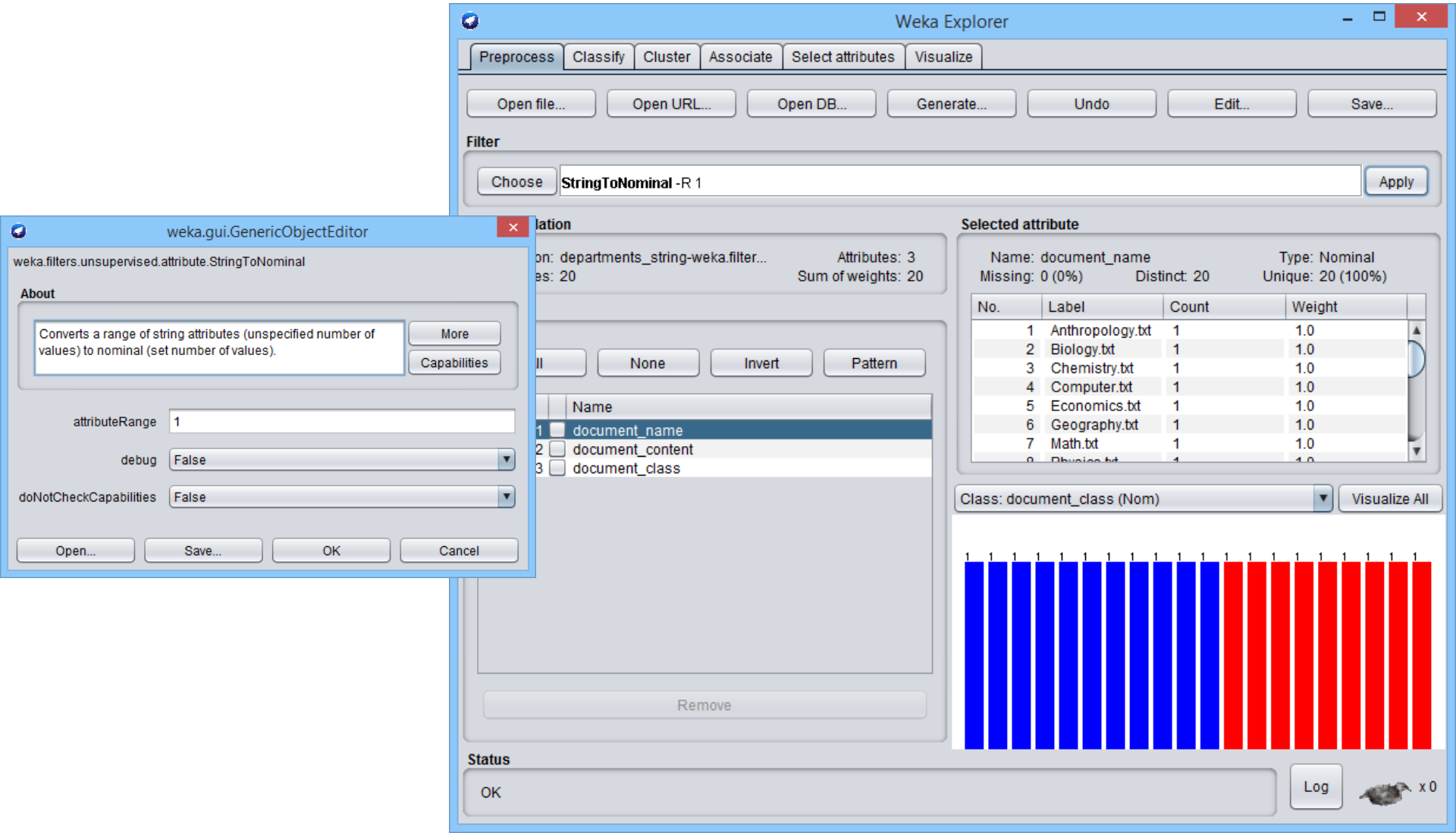




Data preprocessing and visualization

Converting a string attribute into nominal

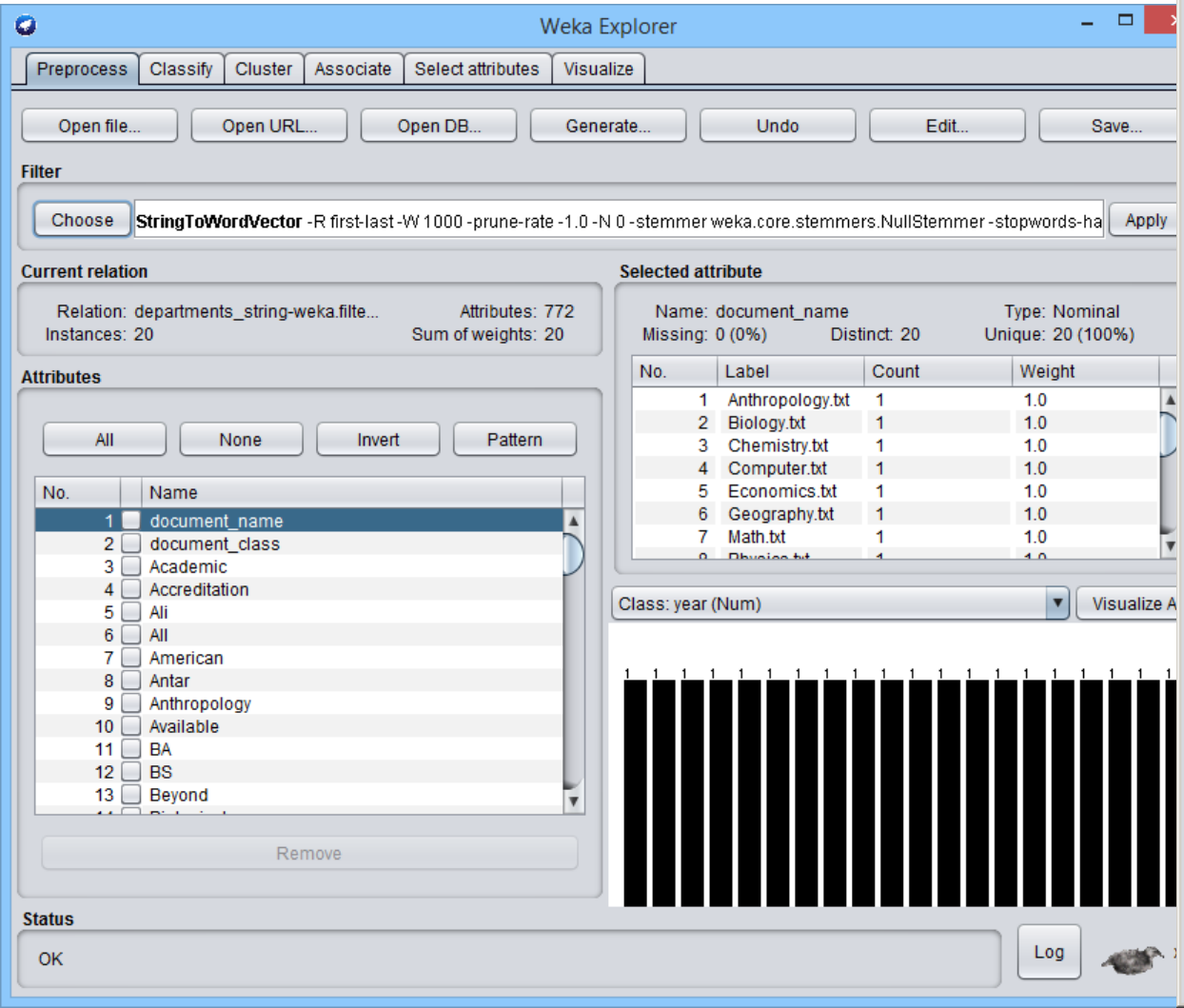
Choose filters/unsupervised/attribute/StringToNominal, set attributeRange to 1, click on Apply

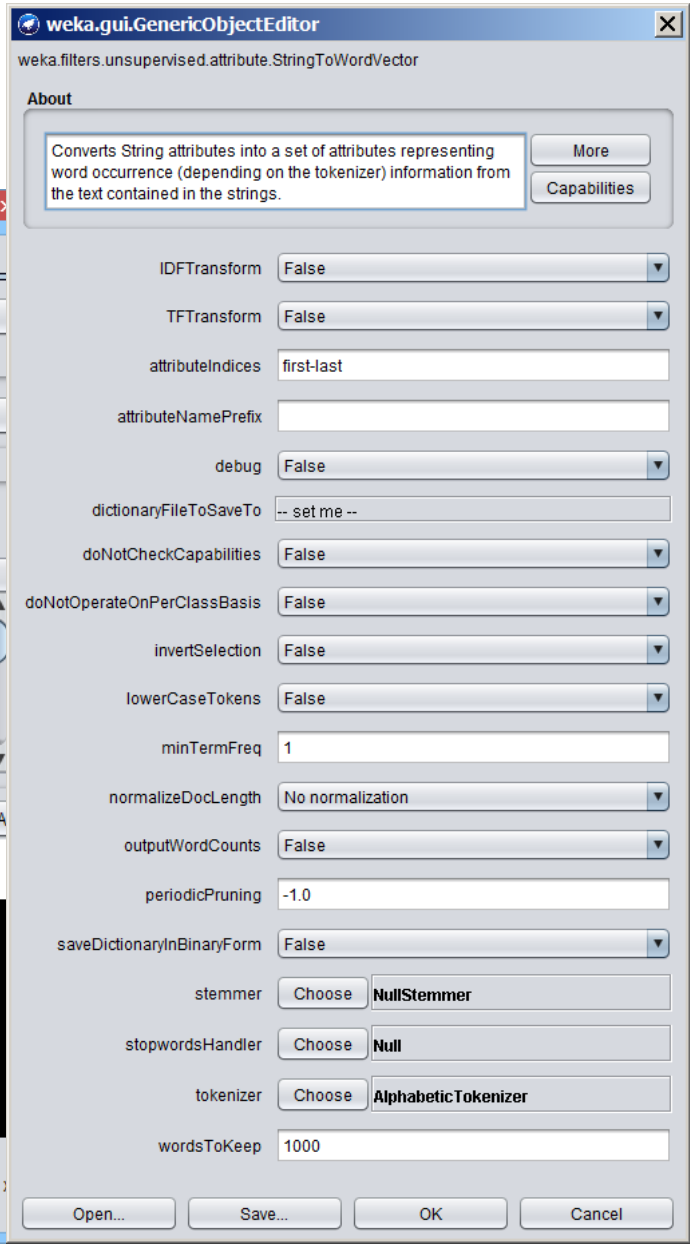


Data preprocessing and visualization

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”

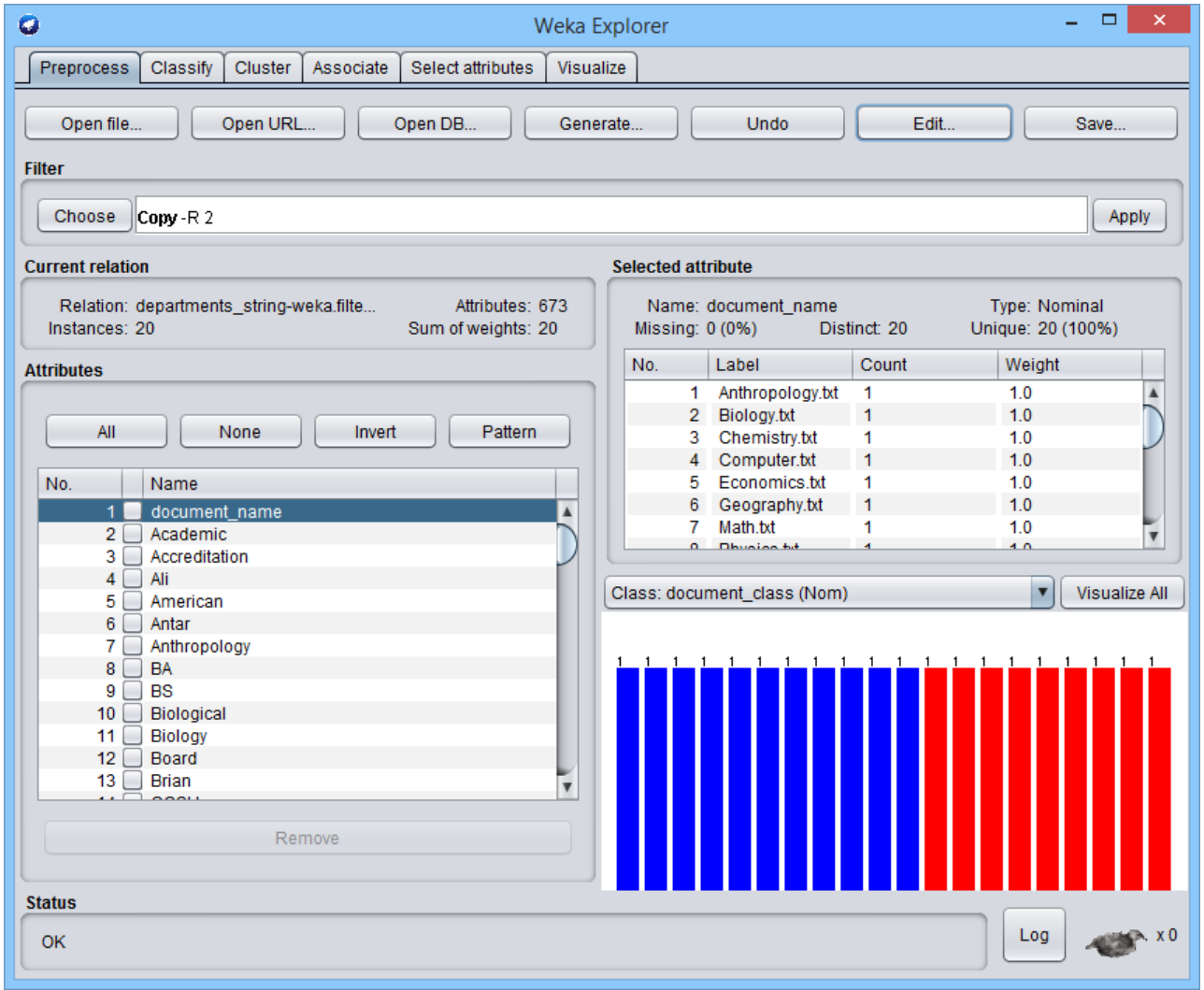




Data preprocessing and visualization

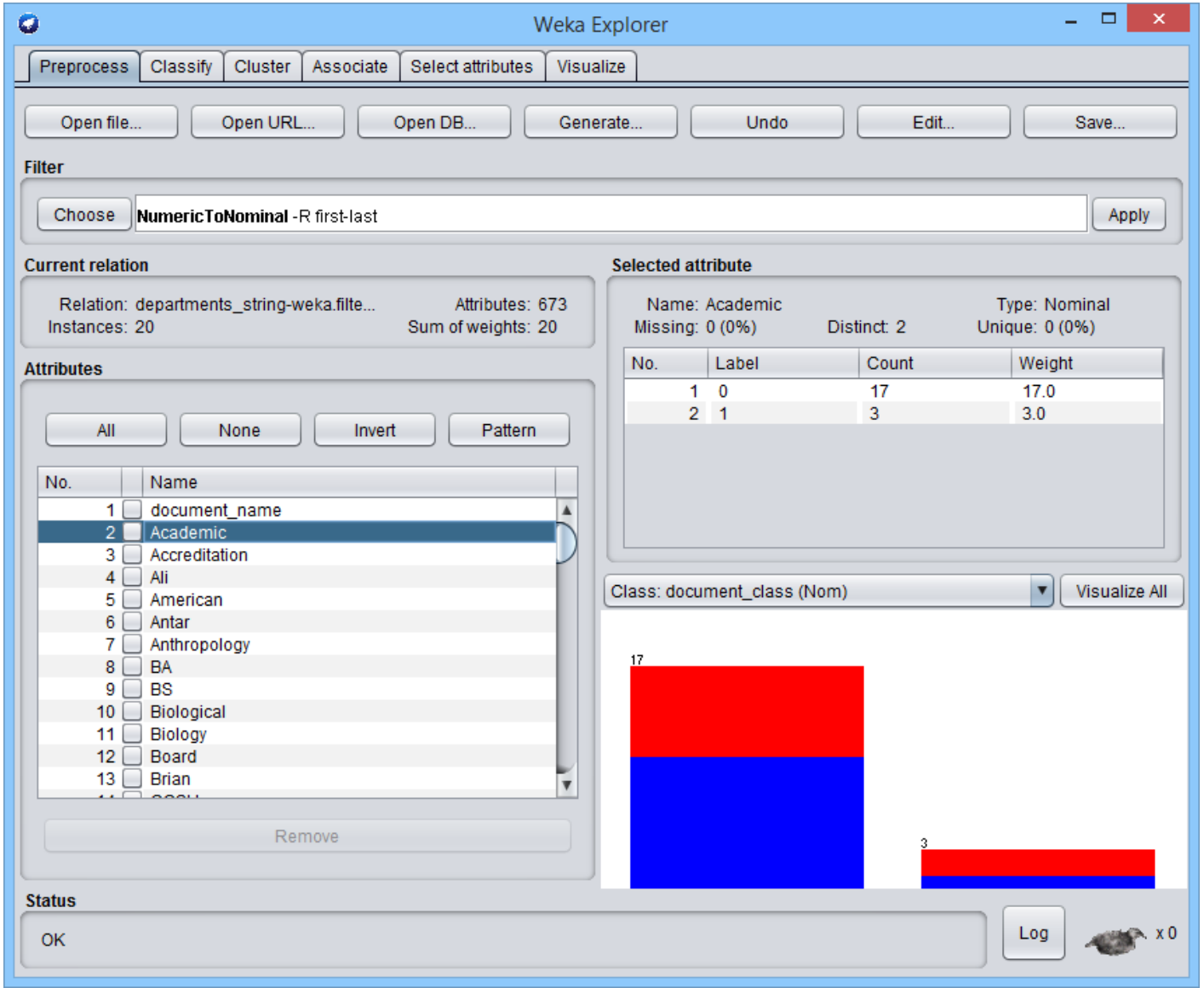
Make document_class last attribute

- 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 NumericToNominal filter)
- Save data on a file for further use



