CS685 : Course Project
Vivaad

We have used Weka for the classification purposes. As we used the GUI model, we do not have any specific code for classification purpose, so we are providing our model and features file as a part of the code.

The structure of our code is as follows :
1) **src** - It contains the main code which was used to obtain the data, clean the data, and preprocess it into text format from xml format. It also contains the code to get the edit-count of each article which we used as a feature. The main files are:

crawler - used to obtain the data

remove_duplicates.py - used to remove any duplicate files downloaded in both controversial and non-controversial category.

xml_to_text.py - used to convert the data from xml format to text format

get_edit_history.rb - used to obtain number of edit history count of each articles

2) **features** - It contains the features file for both category - history and politics-economics. Features file are in .csv format. Files are named as : "<Category>_<feature>.csv" where Category can be History or Politics_Economics and feature can be Basic or Article or Edit. Basic means only bag of words, Article means article length is added as feature, Edit means number of edit count is added as features.

3) **model** - It contains all the model for both category. By model we mean the classifier. They are named as "<Category>_<feature>_<kernel>.model" where, Category can be History or Politics-Economics, feature can be Basic,Article or Edit with the same meaning as in feature, kernel represents the kernel of the svm classifier used. It can be either "linear" specifying linear kernel or can be "radial" specifying that the kernel is radial basis function. We have 6 kernel for each category given by various combination of features and kernel.

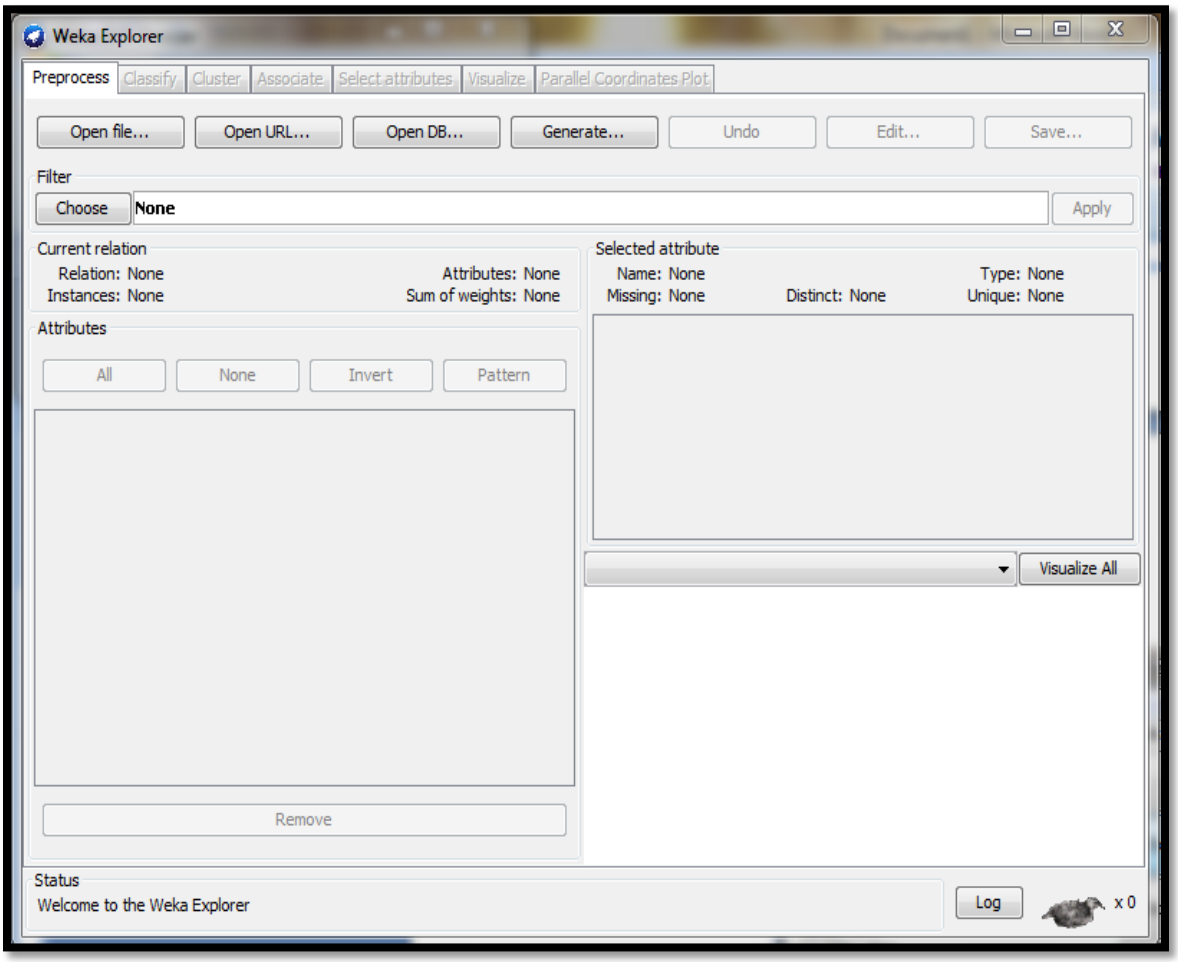
4) **results** - It contains the results obtained from the classifier and accuracy vs model plot for each category.

