Spring 2019: CSCI 6990 Programming Assignment #2

DUE: Wednesday, April 17, 2019 (**Softcopy** @ 4 PM; **Hardcopy** @ 6:30 PM/in class)

Instructions

All work must be your own other than the instructor provided data/code and hints to be used. You are not to work in teams on this assignment.

Description

Training Dataset: Movie review dataset has been collected for sentiment analysis (see http://www.cs.cornell.edu/people/pabo/movie-review-data/). The dataset has been grouped into positive and negation classes (check Moodle for dataset).

Task [Marks 100]: Develop the **analysis report** as described below and submit it: As demonstrated and discussed, the development of NASA's patent-classifier for fifteen-class classification problem, here similarly for the assignments, we will need to do the following steps and develop the **analysis report**:

i) [**10** points] Given the dataset, use Weka's Simple-CLI to build up the initial ARFF file, which will contain the movie-review-text as a string and the output class {positive, negative}. Add the initial class distribution in the report. Also, report the required conversion time in this step.

ii) [**15** points] Convert the text-string to most useful vector using Weka's unsupervised filtering tool: StringToWordVector. Report the parameters you have chosen and explain their roles and justify your selections. Also, report how many words you have collected in this step.

iii) [**15** points] Using Weka's supervised filter, apply '**infoGainAttributeEval**' with '**Ranker'** having threshold value set to 0.0. Now, report the total words remains for classification and report the first 10 words with their information-gain values.

iv) [**20** points] Run 10 different classifiers and measure their performances using 10 FCV. Report all their performances (accuracy in %) including the confusion matrices. You must include Naïve-Bayes approach as one of the 10 classifiers.

v) [**20** points] Report the best method with its parameter(s) you have found including the performance-evaluation matrices. Explain, why do you think your selected top method is the best method out of the 10 methods you tried.

vi) [**20** points] Review literature to explain ‘**infoGainAttributeEval**’ in details and cite the relevant reference(s). Submit the copies of the paper(s) that you have cited to explain ‘**infoGainAttributeEval**’.

i) [**10** points] Given the dataset, use Weka's Simple-CLI to build up the initial ARFF file, which will contain the movie-review-text as a string and the output class {positive, negative}. Add the initial class distribution in the report. Also, report the required conversion time in this step.

STEP 1: DATA IMPORTATION INTO WEKA:

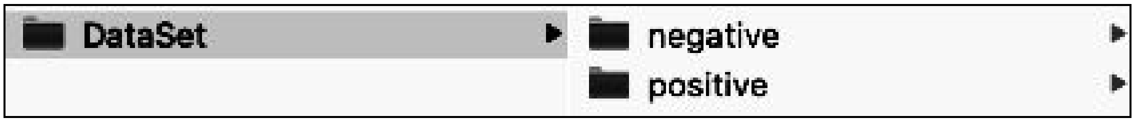
The given '*Movie Review*' dataset is contained across two separate directories (labeled: positive, negative). The directory labeled *positive* contains 1000 text documents (.txt), where each file is a positive movie review. The directory labeled as *negative* contains 1000 text files (.txt), where each file is a negative movie review.

This dataset must be preprocessed and reformatted before any analysis may begin. Since Weka is the tools selected analysis and classification, then the dataset must be converted into an ARFF format. ARFF stands for Attribute-Relation File Format. It is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

Weka provides the necessary converters to reformat the 'Movie Review' dataset to ARFF. As Weka is a Java-based application, its data conversion tools may execute from the command line via calls to the Weka.jar file, package: weka.core.converters, class: TextDirectoryLoader. This is optimal, as it allows data preprocessing into the ARFF format to be automated via system-level scripts. For the purposes of this project, the dataset conversion was accomplished using a python script. See *Appendix 1: Weka Preprocess Data - Time Capture.*

The Weka TextDirectoryLoader requires the filepath for the dataset. For this project, the dataset filepath should be the parent directory of the two subdirectories: negative and positive.

Path to DataSet

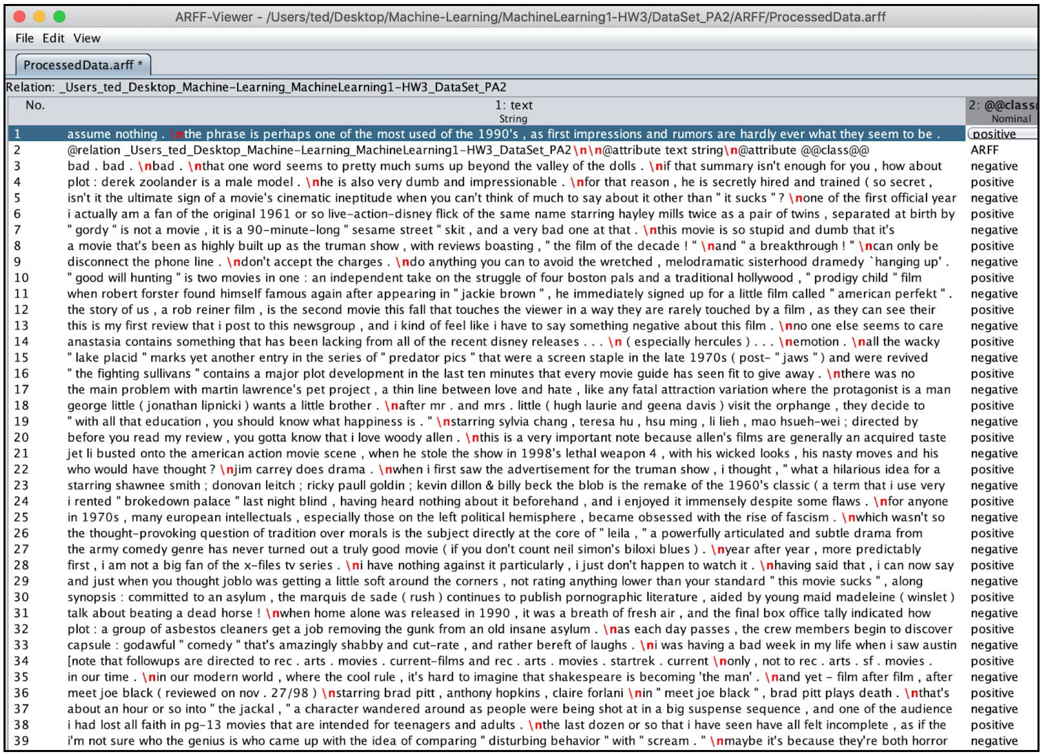


Weka will then use the directory labels to produce a classifier for the text files within those directories. Thus the text files within the negative directory are labeled negative and the text files within the positive directory are labeled with as positive in the resulting ARFF file.

