SaveTheDate

Generated by Doxygen 1.8.13

Contents

Chapter 1

Namespace Index

1.1 Namespace List

Here is a list of all namespaces with brief descriptions:

Chapter 2

Class Index

2.1 Class List

Here are the classes, structs, unions and interfaces with brief descriptions:

Chapter 3

File Index

3.1 File List

Here is a list of all files with brief descriptions:

Chapter 4

Namespace Documentation

4.1 scraper Namespace Reference

Classes

- class PDFScraper

  This class handles a PDF file and determines the tasks and deadlines.

Chapter 5

Class Documentation

5.1 scraper.PDFScraper Class Reference

This class handles a PDF file and determines the tasks and deadlines.

Public Member Functions

- def __init__ (self)

  This function is a constructor for insantiating a PDFScraper object.
- def importFile (self, filePath)

  This function sets the scraper's file path.
- def isDateHeading (self, heading)

  This function determines whether a header is a relevant identifier for dates.
- def getDataFrames (self, startPage=None, endPage=None)

  This function reads the PDF file and returns the structured data in the form of data frames (from the pandas library)
- def getDeadlines (self, dfs)

  This function finds the deadlines and corresponding tasks from DataFrame objects.
- def generateOutput (self, dfs, numDateTables)

  This function generates final output based on DataFrame objects and the number of date tables found.
- def scrape (self, filePath, startPage=None, endPage=None)

  This function drives the scraping function by connecting the other class functions together.

Public Attributes

- dateStrings
- dates
- tasks
- filePath

5.1.1 Detailed Description

This class handles a PDF file and determines the tasks and deadlines.

The file being scraped must have the .pdf extension

5.1.2 Constructor & Destructor Documentation

5.1.2.1 __init__()

```
def scraper.PDFScraper.__init__ (
              self )
```

This function is a constructor for insantiating a PDFScraper object.

The dateStrings variable is a list of strings that are typically used as headers for date columns in a table

The dates and tasks variables are lists that are appended to after scraping is performed

There should only be one instance of the PDFScraper class (singleton)

5.1.3 Member Function Documentation

5.1.3.1 generateOutput()

```
def scraper.PDFScraper.generateOutput (
              self,
              dfs,
              numDateTables )
```

This function generates final output based on DataFrame objects and the number of date tables found.

This function reports the total number of tables and found and how many of those included relevant information

Parameters

dfs	List of DataFrame objects
numDateTables	Integer value representing how many tables in the PDF include deadline information

Returns

2D List representing deadlines and their corresponding tasks in the form of [(date1, task1), (date2, task2), ...]

5.1.3.2 getDataFrames()

```
def scraper.PDFScraper.getDataFrames (
              self,
              startPage = None,
              endPage = None )
```

This function reads the PDF file and returns the structured data in the form of data frames (from the pandas library)

If no start or end pages are provided, all pages are scraped

If only one page is provided, it is assumed to be the starting page and the document is scraped from that page to the end

Parameters

startPage	Integer representing the first PDF page number to begin scraping from
endPage	Integer representing the last PDF page number to scrape

Returns

List of DataFrame objects representing the scraped structured data from the currently imported PDF, with each index representing a separate table

Exceptions

ValueError	if startPage comes after endPage or page numbers less than 1 are provided

5.1.3.3 getDeadlines()

```
def scraper.PDFScraper.getDeadlines (
              self,
              dfs )
```

This function finds the deadlines and corresponding tasks from DataFrame objects.

This function iterates through all data frames (tables) and further iterates the column headings and rows to create deadline/task tuples

The 'dates' and 'tasks' state variables are updated accordingly as found

Parameters

dfs	List of DataFrame objects

Returns

Integer value representing how many tables in the PDF were found that included deadline information Combine dates and tasks lists for output 2D list

5.1.3.4 importFile()

```
def scraper.PDFScraper.importFile (
              self,
              filePath )
```

This function sets the scraper's file path.

Parameters

filePath	A string representing the path to the PDF file

Exceptions

ValueError	if the file path does not end in .pdf

5.1.3.5 isDateHeading()

```
def scraper.PDFScraper.isDateHeading (
              self,
              heading )
```

This function determines whether a header is a relevant identifier for dates.

A relevant identifier is determined by comparing to the dateStrings constant

Parameters

heading	A string representing a table column heading

Returns

True if the heading is a date identifier, False otherwise

5.1.3.6 scrape()

```
def scraper.PDFScraper.scrape (
              self,
              filePath,
              startPage = None,
              endPage = None )
```

This function drives the scraping function by connecting the other class functions together.

This is the only PDFScraper function that is called publicly to scrape the file

This function allows the frontend to communicate with the backend

Parameters

filePath	A string representing the path to the PDF file
startPage	Integer representing the first PDF page number to begin scraping from
endPage	Integer representing the last PDF page number to scrape

Returns

2D List representing deadlines and their corresponding tasks in the form of [(date1, task1), (date2, task2), ...]

5.1.4 Member Data Documentation

5.1.4.1 dates

```
scraper.PDFScraper.dates
```

5.1.4.2 dateStrings

```
scraper.PDFScraper.dateStrings
```

5.1.4.3 filePath

```
scraper.PDFScraper.filePath
```

5.1.4.4 tasks

```
scraper.PDFScraper.tasks
```

The documentation for this class was generated from the following file:

- scraper.py

Chapter 6

File Documentation

6.1 scraper.py File Reference

Contains a class for scraping a course PDF file and determining the tasks and deadlines.

Classes

- class scraper.PDFScraper

  This class handles a PDF file and determines the tasks and deadlines.

Namespaces

- scraper

6.1.1 Detailed Description

Contains a class for scraping a course PDF file and determining the tasks and deadlines.

Author

Samarth Kumar (kumars38)

Date

Mar. 15th, 2022

Index