

ML Toolkits – LightSIDE and Weka

Sandeep Avula
asandeep@live.unc.edu

Outline

Toolkits: LightSIDE and Weka

Workflow

Tips to become comfortable with scikit

Learn how to print “Hello world!”

Learn how to write a “function” that can return the sum of two numbers. (This means you should also check what a main method is)

Learn how to open and read csv and json files.

Learn how to read each line and column.

Where can you do all of this? **YouTube or StackOverflow!**

LightSIDE

Download ZIP file from the course website.



Unpack and open the LightSIDE application.

Go to your terminal when you extracted the file and type
`./run.sh` (for mac)

LightSIDE

LightSIDE

Extract Features | Restructure Data | Build Models | Explore Results | Compare Models | Predict Labels

CSV Files:  

Class:

Type:

Text Fields:


Differentiate Text Fields



Feature Extractor Plugins:

- Basic Features
- Character N-Grams
- Column Features
- Parse Features
- Regular Expressions
- Stretchy Patterns

Configure Basic Features

- Unigrams
- Bigrams
- Trigrams
- POS Bigrams
- POS Trigrams
- Word/POS Pairs
- Line Length
- Count Occurences
- Normalize N-Gram Counts
- Include Punctuation
- Stem N-Grams

 Name: Rare Threshold:


Feature Table:  



Evaluations to Display: Target:

Basic Table Statistics

- Correlation
- F-Score
- Kappa
- Precision
- Recall
- Target Hits
- Total Hits

Features in Table: Search:



[Get Support](#)  Multithreaded 0.1 GB used, 4.0 GB max 

LightSIDE Workflow

Prepare the data

Extract the features

Build models

Make predictions

Error analysis

Extract the features

Load data

Extract Features Restructure Data Build Models Explore Results Compare Models Predict Labels

CSV Files:
sentiment_documents.c...
DOCUMENT_LIST
Documents: sentiment_documents
Class: class
Type: NOMINAL
Text Fields:
text
Differentiate Text Fields

Feature Extractor Plugins:
 Basic Features
 Character N-Grams
 Column Features
 Parse Features
 Regular Expressions
 Stretchy Patterns

Configure Basic Features
 Unigrams
 Bigrams
 Trigrams
 POS Bigrams
 POS Trigrams
 Word/POS Pairs
 Line Length
 Count Occurences
 Normalize N-Gram Counts
 Include Punctuation
 Stem N-Grams

Select extractor

Configuration options

Execute

Extract Name: 1grams_1 Rare Threshold: 5

Performance of features

Feature Table:
1grams
FEATURE_TABLE
Documents: sentiment_documents.cs
Feature Plugins: basic
Feature Table: 1grams
13444 features
Class: class
Type: nominal

Evaluations to Display:
Target: pos
Basic Table Statistics
 Correlation
 F-Score
 Kappa
 Precision
 Recall
 Target Hits
 Total Hits

Features in Table:

Feature	Correlation	F-Score	Kappa	Precision	Recall	Target Hits	Total Hits
frothy	0.0501	0.01	0.005	1	0.005	5	5
gattaca	0.0777	0.0237	0.012	1	0.012	12	12
gingerbread	0.0501	0.01	0.005	1	0.005	5	5
giorgio	0.0593	0.0139	0.007	1	0.007	7	7
goldwyn	0.0549	0.0119	0.006	1	0.006	6	6
governments	0.0549	0.0119	0.006	1	0.006	6	6
gretchen	0.0634	0.0159	0.008	1	0.008	8	8
griffiths	0.0501	0.01	0.005	1	0.005	5	5
guardians	0.0501	0.01	0.005	1	0.005	5	5

Feature Representation

Instance	class	abandone	able	about	above	absence	absolute	absolutely	absurd	accent
1	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	pos	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	neg	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	pos	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	neg	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
11	pos	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
12	neg	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
14	neg	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
15	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
17	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
19	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
20	neg	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
21	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
22	pos	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Model Building + Predictions

Select the model: For example, select Naive Bayes

Evaluation: Feed independent test data set or do n-fold cross validation.

Model Building + Predictions

The screenshot displays the LightSide software interface with several key components highlighted by blue boxes and labels:

- Feature Tables:** Shows a table named 'unigram' with 2038 features and a class of 'class' (nominal type).
- Learning Plugin:** A list of machine learning algorithms where 'Naive Bayes' is selected.
- Evaluation Options:** 'Cross-Validation' is selected, with 'Random' fold assignment and 'Auto' number of folds.
- Configure Naive Bayes:** Options for 'Use Kernel Estimator' and 'Use Supervised Discretization' are shown as unchecked.
- Train Button:** A button labeled 'Train' with a play icon, used to execute the model building process.
- Model Evaluation Metrics:** A table showing the performance of the trained model.
- Model Confusion Matrix:** A table showing the relationship between actual and predicted values.

