Text Analytics 95-865 TA: Bob Fang

Homework 2 & 3 WEKA Automation Workflow Guide

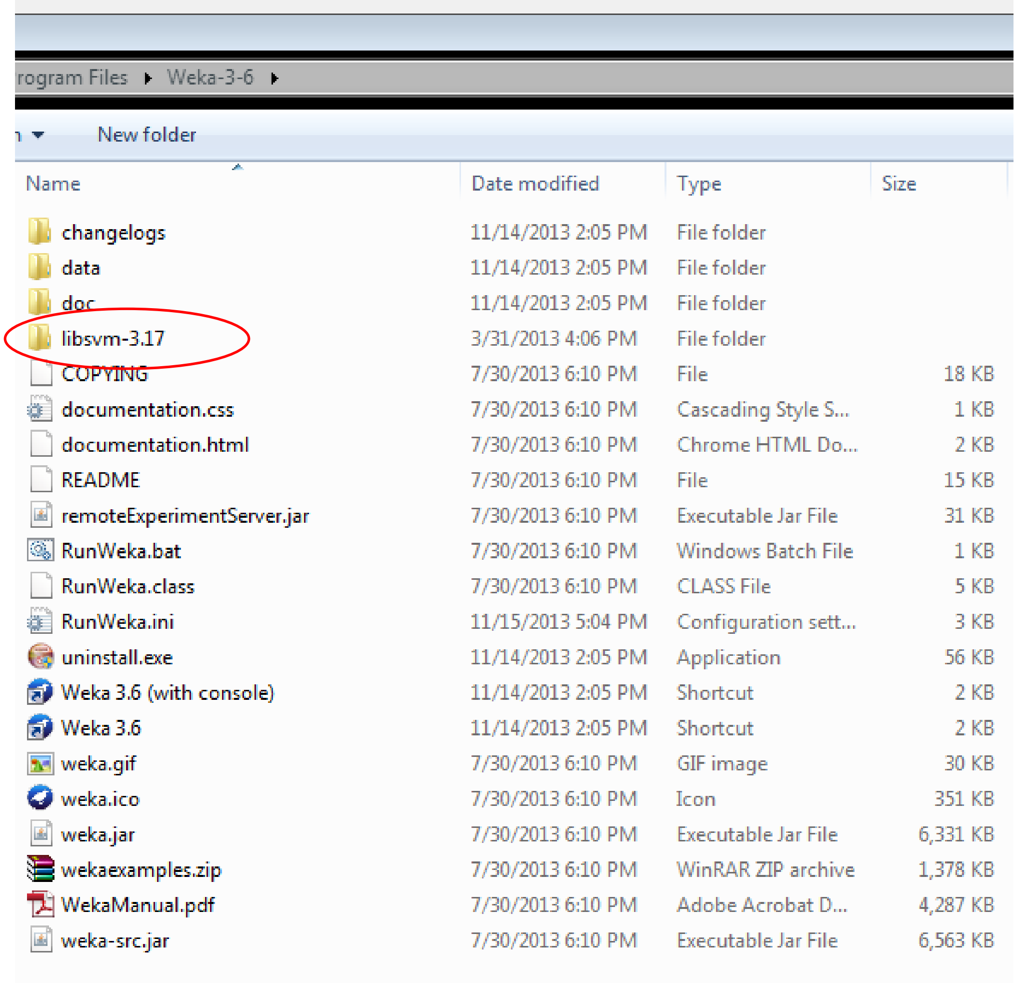
For homework 2 and 3, you will be running different machine learning algorithms against the datasets that will be provided to you.

Tools

WEKA is a useful machine learning work bench that allows you to quickly run various machine learning algorithms against your datasets. We use this because you do not need to implement (program) the various algorithms for your homework assignments. For this class, we will use the latest version of WEKA available here: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

LightSIDE is another machine learning program and we will use it to generate ARFF files which WEKA will read. This is the place where you will be able to determine which features you are going to pass into WEKA for the model building process. We are going to be using an older version of lightSIDE for the purposes of these 2 homework assignments for consistency. <lightsideLink>

libSVM is a support vector machine library add-on for Weka. We will be using it for this class. We will bundle it with DNDW but you have to **install it manually by copying it into your Program Files/Weka-3-6/** folder.



