Task:

1 实现 PDF 文件转成 TXT 文件

2 实现识别图片中的文字，并输出 TXT 文件

3 基于有道词典进行查词

Implement:

1 基础环境配置

Operation system：Centos 7

Python Version: python 3.6.1

Linux Command:

yum update

yum groupinstall “Development tools”

yum -y install automake autoconf libtool zlib-devel libjpeg-devel giflib libtiff-devel libwebp libwebp-devel libicu-devel openjpeg-devel cairo-devel

yum install gcc

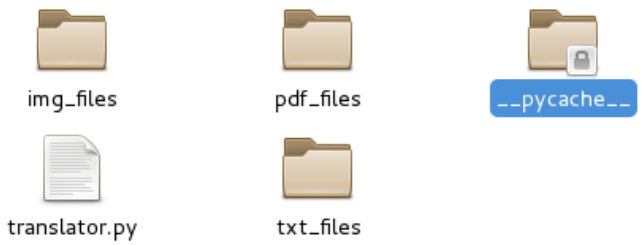
pip3 install wand

pip3 install pytesseract

pip3 install pillow

```
pip3 install tesseract

wget https://github.com/tesseract-ocr/tesseract/archive/3.04.01.tar.gz

mv 3.04.01.tar.gz tesseract-3.04.01.tar.gz

tar xzvf tesseract-3.04.01.tar.gz

cd tesseract-3.04.01/

./autogen.sh

./configure

make

make install

ldconfig


pip3 install pyocr


yum install python-imaging

yum install ImageMagick-devel


export TESSDATA_PREFIX=/usr/local/share/tessdata


pip3    install pdfminer.six


pip3 install urllib
```

2 Project File Structure:



Remark: 顾名思义

3 Python Code Structure：

```
# -*- encoding: utf-8 -*-
import ...

#获取指定文件夹下的所有文件名
def get_files(file_folder):...

#获取图片式PDF文件的内容
def pdf_image_to_string(pdf_file, lang_code=0):...

#获取可读PDF的文本内容
def pdf_text_to_string(pdf_file):...

#使用pyocr 提取图片中的文字
def pyocr_image_to_string(img_file, lang_code=0):...

#使用pytesseract 提取图片中的文字
def pytess_imgae_to_string(img_file):...

#使用ImageEnhance增强提取图片中的文字
def enhance_image_to_string(img_file):...

#模拟浏览器使用有道进行翻译
def youdao_html_translate(text, url = "http://fanyi.youdao.com/translate?smartresult=dict&smartresult=rule&sessionFrom="):...

#写入txt文本
def string_to_text(file, text):...

def translator(file_folder, file_type):...
```

4Reference:

http://www.wisedream.net/2016/07/18/imgProcessing/ocr-with-pytesseract/

https://ivanzz1001.github.io/records/post/ocr/2017/09/08/tesseract-install

http://blog.topspeedsnail.com/archives/3571

http://blog.csdn.net/fighting_no1/article/details/51038942

http://blog.csdn.net/qq_21905401/article/details/77620561