PDF转换相关过程

1、PDF转html

参照pdf2htmlEX的使用方法： **https://github.com/pdf2htmlEX/pdf2htmlEX**

最便捷的为docker启动，mac系统也可通过homebrew安装，亲测可用： **https://github.com/Homebrew/homebrew-core/blob/master/Formula/pdf2htmlex.rb**

使用pdf2htmlEX将pdf文件转换为html

2、html转txt

1.需要安装node.js运行环境

2.然后安装jsdom依赖：

```
npm install jsdom
```

3.然后运行convert.js脚本：

```
node convert.js a.html a.txt
```

> convert.js 脚本的两个参数：
>
> - a.html: html文件的路径
> - a.txt: 输出txt文件的路径
>
> convert.js 脚本做的事情：
>
> - 修改输入的html文件：调整css样式保证页面文字可以选择，删除完全无用的空div标签
> - 根据修改后的html文件，输出txt文本