**Web Page Preprocessing, Learning and Classification**

How to Use Weka for Machine Learning

Opening the data set:

* Open the folder in Weka Preprocess using the ‘Open file’ button.. Select the destination folder.
* Weka will automatically pick the file if it is in ARFF file format. If not, there will be an error message with an open dialogue box. Within that command box select “choose’. Select ‘Text Directory Model”. Then select ‘ok’.

Preprocessing:

* Use the choose button to filter the data. Using NominaltoBinary will convert nominal attributes into binary numeric attributes. Similarly, depending on the data type and the classification to be done, choose the appropriate form of the filter method.

Classifying

* Once the preprocessing is done, we can train the dataset using different algorithms.
* Choose one of the pre programmed machine learning algorithms.
* For test options, you can use the training data and also with test data.

Testing:

* Use the supplied Test Set to test the test dataset. And press ‘start’

#PYTHON#

Our python program has two files to it:

* A1Preprocessing.py→ this file will preprocess the provided dataset
* Classifiers.py → this file will run different types of classification algorithms on the preprocessed dataset and provide an accuracy measure.

The A1Preprocessing.py works by setting up the directories of where you extracted the dataset and have the test and train set folders placed. Examples are shown in the code itself.

Once you have set the folder directories of the train and test course, student, and faculty, you can run the program, it will take a couple of seconds for the preprocessing to complete, as it runs multiple cleaning methods, such as stop word removal, stemming and more.

When the program finishes, you can now run the Classifier python program.

The Classifiers.py requires you to set the directory of the cleaned dataset folder. This would involve moving the .txt files into a separate “Cleaned” folder (or whatever folder you would like)

Once you have set the directories in place, you can run the code. The python code runs three different classifiers, those being naive bayesian, KNN and decision tree. When the code finishes running, it will print various metrics to show how well the classifier algorithms done.