平台需求——高温合金

本需求主要涉及文本挖掘部分的内容。主要分为四个部分：文献上传、预处理、抽取、图谱构建。图1是整体框架结构。

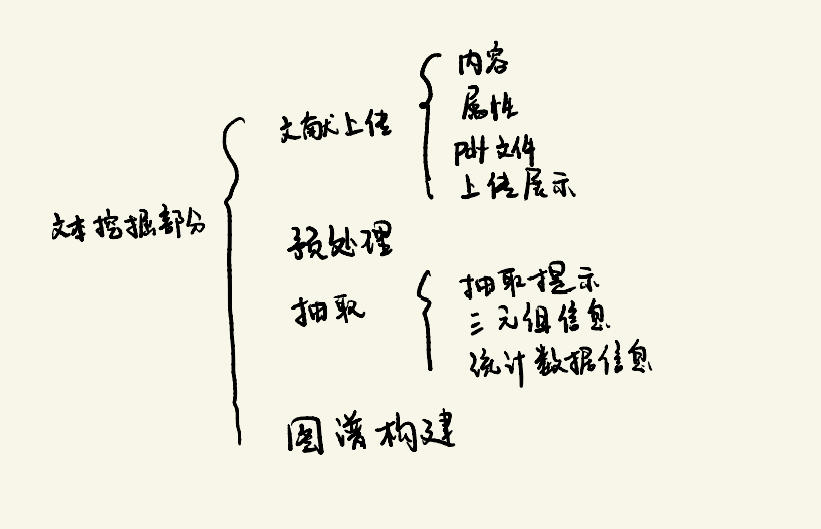


图1：文本挖掘需求整体框架

1. 文献上传
   1. 文献基本信息上传

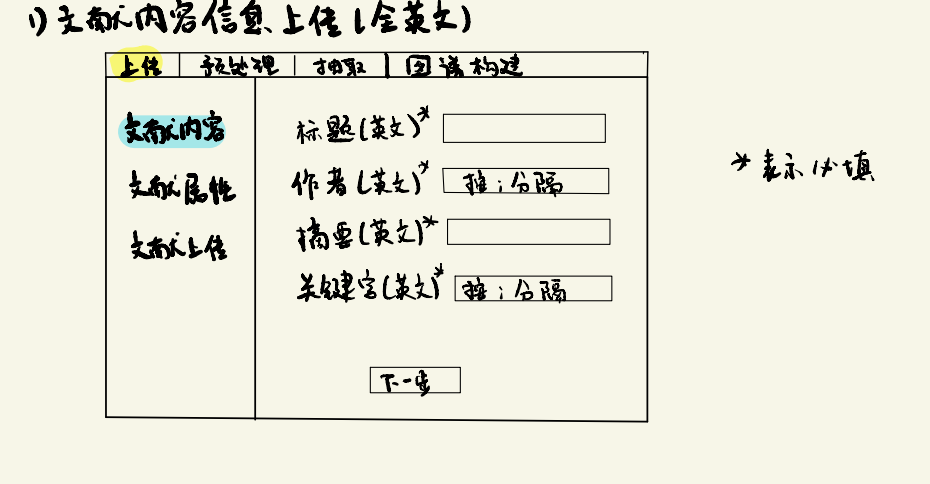


图1-1-1 文献内容上传

表1-1-1 文献内容字段要求

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 输入形式 | 字段类型 | 字段约束 |
| 标题（英文） | 文本框 | titleE：String | 非空，最大长度255，英文校验 |
| 作者（英文） | 文本框 | authorE：String | 非空，最大长度 255，英文校验，多个作者按(';'分隔) |
| 摘要（英文） | 文本框 | l\_abstractE：String | 非空，最大长度 800， 英文校验 |
| 关键字（英文） | 文本框 | keywordsE：String | 非空，最大长度 255， 英文校验，多个按(';' 分隔) |

