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!

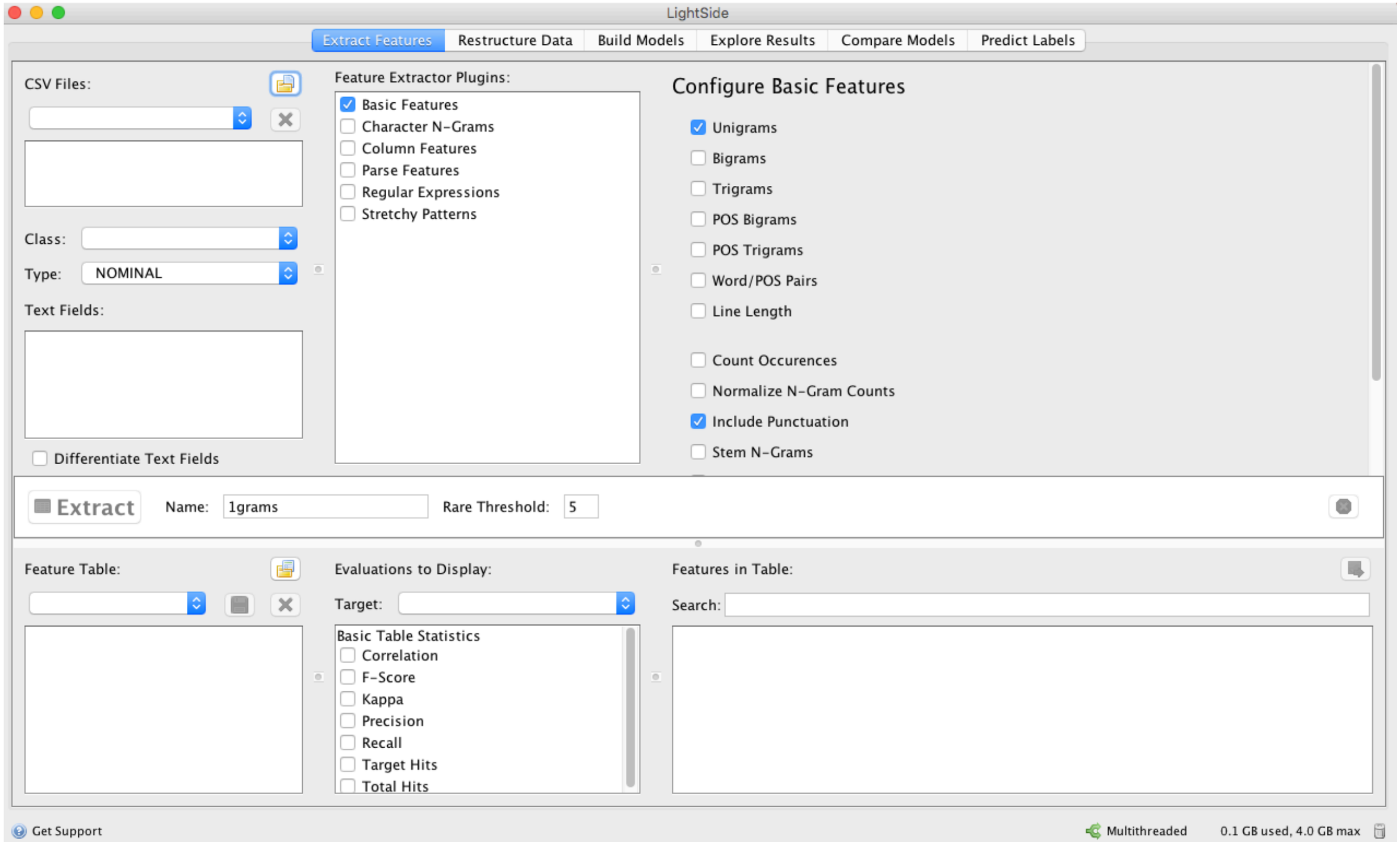
Do you absolutely need to know this for this course? No.

LightSIDE

Download ZIP file from the course website.

Unpack and open the LightSIDE application.

