
Abraca-Data Project

Release 0.1

Abraca-Data

Dec 10, 2021

CONTENTS:

1	Data Preprocessing Module Functions	1
2	Indices and tables	5
	Python Module Index	7
	Index	9

DATA PREPROCESSING MODULE FUNCTIONS

`data_preprocessing.data_preprocessing.expand_contractions(text)`
function to expand the contractions from the text

Parameters *text* – given text

Returns expanded contractions

`data_preprocessing.data_preprocessing.extract_urls(text)`
function for extracting the url from text

Parameters *text* – given text

Returns transformed text

`data_preprocessing.data_preprocessing.find_persons(text)`
function for extracting the person names from text

Parameters *text* – given text

Returns transformed text

`data_preprocessing.data_preprocessing.lemmatize_text(text)`
function for lemmatize text

Parameters *text* – given text

Returns transformed text

`data_preprocessing.data_preprocessing.lower_text(text)`
function for converting the text to lower case

Parameters *text* – given text

Returns transformed text

`data_preprocessing.data_preprocessing.named_entity_recognition(text)`
function for named entity recognition

Parameters *text* – given text

Returns dict with named entity entries

`data_preprocessing.data_preprocessing.remove_accented_chars(text)`
function to remove accented characters from text, e.g. café

Parameters *text* – given text

Returns filtered text with accented text

`data_preprocessing.data_preprocessing.remove_html_tags(text)`
function to remove html tags

Parameters **text** – given text

Returns filtered text with removed html components

`data_preprocessing.data_preprocessing.remove_special_characters(text, remove_digits=False)`
function to remove special characters from the given text

Parameters

- **text** – given text
- **remove_digits** – boolean parameter for removing digits

Returns filtered text

`data_preprocessing.data_preprocessing.remove_stopwords(text)`
function to remove stop words from the given text

Parameters **text** – given text

Returns filtered text with stop words

`data_preprocessing.data_preprocessing.remove_urls(text)`
function to remove url from given text

Parameters **text** – given text

Returns filtered text

`data_preprocessing.data_preprocessing.stemming_text(text)`
function for stemming the text

Parameters **text** – given text

Returns transformed text

`data_preprocessing.data_collection_scripts.sitemap_crawlers.cnbc_sitemap()`
Functions return article links and save them in excel file

Returns excel file

`data_preprocessing.data_collection_scripts.sitemap_crawlers.new_york_times_sitemap()`
Functions return article links and save them in excel file

Returns excel file

`class data_preprocessing.database_records.NewsArticles(*args, **values)`

exception **DoesNotExist**

exception **MultipleObjectsReturned**

`class data_preprocessing.database_records.ProcessedNewsArticle(*args, **values)`

exception **DoesNotExist**

exception **MultipleObjectsReturned**

`data_preprocessing.news_articles.extract_rss_feeds(xml_url, *header_value)`
function to extract all article web links from an xml file :param xml_url: xml link

`data_preprocessing.news_articles.parse_article(article_url)`
function which extracts information given a web url

Parameters **article_url** – article url

Returns json record

INDICES AND TABLES

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

PYTHON MODULE INDEX

d

`data_preprocessing.data_collection_scripts.sitemap_crawlers,`
[2](#)
`data_preprocessing.data_preprocessing,` [1](#)
`data_preprocessing.database_records,` [2](#)
`data_preprocessing.news_articles,` [2](#)

INDEX

C

`cnbc_sitemap()` (in module `data_preprocessing.data_preprocessing`, 1
`data_preprocessing.data_collection_scripts.sitemap_crawlers`, 2
`data_preprocessing.database_records`, 2
`data_preprocessing.news_articles`, 2)

D

`data_preprocessing.data_collection_scripts.sitemap_crawlers` (in module `data_preprocessing.data_preprocessing`, 1)
`data_preprocessing.data_preprocessing` (module, 1)
`data_preprocessing.database_records` (module, 2)
`data_preprocessing.news_articles` (module, 2)

E

`expand_contractions()` (in module `data_preprocessing.data_preprocessing`, 1)
`extract_rss_feeds()` (in module `data_preprocessing.news_articles`, 2)
`extract_urls()` (in module `data_preprocessing.data_preprocessing`, 1)

F

`find_persons()` (in module `data_preprocessing.data_preprocessing`, 1)

L

`lemmatize_text()` (in module `data_preprocessing.data_preprocessing`, 1)
`lower_text()` (in module `data_preprocessing.data_preprocessing`, 1)

M

module
`data_preprocessing.data_collection_scripts.sitemap_crawlers`, 2

N

`named_entity_recognition()` (in module `data_preprocessing.data_preprocessing`, 1)
`new_york_times_sitemap()` (in module `data_preprocessing.data_collection_scripts.sitemap_crawlers`, 2)
`NewsArticles` (class in `data_preprocessing.database_records`, 2)
`NewsArticles.DoesNotExist`, 2
`NewsArticles.MultipleObjectsReturned`, 2

P

`parse_article()` (in module `data_preprocessing.news_articles`, 2)
`ProcessedNewsArticle` (class in `data_preprocessing.database_records`, 2)
`ProcessedNewsArticle.DoesNotExist`, 2
`ProcessedNewsArticle.MultipleObjectsReturned`, 2

R

`remove_accented_chars()` (in module `data_preprocessing.data_preprocessing`, 1)
`remove_html_tags()` (in module `data_preprocessing.data_preprocessing`, 1)
`remove_special_characters()` (in module `data_preprocessing.data_preprocessing`, 2)
`remove_stopwords()` (in module `data_preprocessing.data_preprocessing`, 2)
`remove_urls()` (in module `data_preprocessing.data_preprocessing`, 2)

S

`stemming_text()` (in *module*
data_preprocessing.data_preprocessing),
[2](#)