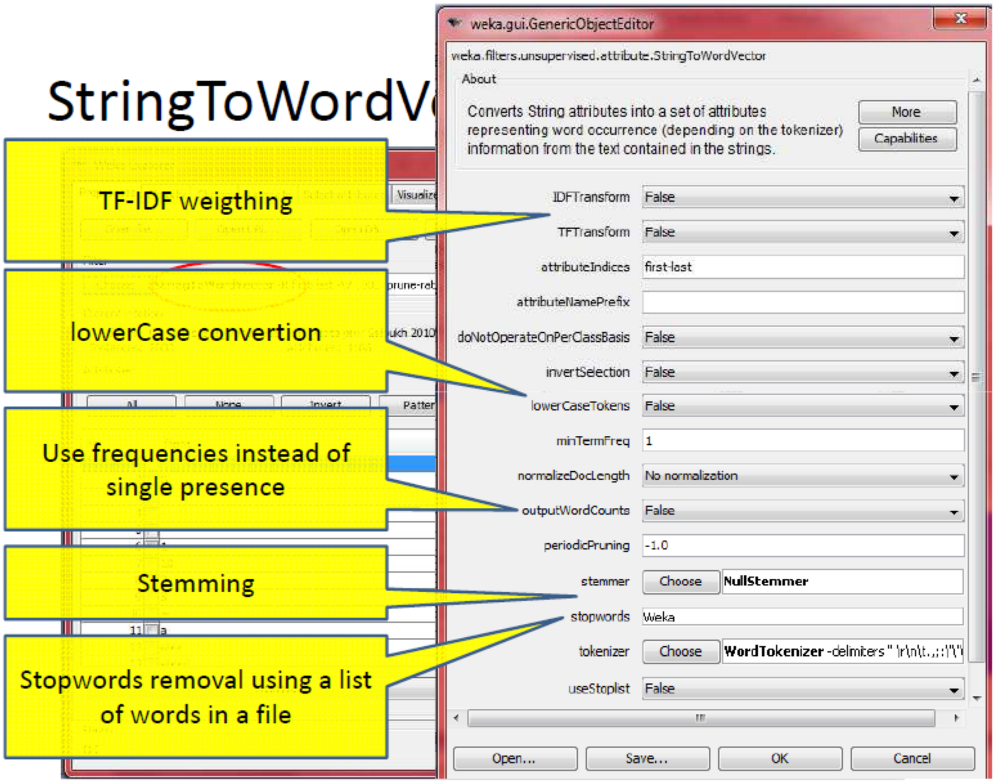
Data Mining II: Advanced Methods and Techniques

Assignment #5: Text Mining

1. Upload the Upload either Movie_Reviews.arff (under the Recourses section) or ReutersCorn-Train.arff (comes with Weka's Data folder) (for Extra credit do both!). Under the Preprocess tab click on Filters and choose the Unsupervised->Attribute->StringToWordVector. This filter will enable us to perform all of the Text mining Data preparation tasks.


StringToWordV

TF-IDF weigthing

lowerCase convertion

Use frequencies instead of single presence

Stemming

Stopwords removal using a list of words in a file

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.StringToWordVector

About

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

IDFTransform: False

TFTransform: False

attributeIndices: first-last

attributeNamePrefix:

doNotOperateOnPerClassBasis: False

invertSelection: False

lowerCaseTokens: False

minTermFreq: 1

normalizeDocLength: No normalization

outputWordCounts: False

periodicPruning: -1.0

stemmer: Choose NullStemmer

stopwords: Weka

tokenizer: Choose WordTokenizer -delmiters " \r\n\t.,;:'\"\

useStoplist: False

Apply String To Vector Filter one step at the time! After you apply the particular set of parameters - save that particular version of the training data set after each set with a descriptive name. Then click Undo and start a new set of steps and a new file to be saved. We will use all these different versions later to build and compare diverse models.

1. Inspect data via Edit Tab

2. Apply default filter first – don’t forget to click Apply button!! Observe the results; save the training data set
3. Inspect data via Edit Tab – how is it different?
4. Undo
5. Set Words to Keep to 100
6. Inspect data via Edit Tab; save this version of the training data set;
7. Undo
8. Perform the same(save training data set and then Undo changes) with DoNotOperateonPerClassBasis
9. Change IDFTransform and TFTransform to True + OutputWordCount should be True as well
10. Utilize the NormilizeDocLength
11. Use the Stemmer – LovinStemmer - break words down in a shorter form
12. StopWords – English stop words list in Weka
13. Tokenizer – use the ngrams
14. MinTermFrequency; periodicPruning – keep only words with term frequency of predetermined min value

As you perform all these different experiments I suggest keeping track of them in the table similar to the one presented below

