Documentation Actes IA

version

Starclay

April 15, 2020

Contents

Welcome to DGCL Actes IA’s documentation!

Document Pipeline

code.utils module

utils.py

Module to manage PDF files

`code.utils.`**`get_file_paths`** (`path`, `recursive=False`, `extension='.pdf'`)

Returns paths for all the files contained in the folder

Parameters:

- **path** – Path to the folder
- **recursive** – if True, will look for pdf files in directory and subdirectories
- **extension** – Extension that files need to have

Returns: A list with the paths to all files with a specific extension in the folder

`code.utils.`**`get_folder`** (`path`)

Get the folder where the file is

Parameters: **path** – Path to the file

Returns: Folder name where the file is

`code.utils.`**`multiprocessing`** (`arr`, `func`, `workers=-1`)

Applies function to array with multiprocessing

Parameters:

- **arr** – List or np.ndarray
- **func** – Function applied to the array
- **workers** – Number of processes in the pool, by default uses the maximum number of cores - 1

Returns: List with the function applied to each element of the input list

`code.utils.`**`multithreading`** (`arr`, `func`, `workers=4`)

Applies function to array with multithreading

Parameters:

- **arr** – List or np.ndarray
- **func** – Function applied to the array
- **workers** – Number of threads in the pool

Returns: List with the function applied to each element of the input list

`code.utils.`**`remove_extension`** (`path`)

Removes the extension of the file

Parameters: **path** – Path to the file

Returns: Path without the extension

`code.utils.`**`unzip`** (`input_path`, `output_path=None`)

Unzip a compressed folder

Parameters:

- **input_path** – Path to the compressed folder
- **output_path** – Path where to uncompress the folder (default will uncompress in the same folder with the same name as the compressed file)

Returns: True if the folder was unzipped successfully, False otherwise

code.extract module

extract.py

Module to extract text from PDF files

`code.extract.`**`text_from_pdf`** (`path`, `ignore_image=False`)

Extracts text from a pdf

Parameters:

- **path** – Path to the pdf file
- **ignore_image** – If True, ignore pdf with images for faster processing

Returns: A String corresponding to the text in the PDF

`code.extract.`**`text_from_pdf_image`** (`path`)

Extracts text from a pdf image

Parameters: **path** – Path to the pdf image file

Returns: A String corresponding to the text in the PDF

`code.extract.`**`text_from_pdf_text`** (`path`)

Extracts text from a pdf text

Parameters: **path** – Path to the pdf text file

Returns: A String corresponding to the text in the PDF

code.pipeline module

class `code.pipeline.`**`Pipeline`**

code.bdd module

bdd.py

Module to manage PostgreSQL database

class `code.bdd.`**`PostgreSQL_DB`** (`enable_mail=True`)

`drop_table` **(**`table_name`**)**

Drop a table from the database

Parameters: **table_name** – Name of the table to drop

Returns: True if the table was successfully dropped, False otherwise

`get_table` **(**`sql_query`**)**

Query the database and store it in a dataframe

Parameters: **sql_query** – Query in the database

Returns: Dataframe containing the result of the query

`save_table` **(**`df`**,** `table_name`**,** `if_exists='replace'`**)**

Save table to postgreSQL database

Parameters:

- **df** – Dataframe to save
- **table_name** – Name for the table in the database
- **if_exists** – ‘replace’ to replace the table or ‘append’ to append to the existing table

Returns: True if the table was saved successfully, False otherwise

send_error_data_by_mail (error_traceback)

Display the last error traceback and send alert email.

Parameters:

- **email_infos** – a dictionary like {'sender':sender@domain.com, 'receivers':['receive1@domain.com', 'receive2@domain.com'], 'server':'mail.part.net'}
- **error_traceback** – error traceback text

Indices and tables

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

Index

Python Module Index