Name	Date modified	Type	Size
changelogs	11/14/2013 2:05 PM	File folder	
data	11/14/2013 2:05 PM	File folder	
doc	11/14/2013 2:05 PM	File folder	
libsvm-3.17	3/31/2013 4:06 PM	File folder	
COPYING	7/30/2013 6:10 PM	File	18 KB
documentation.css	7/30/2013 6:10 PM	Cascading Style S...	1 KB
documentation.html	7/30/2013 6:10 PM	Chrome HTML Do...	2 KB
README	7/30/2013 6:10 PM	File	15 KB
remoteExperimentServer.jar	7/30/2013 6:10 PM	Executable Jar File	31 KB
RunWeka.bat	7/30/2013 6:10 PM	Windows Batch File	1 KB
RunWeka.class	7/30/2013 6:10 PM	CLASS File	5 KB
RunWeka.ini	11/15/2013 5:04 PM	Configuration sett...	3 KB
uninstall.exe	11/14/2013 2:05 PM	Application	56 KB
Weka 3.6 (with console)	11/14/2013 2:05 PM	Shortcut	2 KB
Weka 3.6	11/14/2013 2:05 PM	Shortcut	2 KB
weka.gif	7/30/2013 6:10 PM	GIF image	30 KB
weka.ico	7/30/2013 6:10 PM	Icon	351 KB
weka.jar	7/30/2013 6:10 PM	Executable Jar File	6,331 KB
wekaexamples.zip	7/30/2013 6:10 PM	WinRAR ZIP archive	1,378 KB
WekaManual.pdf	7/30/2013 6:10 PM	Adobe Acrobat D...	4,287 KB
weka-src.jar	7/30/2013 6:10 PM	Executable Jar File	6,563 KB

Next, edit your RunWeka.ini file and make the following modification:

cp=%CLASSPATH%;C:/Program Files/Weka-3-6/libsvm-3.17/java/libsvm.jar

This will add the libsvm libraries to the Classpath used by Weka.

You may also take this time to change this parameter:

maxheap=2048M

to something like m*axheap=4096M*

This doubles the memory that Weka can use and will speed things up. Only do this if you have more than 4GB of RAM to devote to running Weka.

DNDW (Drag & Drop Weka) is a set of windows batch scripts that automate the laborious process of running Weka experiments and collecting experiment results. These were written by a former student of the class (now current TA) to help save time on homework assignments 2 & 3. Due to the lack of time resources and hardware (TA doesn't have a mac), we are unable to offer a native mac version of the scripts. If you have a mac and would like to port the scripts to run effectively on a mac, please email the TA. Otherwise, this document will offer suggestions on what to do if you have a mac.

Goals

The goal of the homework is to get you thinking about feature selection and identify generally what features work well with certain kinds of algorithms. We also want you to be exposed to different datasets and demonstrate that there is no "winning" algorithm that works on all datasets.

Real world tasks

Keep in mind that there was a lot of pre-processing done on the datasets you are receiving so that they are in formats that can be easily accessed by LightSIDE and WEKA. A real-world problem would involve much more of that labor intensive work and you would probably have to write your own implementations of the various machine learning algorithms as well. However, in this class we have relieved you of those duties ☺ so that you can focus on learning the concepts behind Text Analytics.