The runtime to convert the original dataset of 2000 text files into ARFF format took precisely **4790.44116211 milliseconds** or **~4.8 seconds**. This time calculation was achieved by executing the weka jar from a bash command via a python script. The python script captured a before and after timestamps to determine the total runtime. See *Appendix 1: Weka Preprocess Data - Time Capture* for the python implementation. Instructions for running this python script is included within the implementation.

After converting the dataset into ARFF we can examine the contents using Weka as a viewer. The converted ARFF dataset initially has three columns, row number, text content, anda positive/negative label.

Movie Review' Dataset (via Weka ARFF Viewer)

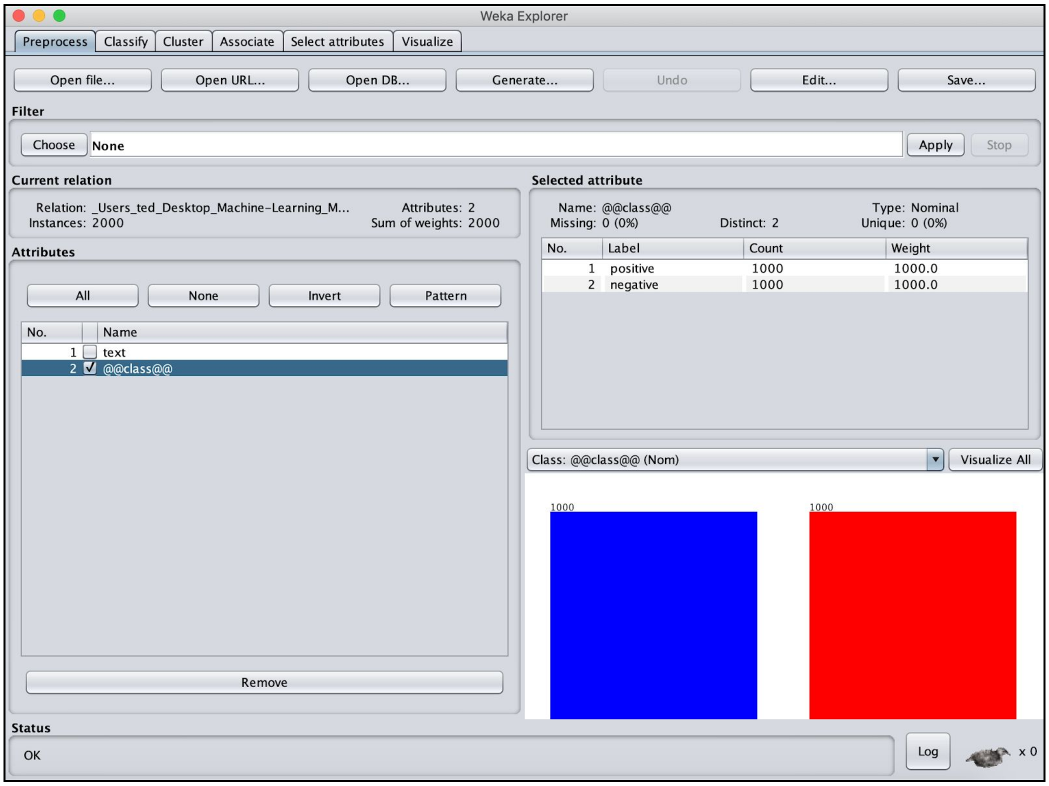

ARFF-Viewer - /Users/ted/Desktop/Machine-Learning/MachineLearning1-HW3/DataSet_PA2/ARFF/ProcessedData.arff
File
Edit
View
ProcessedData.arff *