Data preprocessing and visualization

ARFF department data in binary format (NonSparse and Sparse format, see SparseToNonSparse filter)

departments.arff - Notepad

File Edit Format View Help

```
@relation departments_string-weka.filters.unsupervised.attribute.StringToNominal-R1-
weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-
stemmerweka.core.stemmers.SnowballStemmer-stopwords-handlerweka.core.stopwords.Rainbow-M1-
tokenizerweka.core.tokenizers.AlphabeticTokenizer-weka.filters.unsupervised.attribute.Copy-R2-
weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

@attribute document_name
{Anthropology.txt,Biology.txt,Chemistry.txt,Computer.txt,Economics.txt,Geography.txt,Math.txt,Physics.txt,Political.t
xt,Psychology.txt,Sociology.txt,Art.txt,Communication.txt,English.txt,History.txt,Justice.txt,Languages.txt,Music.txt
,Philosophy.txt,Theatre.txt}
```

depa

File Edit Format View Help

```
@attribute writing {0,1}
@attribute year {0,1}
@attribute document_class {A,B}

@data
{6 1,7 1,31 1,44 1,60 1,64 1,67 1,87 1,89 1,121 1,12
1,171 1,176 1,177 1,183 1,187 1,189 1,201 1,210 1,21
1,302 1,312 1,319 1,329 1,335 1,339 1,347 1,352 1,35
1,395 1,405 1,416 1}
{0 Biology.txt,8 1,9 1,10 1,26 1,44 1,46 1,54 1,55 1
1,166 1,172 1,178 1,182 1,183 1,187 1,199 1,204 1,20
1,264 1,265 1,289 1,298 1,302 1,318 1,328 1,330 1,33
1,383 1,399 1,409 1}
{0 Chemistry.txt,4 1,8 1,18 1,19 1,21 1,25 1,26 1,43
1,145 1,152 1,153 1,159 1,161 1,166 1,181 1,195 1,19
1,393 1}
{0 Computer.txt,2 1,8 1,11 1,15 1,22 1,44 1,48 1,58
1,128 1,130 1,154 1,166 1,182 1,185 1,188 1,196 1,20
1,279 1,295 1,296 1,310 1,334 1,337 1,338 1,363 1,36
{0 Economics.txt,1 1,7 1,17 1,34 1,50 1,56 1,64 1,66
1,161 1,164 1,173 1,175 1,180 1,181 1,190 1,191 1,19
1,252 1,256 1,258 1,275 1,278 1,280 1,282 1,293 1,30
1,416 1}
{0 Geography.txt,7 1,12 1,16 1,24 1,31 1,40 1,55 1,6
1,242 1,248 1,266 1,269 1,273 1,276 1,281 1,282 1,28
1,367 1,391 1,392 1,397 1,403 1,407 1,408 1,410 1,41
{0 Math.txt,7 1,23 1,27 1,30 1,44 1,54 1,55 1,57 1,5
1,182 1,192 1,196 1,203 1,245 1,258 1,275 1,280 1,28
1,387 1,390 1,391 1,396 1}
{0 Physics.txt,3 1,5 1,8 1,13 1,17 1,26 1,33 1,38 1,
1,131 1,132 1,133 1,150 1,158 1,174 1,179 1,194 1,19
```

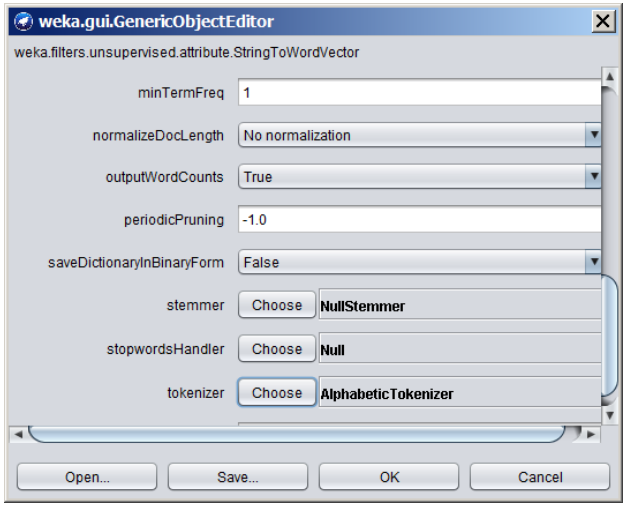
departments.arff - Notepad

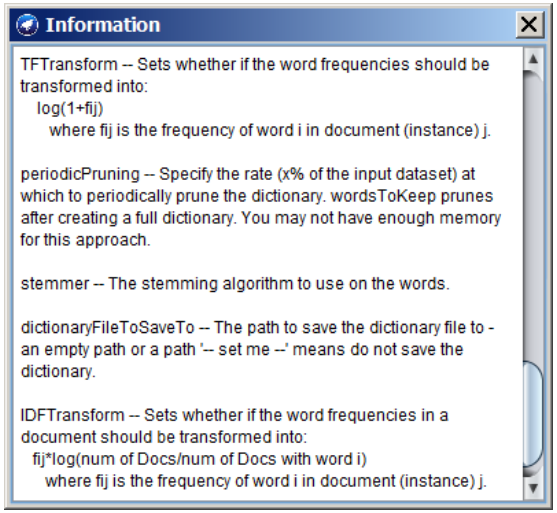
File Edit Format View Help

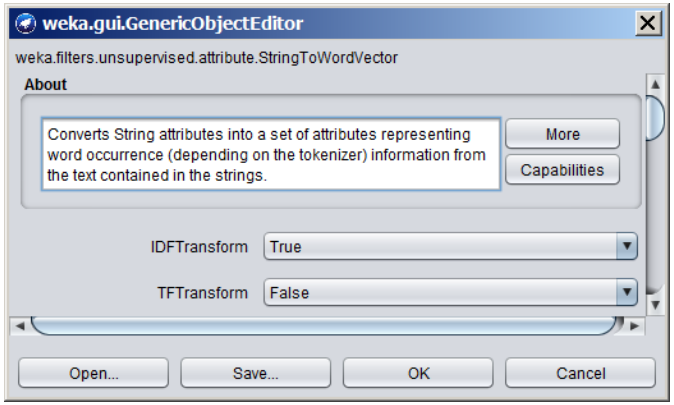
```
@attribute works {0,1}
@attribute writing {0,1}
@attribute year {0,1}
@attribute document_class {A,B}

@data
Anthropology.txt,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Biology.txt,0,0,0,0,0,0,0,0,1,1,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,0,0,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Chemistry.txt,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,1,0,1,0,0,0,1,1,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Computer.txt,0,1,0,0,0,0,0,1,0,0,1,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Economics.txt,1,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Geography.txt,0,0,0,0,0,0,1,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,1,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Math.txt,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Physics.txt,0,0,1,0,1,0,0,1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,0,0,1,0,1,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,1,0,0,0,0,0,0,0,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Political.txt,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,1,1,0,0,0,0,1,0,0,1,0,0,0,0,0,0,1,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Psychology.txt,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,0,0,1,0,0,0,0,1,0,1,0,0,0,0,1,1,0,1,0,0,0,0,1
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Sociology.txt,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
Art.txt,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,
1,0,0,0,0,0,1,1,1,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,0,1,0,0,0,0,0,1,1,0,0,1,0,0,1,0,0,0,1,0,0,1,1,1,1,0,
```

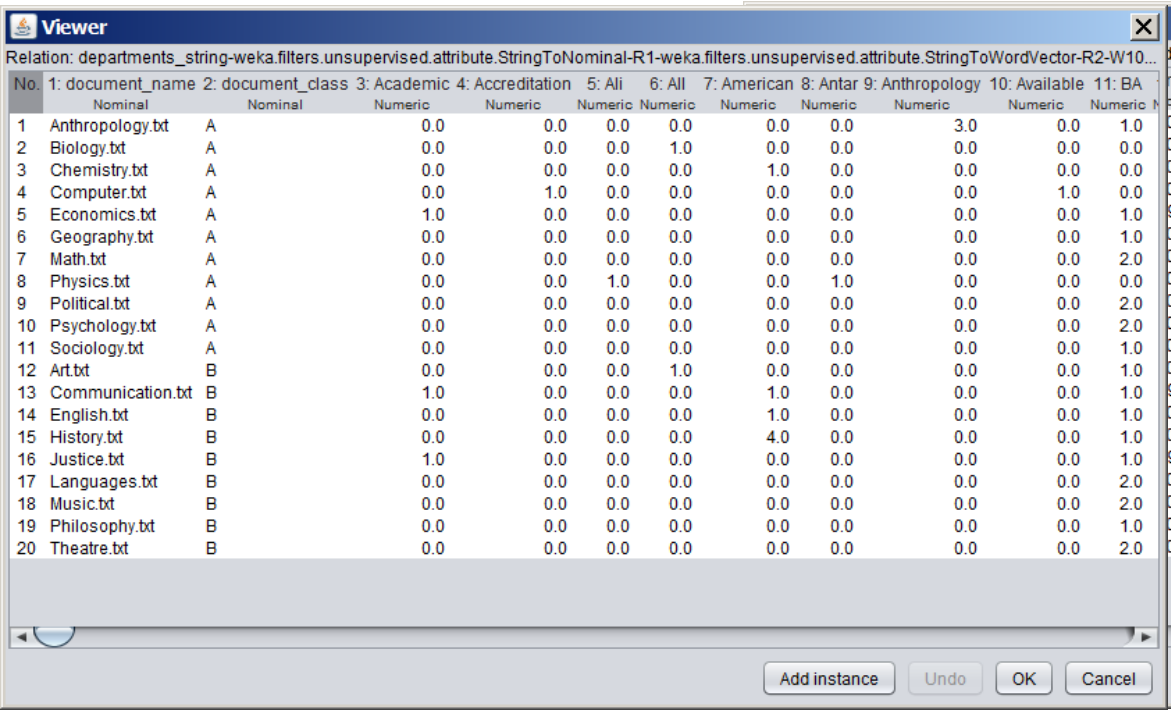
Data preprocessing and visualization

ARFF department data in TF and TFIDF format









Data preprocessing and visualization

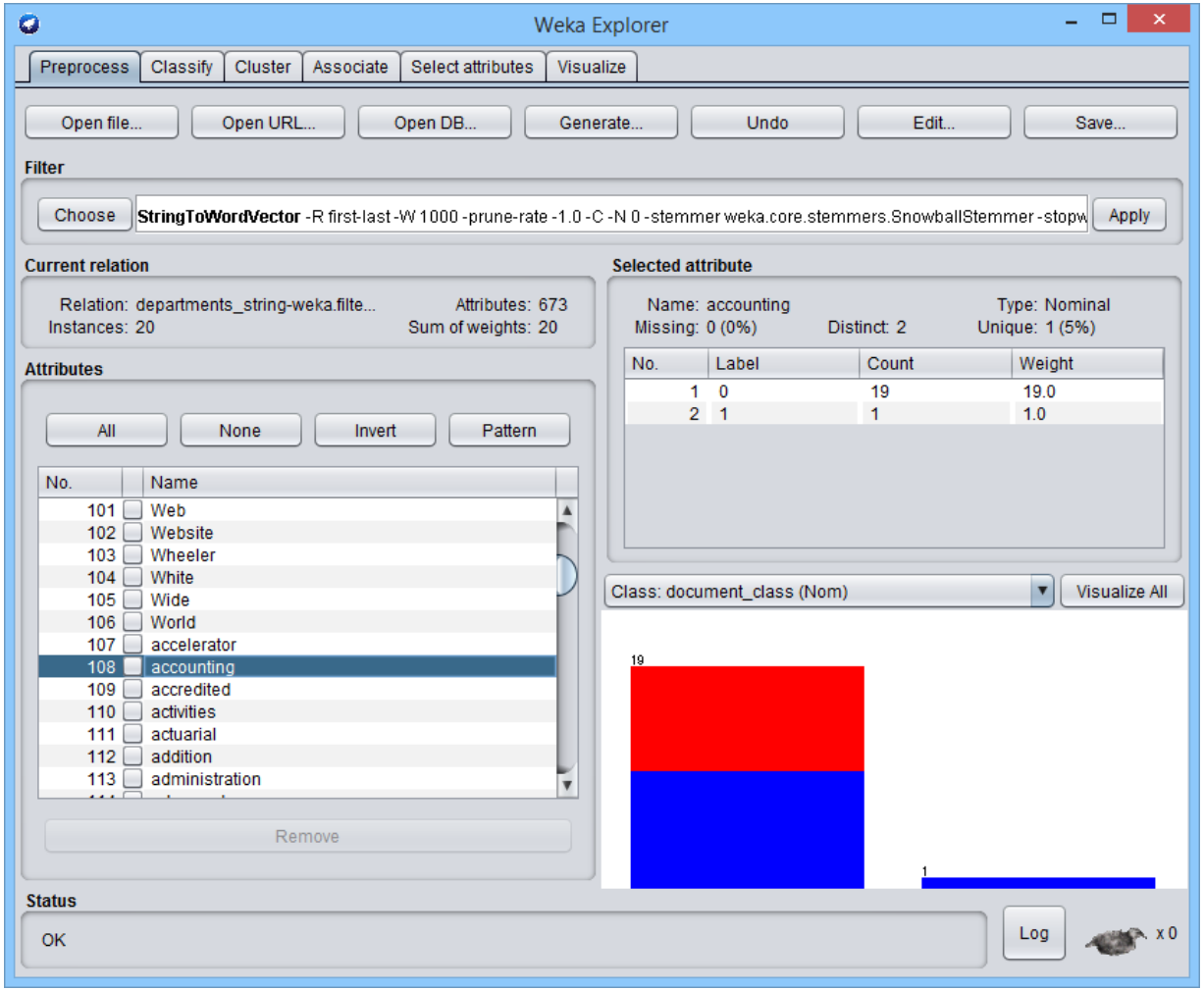
Student Projects

- Preprocess.html
- Visualization.html

Attribute Selection

Finding a minimal set of attributes that preserve the class distribution

Attribute relevance with respect to the class – irrelevant attribute (*accounting*)

