Abraca-Data Project

Release 0.1

Abraca-Data

Dec 10, 2021

CONTENTS:

CHAPTER

ONE

DATA PREPROCESSING MODULE FUNCTIONS

`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

`data_preprocessing.contractions.`**`get_all_contractions`**`()`

Method to get all contraction

Returns dict

`data_preprocessing.data_collection.`**`add_records_to_database`**`(`*database_config_keys*, *source_name*, *article_dataframe*`)`

function to add records in to database

Parameters

- **`database_config_keys`** – database config yaml file
- **`source_name`** – newspaper name
- **`article_dataframe`** – newspaper data

`data_preprocessing.data_preprocessing_functions.`**`expand_contractions`**`(`*text*`)`

function to expand the contractions from the text

Parameters **`text`** – given text

Returns expanded contractions

`data_preprocessing.data_preprocessing_functions.`**`extract_urls`**`(`*text*`)`

function for extracting the url from text

Parameters **`text`** – given text

Returns transformed text

`data_preprocessing.data_preprocessing_functions.`**`find_persons`**`(`*text*`)`

function for extracting the person names from text

Parameters **`text`** – given text

Returns transformed text

`data_preprocessing.data_preprocessing_functions.`**`lemmatize_text`**`(`*text*`)`

function for lemmatize text

Parameters `text` – given text

Returns transformed text

`data_preprocessing.data_preprocessing_functions.`**`lower_text`**`(`*text*`)`

function for converting the text to lower case

Parameters `text` – given text

Returns transformed text

`data_preprocessing.data_preprocessing_functions.`**`named_entity_recognition`**`(`*text*`)`

function for named entity recognition

Parameters `text` – given text

Returns dict with named entity entries

`data_preprocessing.data_preprocessing_functions.`**`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_functions.`**`remove_html_tags`**`(`*text*`)`

function to remove html tags

Parameters `text` – given text

Returns filtered text with removed html components

`data_preprocessing.data_preprocessing_functions.`**`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_functions.`**`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_functions.`**`remove_urls`**`(`*text*`)`

function to remove url from given text

Parameters `text` – given text

Returns filtered text

`data_preprocessing.data_preprocessing_functions.`**`stemming_text`**`(`*text*`)`

function for stemming the text

Parameters `text` – given text

Returns transformed text

`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

Parameters `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

data_preprocessing.processing_records.**process_records**(*database_connection_params*)
function to process database records

Parameters `database_connection_params` – database connection strings

data_preprocessing.rule_based_sentimental_analysis.**text_blob_sentiment**(*given_sentence*)
function to analyze the text blob sentiment

Parameters `given_sentence` – sentence

Returns sentiment

data_preprocessing.rule_based_sentimental_analysis.**vader_sentiment**(*given_sentence*)
function to analyze the vader sentiment

Parameters `given_sentence` – sentence

Returns sentiment

data_preprocessing.utils.**check_record_exist**(*database_credentials*, *article_url*, *processed=False*)

function to check if a record exists in a database or not

Parameters

- `database_credentials` – database credentials
- `article_url` – article url
- `processed` – boolean value to process or not

Returns boolean value

CHAPTER TWO

INDICES AND TABLES

- genindex
- modindex
- search

PYTHON MODULE INDEX

INDEX