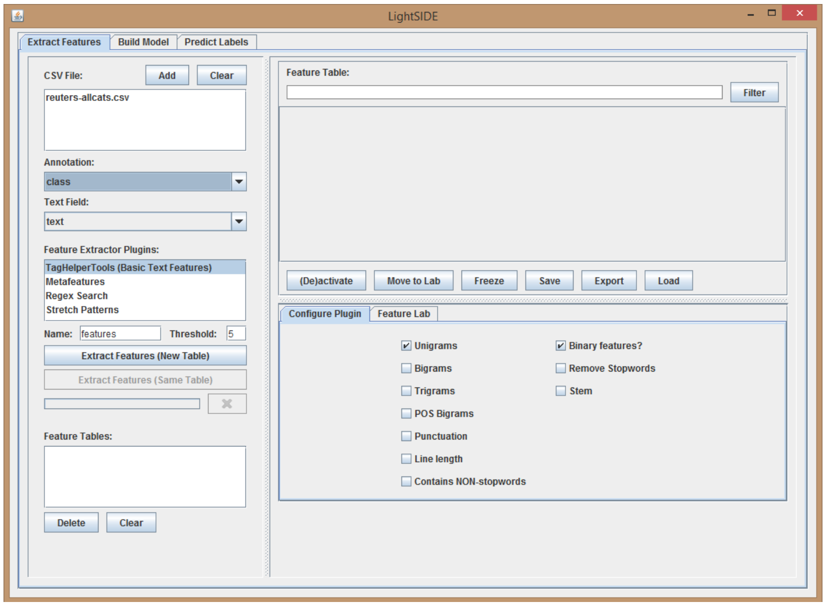
Text Analytics 95-865 TA: Bob Fang
Homework 2 & 3 WEKA Automation Workflow Guide

PC & Mac Users

1. Download & install WEKA (PC & Mac)
2. Download & install older version of LightSide (extract it anywhere you want) (PC & Mac)
3. Download & extract DNDW (put DNDW anywhere you want) (PC Only)
4. Copy libSVM folder into Program Files/Weka-3-6/ (PC Only) This means the path Program Files/Weka-3-6/libsvm-3.17/java/libsvm.jar is valid.

1 should be rather trivial. For 2 & 3, you can put it anywhere as the batch scripts will handle the file paths.

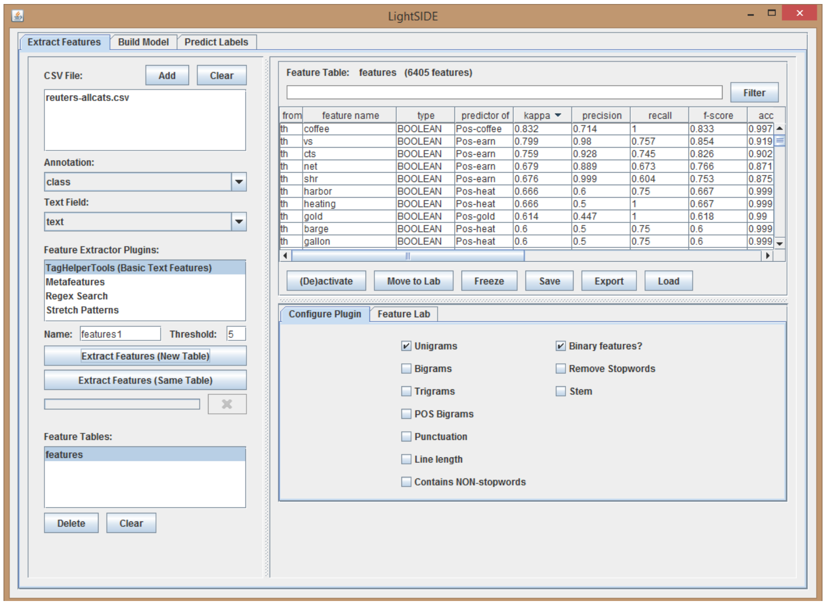
5. Run LightSide and add a CSV file (reuters-allcats.csv in this example)



CSV File: reuters-allcats.csv
Annotation: class
Text Field: text
Name: features Threshold: 5
Feature Table:

