A Model for Age and Gender Profiling of Social Media Accounts Based on Post Contents Documentation

Release

Cheng, Fernandez, Quindoza, Tan

August 15, 2017

Table of Contents

CHAPTER 1

thesis

1.1 Driver50 module

Driver50.**clean** (*x*)

Parameters **x** -- data to be cleaned

Returns cleaned email and links from the data

Driver50.**dimensionReduction** (*X*, *y*, *source*, *mindf*, *maxdf*, *data=None*)

perform dimension reduction

Parameters

- **X** -- text data
- **y** -- classes (gender and age)
- **source** -- twitter, facebook, or merged
- **mindf** -- lower threshold for term frequency filter
- **maxdf** -- upper threshold for term frequency filter
- **data** -- features of the data

Driver50.**evaluate** (*age_data*, *gen_data*, *both_data*, *model*)

Evaluates the age and gender profiling performance of the model (various model structures)

Parameters

- **age_data** -- data feature selected based on age
- **gen_data** -- data feature selected based on gender
- **both_data** -- data feature selected based on age and gender
- **model** -- classifier to be used

Returns

Driver50.**execute** ( )

entire process to execute. Includes feature extraction, dimension reduction, and evaluation

Driver50.**getSpecificFeatures** (*data*, *features*)

filters the features

Parameters

- **data** -- features of the data

- **features** -- specific features to be retrieved

Returns specified features

Driver50.**get_Data_from_CSV** (*source*, *mindf*, *maxdf*, *fs*, *param=None*)

Parameters
- **source** -- twitter, facebook, or merged
- **mindf** -- lower threshold for term frequency filter
- **maxdf** -- upper threshold for term frequency filter
- **fs** -- feature selection method
- **param** -- parameter used by the feature selection

Returns age, gender, and both (combined structure) data

Driver50.**writeToExcel** (*book*, *sheet*, *classifier*, *features*, *row*)

1.2 addEngPOS module

class addEngPOS.**ConnectionFactory**

Bases: object

getConnectionThesis ( )

addEngPOS.**add_english_pos** ( )

adds the english pos to the d :return:

1.3 batchprocessing module

batchprocessing.**getPosts** ( )

batchprocessing.**getPostsFromFile** (*filepath*)

batchprocessing.**updateEngPOS** (*ids*, *texts*)

batchprocessing.**updatePosts** (*ids*, *posts*)

batchprocessing.**writePostsToFile** (*posts*, *filepath*)

1.4 combinepos module

combinepos.**combinePOS** ( )

populate the texts' combined POS

1.5 features package

1.5.1 Submodules

1.5.2 features.CharacterFeatures module

class features.CharacterFeatures.**CharacterFeatures**

Bases: object

Returns the character features of a text

getNumberOfRepeatedPunctuationMarks (*text*)

Parameters **text** -- text to be processed

Returns total number of instances of consecutive punctuation marks

getNumberOfRepetitiveAlphaCharacters (*text*)

Parameters **text** -- text to be processed

Returns total number of instances that alpha characters are repeated more than twice consecutively

getNumberOfSpecialChars (*text*)

Parameters **text** -- text to be processed

Returns total number of special characters besides punctuation marks

getNumberOfWhiteSpaces (*text*)

Parameters **text** -- text to be processed

Returns total number of white spaces

getTotalNumberOfCharacters (*text*)

Parameters **text** -- text to be processed

Returns total number ofcharacters

getTotalNumberOfDigitalNumbers (*text*)

Parameters **text** -- text to be processed

Returns total number of digital numbers

getTotalNumberOfLetters (*text*)

Parameters **text** -- text to be processed

Returns total number of letters

getTotalNumberOfUppercase (*text*)

Parameters **text** -- text to be processed

Returns total number of uppercase letters

1.5.3 features.Context module

class features.Context.**Context**
Bases: object
Returns the contextual features (words after 'my') in a text

process (*s*)

Parameters **s** -- text to be processed

Returns text containing the contextual features

1.5.4 features.EmojisEmoticons module

class features.EmojisEmoticons.**EmojisEmoticons**
Bases: object

getEmojiTFIDF (*data*)

getLabels ( )

1.5.5 features.Feature module

class features.Feature.**Feature** (*X*, *y*, *source*, *data=None*)
Bases: object
Applies dimension reduction to data

applyExtraction (*selection*)
applies feature selection

Parameters
- **selection** -- feature extraction technique
- **type** -- Gender, Age, or Both

Returns feature extracted data

applySelection (*selection*, *type*)
applies feature selection

Parameters
- **selection** -- feature selection technique
- **type** -- Gender, Age, or Both

