

ML Toolkits - I

Sandeep Avula
asandeep@live.unc.edu

Outline

Toolkits: LightSIDE and Weka

Installing toolkits

Workflow

Before we start Scikit

Install Python 3

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!**

Do you absolutely need to know this for this course? **No.**

LightSIDE

Download ZIP file from the course website.

Unpack and open the LightSIDE application.

LightSIDE

LightSide

Extract FeaturesRestructure DataBuild ModelsExplore ResultsCompare ModelsPredict Labels

CSV Files:

Class:

Type:

NOMINAL

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

Extract

Name:

1grams

Rare Threshold:

5

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

Multithreaded0.1 GB used, 4.0 GB max

LightSIDE Workflow

Prepare the data

Extract the features

Build models

Make predictions

Error analysis

Prepare the data

Csv file where one column is the text and the other is the label.

	class	text							
1	neg	plot : two teen couples go to a church party , drink and then drive . they get into an accident . one of							
2	neg	the happy bastard's quick movie review damn that y2k bug . it's got a head start in this movie starring							
3	neg	it is movies like these that make a jaded movie viewer thankful for the invention of the timex indiglo w							
4	neg	quest for camelot is warner bros . ' first feature-length , fully-animated attempt to steal clout from d							
5	neg	synopsis : a mentally unstable man undergoing psychotherapy saves a boy from a potentially fatal acci							
6	neg	capsule : in 2176 on the planet mars police taking into custody an accused murderer face the title mer							
7	neg	so ask yourself what 8mm (eight millimeter) is really all about . is it about a wholesome surveillanc							
8	neg	that's exactly how long the movie felt to me . there weren't even nine laughs in nine months . it's a te							
9	neg	call it a road trip for the walking wounded . stellan skarsg ? rd plays such a convincingly zombified dru							
10	neg	plot : a young french boy sees his parents killed before his eyes by tim roth , oops . . . i mean , an evil							
11	neg	best remembered for his understated performance as dr . hannibal lecter in michael mann's forensics							
12	neg	janeane garofalo in a romantic comedy -- it was a good idea a couple years ago with the truth about c							
13	neg	and now the high-flying hong kong style of filmmaking has made its way down to the classics , and it is							
14	neg	a movie like mortal kombat : annihilation works (and must be reviewed on) multiple levels . first , the							
15	neg	she was the femme in la femme nikita . he was the baldwin in backdraft , sliver , and fair game (v							
16	neg	john carpenter makes b-movies . always has (halloween , escape from new york , the thing) and ,							
17	neg	i'm really starting to wonder about alicia silverstone . sure , she is one of the most beautiful creatures							
18	neg	so what do you get when you mix together plot elements from various successful sci-fi films such as cl							
19	neg								

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

Execute

Extract

Name: 1grams_1

Rare Threshold: 5

Feature Extractor Plugins:

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

Select extractor

Configure Basic Features

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

Configuration options

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:

Search:

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	abandon	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, which is used for building and evaluating machine learning models. The interface is divided into several sections:

- Feature Tables:** A list of feature tables, including 'unigram'. The 'unigram' table is selected, showing details like 'Documents: train.csv', 'Feature Plugins: basic', 'Feature Table: unigram', '2038 features', 'Class: class', and 'Type: nominal'.
- Learning Plugin:** A section for selecting the learning plugin. The 'Naive Bayes' plugin is selected, with other options like 'Logistic Regression', 'Linear Regression', 'Support Vector Machines', 'Decision Trees', and 'Weka (All)' available.
- Evaluation Options:** A section for configuring evaluation options. 'Cross-Validation' is selected, and 'Random' is chosen for 'Fold Assignment'. The 'Number of Folds' is set to 'Auto'.
- Train Button:** A large blue button labeled 'Train' is located at the bottom left, with an arrow pointing to it from the 'Execute' label.
- Model Evaluation Metrics:** A table showing the performance of the trained model. The metrics are Accuracy (0.772) and Kappa (0.544).
- Model Confusion Matrix:** A table showing the confusion matrix for the trained model. The matrix is as follows:

Act \ Pred	neg	pos
neg	195	56
pos	58	191

The interface also includes a 'Trained Models' section at the bottom left, showing the list of trained models, and a 'Get Support' link at the bottom left. The status bar at the bottom right indicates 'Multithreaded' and '0.1 GB used, 4.0 GB max'.