☑ Unigrams ☑ Binary features?
□ Bigrams □ Remove Stopwords
□ Trigrams □ Stem
□ POS Bigrams
□ Punctuation
□ Line length
□ Contains NON-stopwords

6. A bunch of features pop up according to the threshold and features (Unigrams, Binary) we selected.



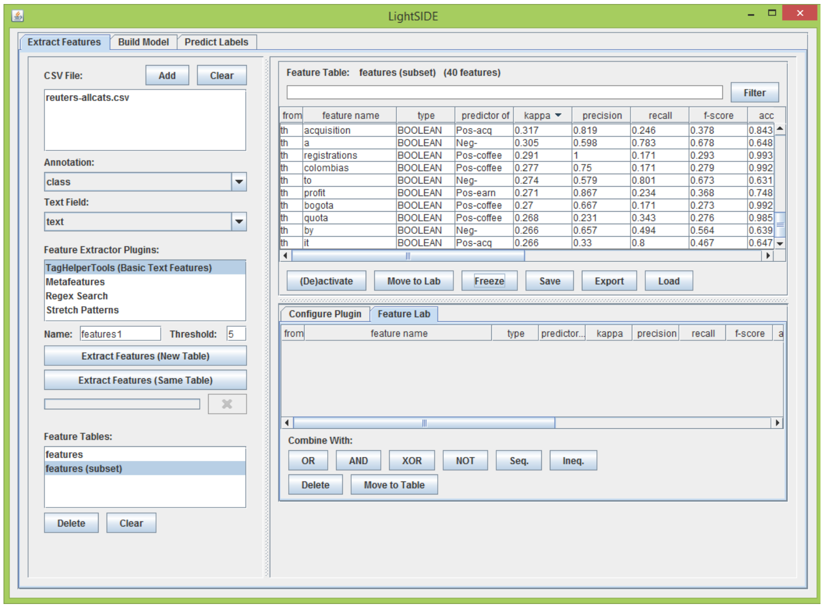
CSV File: reuters-allcats.csv
Annotation: class
Text Field: text
Name: features1 Threshold: 5
Feature Tables: features

Feature Table: features (6405 features)

from	feature name	type	predictor of	kappa	precision	recall	f-score	acc
th	coffee	BOOLEAN	Pos-coffee	0.832	0.714	1	0.833	0.997
th	vs	BOOLEAN	Pos-earn	0.799	0.98	0.757	0.854	0.919
th	cts	BOOLEAN	Pos-earn	0.759	0.928	0.745	0.826	0.902
th	net	BOOLEAN	Pos-earn	0.679	0.889	0.673	0.766	0.871
th	shr	BOOLEAN	Pos-earn	0.676	0.999	0.604	0.753	0.875
th	harbor	BOOLEAN	Pos-heat	0.666	0.6	0.75	0.667	0.999
th	heating	BOOLEAN	Pos-heat	0.666	0.5	1	0.667	0.999
th	gold	BOOLEAN	Pos-gold	0.614	0.447	1	0.618	0.99
th	barge	BOOLEAN	Pos-heat	0.6	0.5	0.75	0.6	0.999
th	gallon	BOOLEAN	Pos-heat	0.6	0.5	0.75	0.6	0.999

☑ Unigrams ☑ Binary features?
□ Bigrams □ Remove Stopwords
□ Trigrams □ Stem
□ POS Bigrams
□ Punctuation
□ Line length
□ Contains NON-stopwords

7. For the homework a typical task is to get top 40 counts by kappa. The lightSide interface is a little clumsy for doing that. Notice that LightSide sorts kappa for you in descending order. In order to get top 40, we have to skip the first 40 features, select the 41st features, scroll till the last feature, HOLD DOWN SHIFT KEY & click on the last feature – highlighting 41st to last features, click on deactivate, and then click on freeze.
8. If all goes well in step 6, we get this:



CSV File: reuters-allcats.csv

Annotation: class

Text Field: text

Feature Extractor Plugins:
TagHelperTools (Basic Text Features)
Metafeatures
Regex Search
Stretch Patterns