Relation: _Users_ted_Desktop_Machine-Learning_MachineLearning1-HW3_DataSet_PA2

No.	1: text String	2: @@class Nominal
1	assume nothing . \nthe phrase is perhaps one of the most used of the 1990's , as first impressions and rumors are hardly ever what they seem to be .	positive
2	@relation _Users_ted_Desktop_Machine-Learning_MachineLearning1-HW3_DataSet_PA2\n\n@attribute text string\n@attribute @@class@@	ARFF
3	bad . bad . \nbad . \nthat one word seems to pretty much sums up beyond the valley of the dolls . \nif that summary isn't enough for you , how about	negative
4	plot : derek zoolander is a male model . \nhe is also very dumb and impressionable . \nfor that reason , he is secretly hired and trained (so secret ,	positive
5	isn't it the ultimate sign of a movie's cinematic ineptitude when you can't think of much to say about it other than " it sucks " ? \none of the first official year	negative
6	i actually am a fan of the original 1961 or so live-action-disney flick of the same name starring hayley mills twice as a pair of twins , separated at birth by	positive
7	" gordy " is not a movie , it is a 90-minute-long " sesame street " skit , and a very bad one at that . \nthis movie is so stupid and dumb that it's	negative
8	a movie that's been as highly built up as the truman show , with reviews boasting , " the film of the decade ! " \nand " a breakthrough ! " \ncan only be	positive
9	disconnect the phone line . \ndon't accept the charges . \ndo anything you can to avoid the wretched , melodramatic sisterhood dramedy `hanging up' .	negative
10	" good will hunting " is two movies in one : an independent take on the struggle of four boston pals and a traditional hollywood , " prodigy child " film	positive
11	when robert forster found himself famous again after appearing in " jackie brown " , he immediately signed up for a little film called " american perfekt " .	negative
12	the story of us , a rob reiner film , is the second movie this fall that touches the viewer in a way they are rarely touched by a film , as they can see their	positive
13	this is my first review that i post to this newsgroup , and i kind of feel like i have to say something negative about this film . \nno one else seems to care	negative
14	anastasia contains something that has been lacking from all of the recent disney releases . . . \n (especially hercules) . . . \nemotion . \nall the wacky	positive
15	" lake placid " marks yet another entry in the series of " predator pics " that were a screen staple in the late 1970s (post- " jaws ") and were revived	negative
16	" the fighting sullivans " contains a major plot development in the last ten minutes that every movie guide has seen fit to give away . \nthere was no	positive
17	the main problem with martin lawrence's pet project , a thin line between love and hate , like any fatal attraction variation where the protagonist is a man	negative
18	george little (jonathan lipnicki) wants a little brother . \nafter mr . and mrs . little (hugh laurie and geena davis) visit the orphange , they decide to	positive
19	" with all that education , you should know what happiness is . " \nstarring sylvia chang , teresa hu , hsu ming , li lieh , mao hsueh-wei ; directed by	negative
20	before you read my review , you gotta know that i love woody allen . \nthis is a very important note because allen's films are generally an acquired taste	positive
21	jet li busted onto the american action movie scene , when he stole the show in 1998's lethal weapon 4 , with his wicked looks , his nasty moves and his	negative
22	who would have thought ? \njim carrey does drama . \nwhen i first saw the advertisement for the truman show , i thought , " what a hilarious idea for a	positive
23	starring shawnee smith ; donovan leitch ; ricky paull goldin ; kevin dillon & billy beck the blob is the remake of the 1960's classic (a term that i use very	negative
24	i rented " brokedown palace " last night blind , having heard nothing about it beforehand , and i enjoyed it immensely despite some flaws . \nfor anyone	positive
25	in 1970s , many european intellectuals , especially those on the left political hemisphere , became obsessed with the rise of fascism . \nwhich wasn't so	negative
26	the thought-provoking question of tradition over morals is the subject directly at the core of " leila , " a powerfully articulated and subtle drama from	positive
27	the army comedy genre has never turned out a truly good movie (if you don't count neil simon's biloxi blues) . \nyear after year , more predictably	negative
28	first , i am not a big fan of the x-files tv series . \ni have nothing against it particularly , i just don't happen to watch it . \nhaving said that , i can now say	positive
29	and just when you thought joblo was getting a little soft around the corners , not rating anything lower than your standard " this movie sucks " , along	negative
30	synopsis : committed to an asylum , the marquis de sade (rush) continues to publish pornographic literature , aided by young maid madeleine (winslet)	positive
31	talk about beating a dead horse ! \nwhen home alone was released in 1990 , it was a breath of fresh air , and the final box office tally indicated how	negative
32	plot : a group of asbestos cleaners get a job removing the gunk from an old insane asylum . \nas each day passes , the crew members begin to discover	positive
33	capsule : godawful " comedy " that's amazingly shabby and cut-rate , and rather bereft of laughs . \ni was having a bad week in my life when i saw austin	negative
34	[note that followups are directed to rec . arts . movies . current-films and rec . arts . movies . startrek . current \nonly , not to rec . arts . sf . movies .	positive
35	in our time . \nin our modern world , where the cool rule , it's hard to imagine that shakespeare is becoming 'the man' . \nand yet - film after film , after	negative
36	meet joe black (reviewed on nov . 27/98) \nstarring brad pitt , anthony hopkins , claire forlani \nin " meet joe black " , brad pitt plays death . \nthat's	positive
37	about an hour or so into " the jackal , " a character wandered around as people were being shot at in a big suspense sequence , and one of the audience	negative
38	i had lost all faith in pg-13 movies that are intended for teenagers and adults . \nthe last dozen or so that i have seen have all felt incomplete , as if the	positive
39	i'm not sure who the genius is who came up with the idea of comparing " disturbing behavior " with " scream . " \nmaybe it's because they're both horror	negative

Loading the 'Movie Review' dataset into Weka. When Weka is launched, it offers the option to open file. Since ARFF is supported by Weka, it preloads information regarding the contents of the dataset such as the initial class distributions in both tabular and graphical formats.

Weka Explorer for the 'Movie Review' Dataset



Filter

None

Current relation

Relation: _Users_ted_Desktop_Machine-Learning_M...   Attributes: 2
Instances: 2000   Sum of weights: 2000

Attributes

No.		Name
1	□	text
2	☑	@@class@@

Selected attribute

Name: @@class@@   Type: Nominal
Missing: 0 (0%)   Distinct: 2   Unique: 0 (0%)

No.	Label	Count	Weight
1	positive	1000	1000.0
2	negative	1000	1000.0

Class: @@class@@ (Nom)

