Matlab案例代码解析

1. 基础函数使用案例

1.2 文件读写

1.2.7 读取pdf文件

- pdfbox.jar下载链接：https://pdfbox.apache.org/download.cgi
- 提取特定字段：contains、strfind 、 split 、strip、deblank
- ASCII码：回车，13；换行，10；空格，32
- 未解决问题：[WARN] PDCIDFontType0 - Using fallback AdobeFangsongStd-Regular for CID-keyed font STSong-Light

```
clear;clc;
% 这里是 pdfbox.jar 所在目录
pdfbox_path = 'D:\MyPrograms\Matlab2016B\toolbox\pdfbox-app-2.0.22.jar';
cleanUp = onCleanup(@()javarmpath(pdfbox_path));
javaaddpath(pdfbox_path);
pdfdoc = org.apache.pdfbox.pdmodel.PDDocument;
% 老版本可以直接文件名
file = java.io.File('data\demo.pdf');
doc = pdfdoc.load(file);
doc.isEncrypted;
reader = org.apache.pdfbox.text.PDFTextStripper;
pdfstr = reader.getText(doc);
```

```
[WARN] FileSystemFontProvider - New fonts found, font cache will be re-built
[WARN] FileSystemFontProvider - Building on-disk font cache, this may take a while
[WARN] FileSystemFontProvider - Finished building on-disk font cache, found 570 fonts
```

```
pdfstr = char(pdfstr);
% 去掉尾部的空白
pdfstr = deblank(pdfstr)
```

```
pdfstr =
    '需求分析：利用 pdfbox 读取 PDF 文件并提取特定字段
     • pdfbox.jar下载链接：https://pdfbox.apache.org/download.cgi
     • 提取特定字段：contains、strfind 、 split 、strip、deblank
     • ASCII码：回车，13；换行，10；空格，32
     • 未解决问题：[WARN] PDCIDFontType0 - Using fallback AdobeFangsongStd-Regular for CID-
     keyed font STSong-Light
     clear;clc;
     % 这里是 pdfbox.jar 所在目录
     pdfbox_path = 'D:\MyPrograms\Matlab2016B\toolbox\pdfbox-app-2.0.22.jar';
     cleanUp = onCleanup(@()javarmpath(pdfbox_path));
     javaaddpath(pdfbox_path);
     pdfdoc = org.apache.pdfbox.pdmodel.PDDocument;
     % 老版本可以直接文件名
     file = java.io.File('DemoPdfRead.pdf');
     doc = pdfdoc.load(file);
     doc.isEncrypted;
     reader = org.apache.pdfbox.text.PDFTextStripper;
     pdfstr = reader.getText(doc);
     pdfstr = char(pdfstr);
     % 去掉尾部的空白
```

```
pdfstr = deblank(pdfstr)
doc.close;
pdfdoc.close;
% 保存成文本
fid = fopen('pdf_text.txt', 'wt');
fprintf(fid, '%s', pdfstr);
fclose(fid);
% 查找指定字段
str = 'pdfbox';
index = strfind(pdfstr, str)
1'
```

```
doc.close;
pdfdoc.close;
% 保存成文本
fid = fopen('data\pdf_text.txt', 'wt');
fprintf(fid, '%s', pdfstr);
fclose(fid);
% 查找指定字段
str = 'pdfbox';
index = strfind(pdfstr, str)
```

```
index = 1×11
     9    36    59   294   311   360   419   447   482   629   883
```