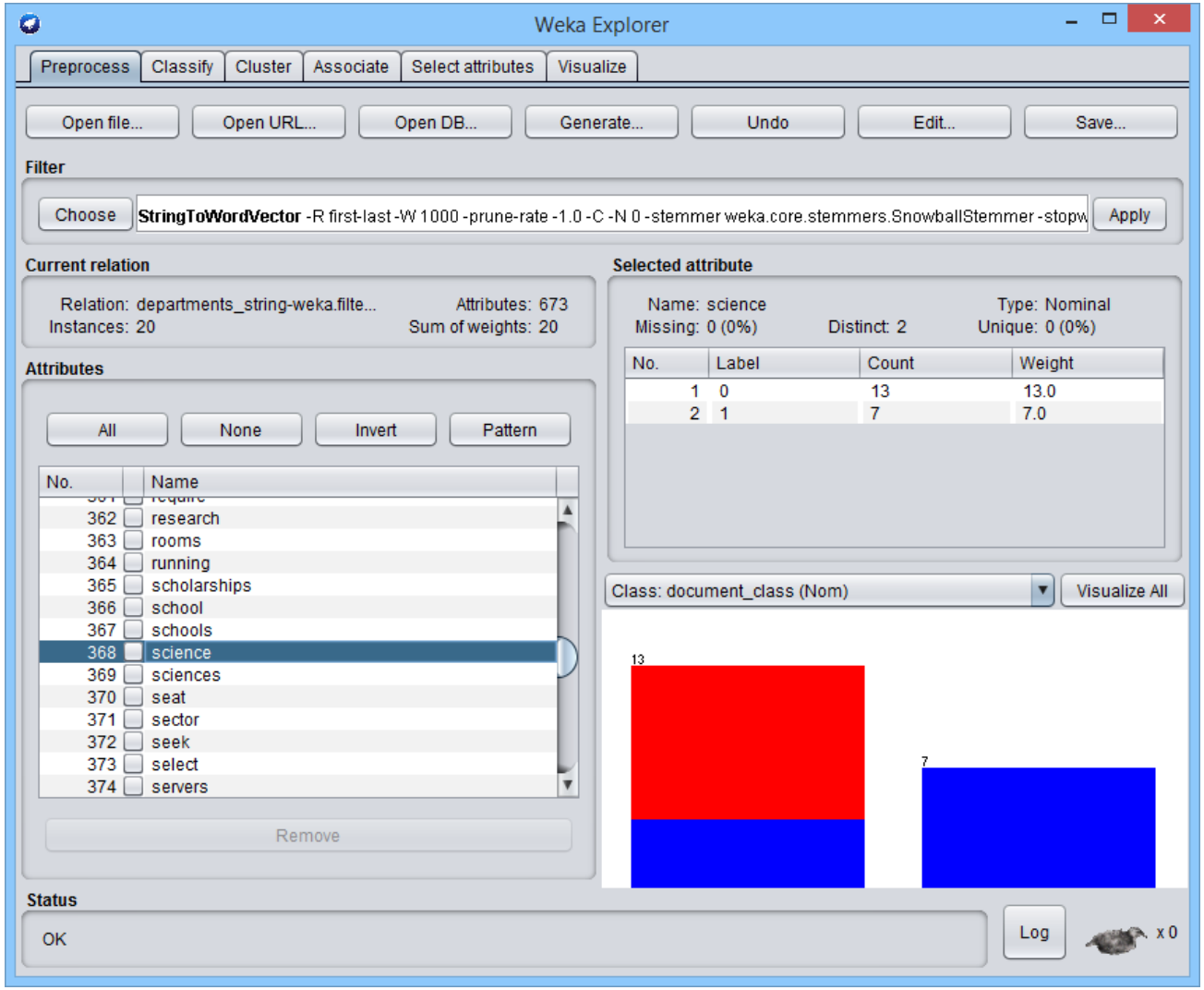


IF accounting=1 THEN class=A (Error=0, Coverage = 1 instance → **overfitting**)

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

Attribute Selection

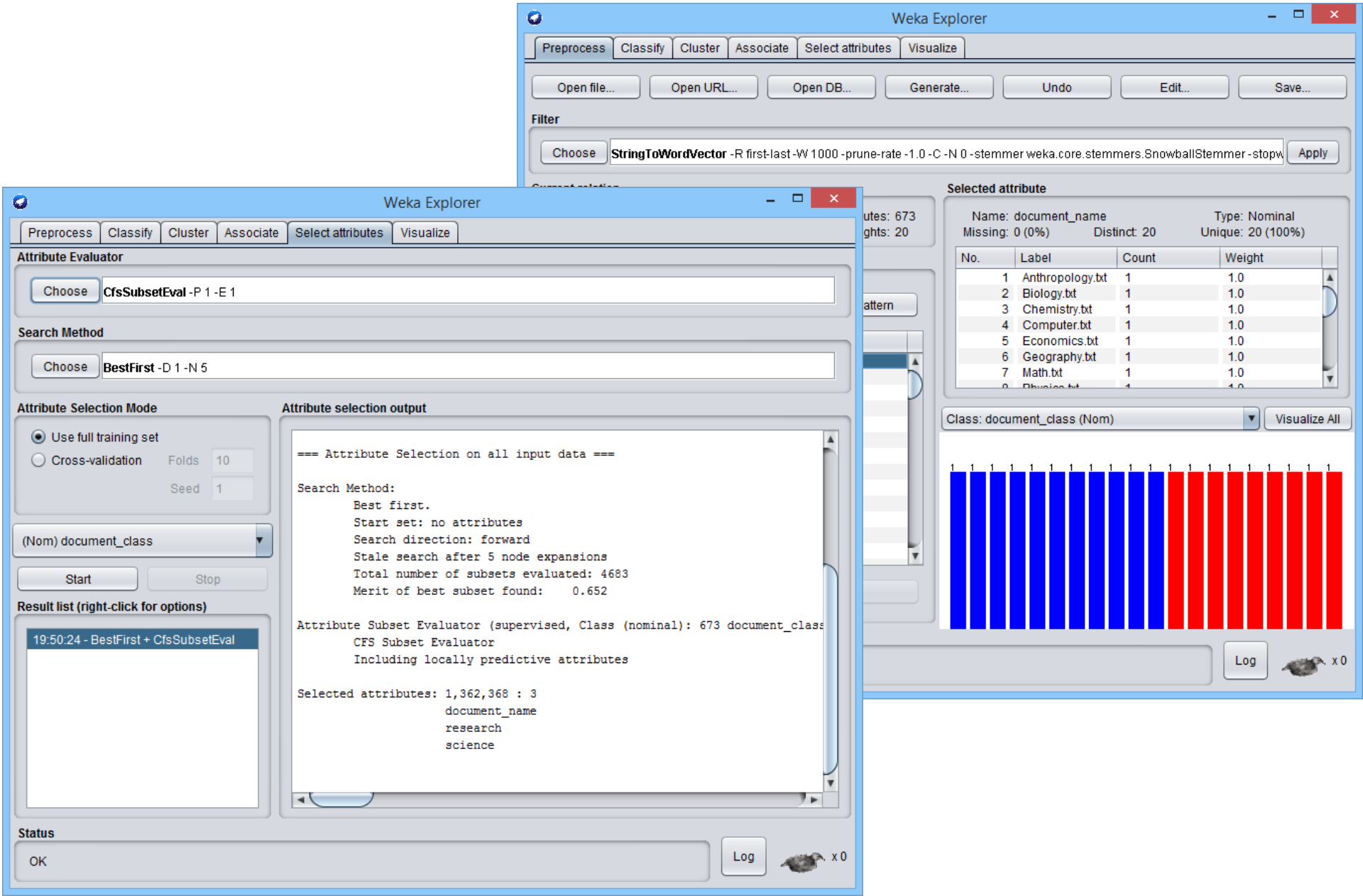
Attribute relevance with respect to the class – relevant attribute (*science*)



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

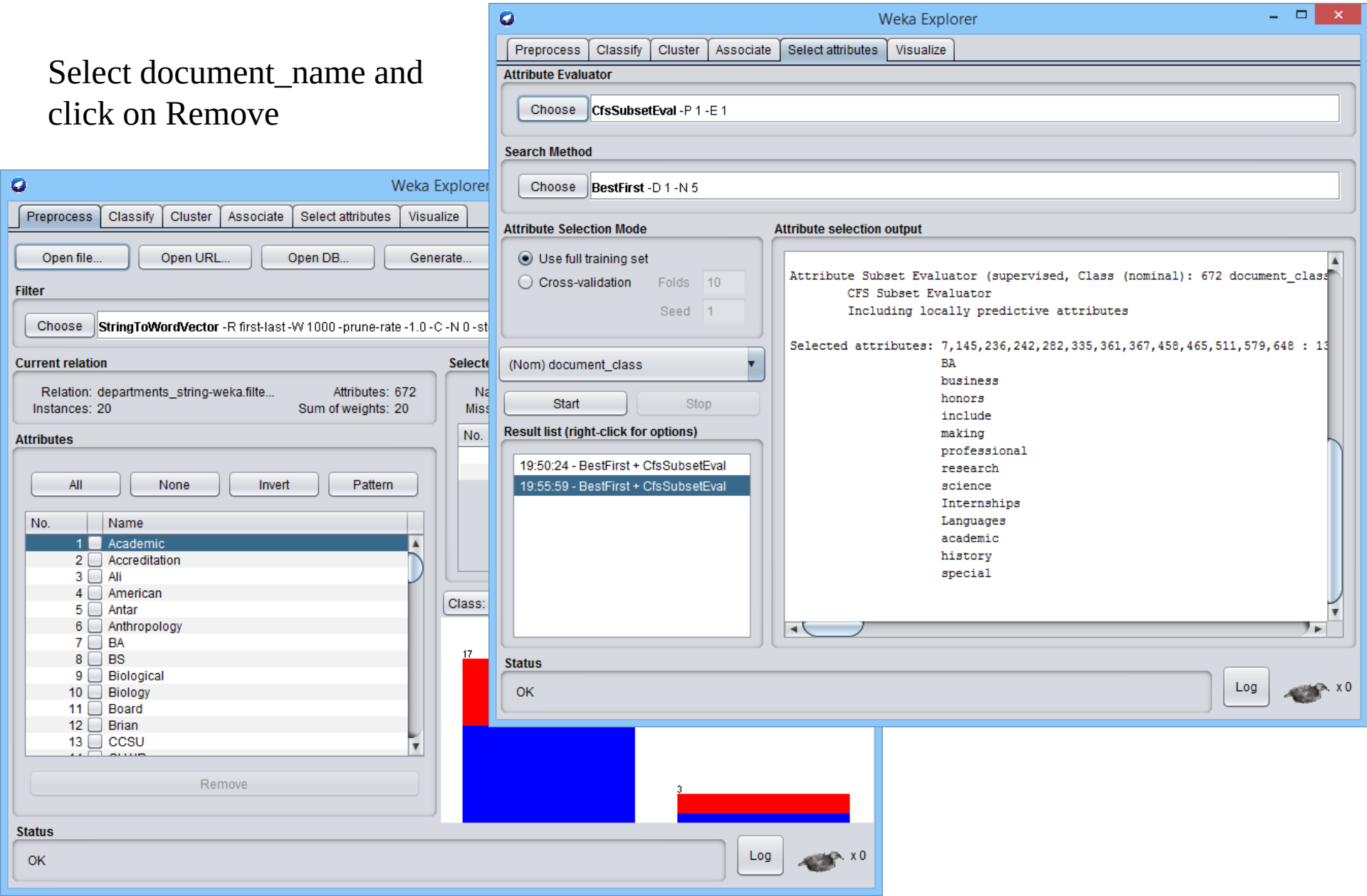
IF science=0 THEN class=B (Error=4/13, Coverage = 13 instances)

Attribute Selection (with document_name)

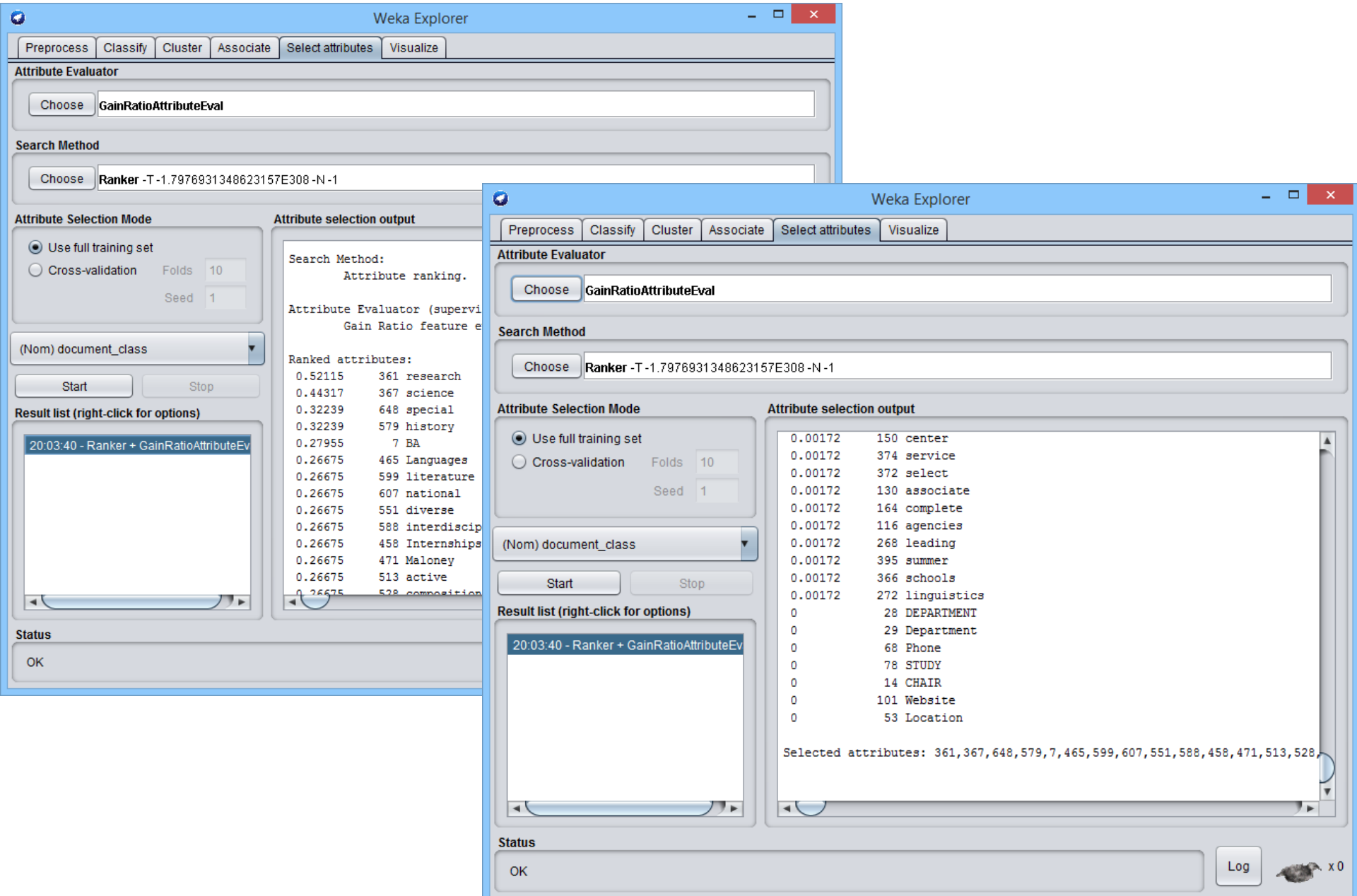


Attribute Selection (without document_name)

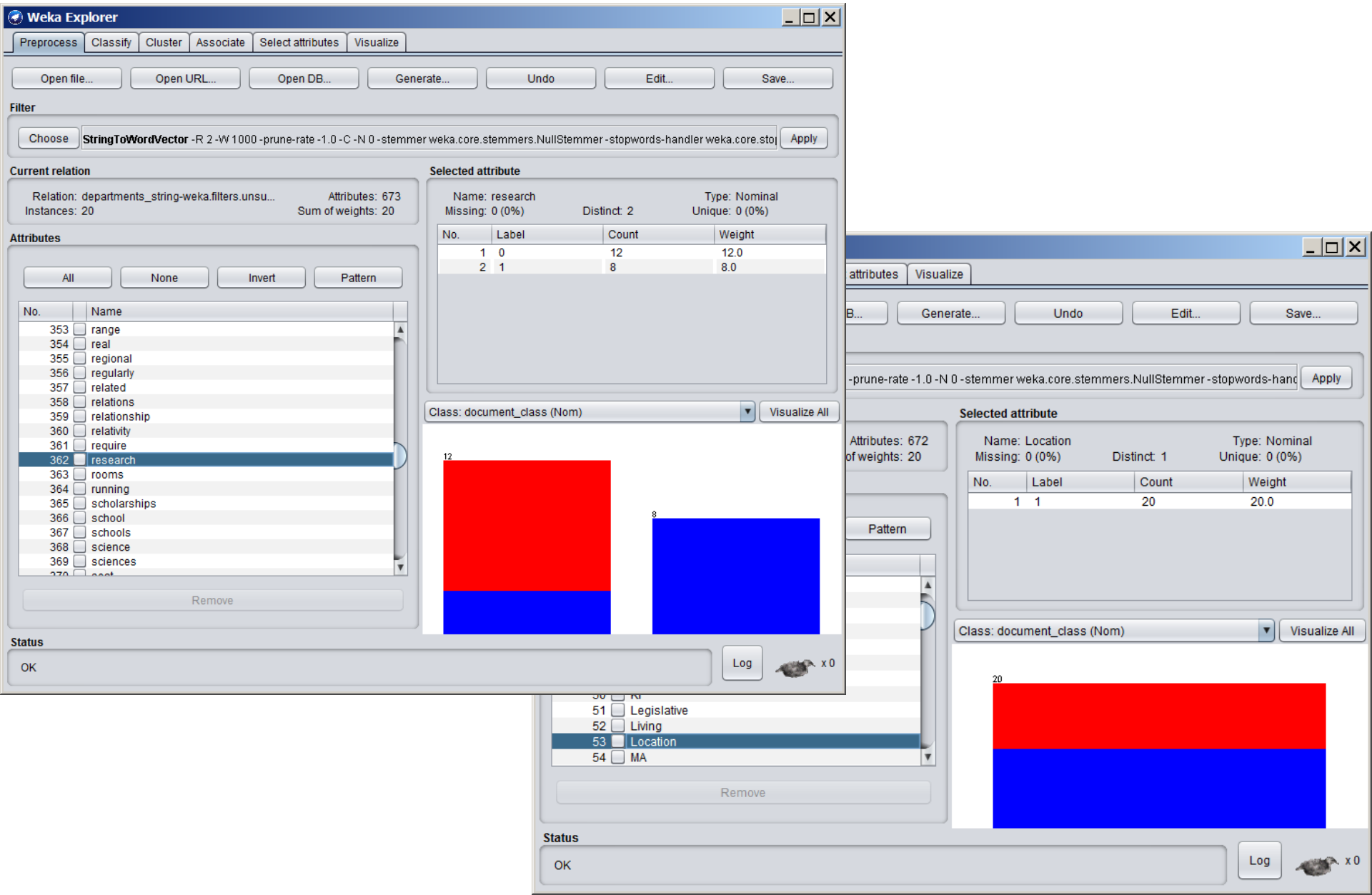
Select document_name and click on Remove



