小组成员：181250065 李邦国、181250071 李昊天

框架

整体搭建使用的是spring boot

相关的技术实现

关于Word文档的解析，使用的是Apache POI提供的相关API，其中doc和wps文档的解析使用的是HWPF，docx文档的解析使用的是XWPF，由于缺乏良好的转换工具，pdf文档的解析方法是先转化为doc，然后使用doc相关api进行操作。

整体效果

获取段落信息


```
{
  "code": 0,
  "msg": "success",
  "data": [
    {
      "paragraphText": "0 天津市北辰区人民法院\r",
      "paragraphId": 1,
      "fontSize": 44,
      "fontName": "宋体",
      "isBold": false,
      "isItalic": false,
      "isInTable": false,
      "lvl": 9,
      "lineSpacing": 800,
      "fontAlignment": 0,
      "isTableRowEnd": false,
      "indentFromLeft": 0,
      "indentFromRight": 0,
      "firstLineIndent": 0
    },
    {
      "paragraphText": "0 民 事 判 决 书\r",
      "paragraphId": 2,
      "fontSize": 52,
      "fontName": "宋体",
      "isBold": false
```


获取图片信息

Response body


```
{
  "code": 0,
  "msg": "success",
  "data": [
    {
      "paragraphBefore": 0,
      "textBefore": null,
      "textAfter": null,
      "height": 217,
      "width": 577,
      "suggestFileExtension": "png",
      "base64Content":
"iVBORw0KGgoAAAANSUhEUgAAAkEAAADZCAYAAAAnruSqAAAAAXNSR0IArs4c6QAAAARnQU1BAACxjwv8YQUAAAAJcEhZcwAAEnQAABJ0Ad5mH3gAAEToS
URBVHhe7b1tbGtbet/317137tjBNVpkpokHu6TayrFUFIjtDwlCk61FBWkacTxxJmMGoRTTnsAxpQkSA2N6SgJ0C7AlY9PT2EAjsQ3shBORSQljEscmG6O
IJAfk0LCDYpI2sTT2QXvIsm6MmXFrf2k/qetZe29yv5KbFKUjHf1/B8Thy9baa69Fcv35rOdl64MPPrhDAFtbGdSucpgUjnF24z8kU79GblLA0dmN9czDs
rW1h8NaBcWEYT4xHaFbLfv6trWXwUk+h6x13FQd11HH9azj9k7baGbVa9MuCsfnuLm7U39ziotmFgam6FrXK8/V8jHEMUG1fIkDdW6zzSlG3SpKjuuWsSg
mrAdOHOcghBBCyNPiHet/H4e1HOJqsQ8SQMKryRRGbMd69LCIADq5aKKIDgrpNPb391HoANlmDZmtLesoS7g1i0hOzOPS6QI6wziKjuNuzo6Qboz0fZu7m
zMcpxtwP3uJwSQOIx5DvpYHWlV93nRjiHi2gtO9+Xl7pX3rNdWCCB+rj/tHZxRAhBBCyBMlUARtZepITaqPZuVZxu5JBVkocVHuz0TF7asJpohje1c/1Bz
WikiMGqrfPX3c3d0Nekr0NEYJFGuH1lHRuLu5Qf/1GDCAQVksTtZY9C8xnJp3CSGEEPJ88YkgEUC17ZZruyeInZiB6eSV9ehh0ecaXrqsKmK9Odo/mlmqx
AqUSgCjQV8/dtIfjIBEymU1isx0Av9VGngkIxghhBBCHgiXCIoqgGR7ajsOjF/fWs88HPa5ojFFoC7TViNCCCGEkDkzEaQdgVMDnwDa29uz7jnZQcwYIcD
osnFkS0t2paIRYqHZicmuFiGEEELIDC2CdHRUBWiV3apGtpjyeb+q2D3JITEaoPdITr/aCTt5gD3Pdpb07/Q0o+/f3fUgu15xp5OQxa6Ykh6xvzbSv0wmS
EQSQggh5E3zjo68qiQxrF4qtbCrLT/2TUeIefaX5PiDJNBtPYIZyOL2vIORkUWldjgTQjoU/iIHXM770S83MM42UXcIj71MHc3sGA2nwJPtMSOJA0sv6bb
kWmEgeaDO4Yj8Wgndbgw7Vru7Jylsv3r4LUNCCCGErM7WN33T993VrooISnMjjBpplHpzC4rk2ang8SPHvPl/FuUJqlVySBjmcdNpF53q+SxPkM0sX5BCH
3P8GqkrJapGY0xaZZzv1HA1S/4zQiNdRh+HcI6Vd2yETL1t5TLy5xMihBBCyNMhNFliEGYCwdcolXrWM4QQQgghz5PAPEFB6G0zyRno8RsihBB
JIoQQQgh5W4hsCSKEEELIi4QiBBCCCEvEp9PkA4xzyM0Aswbgh4Uqi7lN+bh5R6kyvrx+crV1cUx+7BWscLP5bQNVB0FVVdl0
```


Download

Response headers

获取表格信息


```
{
  "code": 0,
  "msg": "success",
  "data": [
    {
      "textBefore": "",
      "textAfter": "",
      "paragraphBefore": {
        "tableTextContent": "被告郭继兴，男，1970年12月15日出生，汉族，住天津市北辰区大张庄镇北孙庄村南新村16号。\r",
        "paragraphId": 8
      },
      "paragraphAfter": {
        "tableTextContent": "被告中国平安财产保险股份有限公司天津分公司，住所地天津市南开区白堤路1号。\r",
        "paragraphId": 21
      },
      "tableContent": [
        {
          "tableTextContent": "a\u0007",
          "paragraphId": 9
        },
        {
          "tableTextContent": "b\u0007",
          "paragraphId": 10
        },
        {
          "tableTextContent": "c\u0007",
```


Download

获取标题信息


```
{
  "code": 0,
  "msg": "success",
  "data": [
    {
      "paragraphText": "委托代理人陈博，广东国晖（天津）律师事务所律师。（特别授权）\r",
      "paragraphId": 0,
      "lineSpacing": -15,
      "indentFromLeft": 0,
      "indentFromRight": 0,
      "firstLineIndent": 0,
      "lvl": 1
    },
    {
      "paragraphText": "被告中国平安财产保险股份有限公司天津分公司，住所地天津市南开区白堤路1号。\r",
      "paragraphId": 0,
      "lineSpacing": -15,
      "indentFromLeft": 0,
      "indentFromRight": 0,
      "firstLineIndent": 0,
      "lvl": 1
    }
  ]
}
```


Download

获取字体信息

Response body


```
{
  "code": 0,
  "msg": "success",
  "data": {
    "color": "0",
    "fontSize": 32,
    "fontName": "仿宋_GB2312",
    "isBold": false,
    "isItalic": false,
    "fontAlignment": 0
  }
}
```


Download

Response headers

代码地址

https://github.com/libanguo/DataScienceApplicationHomework