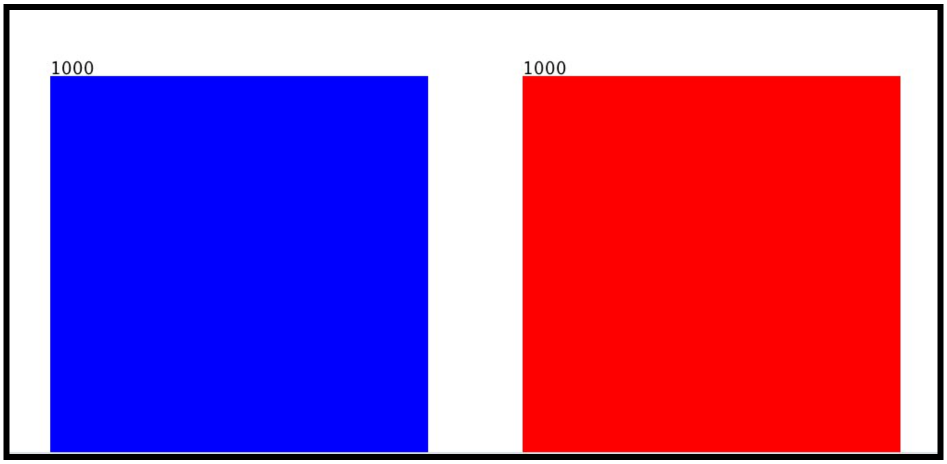
Status

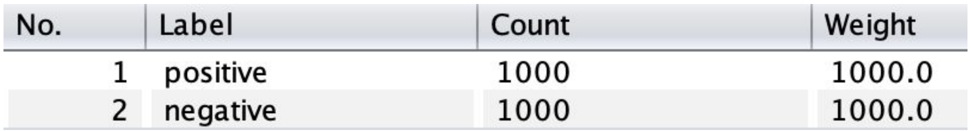
OK

x 0

The class distribution of the 'Movie Review' dataset when the ARFF is initially loaded into Weka amy be graphically displayed.

Initial Class Distribution





No.	Label	Count	Weight
1	positive	1000	1000.0
2	negative	1000	1000.0

ii) [**15** points] Convert the text-string to most useful vector using Weka's unsupervised filtering tool: *StringToWordVector*. Report the parameters you have chosen and explain their roles and justify your selections. Also, report how many words you have collected in this step.

STEP 2: DATA PREPARATION FOR WEKA (TOKENIZING):

Cleaning & Tokenization: The 'movie review' dataset is now imported into Weka, however, it must be optimized before any real analysis may begin. Currently, the input feature data for each row is expressed as a single String containing the full text contents of each review. This is not a practical format for performing any analysis, classifying, or clustering actions. This should instead be converted into a more efficient data structure in the form of a Word Vector. To do this, we must tokenize the text data such that it may be vectorized.

What is Tokenization: To make the provided text document classifiable using Machine Learning we need to do feature extraction that is converting the normal text to a set of features that can then be used by the ML Algorithm to discriminate between negative and positive reviews.

Weka provides built-in preprocessing filters explicitly for this purpose, i.e. converting text data into vector types. According to the Weka documentation, the *StringToWordVector* method performs the following actions with additional options. See *Appendix 2: StringToWordVector API Documentation*

public class StringToWordVector
Converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings. The dictionary is determined from the first batch of data filtered (typically training data). Note that this filter is not strictly unsupervised when a class attribute is set because it creates a separate dictionary for each class and then merges them.

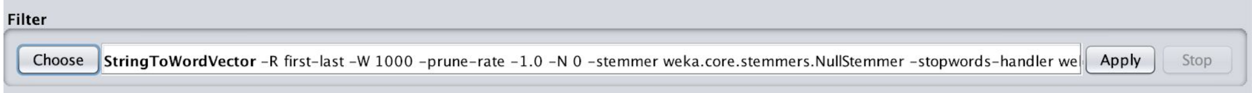
Options Name	**Description**
attributeNamePrefix	Prefix for the created attribute names. (default: "")
stopwordsHandler	The stopwords handler to use (Null means no stopwords are used).
wordsToKeep	The number of words (per class if there is a class attribute assigned) to attempt to keep.

debug	If set to true, filter may output additional info to the console.
outputWordCounts	Output word counts rather than boolean 0 or 1(indicating presence or absence of a word).
lowerCaseTokens	If set then all the word tokens are converted to lowercase before being added to the dictionary.
tokenizer	The tokenizing algorithm to use on the strings.
doNotCheckCapabilities	If set, the filter's capabilities are not checked before it is built. (Use with caution to reduce runtime.)
doNotOperateOnPerClassBasis	If this is set, the maximum number of words and the minimum term frequency is not enforced on a per-class basis but based on the documents in all the classes (even if a class attribute is set).
attributeIndices	Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".
normalizeDocLength	Sets whether if the word frequencies for a document (instance) should be normalized or not.
saveDictionaryInBinaryForm	Save the dictionary as a binary serialized java object instead of in plain text form.
invertSelection	Set attribute selection mode. If false, only selected attributes in the range will be worked on; if true, only non-selected attributes will be processed.
minTermFreq	Sets the minimum term frequency. This is enforced on a per-class basis.
TFTransform	Sets whether if the word frequencies should be transformed into log(1+fij) where fij is the frequency of word i in document (instance) j.
periodicPruning -	Specify the rate (x% of the input dataset) at which to periodically prune the dictionary. wordsToKeep prunes after creating a full dictionary. You may not have enough memory for this approach.
stemmer	The stemming algorithm to use on the words.
dictionaryFileToSaveTo	The path to save the dictionary file to - an empty path or a path '-- set me --' means do not save the dictionary.
IDFTransform	Sets whether if the word frequencies in a document should be transformed into: fij*log(num of Docs/num of Docs with word i) where fij is the frequency of word i in document (instance

To select the *StringToWordVector* filter from in Weka:

1. Click on *Choose* button below Filter
2. Choose: *weka → filters → unsupervised → attribute → StringToWordVector*

Weka GUI with Filter field (with StringToWordVector selected)

