pdf2txt

pdf2txt 使用说明

下载安装 Python

前往 Python 官网下载安装最新版 3.8.3 https://www.python.org/downloads/release/python-383/

拖到网页最下端找到 Windows x86-64 executable installer 点即下载安装即可，注意在安装时勾选 **add python to PATH**

设置环境变量

计算机 右键 属性 - 高级系统设置 - 高级标签 - 环境变量

找到上方用户变量中的 Path 双击，如果上一步中勾选了 add python to PATH，那么下面只需要检查是否存在即可

检查是否存在 `C:\Users\你的用户名\AppData\Local\Programs\Python\Python38`，如没有，点击新建并输入。

检查是否存在 `C:\Users\你的用户名\AppData\Local\Programs\Python\Python38\Scripts`，如没有，点击新建并输入。

设置应用执行别名

按下 Win 徽标搜索 App Execution Aliases （管理应用执行别名）

关闭 python.exe 和 python3.exe 的执行别名

使用 pip 安装 pdfminer.six

Win+R 运行打开 cmd 命令行，

输入下面代码并按回车

```
pip install pdfminer.six
```

安装成功后，应该在 `%USERPROFILE%\AppData\Local\Programs\Python\Python38\Scripts` 文件夹中找到 `pdf2txt.py` 文件。

安装完成后，可能会提示你的 pip 版本过低，可以按照提示说明进行升级，输入以下代码回车等待即可

```
python -m pip install --upgrade pip
```

当然，不升级也是完全可以的~

使用方法

前期准备结束，在 pdf2txt.bat 文件上面右键，选择你常用的编辑器打开，例如 notepad++ 或 vscode，因为这些编辑器有代码高亮功能，方便我们区分不同代码的功能。

前面 8 行代码为注释部分，解释了这个程序的功能，以及可修改的一些参数，即 indir 和 outdir

```
goto start
Converting all pdf files in a folder (including subfolders) to txt files.
Based on pdfminer.six
indir: root directory of pdf files, traversing all subfolders
outdir: output directory of txt files
pydir: location of pdf2txt.py
Directory containing spaces must be enclosed by ""
:start
```

我们需要修改的只有两个部分，即输入和输出文件夹的完整路径：

```
set indir="%USERPROFILE%\Documents\_Reference\Smart Speakers"
set outdir=C:\Users\Tonakai\Desktop\output
```

这个程序**会自动遍历输入文件夹下的所有子文件夹**，建议输出文件夹放在其他位置，避免和 pdf 放在一起。

而 pydir 为上一部分提到的 `pdf2txt.py` 文件的所在位置，应该不需要修改。

这里需要注意的是，**如果路径中存在空格，一定要使用英文双引号括起来。**