Name: features1 Threshold: 5

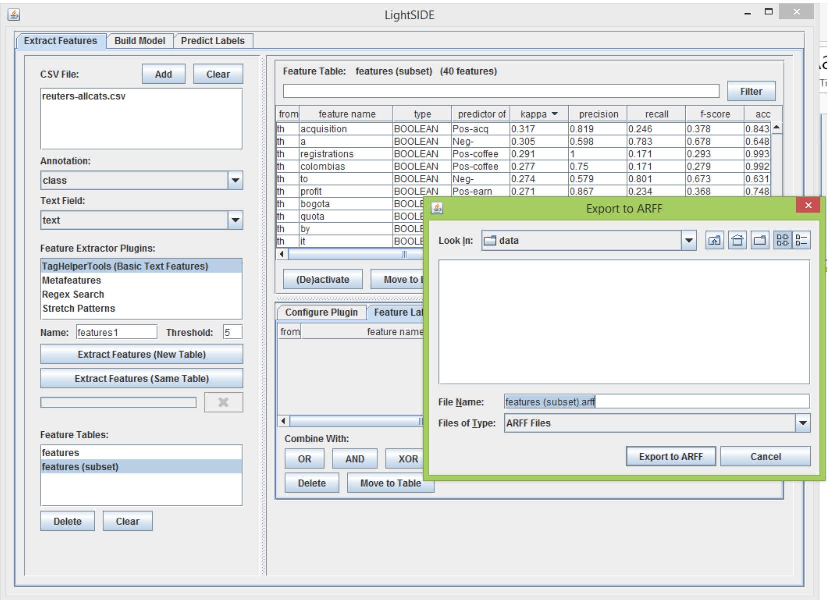
Feature Tables:
features
features (subset)

Feature Table: features (subset) (40 features)

from	feature name	type	predictor of	kappa	precision	recall	f-score	acc
th	acquisition	BOOLEAN	Pos-acq	0.317	0.819	0.246	0.378	0.843
th	a	BOOLEAN	Neg-	0.305	0.598	0.783	0.678	0.648
th	registrations	BOOLEAN	Pos-coffee	0.291	1	0.171	0.293	0.993
th	colombias	BOOLEAN	Pos-coffee	0.277	0.75	0.171	0.279	0.992
th	to	BOOLEAN	Neg-	0.274	0.579	0.801	0.673	0.631
th	profit	BOOLEAN	Pos-earn	0.271	0.867	0.234	0.368	0.748
th	bogota	BOOLEAN	Pos-coffee	0.27	0.667	0.171	0.273	0.992
th	quota	BOOLEAN	Pos-coffee	0.268	0.231	0.343	0.276	0.985
th	by	BOOLEAN	Neg-	0.266	0.657	0.494	0.564	0.639
th	it	BOOLEAN	Pos-acq	0.266	0.33	0.8	0.467	0.647

from	feature name	type	predictor...	kappa	precision	recall	f-score	a

Notice that the 40 features have been placed in the feature tables (features (subsets)). The counter will also reflect the number 40. If you are happy at this point, you can proceed to click on "Export" and save the file in the ARFF format.



CSV File: reuters-allcats.csv

Annotation: class

Text Field: text

Feature Extractor Plugins:
TagHelperTools (Basic Text Features)
Metafeatures
Regex Search
Stretch Patterns

Name: features1 Threshold: 5

Feature Tables:
features
features (subset)

Feature Table: features (subset) (40 features)