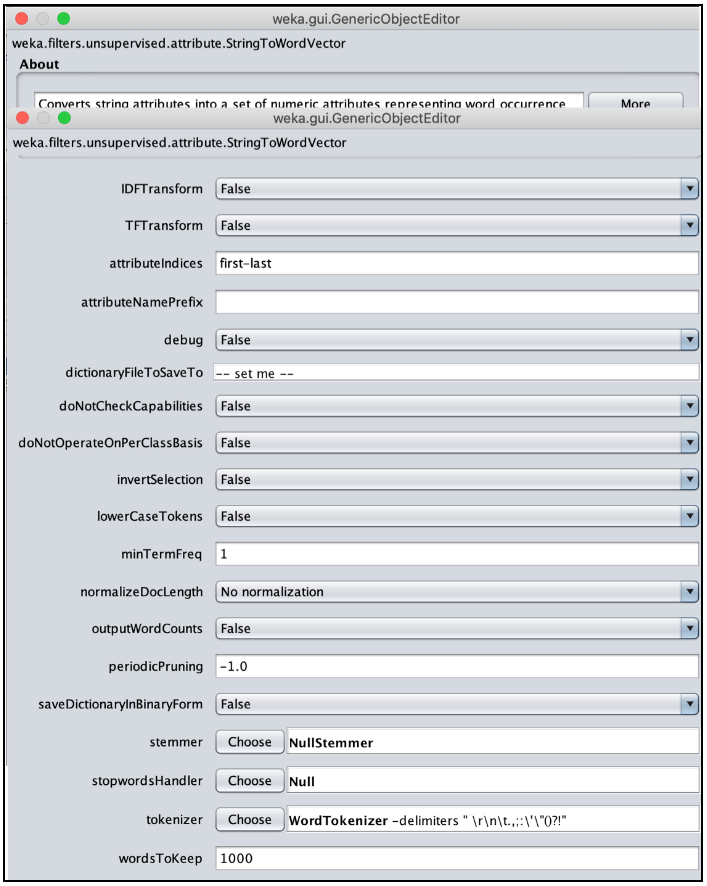


Filter

```
StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler we
```

Options: The StringToWordVector filter has several different options. The default values are shown below. However, adjusting the options can improve the tokenization of the text into features.

Weka GUI - StringToWordVector Options (Default Settings)



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.StringToWordVector

About

Converts string attributes into a set of numeric attributes representing word occurrence

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.StringToWordVector

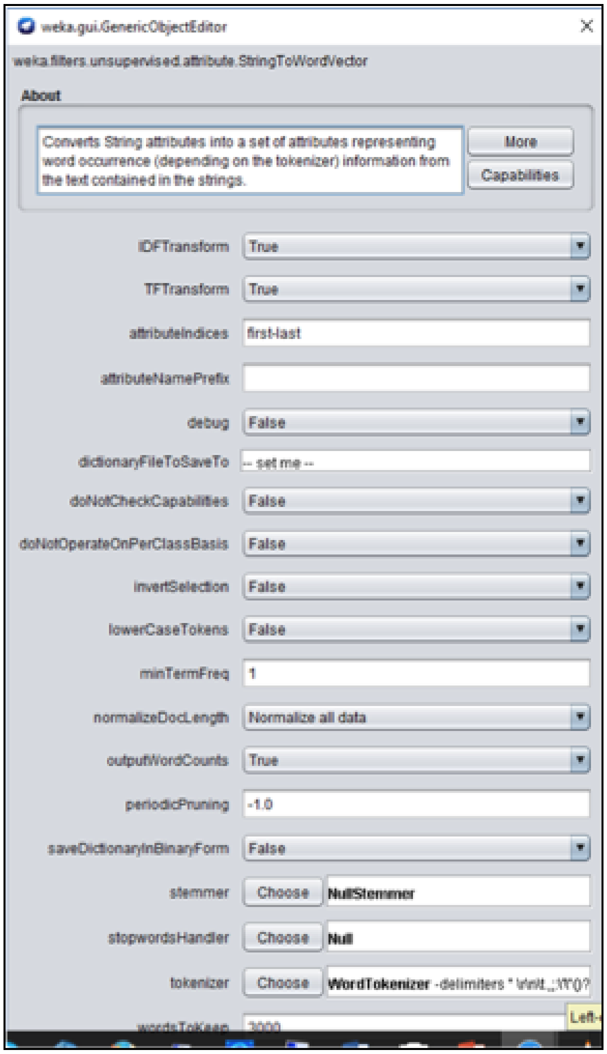
IDFTransform	False
TFTransform	False
attributeIndices	first-last
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	-- set me --
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	False
invertSelection	False
lowerCaseTokens	False
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	False
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	NullStemmer
stopwordsHandler	Null
tokenizer	WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"
wordsToKeep	1000

Default Options: The 'movie review' dataset tokenized using the default options from StringToWordVector filter results in identifying the occurrence of 1165 attributes (i.e. singular words). All 1165 attributes have bimodal distributions in occurences.

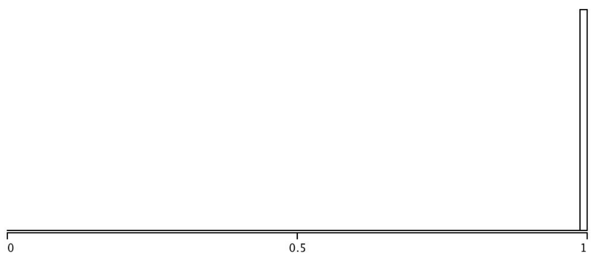
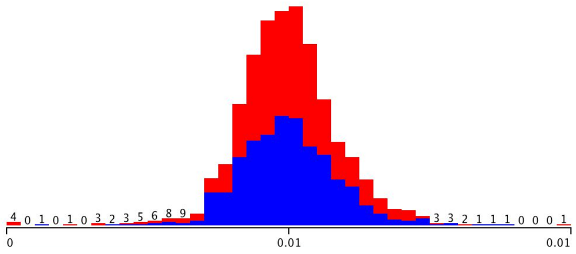
Custom Options: Adjusting StringToWordVector options with the following updates.