Returns feature selected data

getFeatures (*selection*, *mode*)
applies feature selection or extraction

Parameters
- **selection** -- feature selection or extraction technique
- **mode** -- Gender, Age, or Both

Returns feature selected or extracted data

useLasso (*mode*)
applies LASSO feature selection

Parameters **selection** -- feature selection or extraction technique

Returns feature selected data

1.5.6 features.FeatureExtract module

class features.FeatureExtract.**FeatureExtract** (*source*, *mindf*, *maxdf*)

Bases: object

Extracts features from the text and post time

clean (*x*)

cleans the data

Parameters **x** -- text data

Returns cleaned text

fit_transform (*X*)

Parameters **X** -- text data

Returns dataframe containing features extracted

get_liwc ( )

reads the LIWC csv files

Returns dataframe containing the LIWC results

transform (*X*)

The transform is only done after fitting the data, useful for TFIDF features

Parameters **X** -- text data

Returns dataframe containing features extracted

1.5.7 features.FunctionWordCount module

class features.FunctionWordCount.**FunctionWordCount**

Bases: object

FUNCTIONWORDS_FILENAME = 'features/functionwords.txt'

getAdpositionCount (*text*)

getAllFunctionWordCount (*text*)

getArticleCount (*text*)

getAuxillaryCount (*text*)

getConjunctionCount (*text*)

getInterjectionCount (*text*)

getProSentenceCount (*text*)

getPronounCount (*text*)

1.5.8 features.Links module

class features.Links.**Links**
Bases: object

get_keywords (*link*)

get_links (*text*)

get_list_keywords (*text*)

get_title (*link*)

1.5.9 features.POSFeature module

class features.POSFeature.**POSFeature**
Bases: object

ADJECTIVE = 'JJ'

UNKNOWN = 'UNK'

VERB = 'VB'

getCombinedPOSTag (*post*)

getEnglishPOS (*text*)
jvmPath = jpype.getDefaultJVMPath() jpype.startJVM(jvmPath, "-Djava.class.path=dependencies/NormAPI.jar;dependencies/RBPOST.jar") rbpost = JPackage("rbpost").RBPOST result = rbpost.tokenizer_Text(text) tokenizedText = result.split(" ") jpype.shutdownJVM()

getPOSCount (*text*)

populateMappingDictionary ( )

1.5.10 features.POSSequencePattern module

class features.POSSequencePattern.**POSSequencePattern** (*documentList*)
Bases: object

MAX_LENGTH = 7

candidateGen (*fList*)

computeFairSCP (*key*, *count*)

minePOSPatterns (*minsup*, *minadherence*)

retrievePOSTags_docFrequecy ( )

1.5.11 features.Structure module

class features.Structure.**Structure**
Bases: object

ABBREVIATIONS_FILENAME = '../features/abbreviations.txt'

getAvgNCharacterPerParagraph (*text*)

getAvgNSentencePerParagraph (*text*)

getAvgNWordPerParagraph (*text*)

getAvgNWordPerSentence (*text*)

getNParagraphs (*text*)

getNSentenceBegLower (*text*)

getNSentenceBegUpper (*text*)

getNSentences (*text*)

getParagraphs (*text*)

1.5.12 features.TFIDF module

class features.TFIDF.**TFIDF** (*mindf*, *maxdf*)

Bases: object

Processes the TFIDF of text

getFeatureNames ( )

Returns labels of the features

get_testing_TFIDF (*test*)

Parameters **documentList** -- testing text data

Returns tfidf of the text

get_training_TFIDF (*documentList*)

Parameters **documentList** -- training text data

Returns tfidf of the text

1.5.13 features.WordCount module

class features.WordCount.**WordCount**

Bases: object

ABBREVIATIONS_FILENAME = 'features/abbreviations.txt'

getAveLengthWords (*text*)

getDictOfWordsMappedToOccurrence (*text*)

getEntropy (*text*)

getHapaxDislegomena (*text*)

getHapaxLegomena (*text*)

getHonoresR (*text*)

getLolHmmCount (*text*)

getNDifferentWords (*text*)

getNWordsBegCapital (*text*)

getNWordsWithRepLetters (*text*)

getOccurrenceArray (*text*)

getRatioOfHapaxDislegomena (*text*)

getRatioOfHapaxLegomena (*text*)

getRatioOfNetAbbrev (*text*)

getRatioOfShortWords (*text*)

getRatioOfUniqueWords (*text*)

getSichelsS (*text*)

getSimpsonsD (*text*)

getTotalNumberOfWords (*text*)

getWordLengthFreqDist (*text*)

getYulesK (*text*)