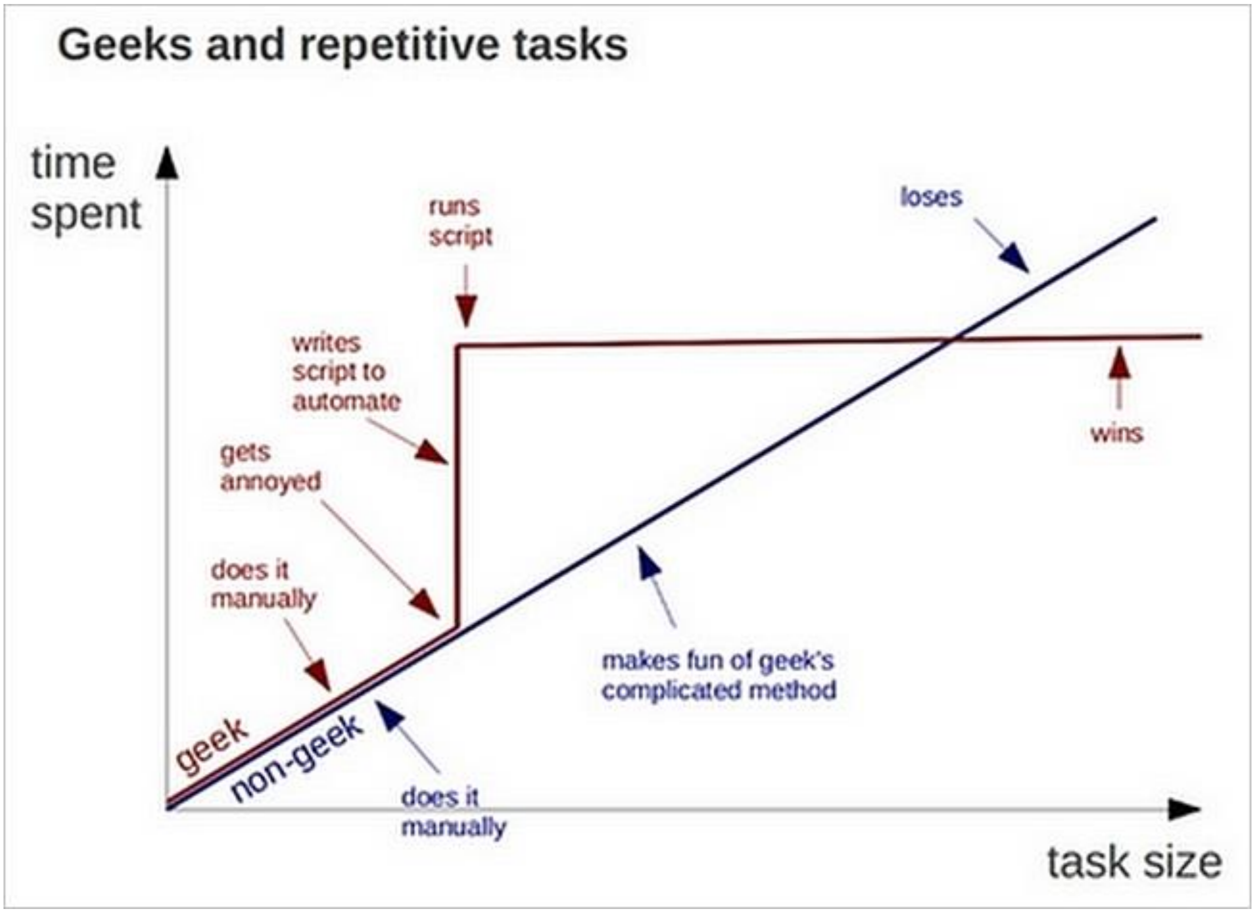
from	feature name	type	predictor of	kappa	precision	recall	f-score	acc
th	acquisition	BOOLEAN	Pos-acq	0.317	0.819	0.246	0.378	0.843
th	a	BOOLEAN	Neg-	0.305	0.598	0.783	0.678	0.648
th	registrations	BOOLEAN	Pos-coffee	0.291	1	0.171	0.293	0.993
th	colombias	BOOLEAN	Pos-coffee	0.277	0.75	0.171	0.279	0.992
th	to	BOOLEAN	Neg-	0.274	0.579	0.801	0.673	0.631
th	profit	BOOLEAN	Pos-earn	0.271	0.867	0.234	0.368	0.748
th	bogota	BOOLE						
th	quota	BOOLE						
th	by	BOOLE						
th	it	BOOLE						