Term Frequency (TF)/Inverse Document Frequency (IDF) options: Models how important a word is to a given document within a collection of documents. It is often used as a weighting factor in searches of information retrieval and text mining. The TF–IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the total collection that contain the word, which helps to adjust for the fact that some words appear more frequently in general. TF–IDF is one of the most popular term-weighting schemes today;

Weka GUI - StringToWordVector filter - (Custom Options)

 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. More Capabilities IDFTransform: True TFTransform: True attributeIndices: first-last attributeNamePrefix: debug: False dictionaryFileToSaveTo: -- set me -- doNotCheckCapabilities: False doNotOperateOnPerClassBasis: False invertSelection: False lowerCaseTokens: False minTermFreq: 1 normalizeDocLength: Normalize all data outputWordCounts: True periodicPruning: -1.0 saveDictionaryInBinaryForm: False stemmer: Choose NullStemmer stopwordsHandler: Choose Null tokenizer: Choose WordTokenizer -delimiters " \r\n\t.,;:\'\"()? wordsToKeep: 3000	*IDFTransform*: **True** Reason: Turn on the weighing factor of Inverse Document Frequency. This helps track how important a word is in a given document
	TFTTransform: **True** Reason: Turn on the weighing factor of Term Frequency. Captures the frequency of the word appearing.
	normalizeDocLength: **Normalize all data** Reason: Normalize all data values between 0-1
	outputWordCounts: **True** Reason: Provides greater granularity in word occurence through counts instead of binary: present(1)/absent(0)

The results of this filtering identifies 1165 attributes, however produces better frequency distributions for performing analysis.

*Frequency of: '**a**' (Default Options)*	*Frequency of: '**a**' (Custom Options)*
	

The statistical algorithms must be applied on these attributes to constructo a predictor will perform better with frequencies similar to the right as opposed to the left.

Total number of attributes identified in both cases is **1165 words**.

iii) [**15** points] Using Weka's supervised filter, apply '**infoGainAttributeEval**' with '**Ranker'** having threshold value set to 0.0. Now, report the total words remains for classification and report the first 10 words with their information-gain values.

STEP 3: FEATURE SELECTION AND RANKING:

The 'Movie Review' dataset has now been prepared into a set of 1165 input features and one output class: 'positive/negative.' However, there are still too many features to perform a meaningful analysis, so more filtering is needed to generate a predictor that uses only the most critical set of attributes. Constructing better predictor models requires the removal of any words (i.e. features) that do not contribute in determining whether a review is negative or positive. So we must rank the features and select the top features that correlate to the output class. Note: It is important to ensure that your @@*class*@@ attribute defining your positive/negative values is assigned as the output class and appears as the last column in the dataset, as all feature ranking must be compared to the output class.

Weka provides built-in tools for performing feature filtering on datasets.

To select the infoGainAttributeEval filter from in Weka:

1. Click on *Choose* button below Filter
2. Choose: *weka* → *filters* → *supervised* → *attribute* → *AttributeSelecton*

iv) [**20** points] Run 10 different classifiers and measure their performances using 10 FCV. Report all their performances (accuracy in %) including the confusion matrices. You must include Naïve-Bayes approach as one of the 10 classifiers.

v) [**20** points] Report the best method with its parameter(s) you have found including the performance-evaluation matrices. Explain, why do you think your selected top method is the best method out of the 10 methods you tried.

vi) [**20** points] Review literature to explain '**infoGainAttributeEval**' in details and cite the relevant reference(s). Submit the copies of the paper(s) that you have cited to explain '**infoGainAttributeEval**'.

According to the official Weka API documentation[1] *infoGainAttributeEva*l is:

> InfoGainAttributeEval :
>
> Evaluates the worth of an attribute by measuring the information gain with respect to the class.
>
> InfoGain(Class,Attribute) = H(Class) - H(Class | Attribute).
> Valid options are:
>
> -M
> treat missing values as a seperate value.
> -B
> just binarize numeric attributes instead
> of properly discretizing them.

Within the official Weka documentation, Mark Hall, author of infoGainAttributeEval cites that this implementation is based on the research from:

Usama M. Fayyad, Keki B. Irani: Multi-interval discretization of continuous valued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence, 1022-1027, 1993.

Igor Kononenko: On Biases in Estimating Multi-Valued Attributes. In: 14th International Joint Conference on Artificial Intelligence, 1034-1040, 1995

[1] InfoGainAttributeEval, Weka API Documentation Revision: 10172 ,Mark Hall http://weka.sourceforge.net/doc.stable-3-8/weka/attributeSelection/InfoGainAttributeEval.html

InfoGainAttributeEval is used for **feature selection tasks**.

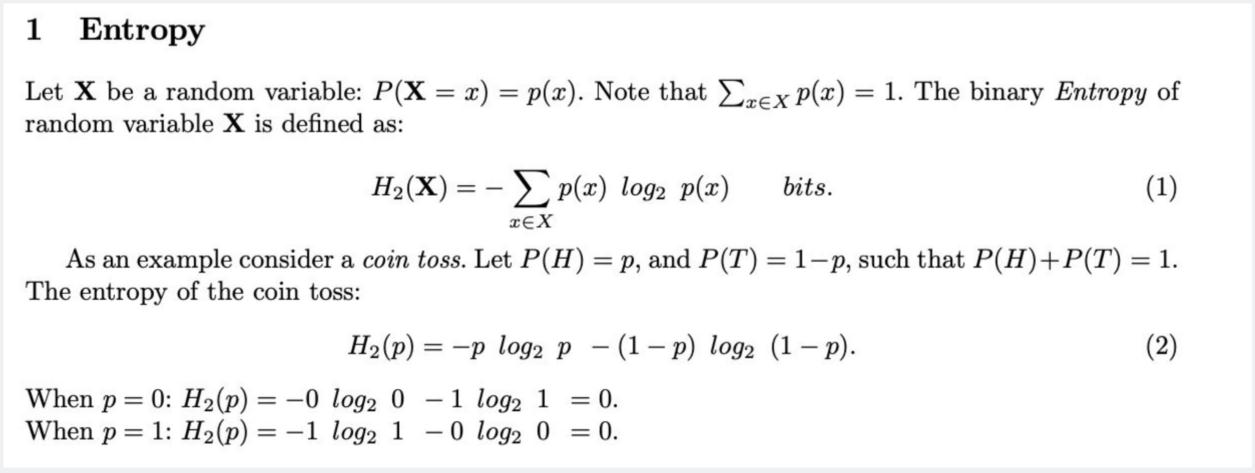
What InfoGainAttributeEval basically does is measuring how each feature contributes in *decreasing the overall entropy*.