Labels and arrows indicate the workflow: 'Execute' points to the 'Train' button, 'Model Selection' points to the 'Learning Plugin' list, 'Evaluation options' points to the 'Evaluation Options' section, and 'Check performance' points to the 'Model Evaluation Metrics' and 'Model Confusion Matrix' tables.

Model Evaluation Metrics:

Metric	Value
Accuracy	0.772
Kappa	0.544

Model Confusion Matrix:

Act \ Pred	neg	pos
neg	195	56
pos	58	191

Weka

Workflow

Prepare the data

Extract the features

Build models

Make predictions

Error analysis

Data

.arff format

```
@relation weather.symbolic
```

Relation name

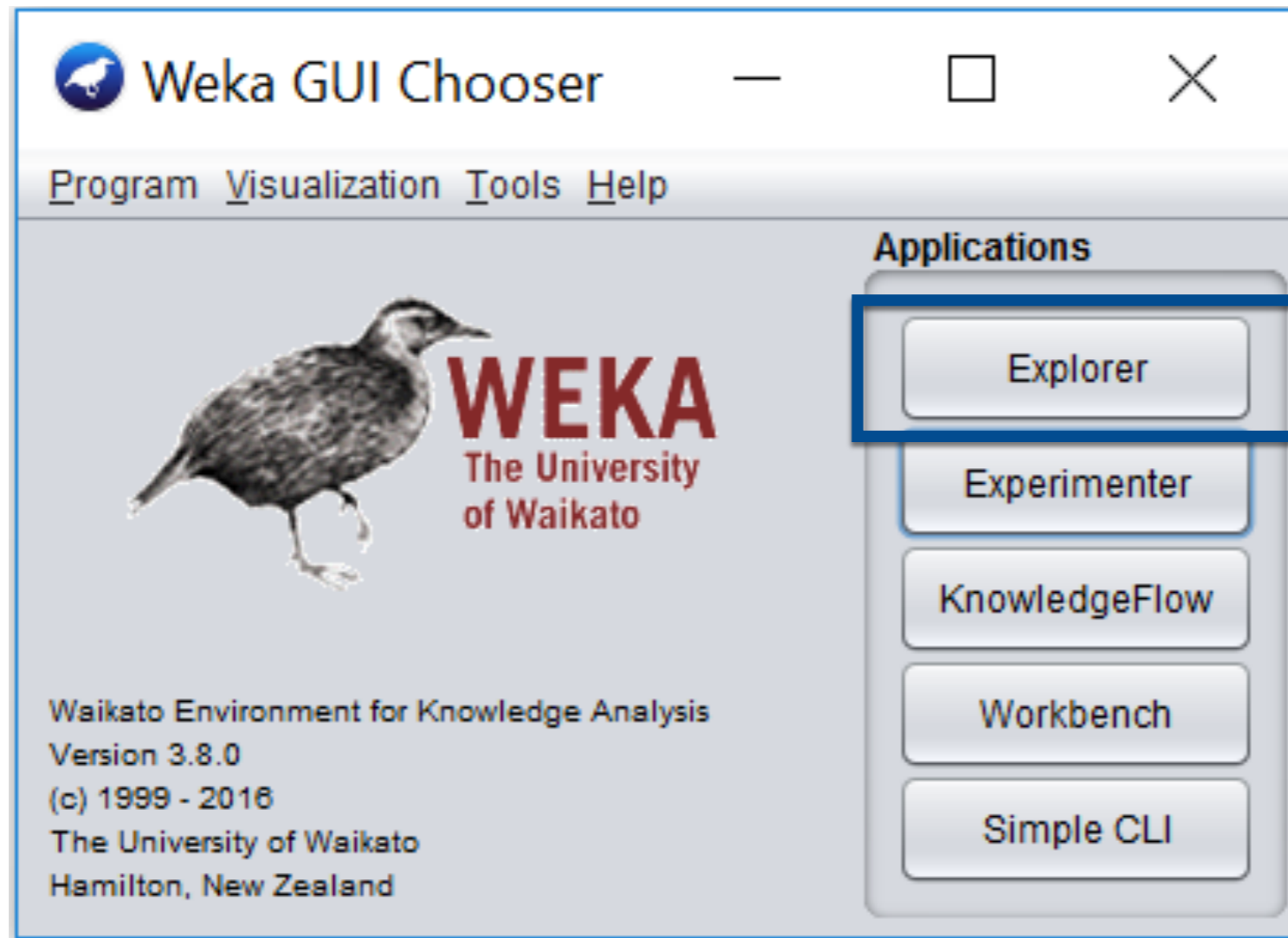
```
@attribute outlook {sunny, overcast, rainy}  
@attribute temperature {hot, mild, cool}  
@attribute humidity {high, normal}  
@attribute windy {TRUE, FALSE}  
@attribute play {yes, no}
```