Look In: data

File Name: features (subset).arff

Files of Type: ARFF Files

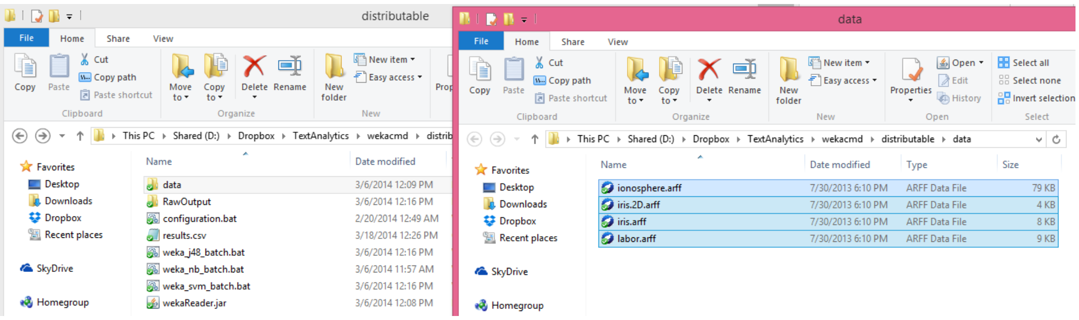
This is probably the most labor-intensive part of the homework because of the manual interaction required with LightSIDE. If you are inclined, you may use something like "Macro Express" to record your mouse & keyboard movements and play them back repeatedly on the different experiments. One key thing to remember when doing that is to have a consistent window size when recording and playing back the macros. Another property of the lightSIDE GUI is that the number of rows shown to you on the features table is controlled by your window size. I won't be responsible for supporting your macros, but if you are familiar with them you should use it since the homework will have you do these tasks a fair number of times.



9. WEKA and DNDW
   Once you have installed Weka and DNDW, you can start processing the ARFF files with DNDW. DNDW will be recording your results in results.csv. You will be required to turn results.csv in along with your homework.

   Mac users will have to install Weka and DNDW on the Heinz virtual lab and process them there. Keep in mind that when using the Heinz virtual lab, your files will be lost after you disconnect from the machine. To avoid any issues relating to that, please put your files on a thumbdrive and connect it to the virtual lab machine.

10. DNDW – Drag and Drop Weka



This PC › Shared (D:) › Dropbox › TextAnalytics › wekacmd › distrib

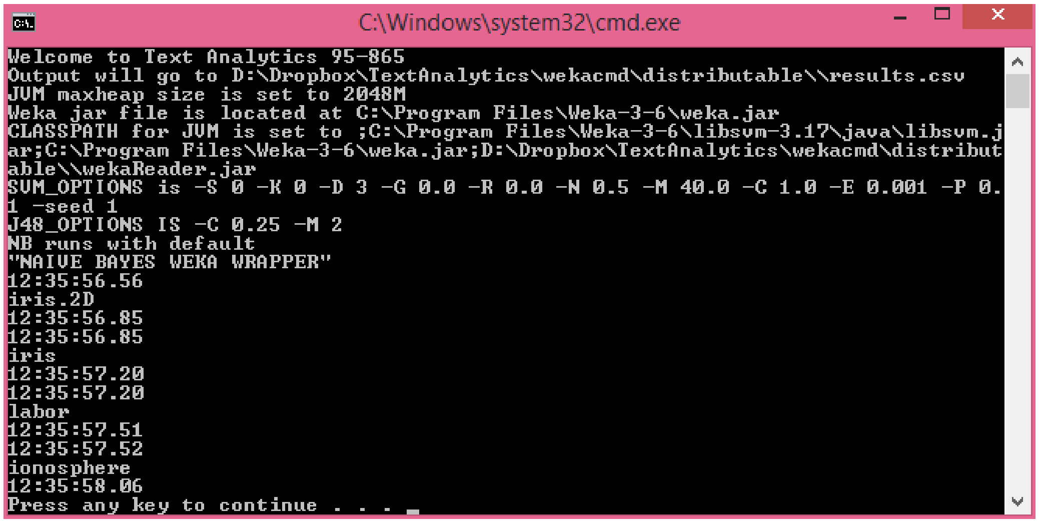
Name	Date modified
data	3/6/2014 12:09 PM
RawOutput	3/6/2014 12:16 PM
configuration.bat	2/20/2014 12:49 AM
results.csv	3/18/2014 12:26 PM
weka_j48_batch.bat	3/6/2014 12:16 PM
weka_nb_batch.bat	3/6/2014 11:57 AM
weka_svm_batch.bat	3/6/2014 12:16 PM
wekaReader.jar	3/6/2014 12:08 PM