LightSIDE


LightSide
Extract Features
Restructure Data
Build Models
Explore Results
Compare Models
Predict 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
Multithreaded
0.1 GB used, 4.0 GB max

LightSIDE Workflow

Prepare the data

Extract the features

Build models

Make predictions

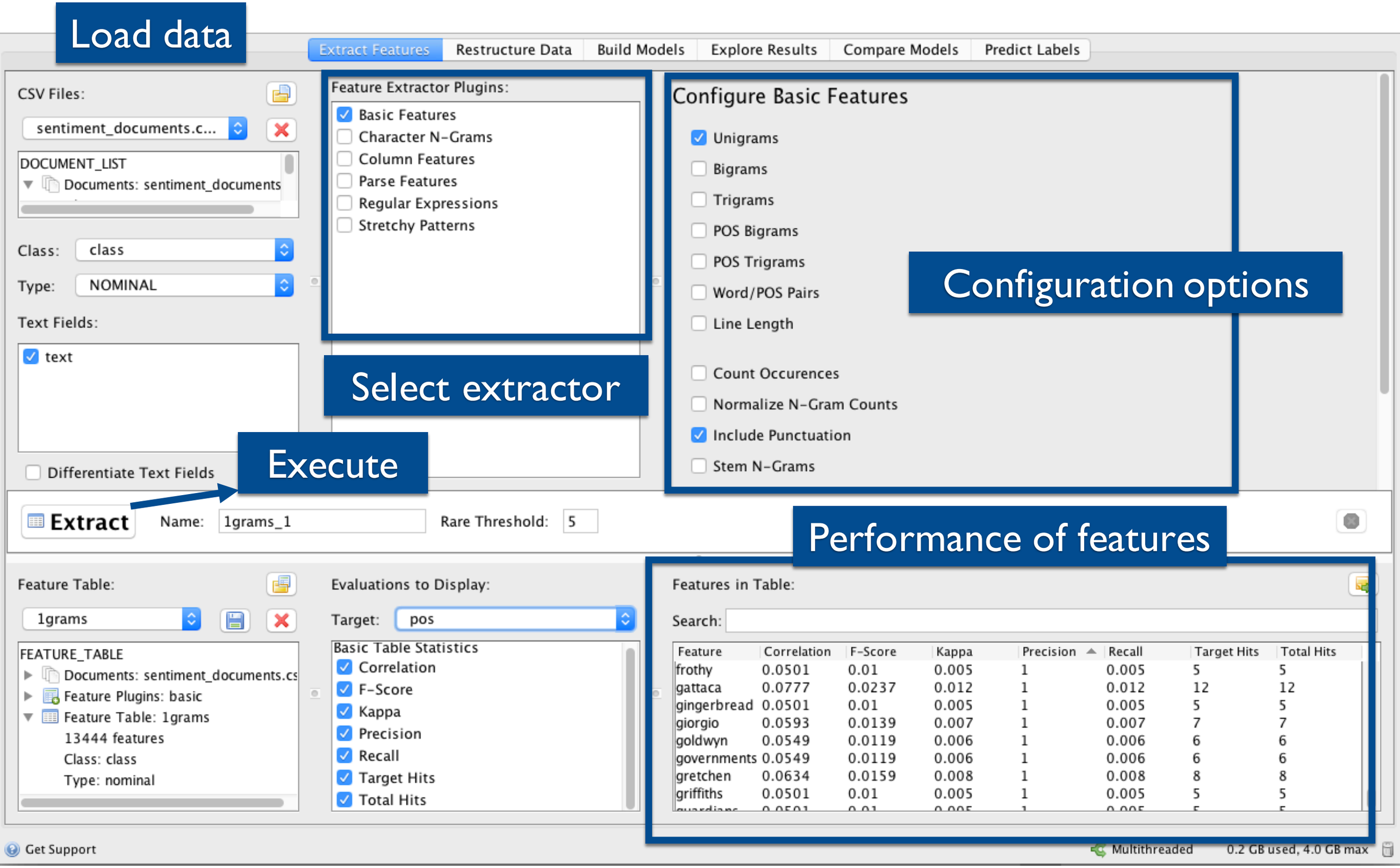
Error analysis

Prepare the data

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

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

