Program do przetwarzania tekstu ze zdjęć

*#===================================określanie używanych modułów==============================*

from urllib.parse import urlparse

from pathlib import Path

import cv2

import pytesseract

import urllib.request

import os

from os import listdir

import pdfkit

path\_wkhtmltopdf = Path("H:/wkhtmltopdf/bin/wkhtmltopdf.exe")   *#Wprowadzić trzeba ścieżkę do własnego*

config = pdfkit.configuration(wkhtmltopdf=path\_wkhtmltopdf)

from os.path import isfile,join

pytesseract.pytesseract.tesseract\_cmd=r'H:\\teseract\\tesseract.exe'   *#Wprowadzić trzeba własną ścieżkę*

*#====================================określanie ścieżek w zmiennych===========================*

\_to\_open=Path("text\_from\_image/links.txt")

\_to\_save=Path("text\_from\_image/images/")

\_to\_pdf=Path("text\_from\_image/pdf/")

\_to\_text=Path("text\_from\_image/txt/")

\_txt\_to\_txt=Path("text\_from\_image/temp.txt")

\_to\_process=Path("text\_from\_image/To Process/")

*#====================================tworzenie inkrementujacych folderow dla zdjec============*

dirlist = [ item for item in os.listdir(\_to\_save) if os.path.isdir(os.path.join(\_to\_save,item))]

dirlist.append("0")

\_max\_dir=max(int(item) for item in dirlist)

\_max\_dir=int(\_max\_dir)+1

\_num=\_max\_dir

os.makedirs(f'{\_to\_save}\{\_num}',exist\_ok=True)

\_to\_save=Path(f"text\_from\_image/images/{\_num}/")

*# ====================================znajdywanie najwyzszego pdf numera===================*

*#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*DOSTĘPNE,ALE CHWILOWO NIE UŻYWANE\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

onlyfiles = [f for f in listdir(\_to\_pdf) if isfile(join(\_to\_pdf,f))]

onlyfiles.append("0.pdf")

\_max\_file=max(int(item[:-4]) for item in onlyfiles)

\_max\_file=int(\_max\_file)+1

*#====================================znajdywanie najwyższego pliku txt numeru=============*

onlyfiles\_txt = [f for f in listdir(\_to\_text) if isfile(join(\_to\_text,f))]

onlyfiles\_txt.append("0.txt")

\_max\_file\_txt=max(int(item[:-4]) for item in onlyfiles\_txt)

\_max\_file\_txt=int(\_max\_file\_txt)+1

*#====================================znajdywanie zdjęć do przetworzenia z folderu=========*

to\_process = [f for f in listdir(\_to\_process) if isfile(join(\_to\_process,f))]

to\_process.insert(0,"\*&%@#!.jpg")

*#==========================================================================================*

links=[]

images=[]

def \_get\_links():

    try:

        f = open(\_to\_open)

        for line in f.readlines():

            if len(line)>3:

                link=urlparse(line)

                links.append(link.geturl())

    except FileNotFoundError:

        print("File [links.txt] not Found")

        f = open(\_to\_open, "w+")

    finally:

        if len(f.readlines())<1:

            \_Clear\_Session()

            print("File [links.txt] is empty")

            f.close()

            return False

        f.close()

        return True

def get\_from\_folder():

    if len(to\_process)>1:

        for item in to\_process[1:]:

            images.append(f'{\_to\_process}\\{item}')

    else:

        print("Folder is empty. Closing program")

        \_Clear\_Session()

def \_Clear\_Session():

    try:

        os.rmdir(f'{\_to\_save}')

        os.remove(f'{\_to\_text}\\{\_max\_file\_txt}.txt')

    except:

        pass

def \_download\_image(url,fullname):

    urllib.request.urlretrieve(url,f'{\_to\_save}\\{fullname}\\')

def images\_from\_links():

    if \_get\_links():

        counter=0

        print("DOWNLOADING IMAGES")

        print("==================")

        for img in links:

            fullname=str(counter)+".jpg"

            \_download\_image(img,fullname)

            images.append(f'{\_to\_save}\\{fullname}')

            print(f'Progress: {round(((counter+1)/len(links))\*100,2)}%')

            counter+=1

        print("==================")

*#========================Pozyskiwanie tekstu z zdjęć=========================*

def \_txt\_from\_img(img):

    with open(f'{\_to\_text}\\{\_max\_file\_txt}.txt','a+',encoding="utf-8") as f:

        f.write((pytesseract.image\_to\_string(img)))

def get\_text():

    try:

        print("PROCESSING IMAGES")

        print("==================")

        for i,img in enumerate(images):

            print(f"Progress: {round((i/len(images))\*100,2)}%")

            \_txt\_from\_img(img)

        print("==================")

    except pytesseract.pytesseract.TesseractError:

        print("Image is not in valid format")

