
KUHERD Documentation

Release 1.0

Joseph St.Amand

May 31, 2017

Contents:

1	KUHERD package	1
1.1	Subpackages	1
1.1.1	KUHERD.Experiments package	1
1.1.1.1	Submodules	1
1.1.1.2	KUHERD.Experiments.MultiDT module	1
1.1.1.3	KUHERD.Experiments.MultiLR module	1
1.1.1.4	KUHERD.Experiments.MultiNB module	1
1.1.1.5	KUHERD.Experiments.MultiRF module	1
1.1.1.6	KUHERD.Experiments.MultiSVM module	2
1.1.1.7	Module contents	2
1.1.2	KUHERD.Tools package	2
1.1.2.1	Submodules	2
1.1.2.2	KUHERD.Tools.makePredictions module	2
1.1.2.3	KUHERD.Tools.trainModel module	2
1.1.2.4	Module contents	2
1.2	Submodules	2
1.3	KUHERD.FeatureSelector module	2
1.4	KUHERD.HerdVectorizer module	2
1.5	KUHERD.LabelTransformer module	4
1.6	KUHERD.Models module	5
1.7	KUHERD.MultiFeatureSelector module	6
1.8	Module contents	6
2	Indices and tables	7
	Python Module Index	9
	Index	11

1.1 Subpackages

1.1.1 KUHERD.Experiments package

1.1.1.1 Submodules

1.1.1.2 KUHERD.Experiments.MultiDT module

`KUHERD.Experiments.MultiDT.MultiDT()`
Program for running an experiment using Decision Tree classifier.

1.1.1.3 KUHERD.Experiments.MultiLR module

`KUHERD.Experiments.MultiLR.MultiLR()`
Program for running an experiment using Logistic Regression classifier.

1.1.1.4 KUHERD.Experiments.MultiNB module

`KUHERD.Experiments.MultiNB.MultiNB()`
Program for running an experiment using Naive Bayes classifier.

1.1.1.5 KUHERD.Experiments.MultiRF module

`KUHERD.Experiments.MultiRF.MultiRF()`
Program for running an experiment using Random Forest classifier.

1.1.1.6 KUHERD.Experiments.MultiSVM module

`KUHERD.Experiments.MultiSVM.MultiSVM()`

Program for running an experiment using Support Vector Machine classifier.

1.1.1.7 Module contents

1.1.2 KUHERD.Tools package

1.1.2.1 Submodules

1.1.2.2 KUHERD.Tools.makePredictions module

`KUHERD.Tools.makePredictions.main()`

1.1.2.3 KUHERD.Tools.trainModel module

`KUHERD.Tools.trainModel.main()`

1.1.2.4 Module contents

1.2 Submodules

1.3 KUHERD.FeatureSelector module

class `KUHERD.FeatureSelector.FeatureSelector` (*scoring_function, kbest*)

Bases: object

fit (*X, Y, label_set*)

Fits the data by training the feature selection model compomnent.

Args: *X* (numpy matrix): The data matrix. *y* (integer list): The labels for the data. *label_set* (str): Either 'purpose' or 'field'.

Return: None

transform (*X*)

Transforms the data, retaining only features learned in the "fit" process.

Args: *X* (numpy matrix): The data matrix. *y* (integer list): The labels for the data.

Return: (numpy matrix): Transformed data matrix.

1.4 KUHERD.HerdVectorizer module

class `KUHERD.HerdVectorizer.HerdVectorizer` (*config*)

Bases: object

Main class responsible for vectorization of the text data. This class is extremely configurable, with many options for each preprocessing behavior, bigrams, stemmers, stopwords, and feature selection. Use of this class is done in the following manner: - Set configuration options for *preproc_config*, *bigram_config*, *stemmer*, *stopwords*,

and feature selection. - Train the vectorizer on a set of documents and their corresponding labels. - After training is complete, the documents may be given to the transform function to convert to TFIDF form.

create_bigram_index_map (*tokenized_docs*)

Creates a mapping from each bigram to a column index.

Args: *tokenized_docs* (list): A list of documents, each document is represented as a single long string

Return: (dictionary): Dictionary where keys are tokens, values are the index into a feature matrix.

create_token_index_map (*tokenized_docs*)

Given the tokenized documents, finds all unique tokens and forms an index map

Args: *tokenized_docs* (list): A list of documents, each document is represented as a single long string

Return: (dictionary): Dictionary where keys are tokens, values are the index into a feature matrix.

static filter_docs (*tokenized_docs*, *tok_index_map*)

Filters tokenized documents, removing all tokens which are not recognized by the specified token index map.

Args: *tokenized_docs* (list): A list of documents, each document is represented as a single long string
tok_index_map (dictionary): A mapping from tokens to their index in the feature matrix.

Return: (list): A list of documents, where each document is a list of (filtered) tokens.

form_bigram_count_matrix (*tokenized_docs*)

Calculates a bigram count matrix from a list of tokenized documents.