Let's take an example. Say we have this dataset :

Temperature	Wind	Class
high	low	play
low	low	play
high	low	play
low	high	cancelled
low	low	play
high	high	canceled
high	low	play

The Entropy, H(X), is defined as follows :

```
H(X) = -sum(Pi*log2(Pi))
```



1 Entropy

Let $\mathbf{X}$ be a random variable: $P(\mathbf{X} = x) = p(x)$. Note that $\sum_{x \in X} p(x) = 1$. The binary *Entropy* of random variable $\mathbf{X}$ is defined as:

$$H_2(\mathbf{X}) = -\sum_{x \in X} p(x) \ \ log_2 \ \ p(x) \qquad bits. \tag{1}$$

As an example consider a *coin toss*. Let $P(H) = p$, and $P(T) = 1-p$, such that $P(H)+P(T) = 1$. The entropy of the coin toss:

$$H_2(p) = -p \ \ log_2 \ \ p \ \ -(1-p) \ \ log_2 \ \ (1-p). \tag{2}$$

When $p = 0$: $H_2(p) = -0 \ \ log_2 \ \ 0 \ \ -1 \ \ log_2 \ \ 1 \ \ = 0$.
When $p = 1$: $H_2(p) = -1 \ \ log_2 \ \ 1 \ \ -0 \ \ log_2 \ \ 0 \ \ = 0$.

, with Pi being the probability of the class i in the dataset, and log2 the base 2 logarithm (in Weka natural logarithm of base *e* is used, but generally we take log2). Entropy basically measures the **degree of "impurity"**. The closest to 0 it is, the less impurity there is in your dataset.

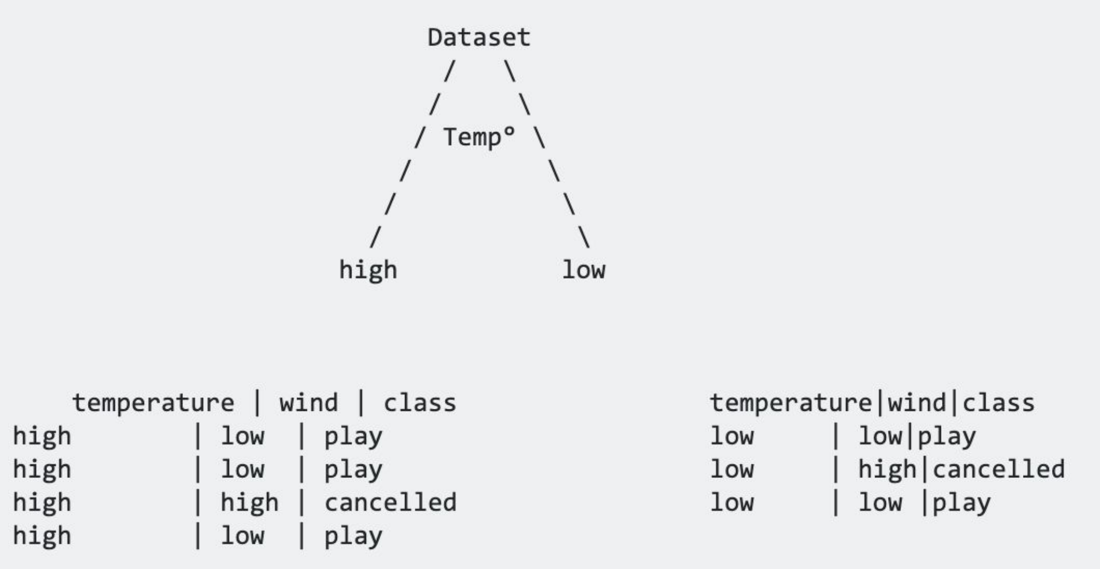
Hence, a good attribute is an attribute that **contains the most information**, i.e, **reduces the most the entropy**. The InfoGainAttributeEval method of Weka is a way of evaluating exactly this.

Now, the entropy of our example is : H(Class) = -(5/7*log2(5/7)+2/7*log(2/7)) = 0,863.

Let's calculate for our example the amount of information carried by the temperature attribute.

InfoGain(Class,Temperature) = H(Class) - H(Class | Temperature).

To get the H(Class | Temperature), we need to split the dataset according to this attribute.



```
                         Dataset
                          /   \
                         /     \
                        / Temp° \
                       /         \
                      /           \
                     /             \
                   high            low



    temperature | wind | class                    temperature|wind|class
high         | low  | play                        low     | low|play
high         | low  | play                        low     | high|cancelled
high         | high | cancelled                   low     | low |play
high         | low  | play
```

Each branch here has its own entropy. We need to first calculate the entropy of each split.

$$H(leftside) \;=\; -\left(\tfrac{3}{4}\, log_2\,(\tfrac{3}{4}) + \tfrac{1}{4}\, log_2\,(\tfrac{1}{4}) \right) \;=\; 0.811$$
$$H(rightside) \;=\; -\left(\tfrac{1}{3}\, log_2\,(\tfrac{1}{3}) + \tfrac{2}{3}\, log_2\,(\tfrac{2}{3}) \right) \;=\; 0.918$$

H(Class | Temperature) is then equals to the sum of both children's entropy, weighted by the proportion of instances that where taken from the parent dataset. In short :

$$H(Class \,|\, Temperature) \;=\; \tfrac{4}{7}\, H(leftside) \;+ \tfrac{3}{7}\, H(rightside)$$

You then have everything to calculate the InfoGain. In this example, it's 0,06 bits. This means that the temperature feature only reduces the global entropy by 0,06 bits, the **feature's contribution to reduce the entropy** (= the **information gain**) is fairly small.