This PC › Shared (D:) › Dropbox › TextAnalytics › wekacmd › distributable › data

Name	Date modified	Type	Size
ionosphere.arff	7/30/2013 6:10 PM	ARFF Data File	79 KB
iris.2D.arff	7/30/2013 6:10 PM	ARFF Data File	4 KB
iris.arff	7/30/2013 6:10 PM	ARFF Data File	8 KB
labor.arff	7/30/2013 6:10 PM	ARFF Data File	9 KB

Select your ARFF files, proceed to drag and drop them on the batch file that contains the algorithm you want to run.

For example, if we want to process all these ARFF files with the Naïve Bayes algorithm, you drop them on to the weka_nb_batch.bat file:



```
Welcome to Text Analytics 95-865
Output will go to D:\Dropbox\TextAnalytics\wekacmd\distributable\\results.csv
JVM maxheap size is set to 2048M
Weka jar file is located at C:\Program Files\Weka-3-6\weka.jar
CLASSPATH for JVM is set to ;C:\Program Files\Weka-3-6\libsvm-3.17\java\libsvm.j
ar;C:\Program Files\Weka-3-6\weka.jar;D:\Dropbox\TextAnalytics\wekacmd\distribut
able\\wekaReader.jar
SVM_OPTIONS is -S 0 -K 0 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.
1 -seed 1
J48_OPTIONS IS -C 0.25 -M 2
NB runs with default
"NAIVE BAYES WEKA WRAPPER"
12:35:56.56
iris.2D
12:35:56.85
12:35:56.85
iris
12:35:57.20
12:35:57.20
labor
12:35:57.51
12:35:57.52
ionosphere
12:35:58.06
Press any key to continue . . . _
```

All the batch files will execute configuration.bat first and display the environment variables in the beginning.

The remaining outputs are the dataset name, start time and end time. All the raw outputs from Weka will be placed inside the ./RawOutput folder where you can examine them.

The program will append the results of your experiment to **results.csv**. You can open this file with excel later to collect your results (precision, recall, f number). This is done for your convenience, otherwise you would have to go through the Weka output and manually copy the values you need for your homework report. Keep in mind that you should not lock the file for editing when the batch files are running, otherwise your results won't be appended to the csv file. If you want to peak inside results.csv, making a copy would be the best way.

Notes for Mac Users

Weka, LightSIDE and libSVM are java based programs. This is good news because they can be run natively on your computer. The automation component (DNDW) is unfortunately written in windows batch files. You have a few options:

1. Use Bootcamp and run everything in windows.
2. Use OSX Lightside and create the ARFF files in Mac, copy them into Bootcamp and run DNDW in windows bootcamp.
3. Use OSX LightSIDE and create ARFF files in mac, copy them on to a USB drive and run DNDW in Heinz Virtual Labs.

Instructions for Heinz Virtual Labs can be found here:

http://www.heinz.cmu.edu/computing-services/virtual-labs/index.aspx