Args: *tokenized_docs* (list): A list of documents, each document is represented as a single long string

Return: (sparse numpy matrix): A sparse count matrix in COO format.

form_count_matrix (*tokenized_docs*)

Forms the count matrix from the tokenized documents

Args: *tokenized_docs* (list): A list of lists representing the tokenized documents. Each document is a list of tokens.

Return: (sparse numpy matrix): A sparse count matrix in COO format.

getConfig ()

Retrieve the complete configuration needed to build a vectorizer.

Returns the complete configuration needed to build a vectorizer. Note that the config returned may only be used to train a new vectorizer. The config does NOT give model persistence.

get_bigram_config ()

Retrieve the bigram configuration.

get_preproc_config ()

Retrieve the preprocessor configuration.

static lancaster_stemmer (*docs*)

Lancaster stemming algorithm

static porter_stemmer (*docs*)

Porter stemming algorithm

set_bigram_config (*name*, *value*)

Set the bigram configuration.

set_bigrams (*bigrams*, *bigram_window_size*, *bigram_filter_size*, *bigram_nbest*)

Set the bigram configuration.

set_feature_selector (*scoring_func, kbest, multi_type*)

Set the feature selection configuration values.

set_preproc_config (*name, value*)

Set the preprocessor configuration value.

set_stemmer (*the_stemmer*)

Set the stemmer configuration values.

static snowball_stemmer (*docs*)

Snowball stemming algorithm

tokenize_docs (*docs*)

Breaks each document down into a list of words(tokens).

Converts a list of documents(each document is given as a single string) and converts them to their tokenized form in the following manner(some steps may be skipped if configured as such in the configuration settings)

- break document into tokens
- remove punctuation
- stem tokens

Args: docs (list): A list of documents, each document is represented as a single long string

Return: (list): The tokenized documents as a list of lists, each item of the outer list is a document, which is represented as a list of words.

train (*docs, y, label_set*)

Takes a list of documents, and the corresponding labels and trains the preprocessor(including feature selection).

Args: docs (list): A list of documents, where each document is represented as a string. y (list): A list of integers representing the label for each document. label_set (str): Specifies the label set so that the input 'y' may be interpreted. Valid entries are either 'purpose' or 'field'.

@param docs The list of documents @param y A vector of labels which correspond to each document

transform_data (*docs*)

Transforms documents into a sparse matrix.

Must be called after the preprocessor has been trained on some data. Process is as follows: -
tokenize documents -search for bigrams -transform to TFIDF representation -select features

Args: docs (list): A list of documents, each document is represented as a string.

Return: (sparse numpy matrix): A sparse CSR formatted matrix, each row corresponds to a document, ordering of documents is preserved.

KUHERD.HerdVectorizer.**main**()

1.5 KUHERD.LabelTransformer module

class KUHERD.LabelTransformer.**LabelTransformer** (*label_set_name, labels*)

Bases: object

default_labels (*target_set*)

label2mat (*x*)

Converts a vector of strings to a matrix of of zero-one valued columns.

Args: *x* (list): A vector containing string values representing the labels.

Return: (numpy matrix): A matrix of zero-one valued columns.

mat2vec (*M*)

Converts a zero-one valued label matrix to an integer valued label vector.

Args: *M* (numpy mat): A zero-one valued label matrix.

vec2mat (*x*)

Converts a vector of integers to a matrix of of zero-one valued columns.

Args: *x* (list): A vector containing integer values mapping to members of the *label_type*.

Return: (numpy matrix): A matrix of zero-one valued columns.

vec2string (*label_vec*)

Converts vector containing integers to a string representation using the label set dictionaries. **Args:**

label_vec (list): A vector containing integer values mapping to members of the *label_type*.

Return: (list): A list of strings that are members of the *label_type*.

1.6 KUHERD.Models module

class KUHERD.Models.**ClassificationModel** (*config*)

Bases: object

fit (*X*, *Y*)

Trains the model.

Fitting or “training” must be done before the model is able to make predictions.

Args: *X* (numpy matrix): Training samples. *Y* (numpy matrix): Training labels.

Returns: None: No return value.

get_config ()

Returns the configuration used to build this model.

Returns: dict: dictionary containing target label set, internal model configuration, and model name.

predict (*X*)

Make predictions.

Args: *X* (numpy matrix): Training samples.

Returns: numpy matrix: predicted label values.

class KUHERD.Models.**PurposeFieldModel** (*config*)

Bases: object

fit (*abstracts*, *Y_purpose*, *Y_field*)

Trains the model.

Input arguments must all be the same length.

Args: *abstracts* (list): A list of documents, each document is represented as a list of words. *Y_purpose* (list): A list of labels of the ‘purpose’ variety. *Y_field* (list): A list of labels of the ‘field’ variety.

get_config()

Returns the configuration used to build this model.

Returns: dict: dictionary containing the following keys, 'purpose_vectorizer', 'field_vectorizer', 'purpose_model', 'field_model'. Each entry is the configuration required to build the model.

predict (*abstracts*)

Make predictions on the input data.

The list of documents input is vectorized and input to the prediction model, which generates label predictions. This process is done separately for generating both purpose and field label predictions.

Args: abstracts (list): A list of documents, each document is represented as a list of words.