1.5.14 Module contents

1.6 model package

1.6.1 Submodules

1.6.2 model.Document module

class model.Document.**Document** (*content*, *posSequence*)
Bases: object

1.6.3 model.Post module

class model.Post.**Post** (*id*, *content*, *epos*, *fpos*)
Bases: object

1.6.4 model.RootModel module

class model.RootModel.**RootModel** (*data*, *type*, *modelType*, *k=10*)
Bases: object

This class represents the parallel and combined structure. Its results can be fed to the StackModel for the stacked model structure.

evaluateKfold (*train_predictions=None*, *test_predictions=None*)

Parameters
- **train_predictions** -- predictions of the model for the training data
- **test_predictions** -- predictions of the model for the testing data

Returns returns the metrics for both training data and testing data

getPredictions ( )

Returns the predictions of the model for training and testing data

getTestingUser (*ind*)

Parameters **ind** -- k-fold index

Returns users for the testing data for the ith k-fold

getTestingX (*ind*)

Parameters **ind** -- k-fold index

Returns testing data for the ith k-fold

getTestingy (*ind*)

Parameters **ind** -- k-fold index

Returns testing results for the ith k-fold

getTrainingUser (*ind*)

Parameters **ind** -- k-fold index

Returns users for the training data for the ith k-fold

getTrainingX (*ind*)

Parameters **ind** -- k-fold index

Returns training data for the ith k-fold

getTrainingy (*ind*)

Parameters **ind** -- k-fold index

Returns training results for the ith k-fold

1.6.5 model.StackModel module

class model.StackModel.**StackModel** (*root*, *modelType*, *data*, *type*, *k=10*)

Bases: object

This class represents the stacked structure.

evaluateKfold (*train_predictions=None*, *test_predictions=None*)

Parameters
- **train_predictions** -- predictions of the model for the training data
- **test_predictions** -- predictions of the model for the testing data

Returns returns the metrics for both training data and testing data

getPredictions ( )

Returns the predictions of the model for training and testing data

getTestingUser (*ind*)

Parameters **ind** -- k-fold index

Returns users for the testing data for the ith k-fold

getTestingX (*ind*)

Parameters **ind** -- k-fold index

Returns testing data for the ith k-fold

getTestingy (*ind*)

Parameters **ind** -- k-fold index

Returns testing results for the ith k-fold

getTrainingUser (*ind*)

Parameters **ind** -- k-fold index

Returns users for the training data for the ith k-fold

getTrainingX (*ind*)

Parameters **ind** -- k-fold index

Returns training data for the ith k-fold

getTrainingy (*ind*)

Parameters **ind** -- k-fold index

Returns training results for the ith k-fold

1.6.6 Module contents

1.7 pipelinewraps package

1.7.1 Submodules

1.7.2 pipelinewraps.AgeRangeWrap module

class pipelinewraps.AgeRangeWrap.**AgeRangeWrap**

Bases: sklearn.base.TransformerMixin

Transforms the age to numerical labels TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X, y=None, **fit_params*)

transform (*X, **transform_params*)

pipelinewraps.AgeRangeWrap.**enrange** (*x*)

Parameters **x** -- age of the user

Returns age range group

pipelinewraps.AgeRangeWrap.**getClasses** ( )

Returns array of the age ranges

1.7.3 pipelinewraps.CharacterWrap module

class pipelinewraps.CharacterWrap.**CharacterWrap**

Bases: sklearn.base.TransformerMixin

Processes all character features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, *y=None*, ***transform_params*)

1.7.4 pipelinewraps.ContextualWrap module

class pipelinewraps.ContextualWrap.**ContextualWrap** (*target=None*)

Bases: sklearn.base.TransformerMixin

Processes all contextual features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, **args*, ***kwargs*)

transform (*X*, *y=None*, ***transform_params*)

1.7.5 pipelinewraps.EmojiWrap module

class pipelinewraps.EmojiWrap.**EmojiWrap** (*target=None*)

Bases: sklearn.base.TransformerMixin

Processes all emoji features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, **args*, ***kwargs*)

transform (*X*, *y=None*, ***transform_params*)

1.7.6 pipelinewraps.ExtractionWrap module

class pipelinewraps.ExtractionWrap.**ExtractionWrap** (*extraction*, *target=None*)

Bases: sklearn.base.TransformerMixin

Performs feature extraction

fit (*X*, **args*, ***kwargs*)

transform (*X*, *y=None*, ***transform_params*)

1.7.7 pipelinewraps.FunctionWrap module

class pipelinewraps.FunctionWrap.**FunctionWrap**

Bases: sklearn.base.TransformerMixin

Processes all function word features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, *y=None*, ***transform_params*)

1.7.8 pipelinewraps.GenderWrap module

class pipelinewraps.GenderWrap.**GenderWrap**

Bases: sklearn.base.TransformerMixin

Transforms the gender to numerical labels TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, ***transform_params*)

pipelinewraps.GenderWrap.**enrange** (*x*)

