

Parameters for running ClassificationBox

Following is a list of input parameters for running ClassificationBox. Each of the parameters is specified in a task specification file (e.g., *text_classification.xml*). A variable in bold face is a mandatory option. All the variables serve some documenting purpose.

<Task>

- **name**: specify the name of task (It is OK to have any arbitrary English name. It will be primarily used as an identifier of this task, naming all intermediate/final results),
- **type**: specify the type of task, {classification | clustering | ...}, At the moment (Oct 25, 2004) there is only one task available: classification.

[Dataset]

- **name**: specify the name of data set (It may use an arbitrary English name),
- **dataset_location**: specify absolute directory name of data set. Note that you should place “file.separator” at the end of this variable.
- **document_index_file**: Specify the name of document index. This file should be placed at *\your_classificationbox_directory\cbox\data\your_index_file_name.idx*.
- **dataset_type**: specify the physical appearance of each data instance in the given data set. For instance, it might be a series of documents in any natural language. In this version, only a data instance in text is considered. Note that this version of ClassificationBox is only able to handle data in text.
- **dataset_format**: specify how the given data set is comprised of data instances. It will be either *consecutive* or *separate*.
- **train_portion**: specify the number of items from the data set that will be used for training.
- **active_classes**: specify class labels which are actively used for a given task. It might happen that you will not use any available class labels in given data set. Write “none” if that is the case.
- **active_attributes**: specify a subset of attributes which are actively used for a given task. Again, it might not be necessary to make use of any available attributes in an instance. Write “none” if that it is the case.
- **positive_class**: Specify the label of positive class. Except for those instances belonging to this specified class label, the remaining instances are regarded as negative examples.

In the primary implementation of ClassificationBox there is only one “positive” class, with the remaining classes regarded as “negative” classes.

[Indexing]

- **remove_stopword**: Specify whether to remove stop-words from a given data set when indexing a given data set. *Default*: *remove_stopword* = “yes”
- **prune_word**: Specify whether to prune a list of words matching a certain criteria
 - o **infrequent**: Specify the lower bound of word occurrences (in a document) to be removed

- **toofrequent:** Specify the upper bound of word occurrences (in a document) to be removed
- *Default:* prune_word = “yes”, infrequent = “2”, toofrequent = “100”
- **stemming:** Specify whether to apply stemming when indexing them, *Default:* stemming = “no”

By removing stop-words and pruning words in certain frequencies, you will be able to remove the noise in a given data set—that which is not relevant in performing a given task.

[Feature_Selection]

The current version of “ClassificationBox” does not support any of the feature selection methods.

- method = “none” as default value, *Default:* method = “none”

[Representation]

- model: Specify which text representational model is to be used, *Default:* model = “vector space,” where “vector space” means that a vector space model is used to represent a text instance.

[Learning]

- **method:** Specify which learning method is to be used for the given task. Since the current version of “ClassificationBox” is only capable of classifying text documents, it would be the name of a specific text classification method. *Default:* method = “wh,” where “wh” refers to Widrow-Hoff. There are three classification methods available: wh, eg, and knn, standing for Widrow-Hoff, Exponentiated-Gradient, and k- Nearest Neighbor, respectively.
- cross_validation: Specify whether to apply cross validation when training a classifier. *Default:* cross_validation = “no”

[Evaluation]

- **method:** Specify the evaluation metrics you wish to see, *Default:* method = “precision, recall, f1, false_alarm, miss”

[Output]

- output_file: Specify the name of output file, *Default:* output_file = “task_name.out”
- output_dir: Specify the name of output directory, which is used to write the intermediate and final results, *Default:* output_dir = “none”

[Other]

- verbosity: Specify whether to turn on verbosity

Example) text_classification.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
<text_learning>
<task name="text_data1_classification" type="classification" />
<dataset name="text_data1" dataset_location="c:\text_data1\"
  document_index_file="c:\ClassificationBox\cbox\data\document_index_text
_data1.idx" dataset_type="text" dataset_format="separate" train_portion="80"
  active_classes="none" active_attributes="none" positive_class="terrorism" />
<indexing remove_stopword="yes" prune_word="yes" infrequent="5"
  toofrequent="100" stemming="no" />
<feature_selection method="none" />
<representation model="vector space" />
<learning method="wh, eg, knn" cross_validation="none" />
<evaluation method="precision, recall, f1, false_alarm, miss" />
<output output_file="text_classification_text_data1.out" output_dir="none" />
<other verbosity="yes" />
</text_learning>
```

1. "task name" The name of this task is called "text_data1_classification." This name will be used to distinguish this task from others. In particular, this name will be used for naming all intermediate results (automatically removed when a task is done) and any final results.
2. "dataset_location" The data set for this task is placed at "c:\text_data1\" Note you must place "file.separator" (e.g., "\" for Windows) at the end of this value
3. "document_index_file" A document index file ("document_index_text_data1.idx") has been placed under the directory of "c:\ClassificationBox\cbox\data\"
4. "positive_class" A category "terrorism" is assigned to this field. There are four categories for "text_data1": terrorism, al_qaeda, arms_proliferation, and narcotics. The other three categories (i.e., al_qaeda, arms_proliferation, and narcotics) are regarded as "negative" classes.
5. "output_file" The experimental result will be printed out as a file entitled "text_classification_text_data1.out" It is a text file and you may want to use any Editor (e.g., Notepad for Windows and Emacs for Unix) to read this file.