Attribute Selection (ranking)

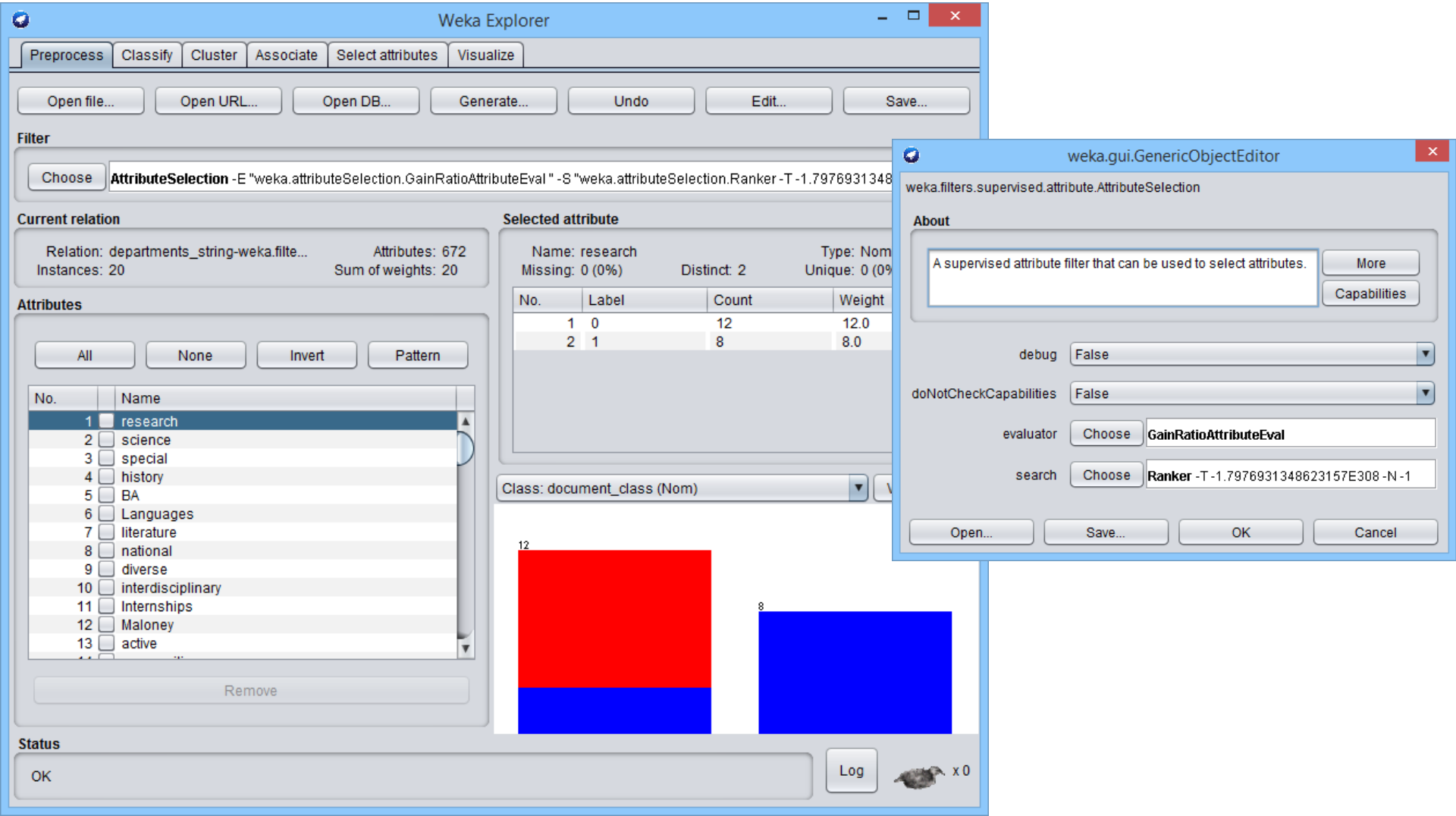


Attribute Selection (explanation of ranking)



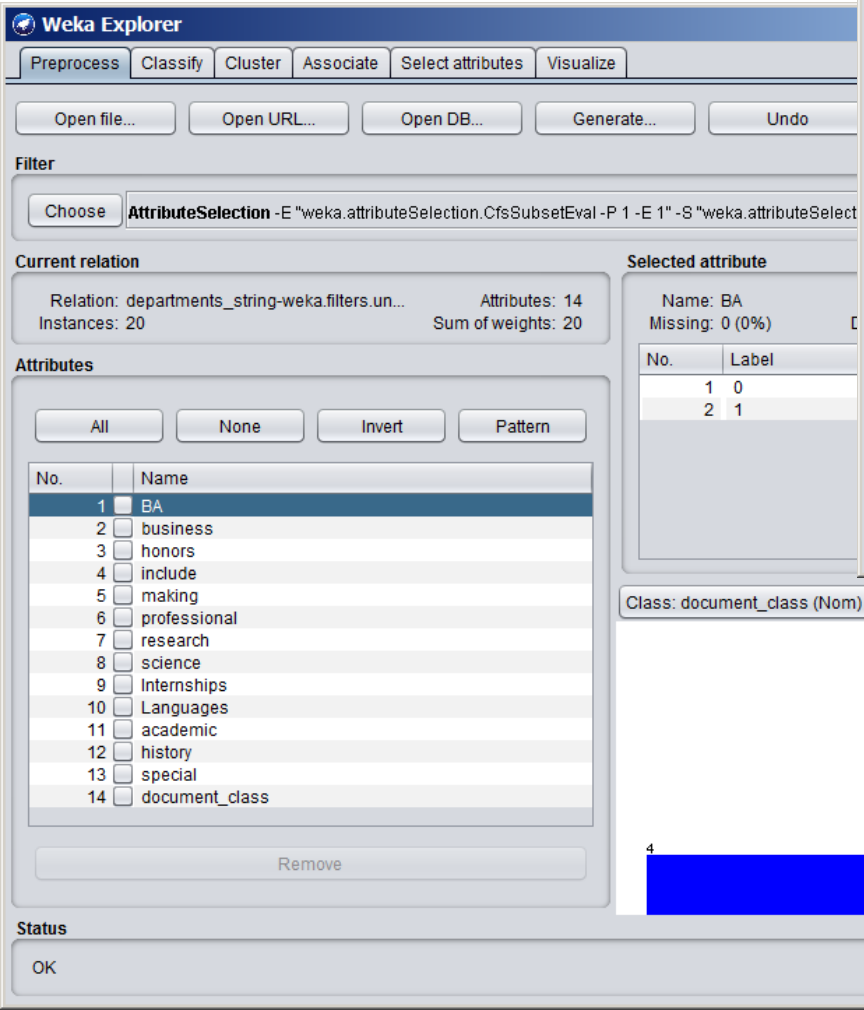
Attribute Selection (using filters)

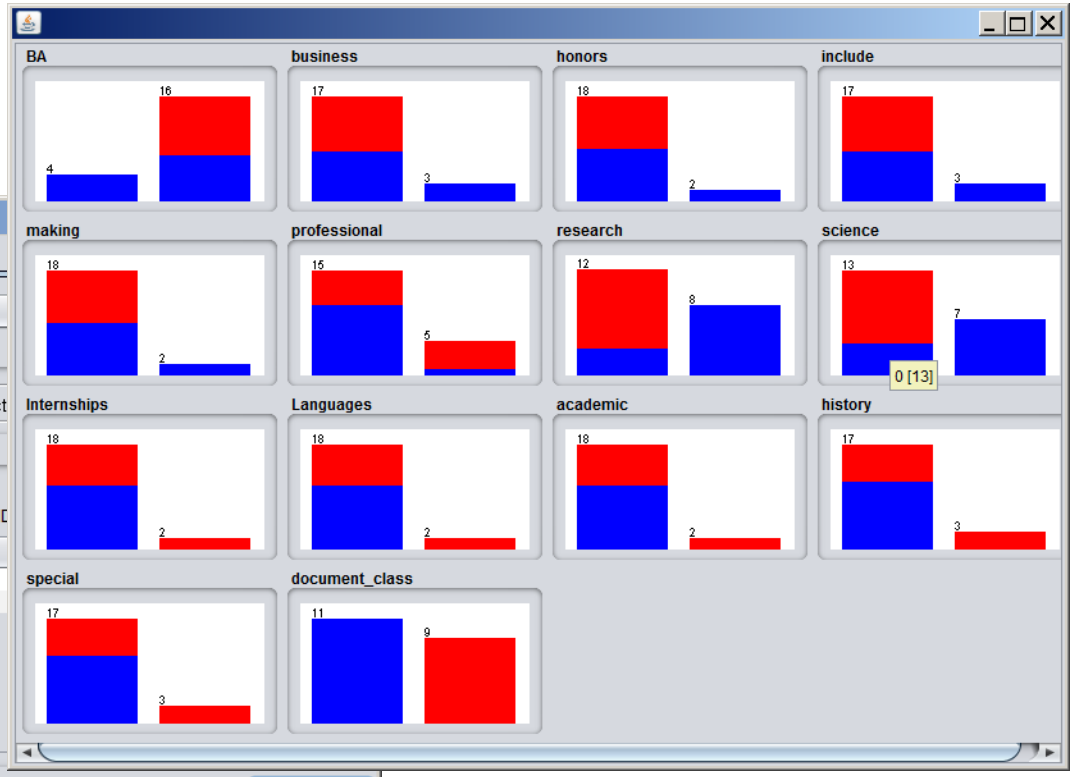
- Choose filters/supervised/attribute/AttributeSelection
- Set parameters to InfoGainAttributeEval and Ranker
- Click on Apply and see the attribute ordering



Attribute Selection (using filters)

Choose filters/supervised/attribute/AttributeSelection and use CfsSubsetEval and BestFirst search. Then click on Visualize All





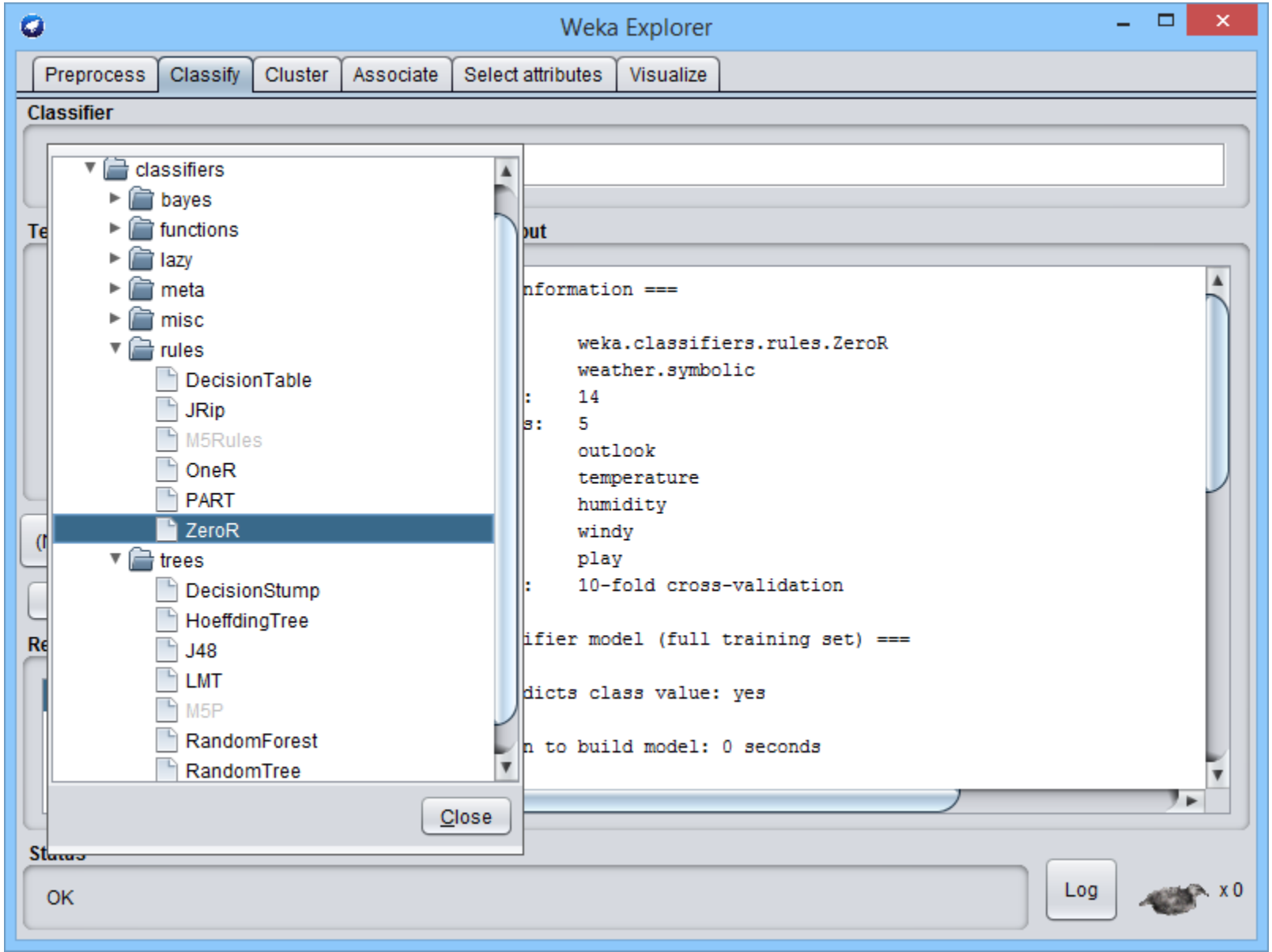
Attribute Selection

Student Projects

- Attribute Selection.html

Classification – creating models (hypotheses)

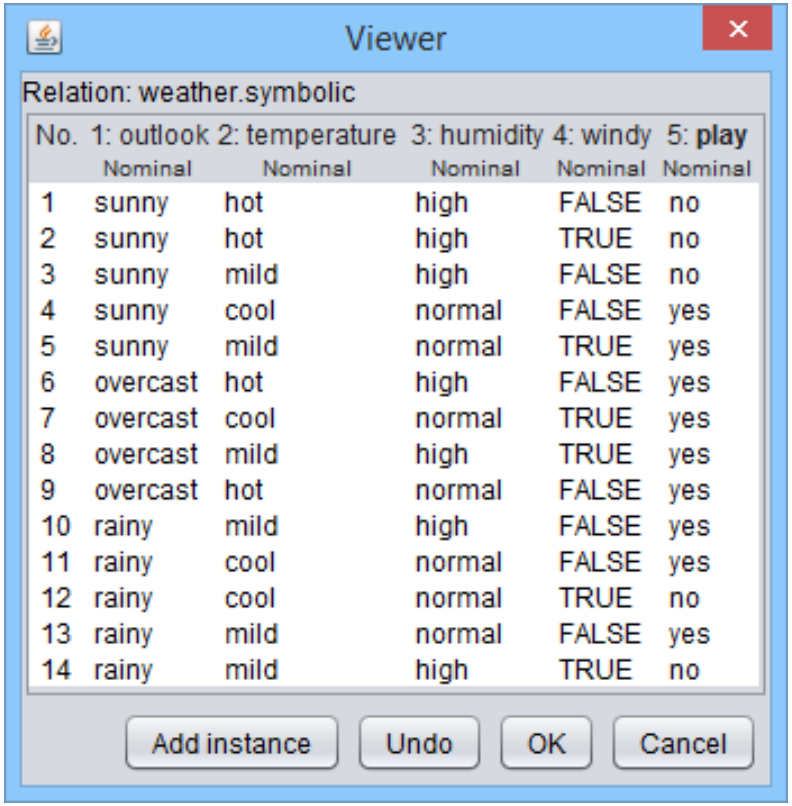
Mapping (independent attributes -> class)