Model Selection

Evaluation options

Execute

Check performance

Weka

Workflow

Prepare the data

Extract the features

Build models

Make predictions

Error analysis

Data

.arff format

```
@relation weather.symbolic
```

Relation name

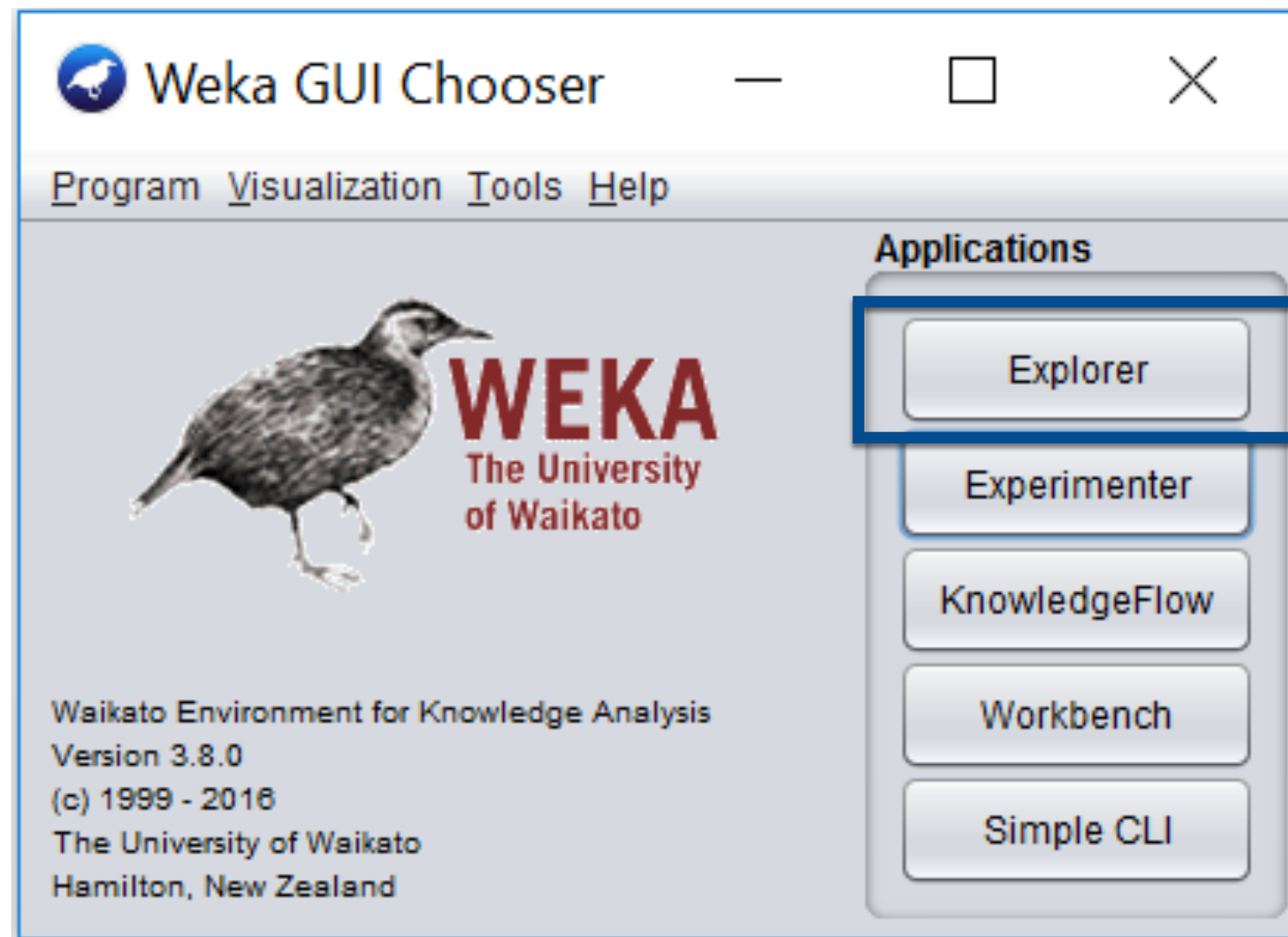
```
@attribute outlook {sunny, overcast, rainy}  
@attribute temperature {hot, mild, cool}  
@attribute humidity {high, normal}  
@attribute windy {TRUE, FALSE}  
@attribute play {yes, no}
```