* 1. 文献属性信息上传
  2. 注：添加多文件上传功能：提供excel填写元信息，再上传（多文件上传单独一个页面）

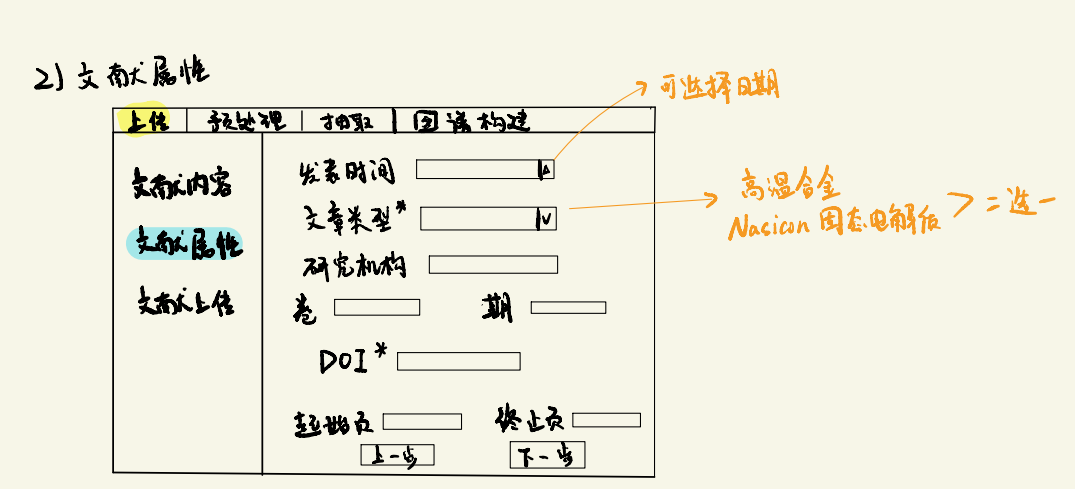


图1-2-1 文献属性上传

表1-2-1 文献属性字段要求

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 输入形式 | 字段类型 | 字段约束 |
| 发表实现 | 时间控件 | publish\_date:String | 非空， yyyy-mm-dd |
| 文章类型 | 选择框 | reference\_type： String | 非空{选项：高温合金，值：高温合金；选项：Nascion固态电解质，值：Nascion固态电解质} |
| 研究机构 | 文本框 | research\_institute： String | 非空，最大长度 255 |
| 卷 | 文本框 | volume：int | 非空，数字，最大长度 10 |
| 期 | 文本框 | issue：int | 非空，数字，最大长度 10 |
| DOI | 文本框 | doi：String | 非空，最大长度 255 |
| 起始页 | 文本框 | start\_page：int | 非空，数字，最大长度 10 |
| 终止页 | 文本框 | end\_page：int | 非空，数字，最大长度 10 |

* 1. PDF格式文献上传

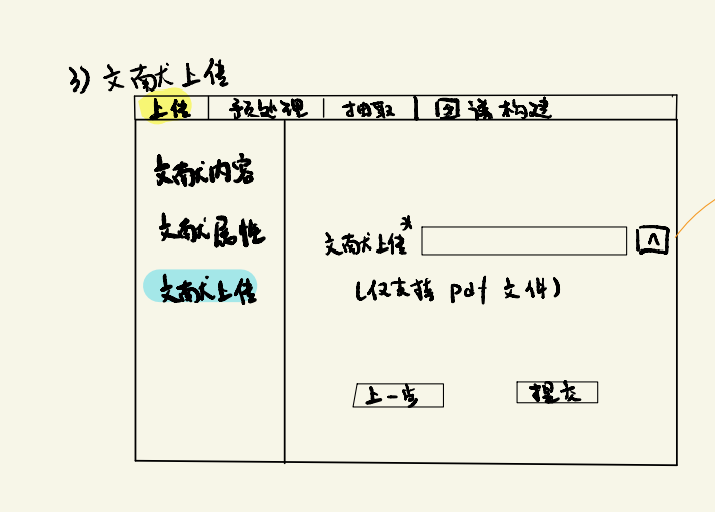


图1-3-1 文献PDF格式上传

表1-3-1 文献PDF格式上传字段要求

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 输入形式 | 字段类型 | 字段约束 |
| 附件上传 | 文件选择控件 | 地址地段 file\_attachment： String | 仅支持PDF格式 |

