Documentation

## preprocessing

preprocessing(df, remove\_useless\_sentences = False, tokenize = False, remove\_one\_characters = False, lemmatize = False, remove\_stop\_words = False)

returns:

* just the df if tokenize is False
* the df and x (tokenized version of the synopsis) if tokenize is True

All the parameters are used to choose the operation to perform on the original data.

Example:

df\_train, x\_train = functions.preprocessing(df=df\_train, remove\_useless\_sentences = True, tokenize=True, remove\_one\_characters=True, lemmatize=True, remove\_stop\_words=True)

## target\_variable

target\_variable(df, col)

Takes as input the df and returns the target variable array y with the target labels. The output is a list-like array.

Example:

y\_train = functions.target\_variable(df\_train, 'genre')

print(y\_train)

## tokenizer\_padding

tokenizer\_padding(x\_train, max\_length, x\_test = False)

Returns the padded sequences and the tokenizer used to perform these tasks. If x\_test is set to True, returns the padded test matrix too.

Example:

padded\_train, padded\_test, tokenizer = functions.tokenizer\_padding(x\_train=x\_train, max\_length=30, x\_test=x\_test)

print(padded\_train, "\n\n", padded\_test, len(tokenizer.word\_index))

## create\_lexicon

create\_lexicon(df\_train, genre\_col, clean\_synopsis\_col)

Creates a lexicon for all the genres in the genre\_col column for the df. The clean\_synopsis\_col is the column with the clean tokenized synopsis to use, it is a **list-like** input ( be careful to not use strings here!)

## merge\_genres

merge\_genres(df, to\_merge, inplace=True)

Given a to\_merge dictionary, returns the df with the (key,value) merged genres. It can be done inplace or not, if not it will return both the actual genre and the mapped genres column. An example:

to\_merge = {'horror':'thriller', 'crime':'action'}

df = functions.merge\_genres(df, to\_merge=to\_merge, inplace=True)

df.head(50)

It maps horror to thriller and crime to action, so there are just 8 out of 10 genres left.