Classification – creating models (hypotheses)

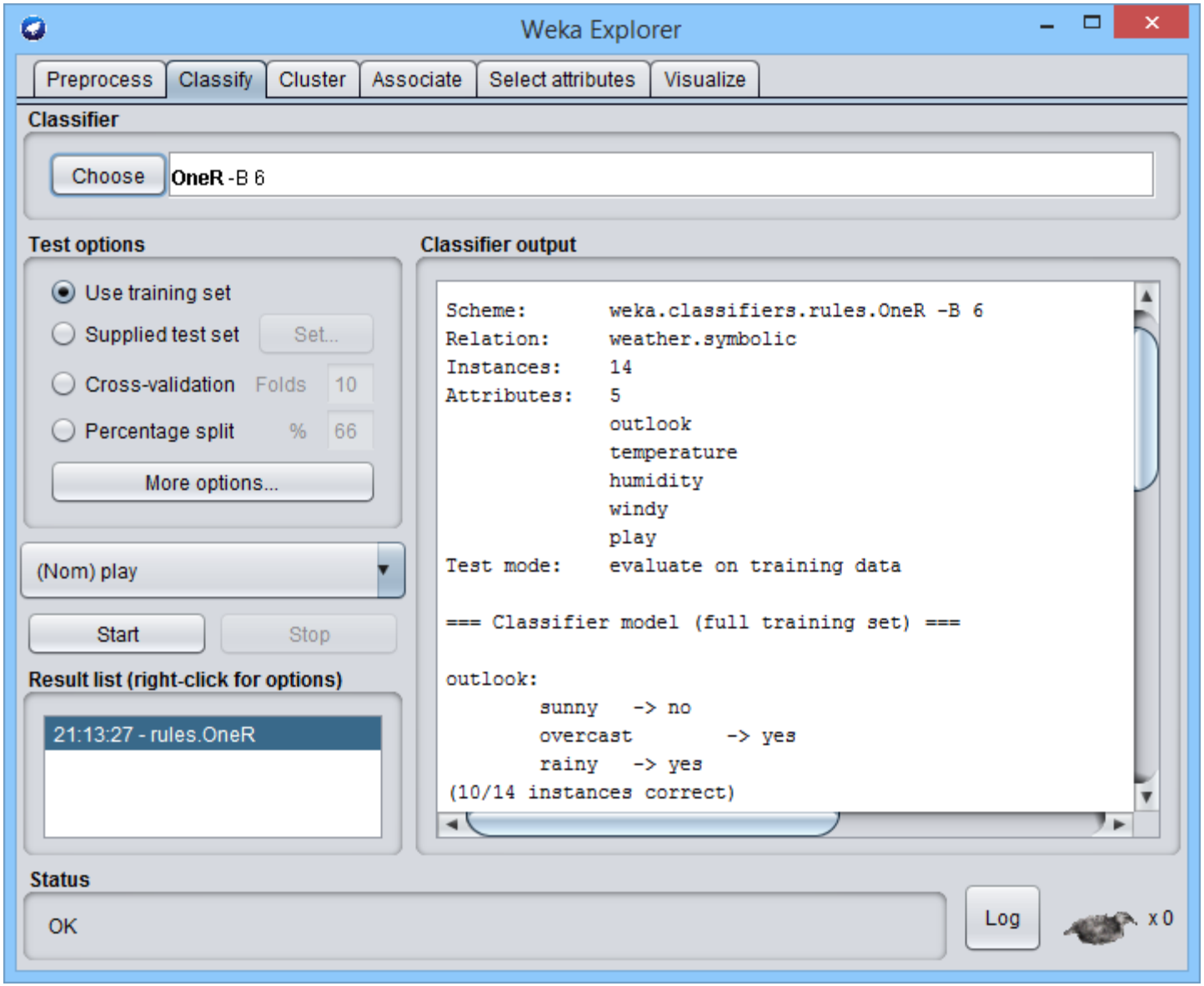
Inferring one-attribute rules - OneR

Weather data (weather.nominal.arff)



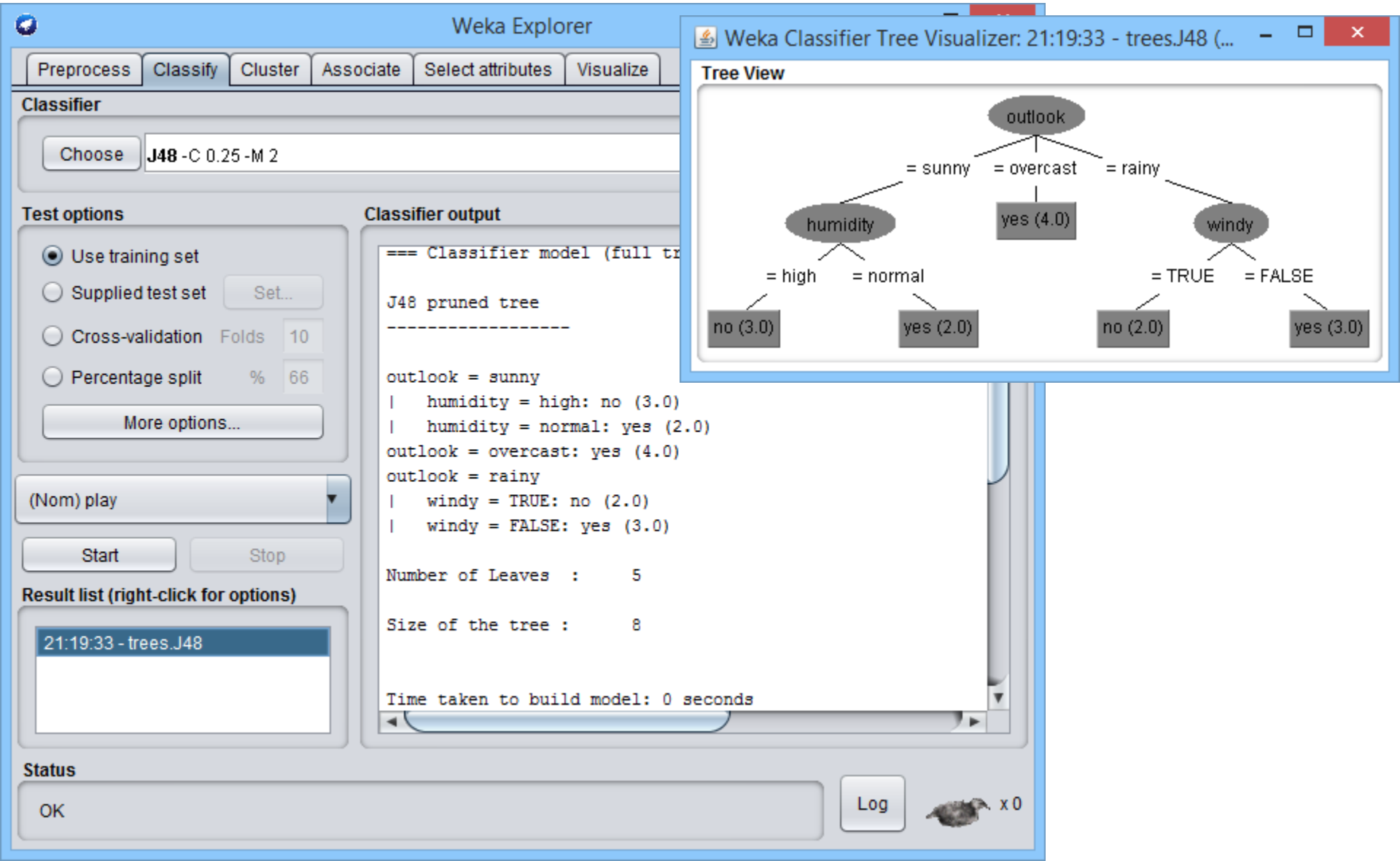
Attribute	Rules	Errors	Total error
outlook	sunny -> no overcast -> yes rainy -> yes	2/5 0/4 2/5	4/14
temperature	hot -> no mild -> yes cool -> yes	2/4 2/6 1/4	5/14
humidity	high -> no normal -> yes	3/7 1/7	4/14
windy	false -> yes true -> no	2/8 3/5	5/14

Classification – OneR



Classification – decision tree

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



Classification – decision tree

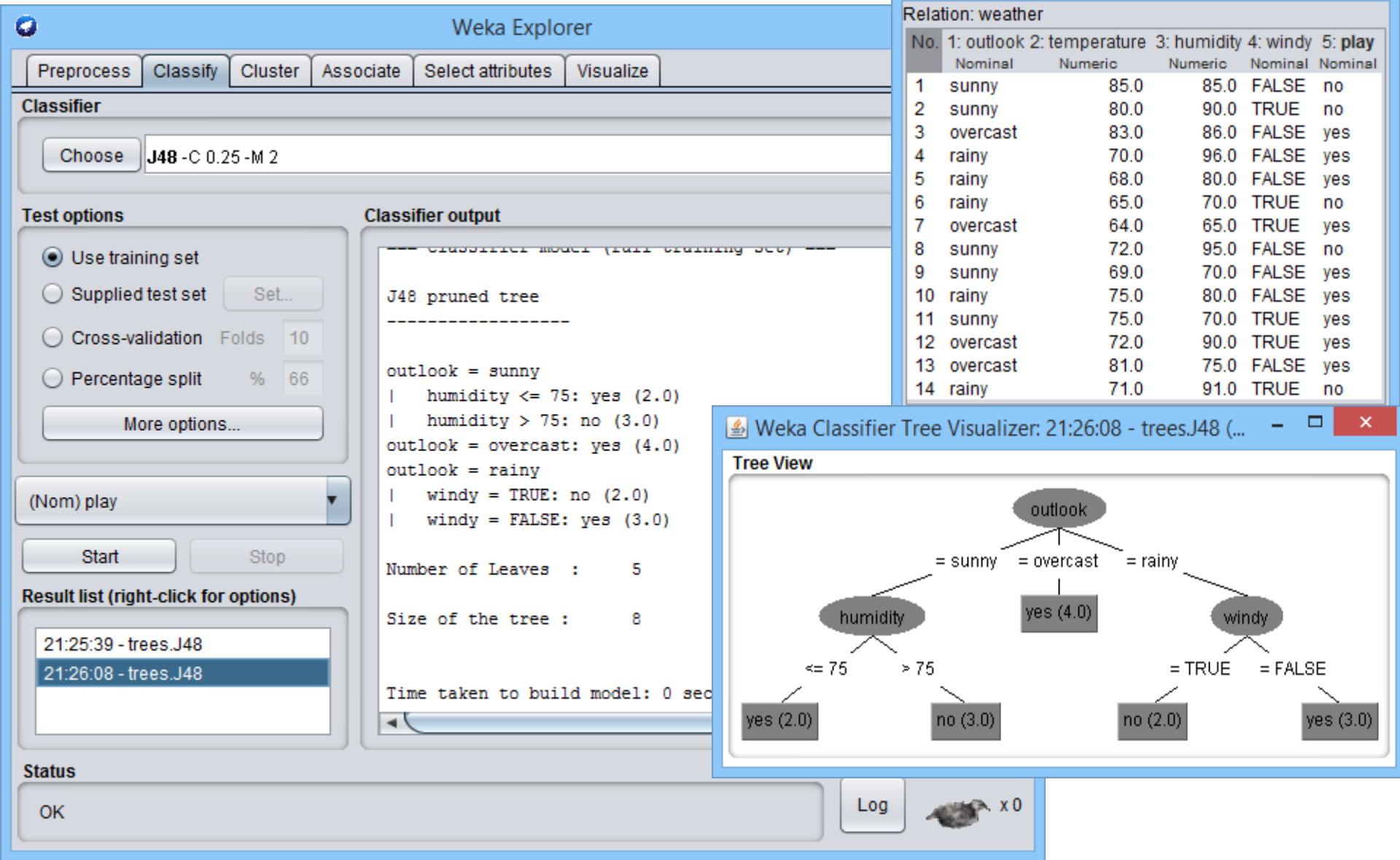
Top-down induction of decision trees (TDIDT, old approach know from pattern recognition):