Features or attributes

```
@data  
sunny,hot,high,FALSE,no  
sunny,hot,high,TRUE,no  
overcast,hot,high,FALSE,yes  
rainy,mild,high,FALSE,yes  
rainy,cool,normal,FALSE,yes  
rainy,cool,normal,TRUE,no  
overcast,cool,normal,TRUE,yes  
sunny,mild,high,FALSE,no  
sunny,cool,normal,FALSE,yes  
rainy,mild,normal,FALSE,yes  
sunny,mild,normal,TRUE,yes  
overcast,mild,high,TRUE,yes  
overcast,hot,normal,FALSE,yes  
rainy,mild,high,TRUE,no
```

Data

Weka Explorer



Select Explorer

Weka Explorer

Open .arff file from here

The screenshot shows the Weka Explorer interface with several key components highlighted by blue boxes and text annotations:

- Open file...:** A button in the top toolbar, highlighted by a blue box and the text "Open .arff file from here".
- Filter:** A section containing a "Choose" button and a dropdown menu currently set to "None", highlighted by a blue box and the text "Filtering options (eg: normalization)".
- Attributes:** A list of 14 attributes with checkboxes, highlighted by a blue box and the text "Feature selection". The attributes are:
 - 1 review_stars_z
 - 2 word_count_z
 - 3 lexical_diversity_z
 - 4 averaged_wordcount_lexicaldiversity_z
 - 5 correct_spell_ratio_z
 - 6 price_included_z
 - 7 pro/con_included_z
 - 8 stars_included_z
 - 9 price_pro_stars_average_z
 - 10 negative_fear_z
 - 11 sadness_z
 - 12 anxiety_z
 - 13 anger_z
 - 14 joy_z
- Selected attribute:** A panel showing statistics for the selected attribute "review_stars_z", highlighted by a blue box and the text "Feature distribution".

Statistic	Value
Minimum	-2.435
Maximum	1.39
Mean	-0.001
StdDev	1.027
- Class:** A dropdown menu set to "useful_class2 (Nom)", highlighted by a blue box and the text "Feature distribution".
- Visualize All:** A button to generate a visualization of the data.
- Bar Chart:** A stacked bar chart showing the distribution of the selected attribute across the class. The x-axis represents the attribute values, and the y-axis represents the count of instances. The bars are stacked with blue at the bottom and red on top.

Attribute Value	Blue Count	Red Count	Total Count
-2.44	0	50	50
-0.52	0	120	120
0	0	0	0
1.39	0	247	247
1.39	401	0	401
1.39	182	0	182

Weka Explorer

The screenshot shows the Weka Explorer interface with several key components highlighted by blue boxes and labels:

- Classifier:** A dropdown menu is set to "Logistic -R 1.0E-8 -M -1 -num-decimal-places 4". A blue box labeled "Pick a model" points to this dropdown.
- Test options:** A panel on the left with radio buttons for "Use training set" (selected), "Supplied test set", "Cross-validation" (with Folds: 10), and "Percentage split" (with %: 66). A blue box labeled "Evaluation options" points to this panel.
- Classifier output:** A large text area on the right displaying performance metrics and a confusion matrix. A blue box labeled "Results" points to this area.

Classifier output details:

Correctly Classified Instances	601	78.2552 %
Incorrectly Classified Instances	167	21.7448 %
Kappa statistic	0.4966	
Mean absolute error	0.3063	
Root mean squared error	0.3908	
Relative absolute error	67.3928 %	
Root relative squared error	81.9907 %	
Total Number of Instances	768	

Confusion Matrix:

		Actual Accuracy By Class ==					
(Nom) class		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
		0.890	0.418	0.799	0.890	0.842	0.504
		0.582	0.110	0.739	0.582	0.651	0.504
	Weighted Avg.	0.783	0.310	0.778	0.783	0.775	0.504

Confusion Matrix Legend:

```
=== Confusion Matrix ===
  a  b  <-- classified as
445 55 | a = tested_negative
112 156 | b = tested_positive
```