        \_Clear\_Session()

*#=======================NIE UŻYWANE. POTRZEBNE LEPSZE ROZWIĄZANIE NA ZACHOWANIE FORMATOWANIA TEKSTU PRZY ROBIENIU W PDF*

def txt\_to\_txt():

    counter=0

    f = open(f'{\_to\_text}\\{\_max\_file\_txt}.txt','r',encoding="utf-8")

    txt=open(\_txt\_to\_txt,'a+',encoding="utf-8")

    for line in f.readlines():

        for i,j in enumerate(line):

            if counter>100:

                if j==" ":

                    txt.write("<br>")

                    counter=0

            txt.write(j)

            counter+=1

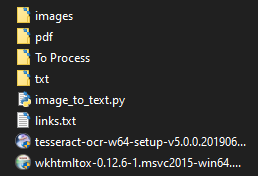
    f.close()

    txt.close()

"""

txt\_to\_txt()

"""



*#CHWILOWO UŻYTKOWE KOMENTY*

*#   images\_from\_links()*

*#           /*

*#   get\_from\_folder()*

*#*

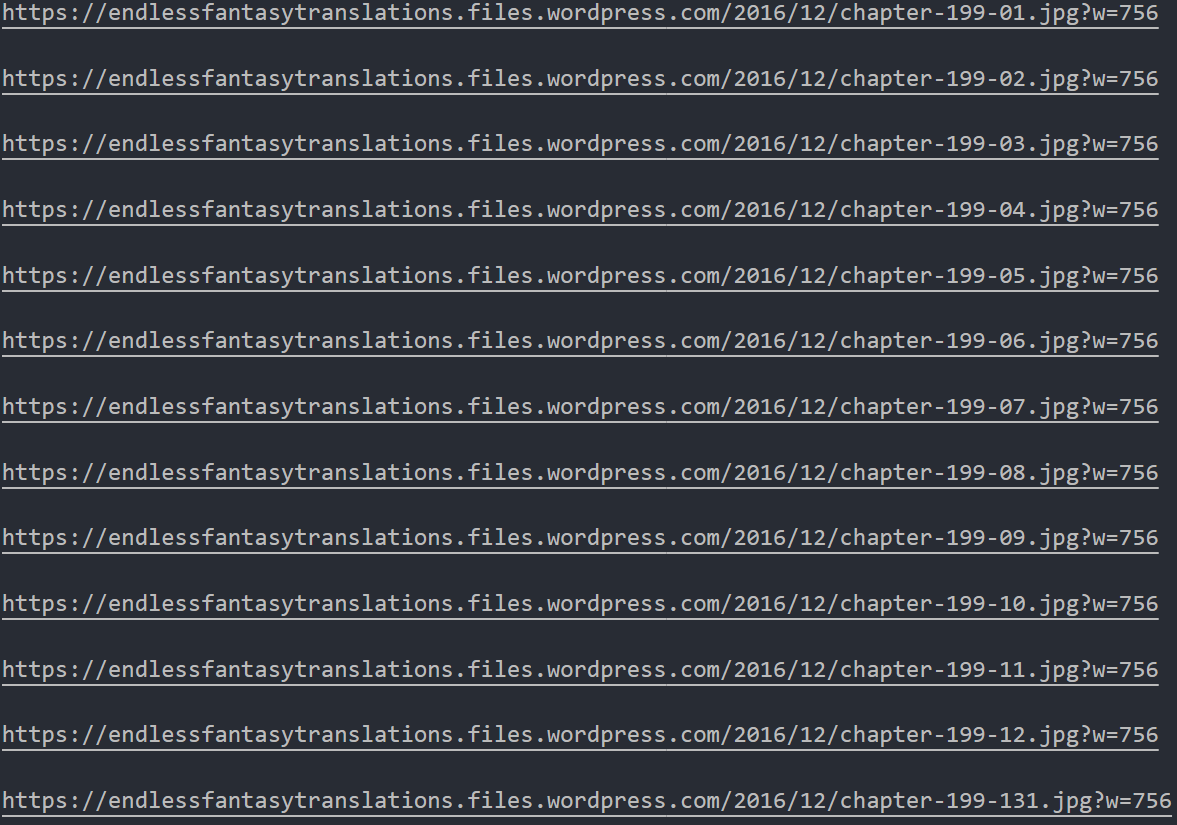
*#       get\_text()*

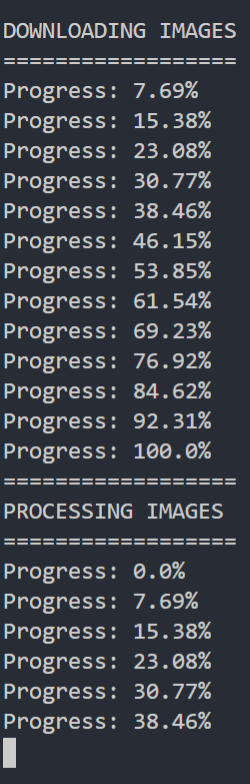
Program przyjmuje Zdjęcia i przetwarza z nich tekst następnie zapisując w pliku tekstowym. Zdjęcia można przetwarzać przez wrzucenie do folderu [To Process],