- 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

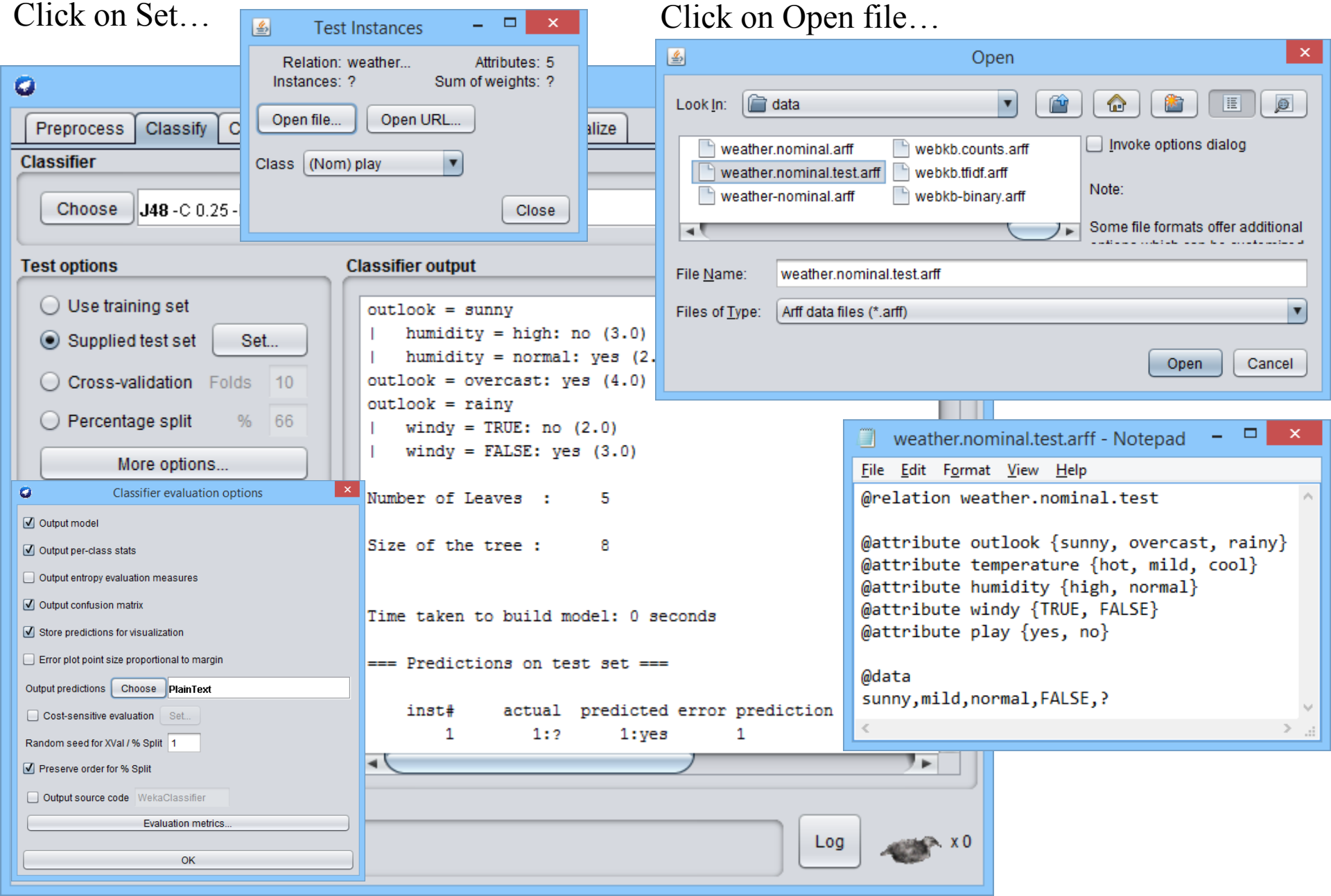
weather.arff



Classification – predicting class

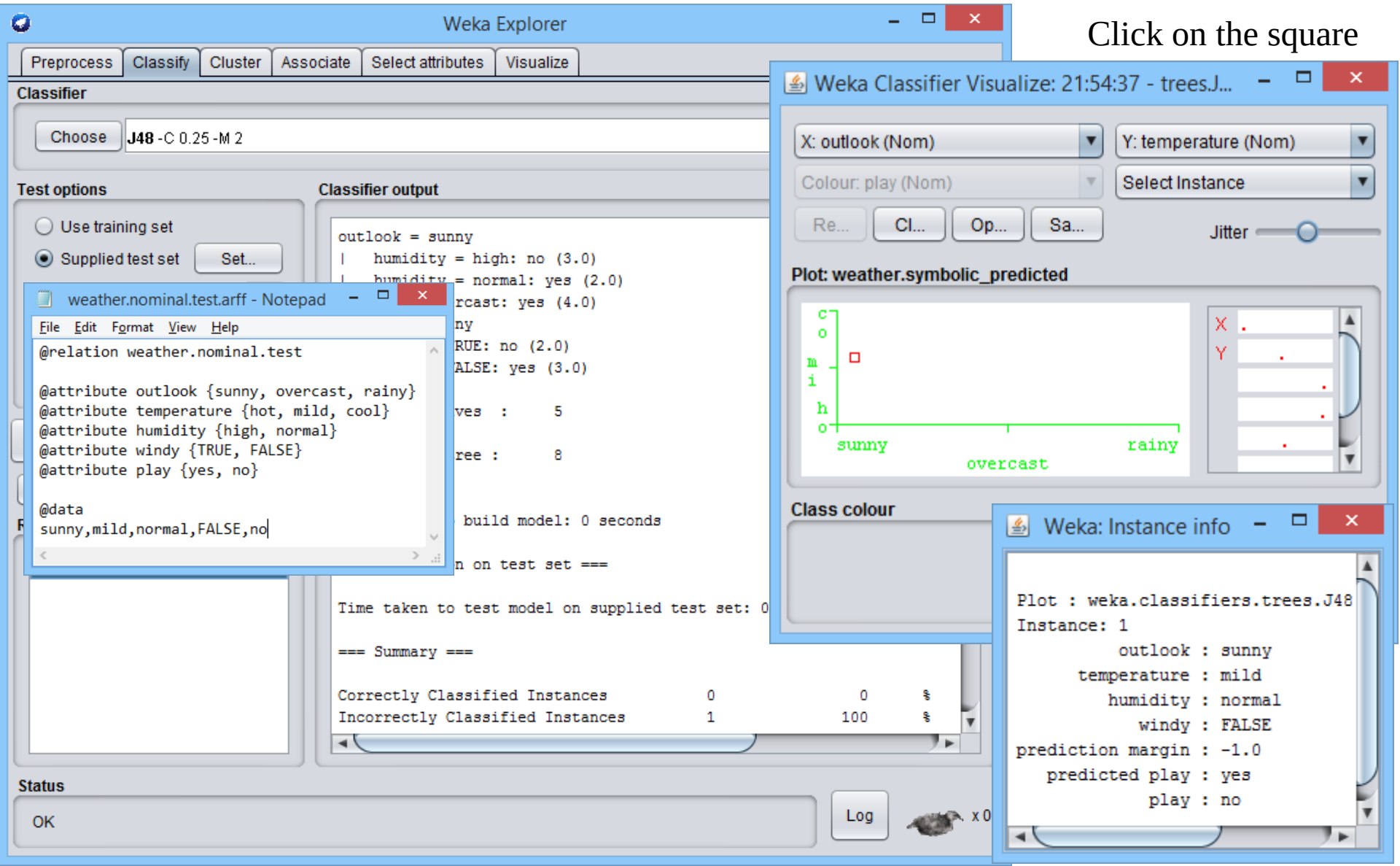
Click on Set…

Click on Open file…



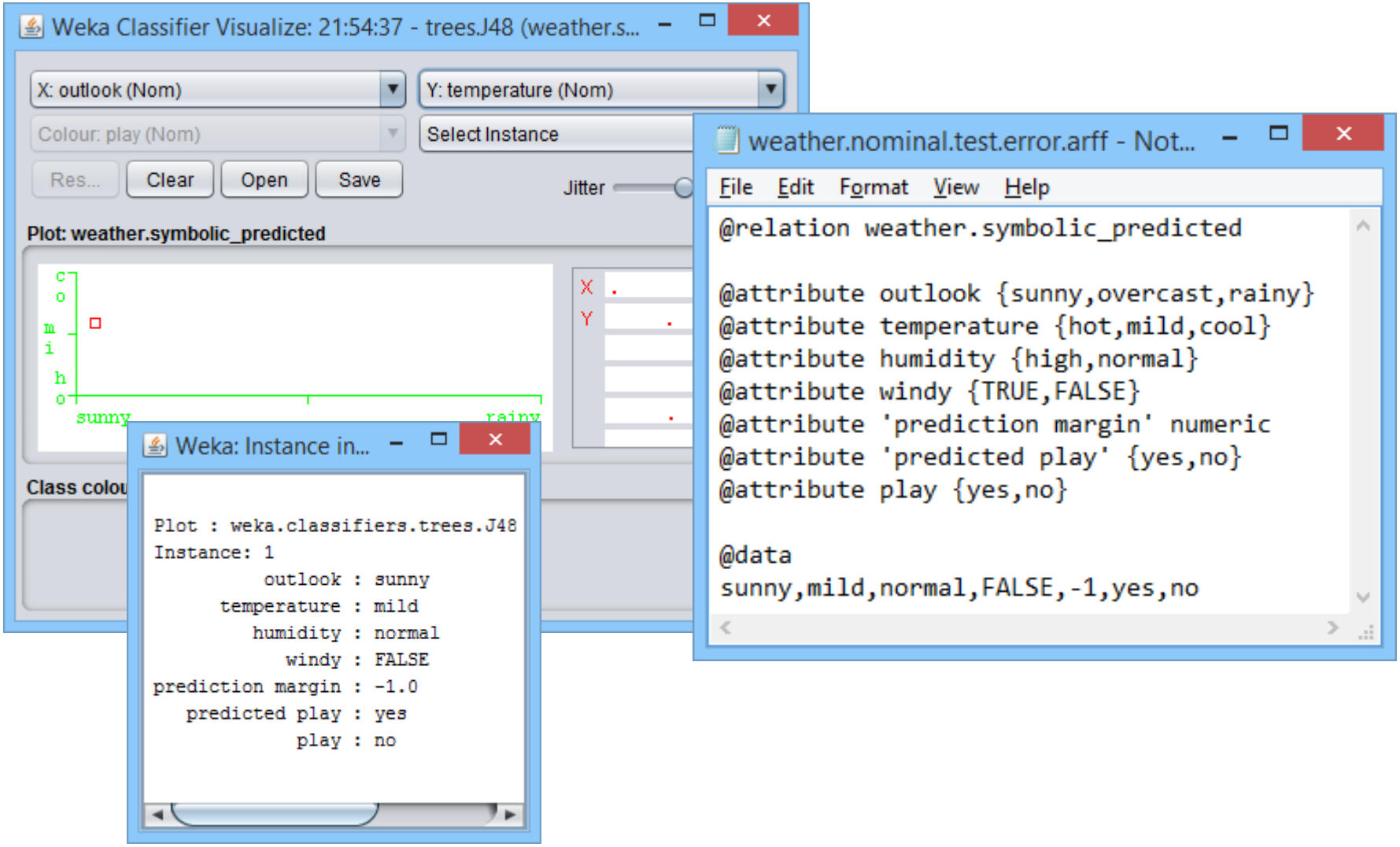
Classification – predicting class

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



Classification – predicting class

Click on Save



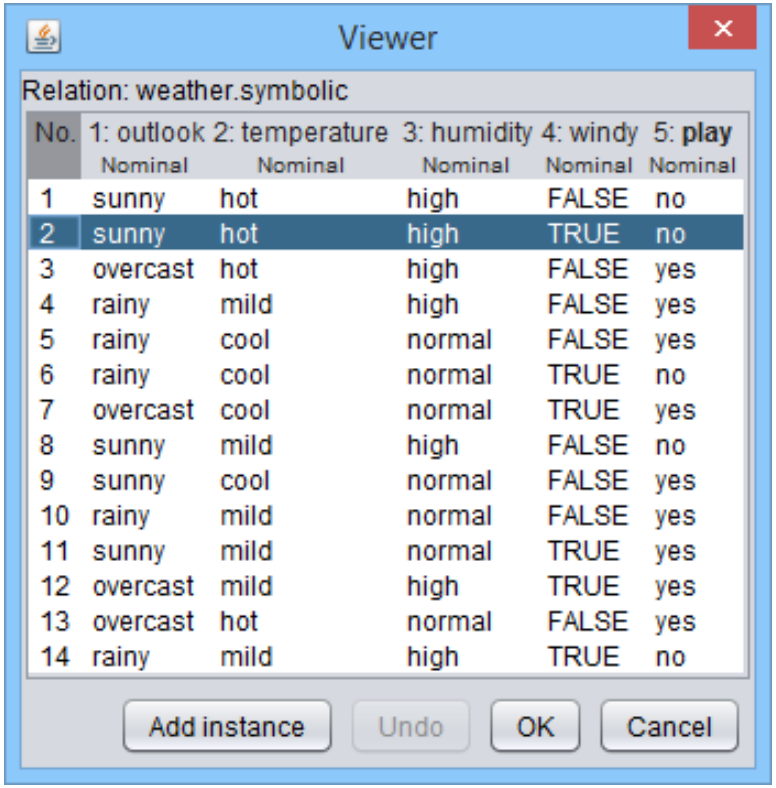
Classification

Student Projects

- Classification.html

Prediction (no model, lazy learning)

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



- K-nearest neighbor (IBk)
  Take the class of the nearest neighbor or the majority class among K neighbors

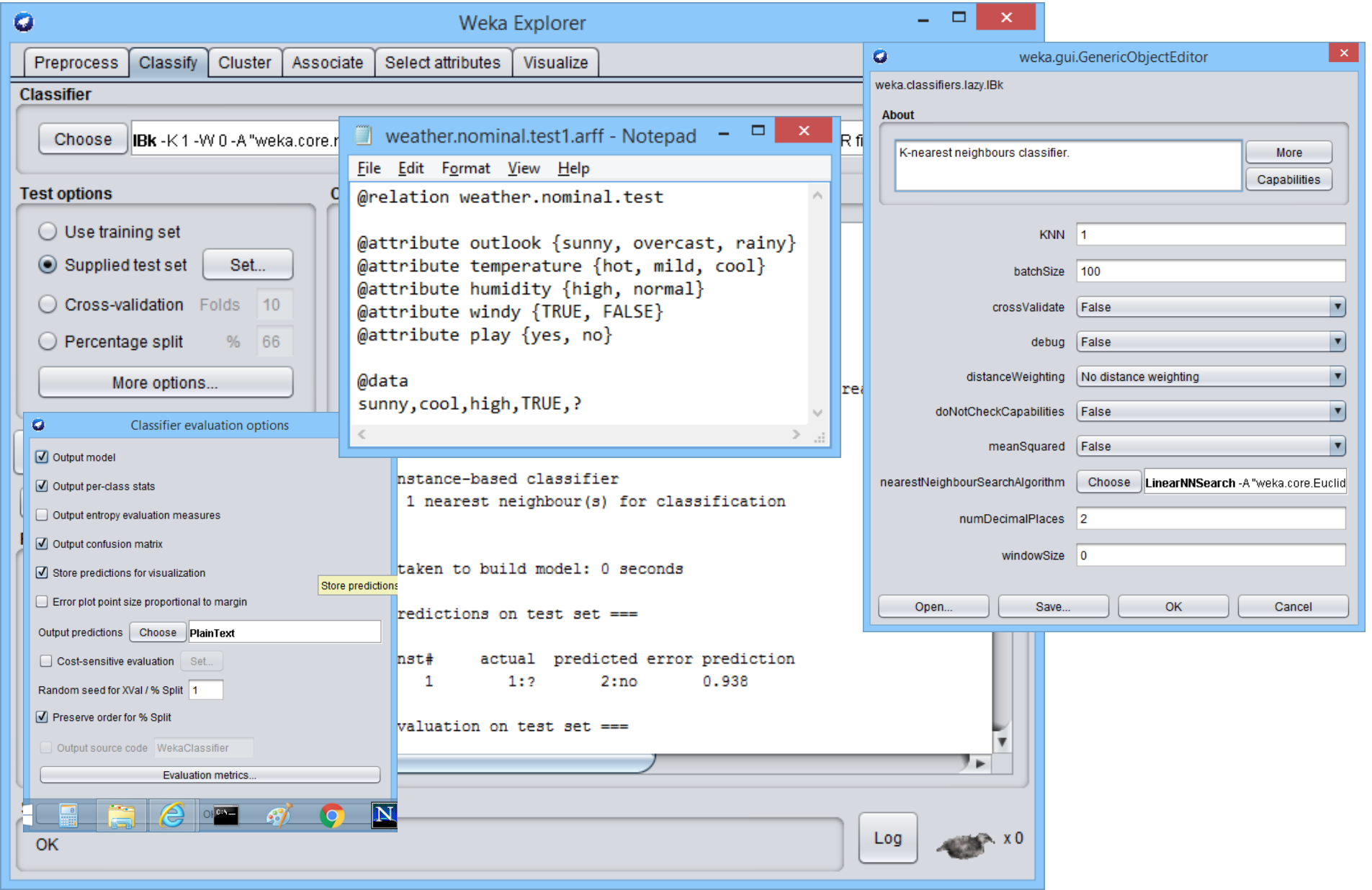
  K=1 -> no
  K=3 -> no
  K=5 -> yes
  K=14 -> yes (Majority predictor, ZeroR)

- 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)

