PDF2TXT环境安装

一、创建一个新的conda环境

```
conda create --name pdf2txt python=3.7
```

二、Pycharm中打开工作目录

工作目录必须包含Json，pdf，text三个文件夹（可以直接复制目录结构）

打开Pycharm选择新的Python解释器（pdf2txt环境），然后打开pdf2txt.py文件

三、安装基本的库

Pycharm会自动显示Python文件中没有安装的第三方包，Pandas、tesserocr和pdfplumber应该没有安装，使用conda自带的命令行工具，切换到环境后直接用pip安装就行（哪个包爆红就安装哪个），其中tesserocr可能报错，如果pip不了试一下以下代码。如果还不行就只能百度。

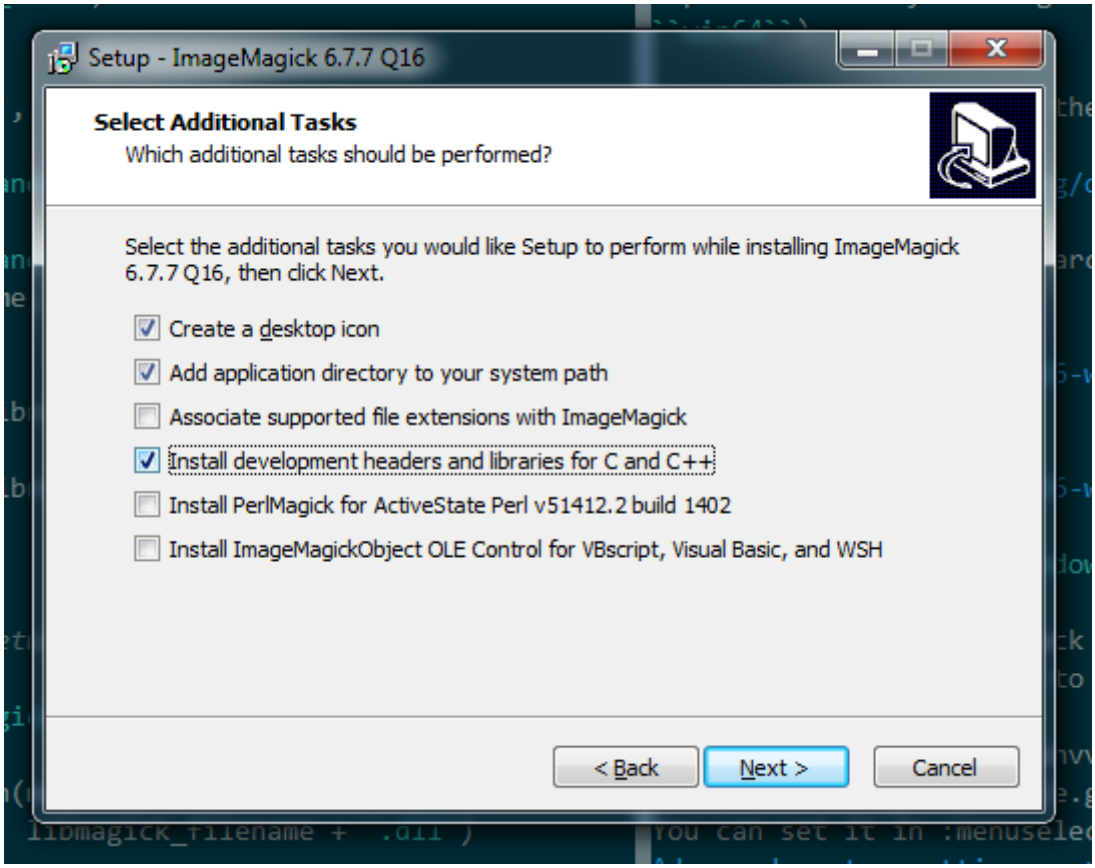
```
conda remove tesserocr
conda install -c simonflueckiger tesserocr
```