* 1. 上传展示

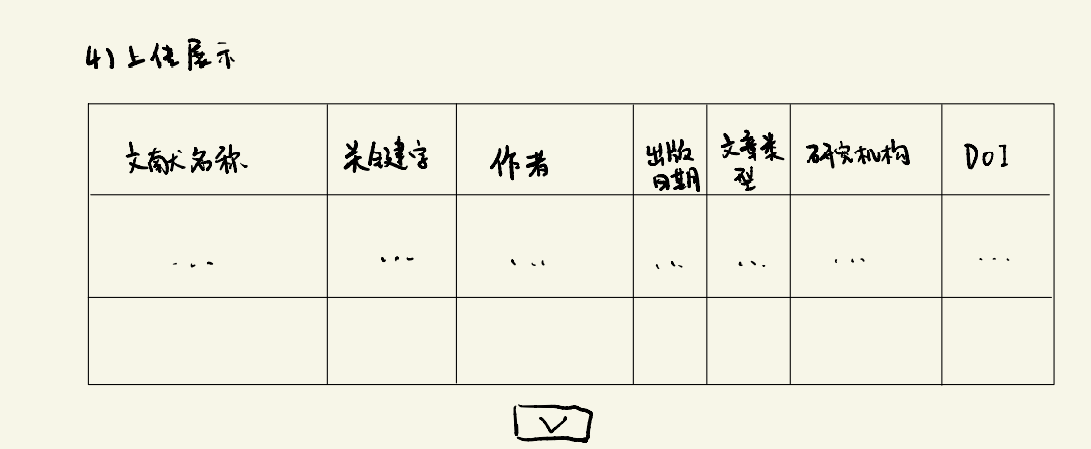


图1-4-1 上传展示

1. 预处理

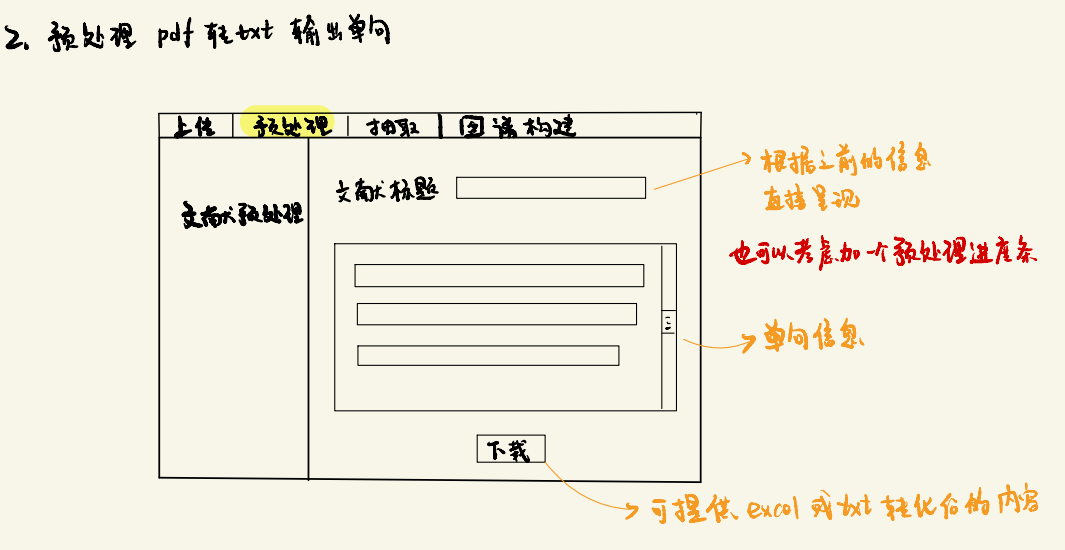


图2-1-1 预处理结果呈现

文献标题部分按之前输入的标题内容的呈现，然后根据PDF2TXT转换的呈现，按照一句一句话的样式输出。下载按钮下载得到txt格式的PDF转文本单句内容

1. 抽取
   1. 抽取提示

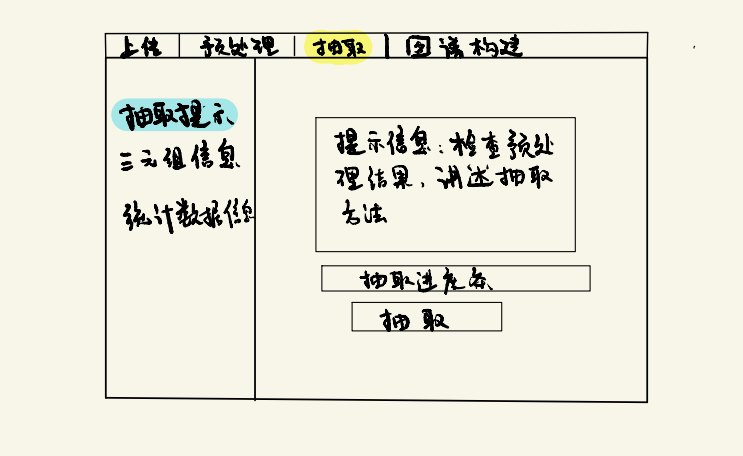


图3-1-1 抽取提示

抽取提示信息主要提示用户检查预处理后的结果，讲述抽取的方法。按下抽取按钮，抽取过程开始，抽取进度条开始走动，走到100%时结束。可以考虑在提取结束后自动跳转到三元组信息的呈现部分。

