Lovely Corpus

Generated by Doxygen 1.8.4

Wed Aug 7 2013 11:23:26

Contents

Chapter 1

LovelyCorpus Copyright (C) 2013 Daniel Leblanc

Corpus management toolkit

This tool is designed to allow for easier data preprocessing in machine learning applications. Minor changes to improve the quality of a corpus can produce large gains the the overall accuracy of the final algorithm. For a complete breakdown of the current features see the documentation section.

Anyone interested in assisting with development, providing feedback, or suggesting new features please contact Daniel Leblanc dleblanc@pdx.com.

Current task list is stored in the file <tasks>

Chapter 2

Namespace Index

2.1 Packages

Here are the packages with brief descriptions (if available):

Chapter 3

File Index

3.1 File List

Here is a list of all files with brief descriptions:

Chapter 4

Namespace Documentation

4.1 FlogCorpus Namespace Reference

Functions

- def runTests

Variables

- string SEPERATE = "**********"
- list ALLTESTS = ["kfold", "setIntRange", "entropyGain", "checkRange"]

4.1.1 Function Documentation

4.1.1.1 def FlogCorpus.runTests (*tests* = ALLTESTS)

4.1.2 Variable Documentation

4.1.2.1 list FlogCorpus.ALLTESTS = ["kfold", "setIntRange", "entropyGain", "checkRange"]

4.1.2.2 string FlogCorpus.SEPERATE = "**********"

4.2 LovelyCorpus Namespace Reference

Functions

- def attributeCount
- def removeOutliers
- def checkDuplicates
- def setIntRange
- def kfold
- def entropyGain
- def egain
- def checkRange

Variables

- int FIRST = 0
- int LAST = -1

4.2.1 Function Documentation

4.2.1.1 def LovelyCorpus.attributeCount (*data*)

```
Verify that all elements in the data have the same number of attributes.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.1.2 def LovelyCorpus.checkDuplicates (*data*, *remove* = `False`, *fname* = `None`)

```
Check the data for an duplicate entries.

If the remove flag is set, duplicates are removed and the data is
written to the file name provided.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.1.3 def LovelyCorpus.checkRange (*data*, *minval*, *maxval*, *checklist* = `[]`, *clposition* = `LAST`, *remove* = `False`, *truncate* = `False`)

```
Verify that the attributes are with the provided range.

If no checklist is provided all attributes except the class identifier
are checked against the provided range.  If the remove flag is set, any line
containing a value outside the range are removed.  If the truncate flag
is set the values will be truncated to fall within the provided range.  If
the remove or truncate flag are not set, a list of the line numbers that contain
out of bounds values is returned.  Non numerical attributes that are included
in the checklist are ignored.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.1.4 def LovelyCorpus.egain (*data*, *attr*, *clposition*)

```
Compute the entropy gain of a given attribute.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.1.5 def LovelyCorpus.entropyGain (*data*, *clposition* = `LAST`)

```
Computes the entropy gain of each attributes and returns a list
of those values.

The list of values match the positions of the attributes in the file.
The entropy gain is weighted by the number of different values that are
possible for the attribute.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.1.6 def LovelyCorpus.kfold (*fname, k*)

```
Creates 'k' training files and 'k' test files.

The file referenced by 'fname' can be relative or absolute path name.
The names of the training and test files will be based on the provided
'fname'.  Any suffix on the provided file name will be discarded and
replaced with either "train" or "test".

Example:
    k = 2
    fname = "data.txt"
    output = ["data0.train", "data0.test", "data1.train", "data1.test"]

A list of filenames will be returned.
```

4.2.1.7 def LovelyCorpus.removeOutliers (*data, amount, clposition* = LAST, *newfname* = None)

```
Remove a given amount of data outliers.

The amount provided can either be an integer value or a decimal value.
If it is an integer that number of outliers are removed.  If it is a
decimal that percentage of outliers are removed.  Outliers are
determined using their Euclidean distance from the mean for their class.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.1.8 def LovelyCorpus.setIntRange (*data, mn, mx, clposition* = LAST, *attr* = [], *newfname* = None)

```
Change attributes so that they fall within the given range.

All attributes are changed to be integers that fall within the range.
The range is defined as inclusive of the min and max values provided.
i.e. [mn, mx]. The index of the classifier for each entry is stored
in the clposition variable.  If only certain attributes should be
changed a list of indices can be included in the attr variable. The
default behavior is to overwrite the existing file, but a new file
name can be included.

The 'data' variable can be a file name (example.txt), a list of comma
separated strings (["1,2,3", "4,5,6", ...]), or a list of already separated
values ([[1, 2, 3], [4, 5, 6], ...]).
```

4.2.2 Variable Documentation

4.2.2.1 int LovelyCorpus.FIRST = 0

4.2.2.2 int LovelyCorpus.LAST = -1

Chapter 5

File Documentation

5.1 /u/dleblanc/LovelyCorpus/FlogCorpus.py File Reference

Namespaces

- FlogCorpus

Functions

- def FlogCorpus.runTests

Variables

- string FlogCorpus.SEPERATE = "**********"
- list FlogCorpus.ALLTESTS = ["kfold", "setIntRange", "entropyGain", "checkRange"]

5.2 /u/dleblanc/LovelyCorpus/LovelyCorpus.py File Reference

Namespaces

- LovelyCorpus

Functions

- def LovelyCorpus.attributeCount
- def LovelyCorpus.removeOutliers
- def LovelyCorpus.checkDuplicates
- def LovelyCorpus.setIntRange
- def LovelyCorpus.kfold
- def LovelyCorpus.entropyGain
- def LovelyCorpus.egain
- def LovelyCorpus.checkRange

Variables

- int LovelyCorpus.FIRST = 0
- int LovelyCorpus.LAST = -1

5.3 /u/dleblanc/LovelyCorpus/README.md File Reference

Index