四、安装内部库

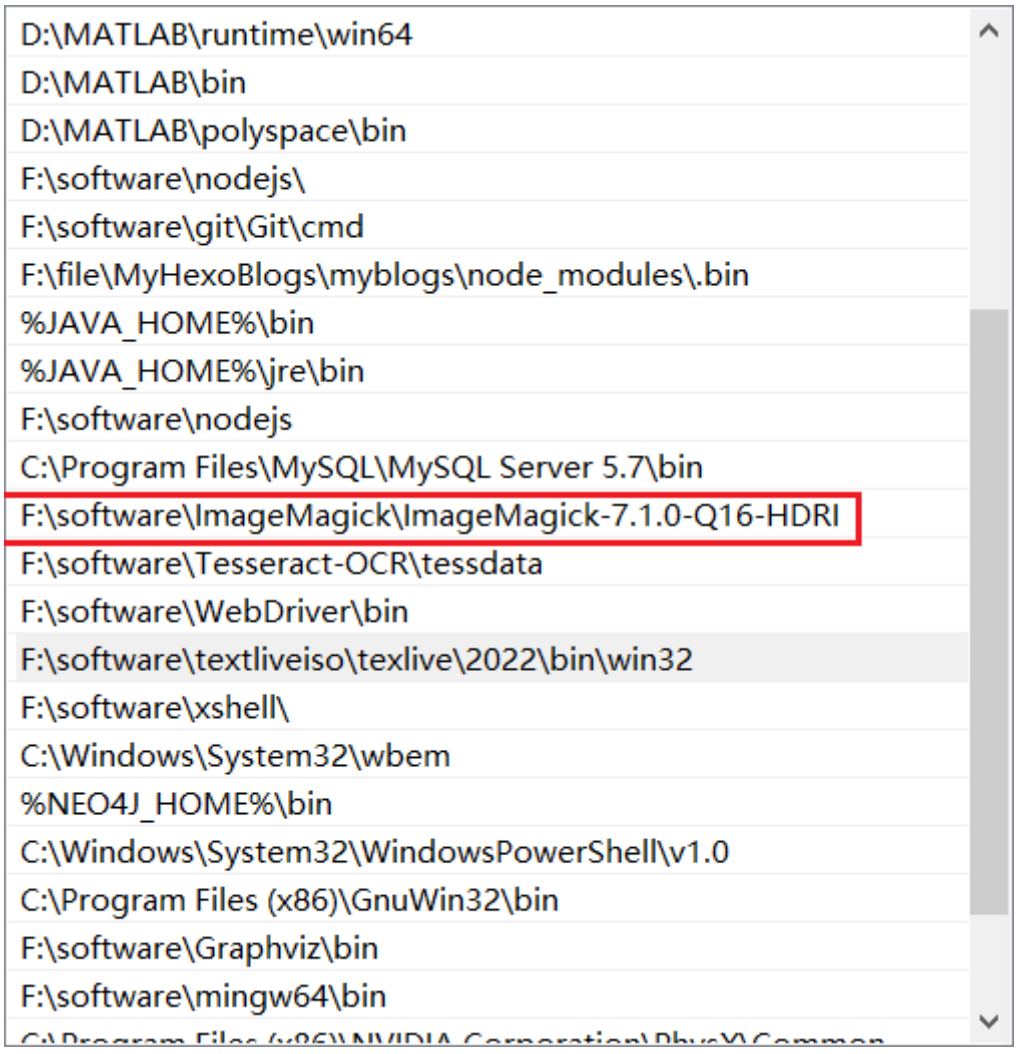
因为pdfplumber里面的函数很多都是包装接口，并不能完成具体的功能，所以还需要安装一些软件。（安装包随本文件一同发送）

ImageMagick安装

点击安装文件，基本执行next就行，除了在中间勾选一个install c/c++，最好不要勾选自动添加环境变量，而是安装完Imagemagick后自己配环境变量（没试过自动配置也可以试一下）。



按照安装路径配置环境变量

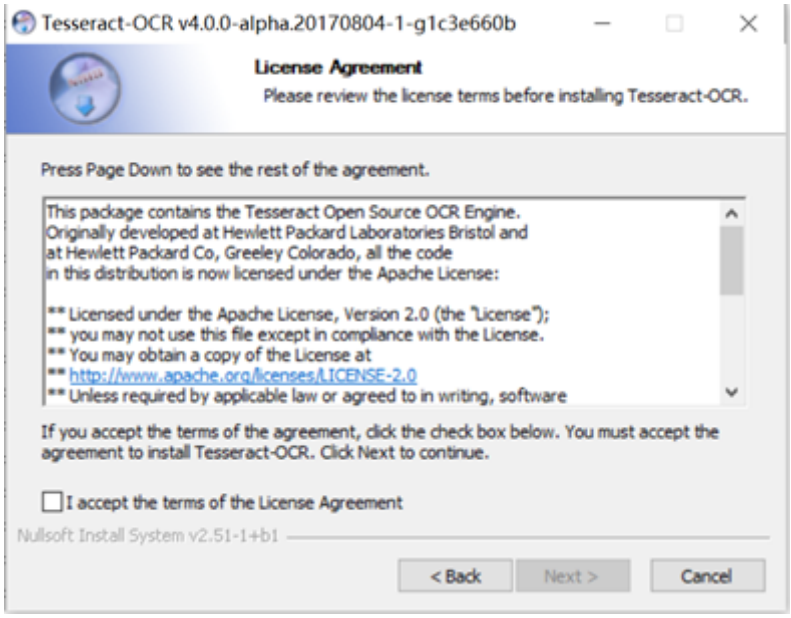


Ghostscript安装

安装完ImageMagick后仍然会报错，需要安装Ghostscript，点击downloadsgs10021w64直接无脑安装就好了。

五、安装tesserocr

1. 运行tesserocr安装程序，无脑安装



2. 完成后进入到tesserocr的安装目录找到tessdata文件夹，将这个文件夹整个复制到使用的conda环境目录下，Anaconda3->envs->pdf2txt 路径下（如果程序还是报错找不到tessdata目录就将复制后的路径添加到环境变量里）。