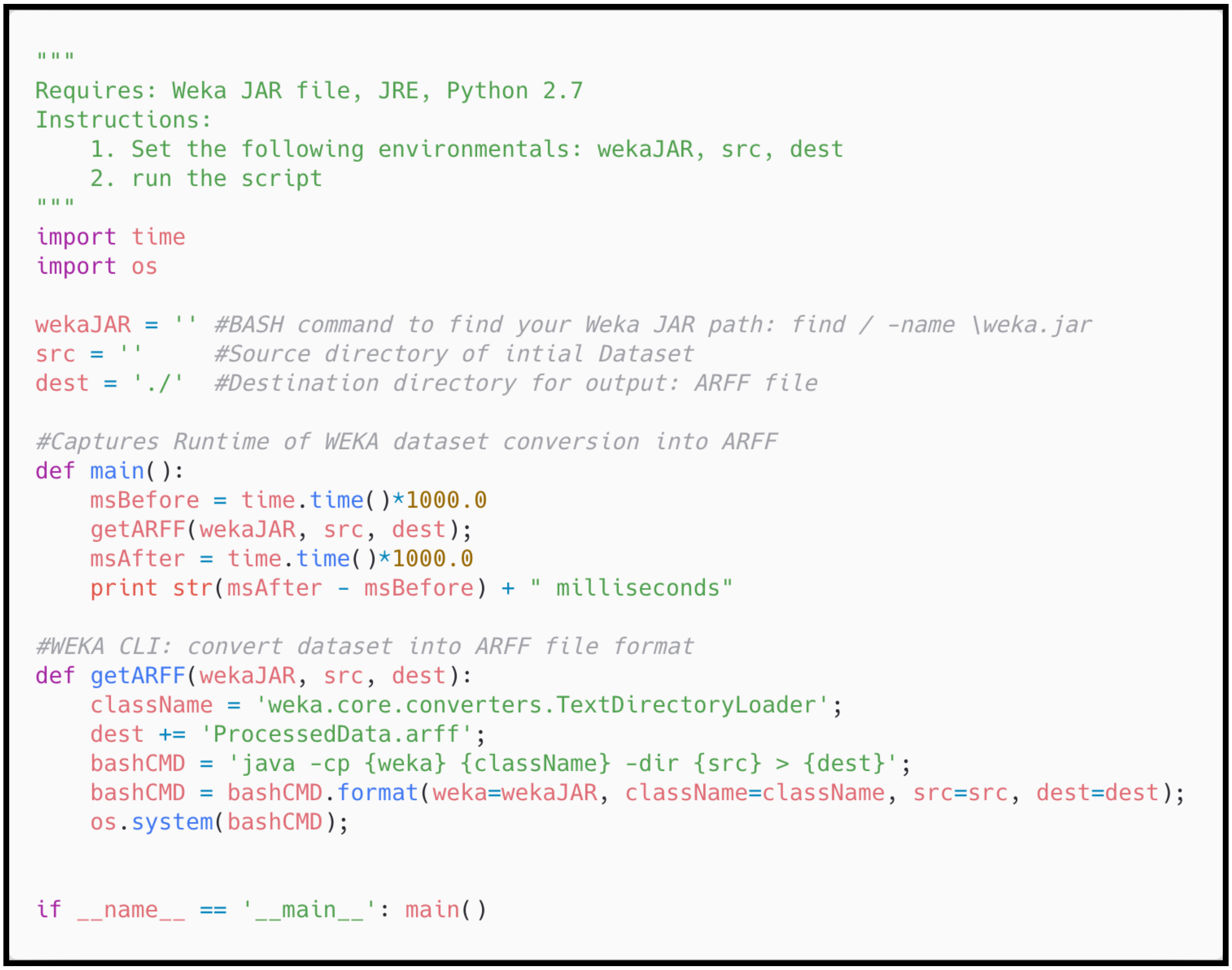
This is pretty obvious looking at the instances in the dataset, as we can see at a first glance that the temperature doesn't affect much the final class, unlike the wind feature.

Source for part of this answer :
Anuj Sharma and Shubhamoy Dey. Article: Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. IJCA Special Issue on Advanced

Computing and Communication Technologies for HPC Applications ACCTHPCA(3):15-20, July 2012

Appendix 1: *Weka Preprocess Data - Time Capture*

System-level Python Script



```python
"""
Requires: Weka JAR file, JRE, Python 2.7
Instructions:
    1. Set the following environmentals: wekaJAR, src, dest
    2. run the script
"""
import time
import os

wekaJAR = '' #BASH command to find your Weka JAR path: find / -name \weka.jar
src = ''     #Source directory of intial Dataset
dest = './'  #Destination directory for output: ARFF file

#Captures Runtime of WEKA dataset conversion into ARFF
def main():
    msBefore = time.time()*1000.0
    getARFF(wekaJAR, src, dest);
    msAfter = time.time()*1000.0
    print str(msAfter - msBefore) + " milliseconds"

#WEKA CLI: convert dataset into ARFF file format
def getARFF(wekaJAR, src, dest):
    className = 'weka.core.converters.TextDirectoryLoader';
    dest += 'ProcessedData.arff';
    bashCMD = 'java -cp {weka} {className} -dir {src} > {dest}';
    bashCMD = bashCMD.format(weka=wekaJAR, className=className, src=src, dest=dest);
    os.system(bashCMD);


if __name__ == '__main__': main()
```

Appendix 2: Sources