Extract the features



Feature Representation

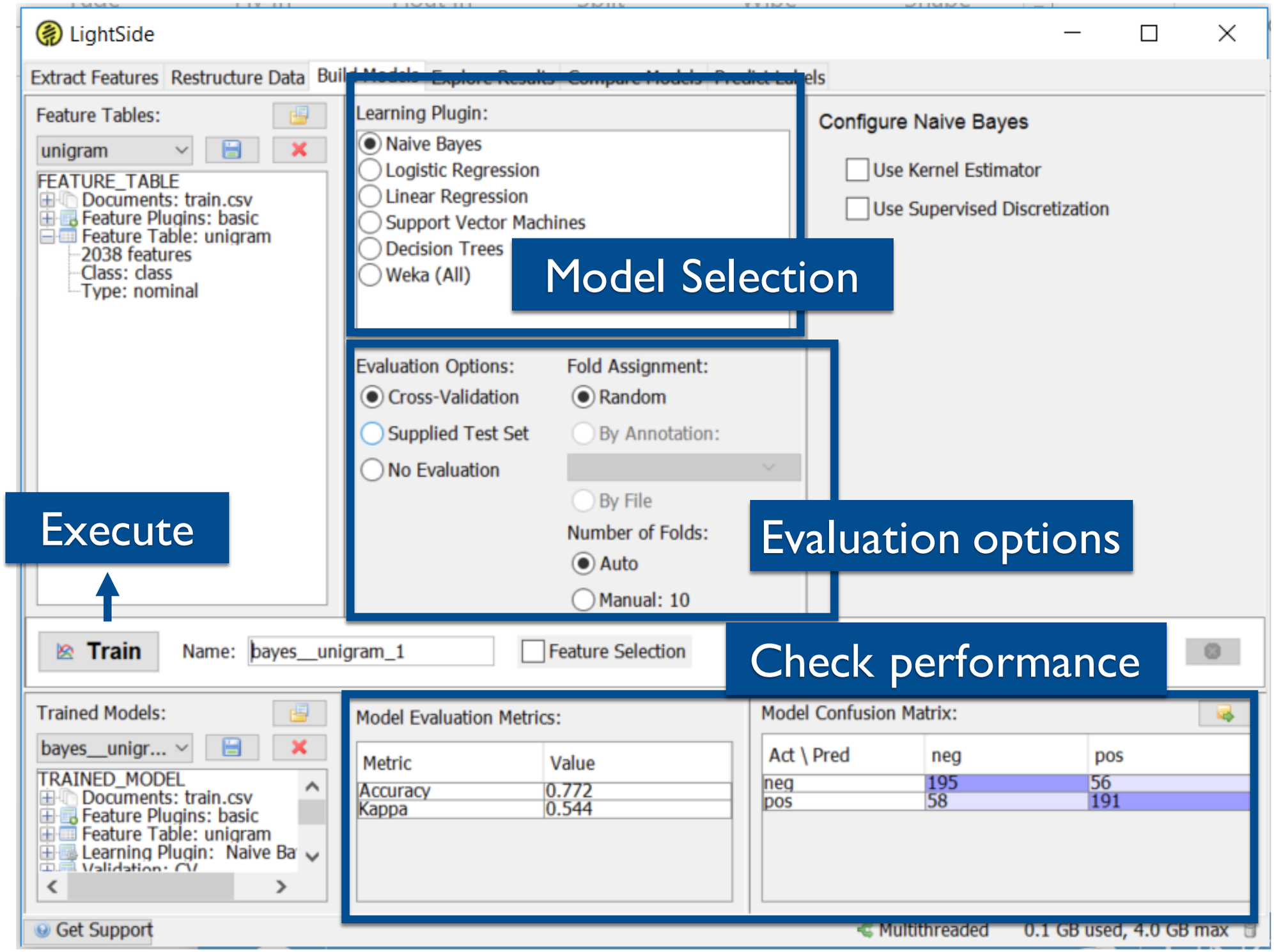
Instance	class	abandone	able	about	above	absence	absolute	absolutely	absurd	accent
1	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	pos	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	neg	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	pos	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	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