5) **data** – It contains the text file of each article in the different category.

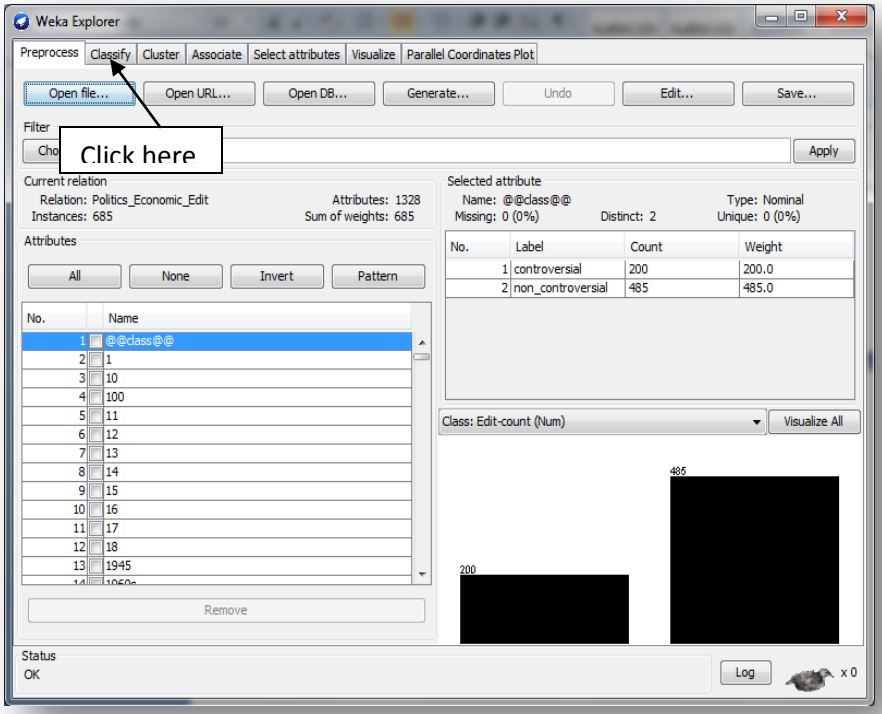
How to run the code :

1) Open Weka in GUI mode and select Explorer.