Lub w pliku [links.txt]. Program obsługuje przetworzenie kilku zdjęć w plik Tekstowy. Przy podaniu linków zdjęcia są zapisywane w inkrementacyjnym folderze (FOLDERY: 1,2,3,4,5,…,n)[Program wyszukuje najwyższy folder i tworzy o jeden wyższy],a w nich pobrane zdjęcia od 0 do ilości zdjęć. Następnie zdjęcia są przetworzone i zapisane w pliku tekstowym o inkrementacyjnej nazwie.

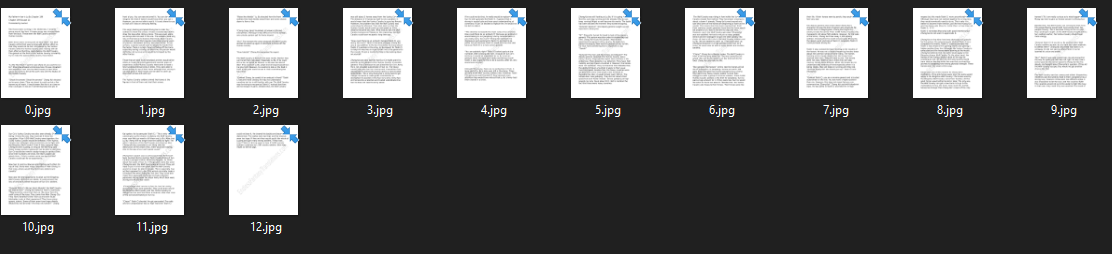
Przykładowe działanie z linków:

* Plik links.txt



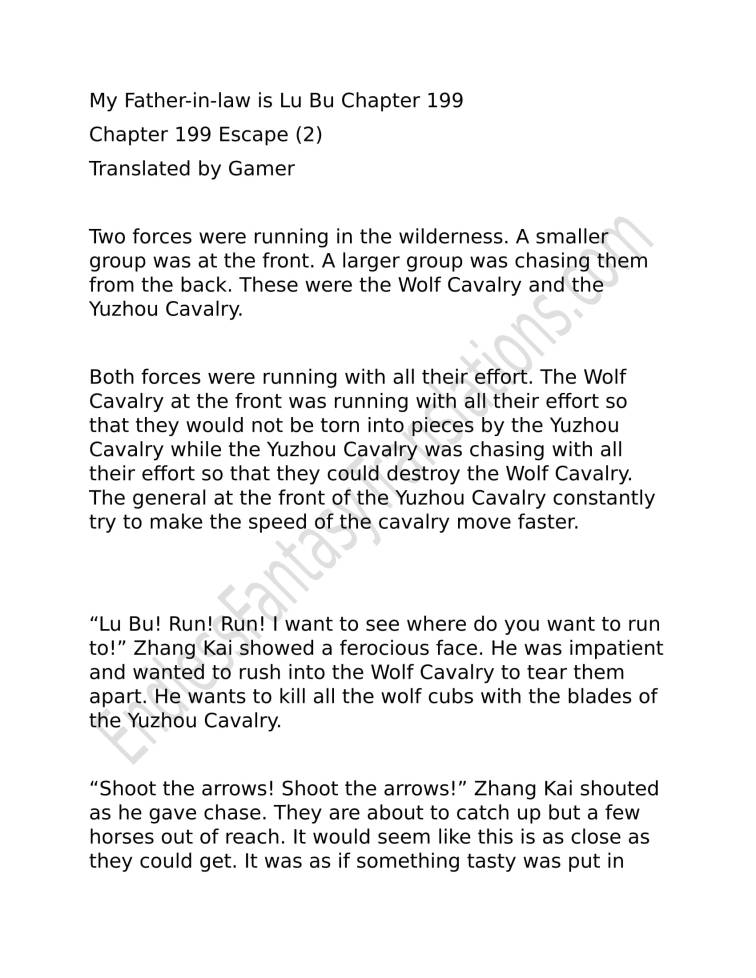


* Pliki zostają pobrane z linku



* Zapisane w folderze (w tym przypadku [3])
* Wynik:

Zdjęcie tekstu zostało przetworzone na tekst w pliku .txt

Zapisany w folderze [txt] w pliku (w tym przypadku 1.txt)