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

Weka Explorer interface showing the 'Preprocess' tab.

Buttons: Preprocess, Classify, Cluster, Associate, Select attributes, Visualize

Open file... (highlighted)

Filter: Choose, None (highlighted)

Filtering options (eg: normalization)

Current relation: Relation: yelp_usefulness_training, Instances: 1000, Attributes: 56, Sum of weights: 1000

Attributes: All, None, Invert, Pattern

Feature selection

No.	Name
1	<input checked="" type="checkbox"/> review_stars_z
2	<input type="checkbox"/> word_count_z
3	<input type="checkbox"/> lexical_diversity_z
4	<input type="checkbox"/> averaged_wordcount_lexicaldiversity_z
5	<input type="checkbox"/> correct_spell_ratio_z
6	<input type="checkbox"/> price_included_z
7	<input type="checkbox"/> pro/con_included_z
8	<input type="checkbox"/> stars_included_z
9	<input type="checkbox"/> price_pro_stars_average_z
10	<input type="checkbox"/> negative_fear_z
11	<input type="checkbox"/> sadness_z
12	<input type="checkbox"/> anxiety_z
13	<input type="checkbox"/> anger_z
14	<input type="checkbox"/> joy_z

Selected attribute: Name: review_stars_z, Type: Numeric, Missing: 0 (0%), Distinct: 5, Unique: 0 (0%)

Statistic	Value
Minimum	-2.435
Maximum	1.39
Mean	-0.001
StdDev	1.027

Feature distribution

Class: useful_class2 (Nom)

Visualize All

Log x 0

Weka Explorer

The screenshot shows the Weka Explorer application window. At the top, there are tabs for Preprocess, Classify, Cluster, Associate, Select attributes, and Visualize. The 'Classify' tab is active. Below the tabs, the 'Classifier' section shows a 'Choose' button and a text field containing 'Logistic -R 1.0E-8 -M -1 -num-decimal-places 4'. A blue box labeled 'Pick a model' points to this text field. To the left, the 'Test options' section has four radio buttons: 'Use training set' (selected), 'Supplied test set' (with a 'Set...' button), 'Cross-validation' (with 'Folds' set to 10), and 'Percentage split' (with '%' set to 66). A 'More options...' button is below these. A blue box labeled 'Evaluation options' points to this section. Below the test options are 'Start' and 'Stop' buttons. The 'Result list (right-click for options)' section shows a single entry: '11:01:54 - functions.Logistic'. The 'Classifier output' section displays various performance metrics and a confusion matrix. A blue box labeled 'Results' points to this section. At the bottom, there is a 'Status' bar with 'OK' and 'Log' buttons, and a small icon with 'x 0'.

Classifier

Choose **Logistic -R 1.0E-8 -M -1 -num-decimal-places 4**

Test options

- ☒ Use training set
- ☐ Supplied test set **Set...**
- ☐ Cross-validation Folds **10**
- ☐ Percentage split % **66**

More options...

Classifier output

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

Evaluation options

(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

Results

=== Confusion Matrix ===

a	b	← classified as
445	55	a = tested_negative
112	156	b = tested_positive

Status

OK Log  x 0