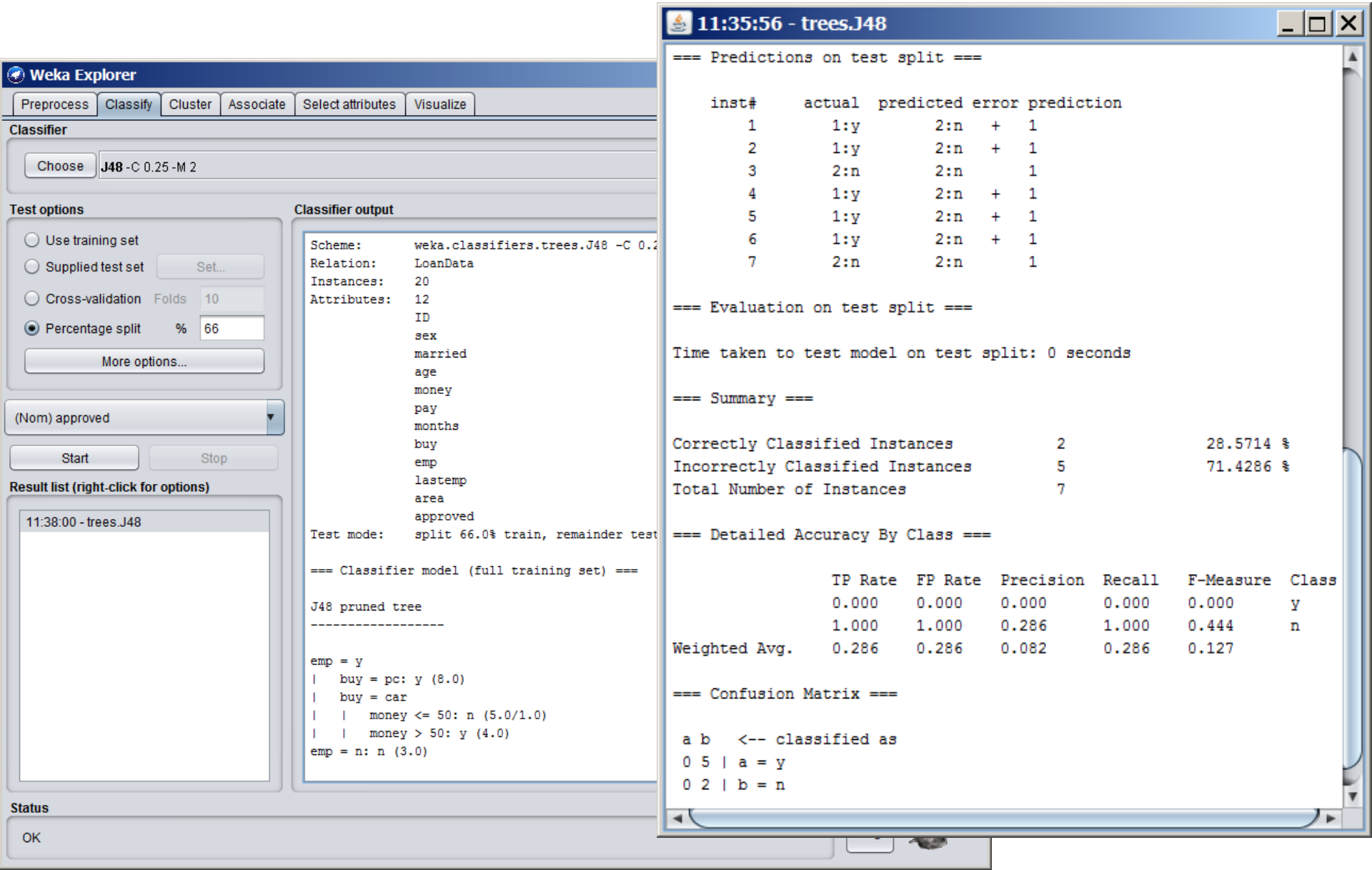


Prediction

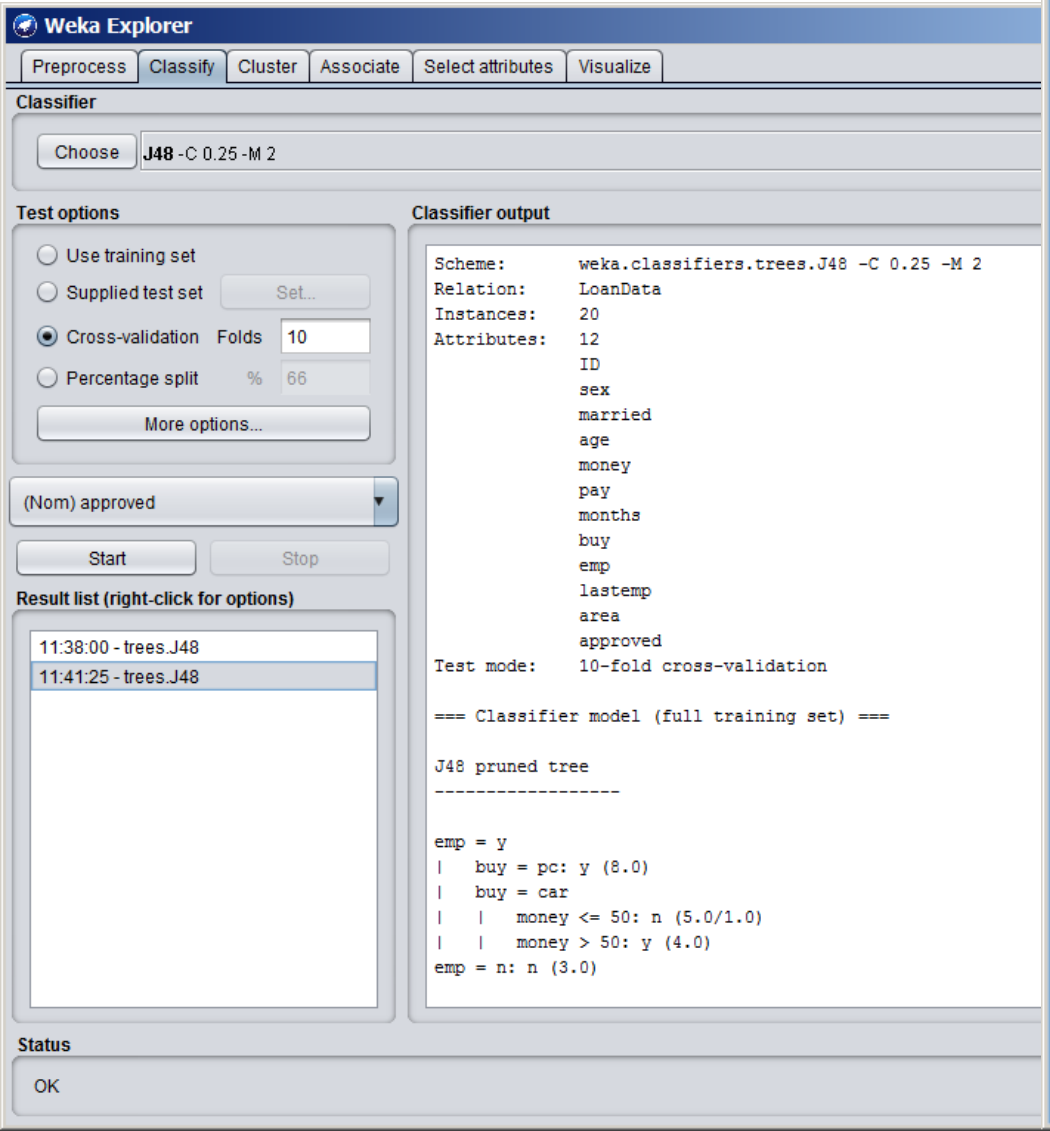
Student Projects

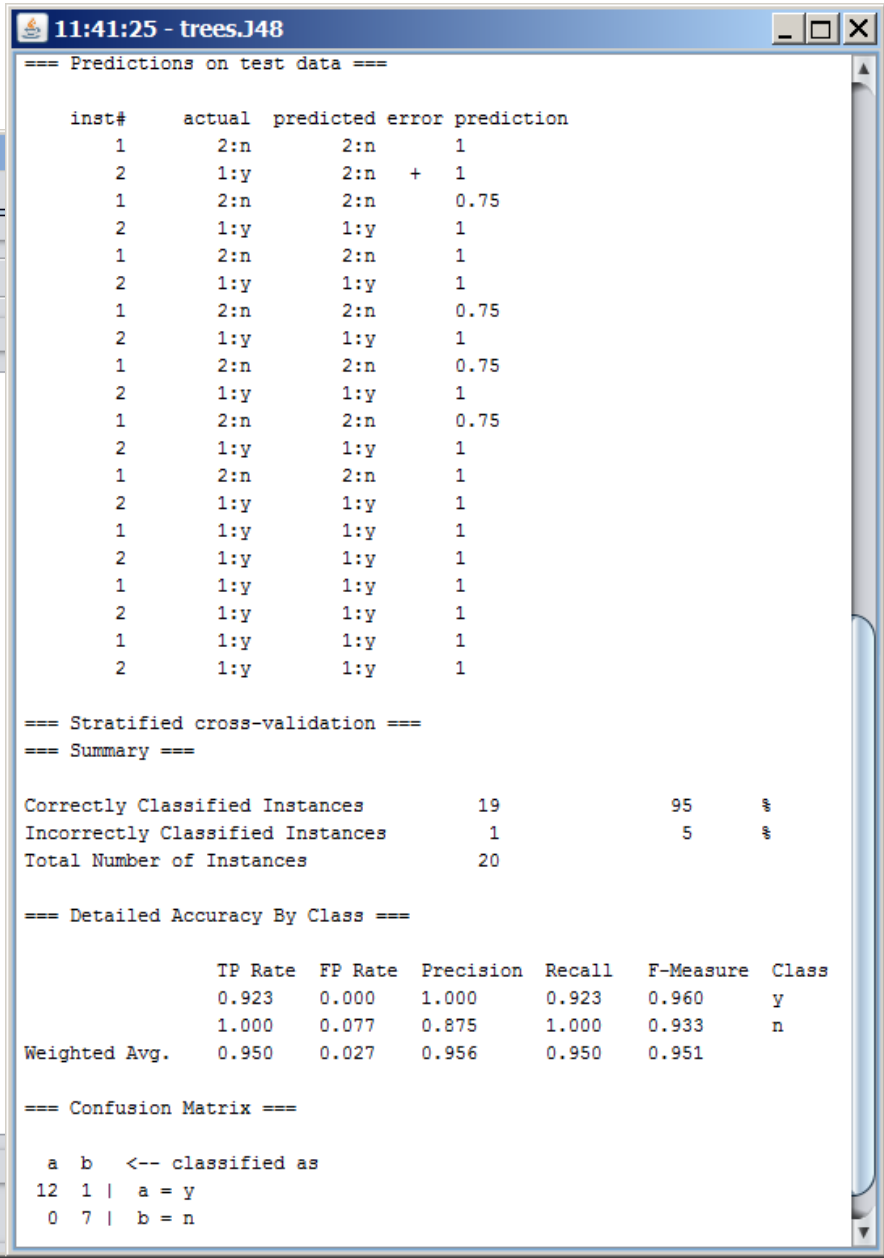
- Prediction.html

Model evaluation – holdout (percentage split)