Weka

Workflow

Prepare the data

Extract the features

Build models

Make predictions

Error analysis

Data

.arff format

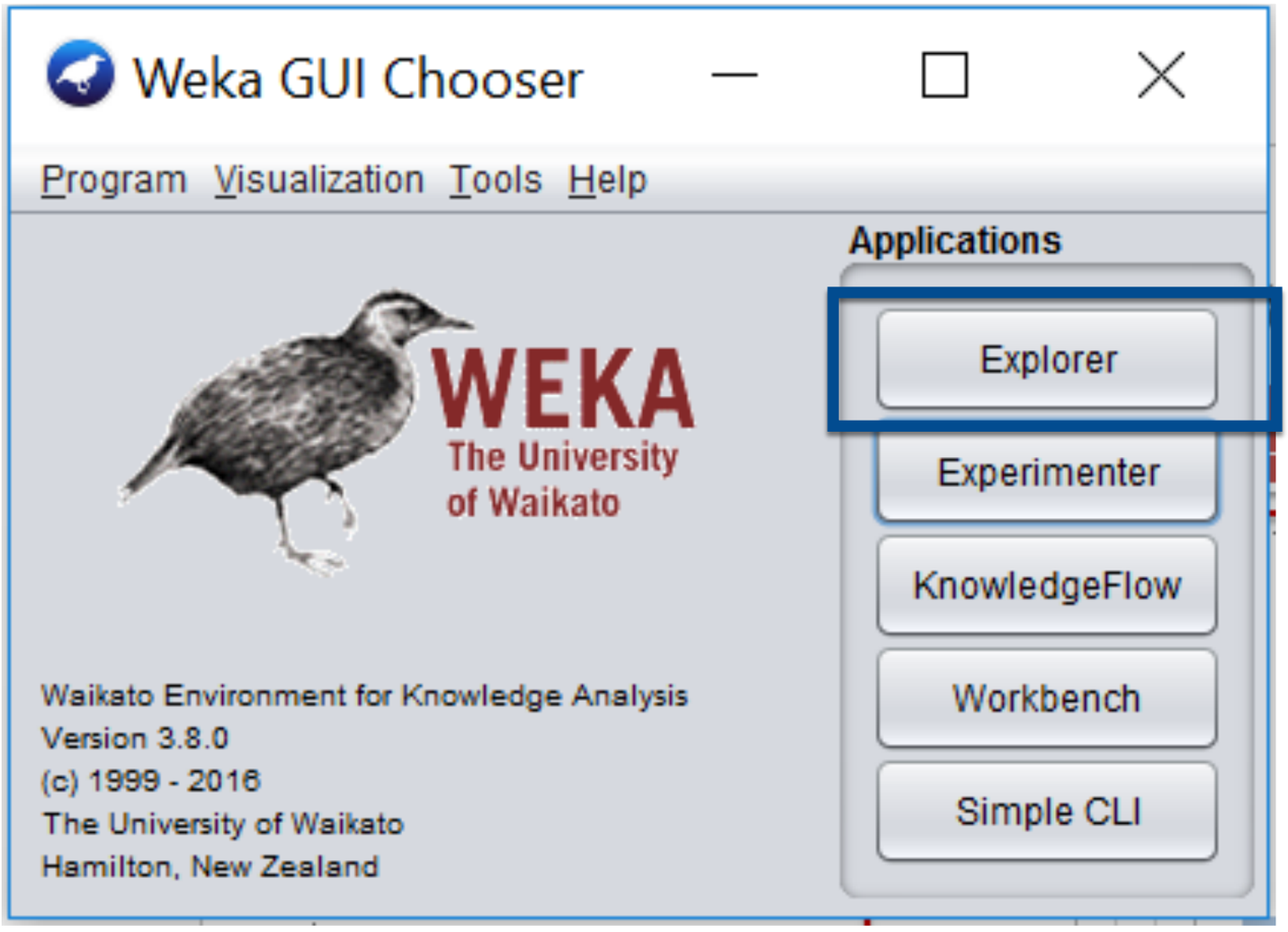
```
@relation weather.symbolic

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

@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
```

Relation name

Features or attributes

Data

Weka Explorer



Select Explorer

Weka Explorer

Open .arff file from here

Filtering options (eg: normalization)

Feature selection

Feature distribution

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

Choose | None | Apply

Current relation

Relation: yelp_usefulness_training — Attributes: 56
Instances: 1000 — Sum of weights: 1000

Attributes

All | None | Invert | Pattern

No.	Name
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

Remove

Status

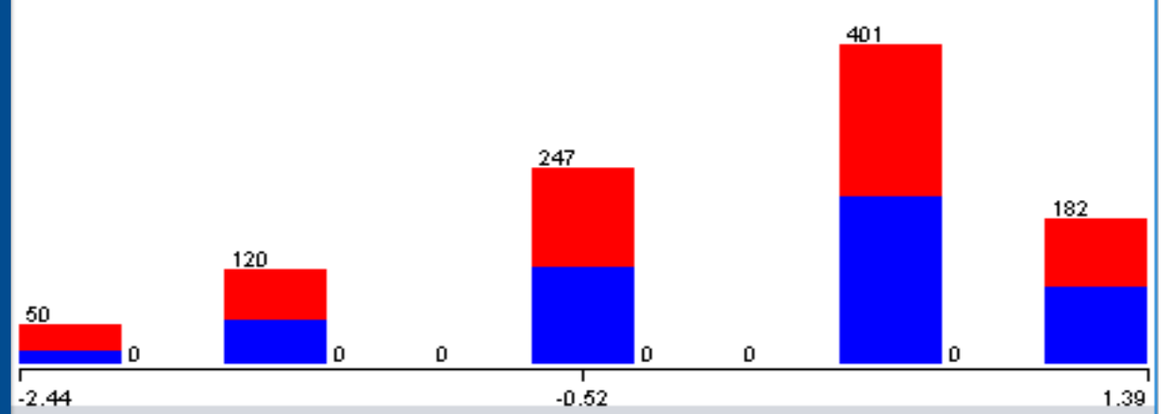
OK

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

Class: useful_class2 (Nom) | Visualize All



Log | x 0

Weka Explorer