Parameters **x** -- gender of the user

Returns 0 for F and 1 for M

pipelinewraps.GenderWrap.**getClasses** ( )

Returns returns the gender classes

1.7.9 pipelinewraps.ItemSelector module

class pipelinewraps.ItemSelector.**ItemSelector** (*key*)

Bases: sklearn.base.BaseEstimator, sklearn.base.TransformerMixin

fit (*x*, *y=None*)

transform (*data_dict*)

1.7.10 pipelinewraps.LinkWrap module

class pipelinewraps.LinkWrap.**LinkWrap** (*target=None*)

Bases: sklearn.base.TransformerMixin

Processes all link features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, **args*, ***kwargs*)

transform (*X*, *y=None*, ***transform_params*)

1.7.11 pipelinewraps.POSSeqWrap module

class pipelinewraps.POSSeqWrap.**POSSeqWrap**

Bases: sklearn.base.TransformerMixin

Processes all POS features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, *y=None*, ***transform_params*)

pipelinewraps.POSSeqWrap.**dfToDocument** (*df*)

1.7.12 pipelinewraps.PostTimeWrap module

class pipelinewraps.PostTimeWrap.**PostTimeWrap**

Bases: sklearn.base.TransformerMixin

Processes all word features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, ***transform_params*)

pipelinewraps.PostTimeWrap.**enrange** (*x*)

Parameters **x** -- exact hour posted

Returns time group

pipelinewraps.PostTimeWrap.**getClasses** ( )

Returns returns the post time classes

1.7.13 pipelinewraps.SelectionWrap module

class pipelinewraps.SelectionWrap.**SelectionWrap** (*selection*, *target=None*)

Bases: sklearn.base.TransformerMixin

Performs feature selection

fit (*X*, *y*, **args*, ***kwargs*)

transform (*X*, *y=None*, ***transform_params*)

1.7.14 pipelinewraps.StackAgeRangeWrap module

class pipelinewraps.StackAgeRangeWrap.**StackAgeRangeWrap**

Bases: sklearn.base.TransformerMixin

Transforms the age multiclass to multilabel binary TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, ***transform_params*)

pipelinewraps.StackAgeRangeWrap.**getClasses** ( )

Returns array of the age ranges

1.7.15 pipelinewraps.StackGenderWrap module

class pipelinewraps.StackGenderWrap.**StackGenderWrap**

Bases: sklearn.base.TransformerMixin

Transforms the gender multiclass to multilabel binary TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, ***transform_params*)

pipelinewraps.StackGenderWrap.**getClasses** ( )

Returns returns the gender classes

1.7.16 pipelinewraps.StructureWrap module

class pipelinewraps.StructureWrap.**StructureWrap**

Bases: sklearn.base.TransformerMixin

Processes all structure features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, *y=None*, ***transform_params*)

1.7.17 pipelinewraps.WordWrap module

class pipelinewraps.WordWrap.**WordWrap**

Bases: sklearn.base.TransformerMixin

Processes all word features of the data. TransformerMixin gives it the standard fit and transform functions to transform the data

fit (*X*, *y=None*, ***fit_params*)

transform (*X*, *y=None*, ***transform_params*)

1.7.18 Module contents

1.8 prepareedsthesis module

class prepareedsthesis.**ConnectionFactory**

Bases: object

getConnectionThesis ( )

prepareedsthesis.**addposts** ( )

prepareedsthesis.**addusers** (*limit=None*)

1.9 utility package

1.9.1 Submodules

1.9.2 utility.DataCleaner module

class utility.DataCleaner.**DataCleaner**
Bases: object

URL = 'URL'

USERNAME = 'USERNAME'

clean_data (*post_content*)

clean_email (*post_content*)

1.9.3 utility.LanguageDetector module

class utility.LanguageDetector.**Language**
Bases: object

ENGLISH = 0

FILIPINO = 1

TAGLISH = 2

UNKNOWN = -1

getLanguage (*code*)

class utility.LanguageDetector.**LanguageDetector**
Bases: object

englishOrTagalog (*string*)

getLanguage (*text*)

getLanguageDetailed (*text*)

1.9.4 utility.PostCleaner module

class utility.PostCleaner.**PostCleaner**
Bases: object

changeEmojisToText (*postContent*)

changeForeignToText (*postContent*)

changeLinkToText (*postContent*)

fixAcronymSpaces (*postContent*)

getEmojis (*postContent*)

insertSpace (*postContent*)

normalizeUnicode (*postContent*)

removeEmojis (*postContent*)

1.9.5 Module contents

- *Index*
- *Module Index*
- *Search Page*

F

G

I

L

M

N

P

R