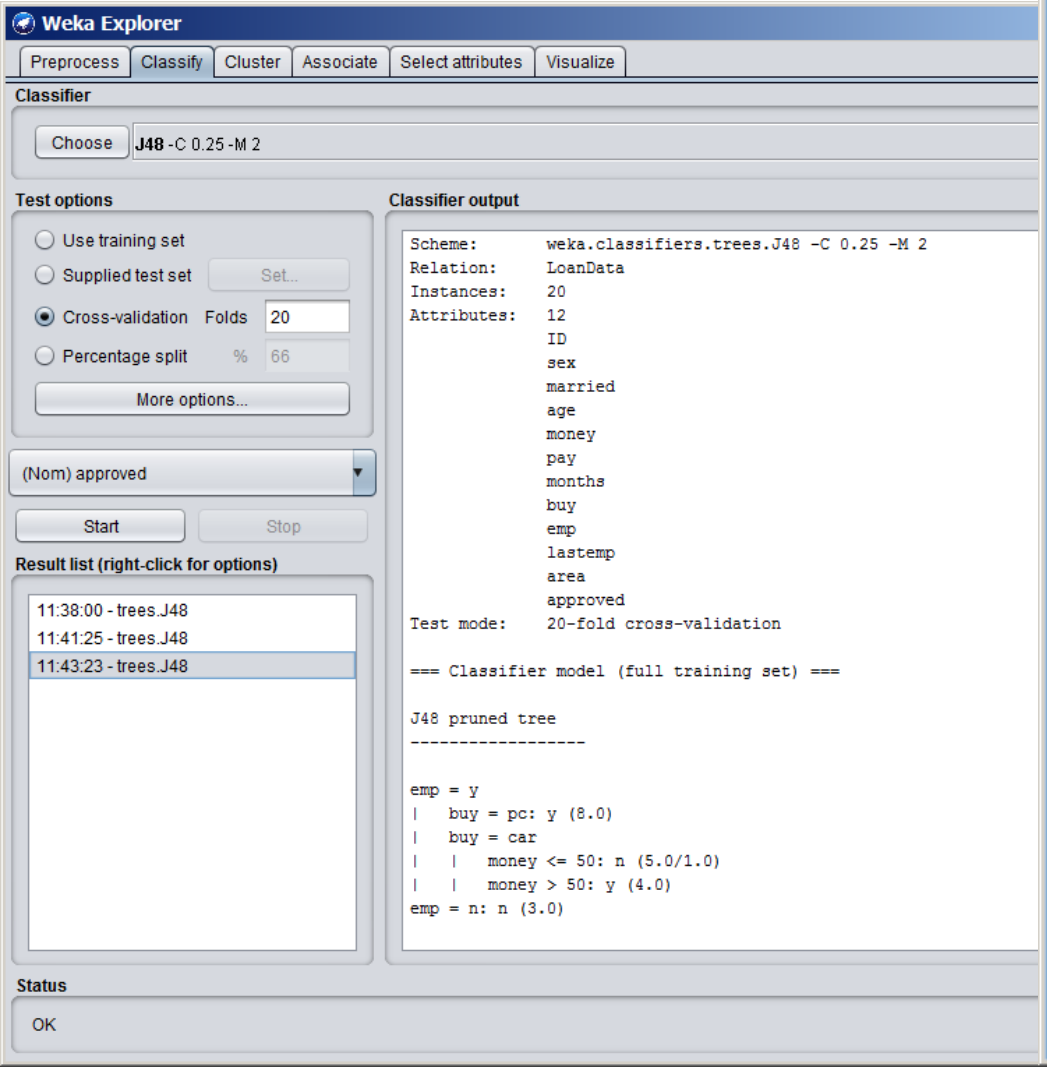


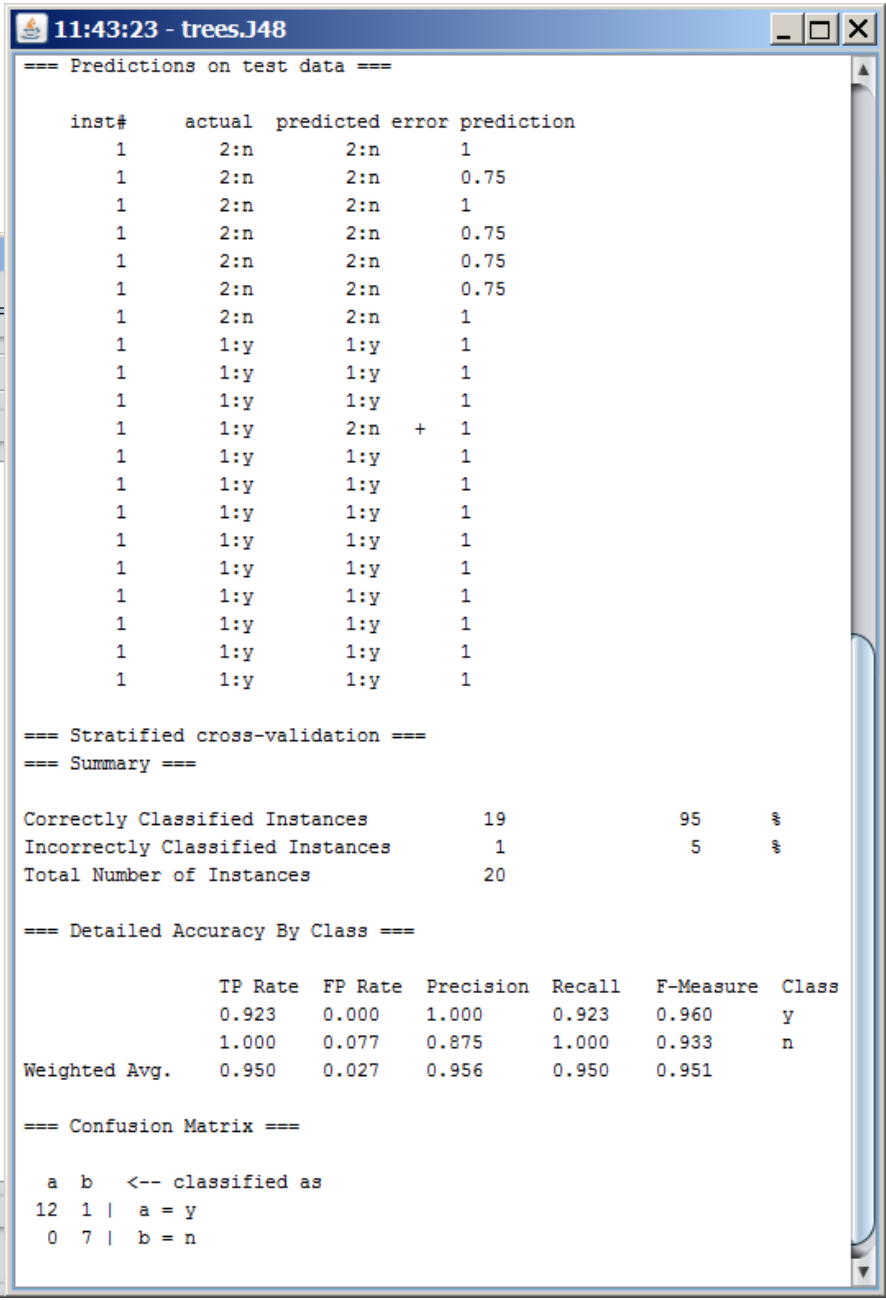
Model evaluation – cross validation



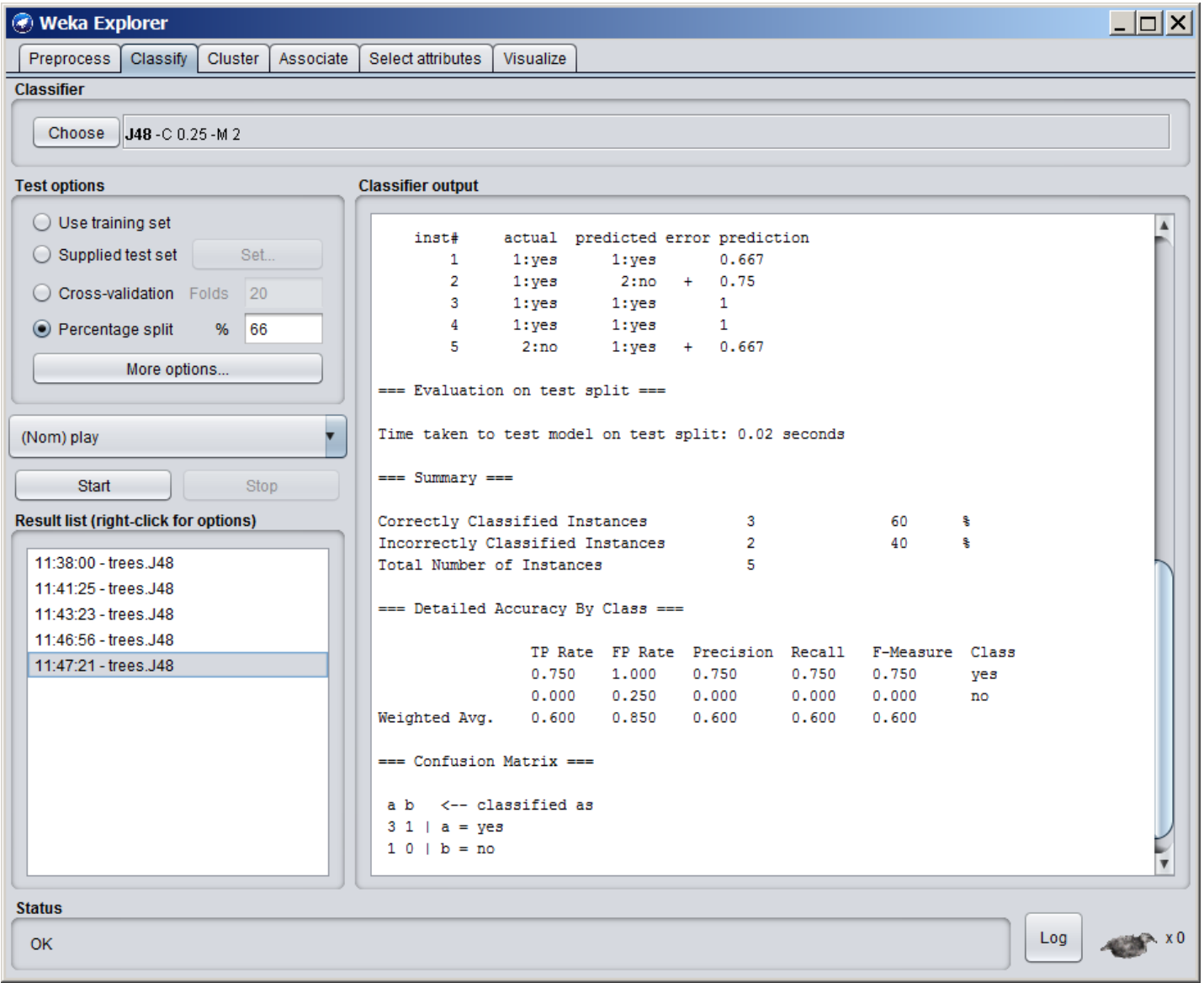


Model evaluation – leave one out cross validation





Model evaluation – confusion (contingency) matrix



predicted

actual	yes	no
yes	3	1
no	1	0

predicted

actual	yes	no
yes	TP	FN
no	FP	TN

Precision = TP/(TP+FP)
Recall = TP/(TP+FN)

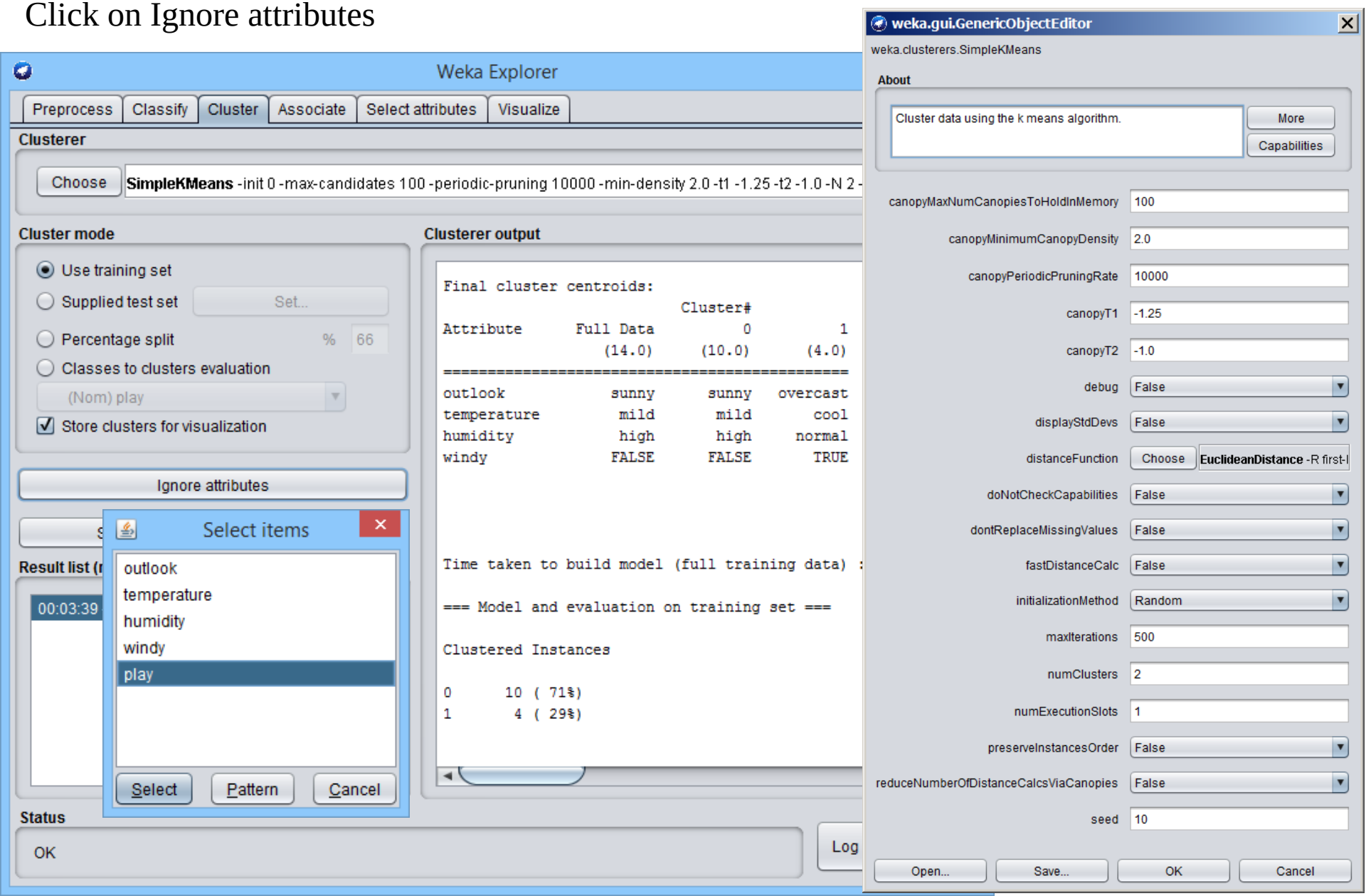
Model evaluation

Student Projects

- Evaluation.html

Clustering – k-means

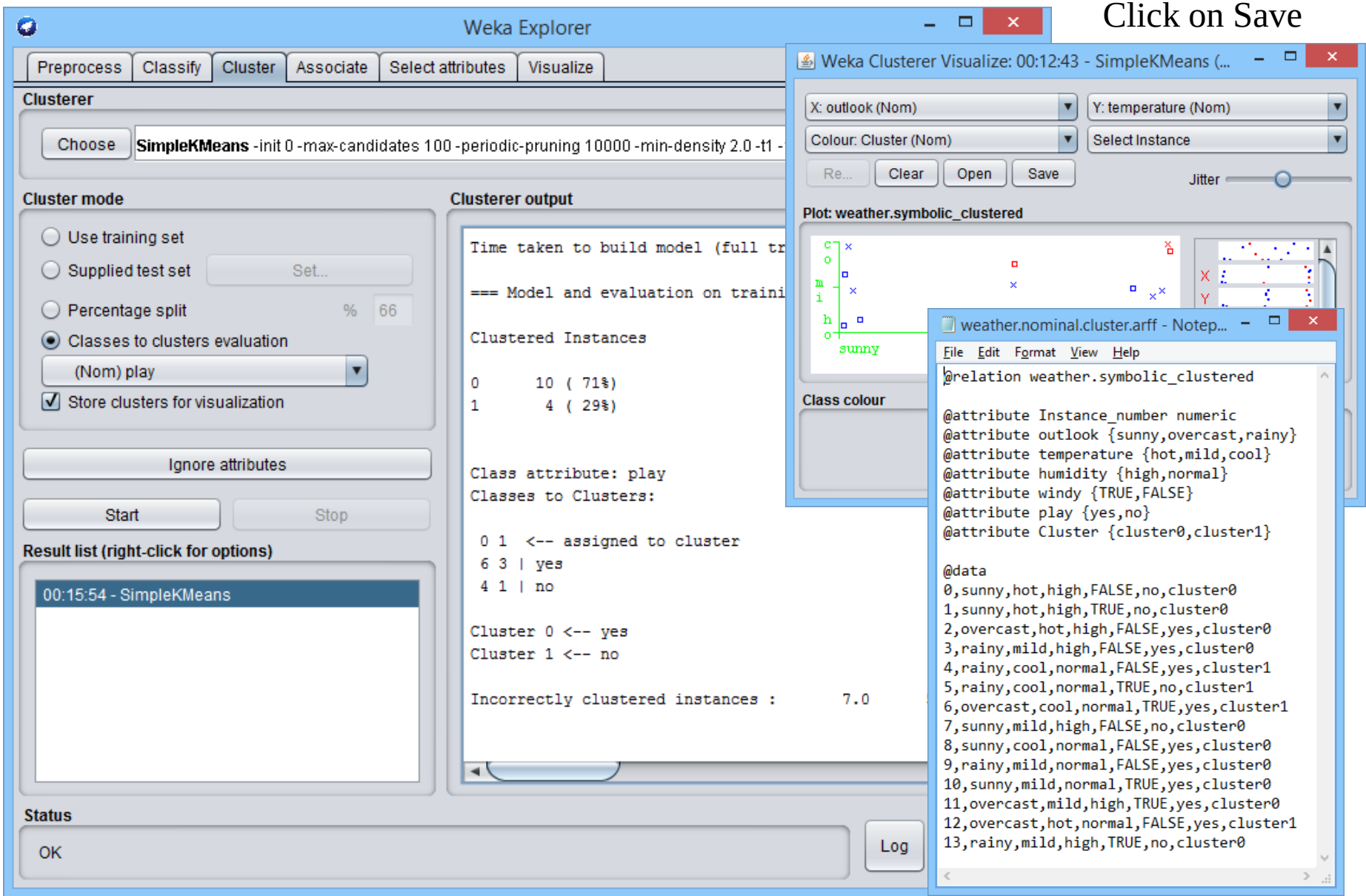
Click on Ignore attributes



Clustering – classes to clusters evaluation

Right click on Result list, select Visualize cluster assignments

Click on Save



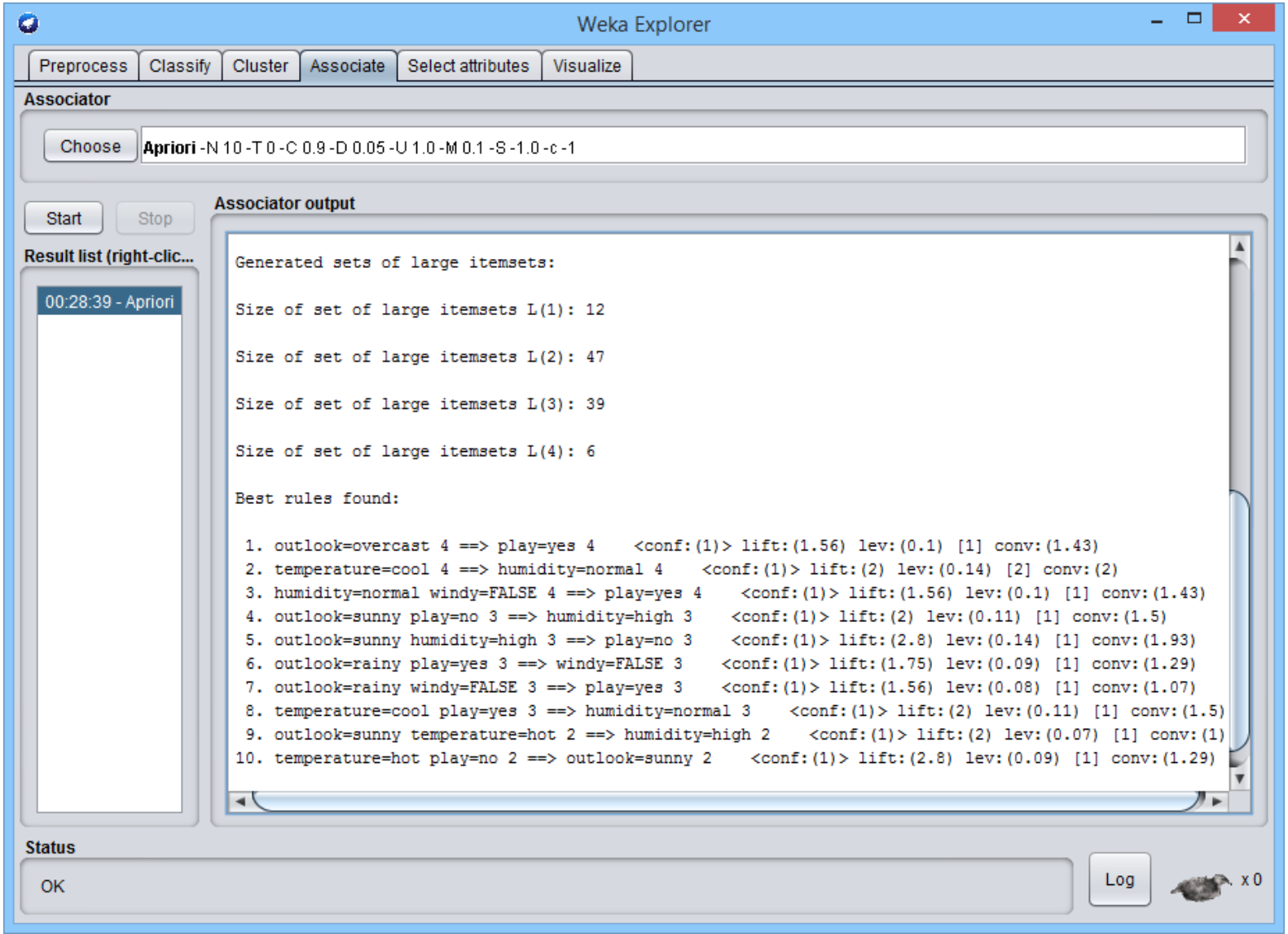
Clustering

Student Projects

- Clustering.html

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

Student Projects

- Association.html

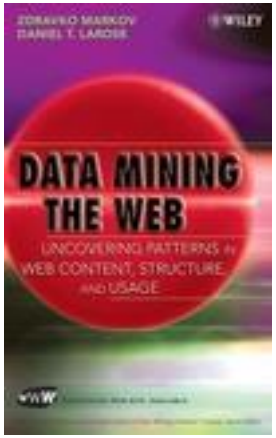
Document classification and clustering

Predict the class of the Theatre document

1. Create a training set – all departments excluding Theatre (data collection)
2. Use Binary, Term Frequency or TFIDF representation (data preprocessing)
3. Select a relevant subset of attributes (attribute selection)
4. Use J48, IBk, and Naïve Bayes (classification)
5. Evaluate all models by cross validation (model evaluation)
6. Choose the best model and predict the class of Theatre (prediction)
7. Cluster the training set with K-means compare the cluster centroids with Theatre

Document classification and clustering

Teaching resources and student projects based on Weka



- Zdravko Markov and Daniel T. Larose, *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, Wiley 2007 (free excerpts: Chapter 1, TOC, Index)
- Lecture slides: dmw1.pdf, dmw2.pdf, dmw3.pdf, dmw4.pdf, dmw5.pdf
- Data sets: http://www.cs.ccsu.edu/~markov/dmwdata.zip
- Clustering.html

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

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