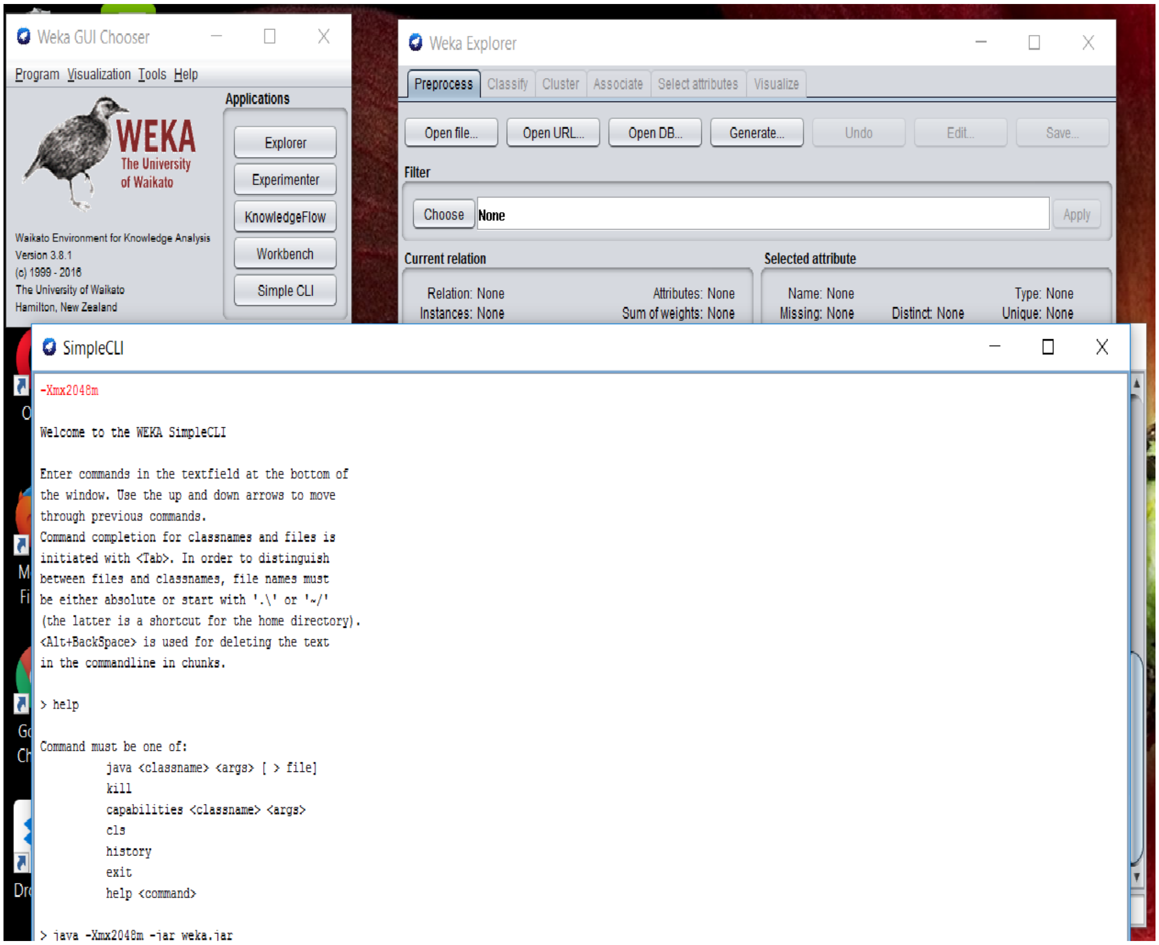
File Name	Stopwords	Stemming	Presence/Frequency	Normalization	MinFreq
TrainMovies1_StopWords_Stemming.arff	y	y	…		
TrainMovies2_NoStopWords_Stemming_IDF.arff	n	y	IDF		
…					

Once you have performed the data prep – utilize each of the training data sets on several different classification models. Compare the % of correctly classified instances – provide in depth discussion on the evaluation, comparison and pros/cons of the different ways of data prep for text mining.

File Name	Naïve Bayes	J48	SVM	…
TrainMovies1_StopWords_Stemming.arff	74.6%			
TrainMovies2_NoStopWords_Stemming_IDF.arff				
…				

If you start having problems with running out of memory you can set the Maxheap=2048m in the CLI (command line interface) by typing the following command:

```
java -Xmx2048m -jar weka.jar
```



Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand

Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Distinct: None
Type: None
Unique: None

```
-Xmx2048m

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
        java <classname> <args> [ > file]
        kill
        capabilities <classname> <args>
        cls
        history
        exit
        help <command>

> java -Xmx2048m -jar weka.jar
```