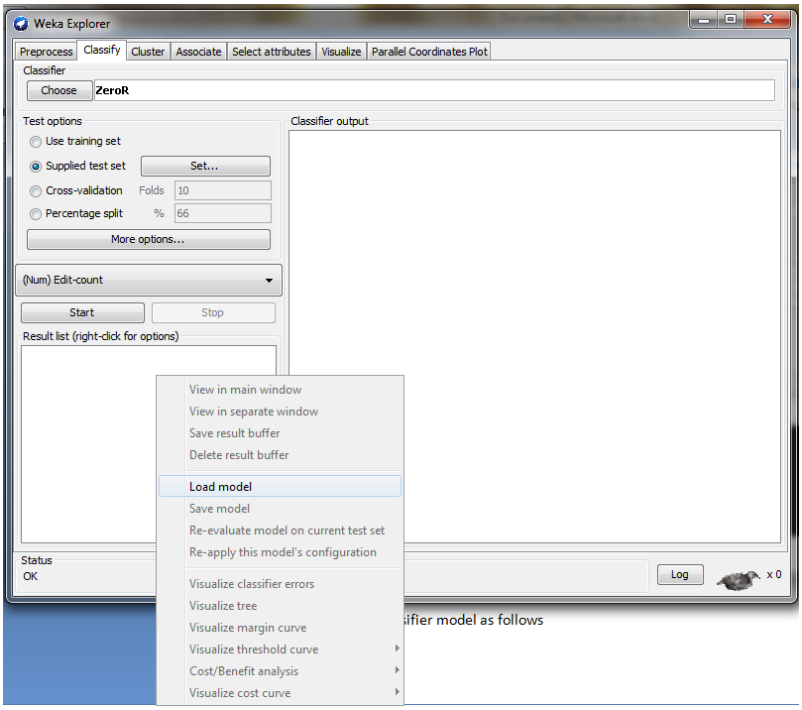
2) Load the features files(.csv files) using Open file… button. A dialog box will open when you click on Open file… , select the feature file you want to use and click Open.



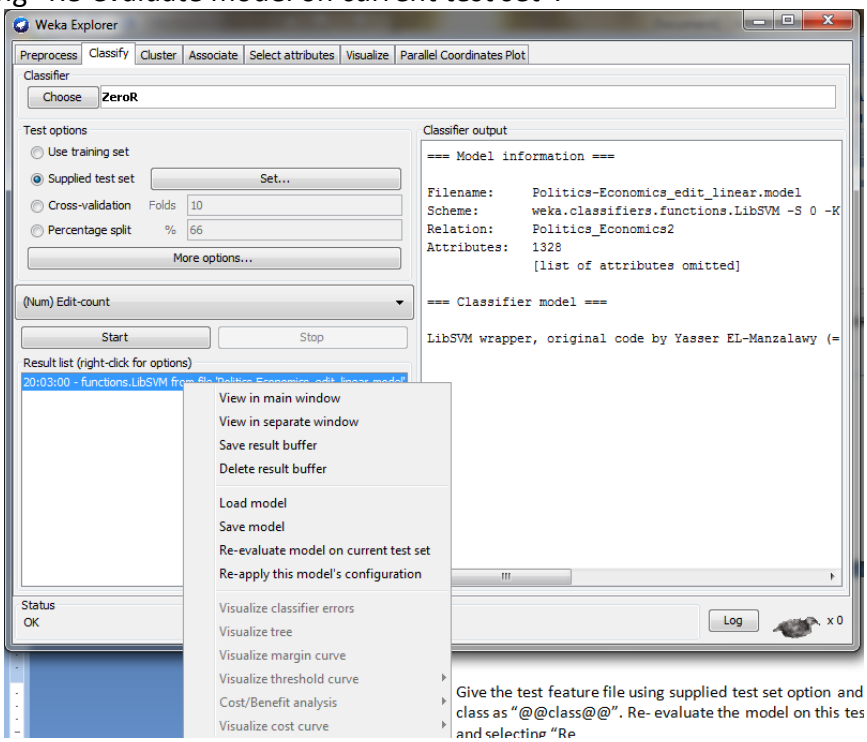
3) After file is loaded go to Classify tab.



4) Load the classifier model.



5) Give the test feature file using supplied test set option and opening the file and selecting class as “@@class@@”. Re- evaluate the model on this test set by right-click on model and selecting “Re-evaluate model on current test set”.



Give the test feature file using supplied test set option and
class as “@@class@@”. Re- evaluate the model on this tes
and selecting “Re

6) If you want to retrain the model, right click on model and select “Re-apply this model configuration” and then select “cross-validation” and class as “(Nom)@@class@@” and click on Start.