Returns: dictionary: dictionary containing two lists of predictions, dictionary keys are 'purpose' and 'field'.

1.7 KUHERD.MultiFeatureSelector module

class KUHERD.MultiFeatureSelector.**MultiFeatureSelector** (*scoring_function*, *kbest*, *multi_integrator*)

Bases: object

fit (*X*, *Y*, *label_set*)

Trains the feature selection process

Args: X (numpy matrix): Training samples. Y (numpy matrix): Training Labels. label_set (str): Denotes if label set is of the 'purpose' or 'field' type.

transform (*X*)

Transforms the data by selecting the features learned in the training or "fit" process.

Args: X (numpy matrix): Data samples to run feature selection on.

Return: (numpy matrix): Data with only the selected features.

1.8 Module contents

CHAPTER 2

Indices and tables

- `genindex`
- `modindex`
- `search`

k

- KUHERD, [6](#)
- KUHERD.FeatureSelector, [2](#)
- KUHERD.HerdVectorizer, [2](#)
- KUHERD.LabelTransformer, [4](#)
- KUHERD.Models, [5](#)
- KUHERD.MultiFeatureSelector, [6](#)

C

ClassificationModel (class in KUHERD.Models), 5
 create_bigram_index_map()
 (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3
 create_token_index_map()
 (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3

D

default_labels() (KUHERD.LabelTransformer.LabelTransformer
 method), 4

F

FeatureSelector (class in KUHERD.FeatureSelector), 2
 filter_docs() (KUHERD.HerdVectorizer.HerdVectorizer
 static method), 3
 fit() (KUHERD.FeatureSelector.FeatureSelector method),
 2
 fit() (KUHERD.Models.ClassificationModel method), 5
 fit() (KUHERD.Models.PurposeFieldModel method), 5
 fit() (KUHERD.MultiFeatureSelector.MultiFeatureSelector
 method), 6
 form_bigram_count_matrix()
 (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3
 form_count_matrix() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3

G

get_bigram_config() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3
 get_config() (KUHERD.Models.ClassificationModel
 method), 5
 get_config() (KUHERD.Models.PurposeFieldModel
 method), 5
 get_preproc_config() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3

getConfig() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3

H

HerdVectorizer (class in KUHERD.HerdVectorizer), 2

K

KUHERD (module), 6
 KUHERD.Experiments (module), 2
 KUHERD.Experiments.MultiDT (module), 1
 KUHERD.Experiments.MultiLR (module), 1
 KUHERD.Experiments.MultiNB (module), 1
 KUHERD.Experiments.MultiRF (module), 1
 KUHERD.Experiments.MultiSVM (module), 2
 KUHERD.FeatureSelector (module), 2
 KUHERD.HerdVectorizer (module), 2
 KUHERD.LabelTransformer (module), 4
 KUHERD.Models (module), 5
 KUHERD.MultiFeatureSelector (module), 6
 KUHERD.Tools (module), 2
 KUHERD.Tools.makePredictions (module), 2
 KUHERD.Tools.trainModel (module), 2

L

label2mat() (KUHERD.LabelTransformer.LabelTransformer
 method), 4

LabelTransformer (class in
 KUHERD.LabelTransformer), 4

lancaster_stemmer() (KUHERD.HerdVectorizer.HerdVectorizer
 static method), 3

M

main() (in module KUHERD.HerdVectorizer), 4
 main() (in module KUHERD.Tools.makePredictions), 2
 main() (in module KUHERD.Tools.trainModel), 2
 mat2vec() (KUHERD.LabelTransformer.LabelTransformer
 method), 5
 MultiDT() (in module KUHERD.Experiments.MultiDT),
 1

MultiFeatureSelector (class in
 KUHERD.MultiFeatureSelector), 6
MultiLR() (in module KUHERD.Experiments.MultiLR),
 1
MultiNB() (in module KUHERD.Experiments.MultiNB),
 1
MultiRF() (in module KUHERD.Experiments.MultiRF),
 1
MultiSVM() (in module
 KUHERD.Experiments.MultiSVM), 2

P

porter_stemmer() (KUHERD.HerdVectorizer.HerdVectorizer
 static method), 3
predict() (KUHERD.Models.ClassificationModel
 method), 5
predict() (KUHERD.Models.PurposeFieldModel
 method), 6
PurposeFieldModel (class in KUHERD.Models), 5

S

set_bigram_config() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3
set_bigrams() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3
set_feature_selector() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 3
set_preproc_config() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 4
set_stemmer() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 4
snowball_stemmer() (KUHERD.HerdVectorizer.HerdVectorizer
 static method), 4

T

tokenize_docs() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 4
train() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 4
transform() (KUHERD.FeatureSelector.FeatureSelector
 method), 2
transform() (KUHERD.MultiFeatureSelector.MultiFeatureSelector
 method), 6
transform_data() (KUHERD.HerdVectorizer.HerdVectorizer
 method), 4

V

vec2mat() (KUHERD.LabelTransformer.LabelTransformer
 method), 5
vec2string() (KUHERD.LabelTransformer.LabelTransformer
 method), 5