* 1. 三元组信息

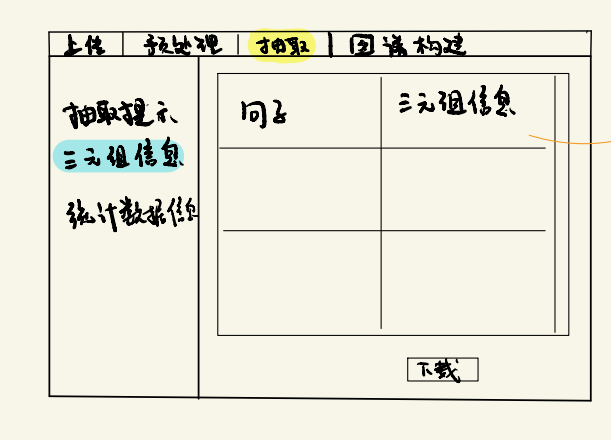


图3-2-1 三元组信息呈现

三元组信息按照表格的形式呈现，表格是两列格式，第一列是原句，第二列是三元组信息，<实体，关系，实体>，这里也可以提供下载，点击下载按钮，可以得到json格式的文件。

* 1. 统计数据信息（暂定）

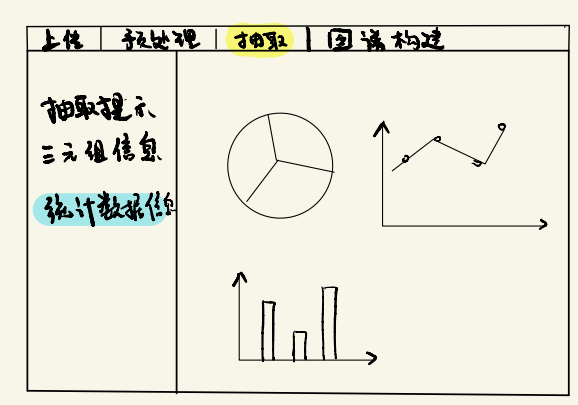


图3-3-1 统计数据信息呈现

这里主要呈现抽取得到的一些信息的统计表格，比如实体类别占比饼图，实体数量柱状图等等。这些图表都要以可交互的形式呈现。

1. 图谱构建
2. 注：添加单篇文献图谱、多篇文献图谱、同一个文献的子图等

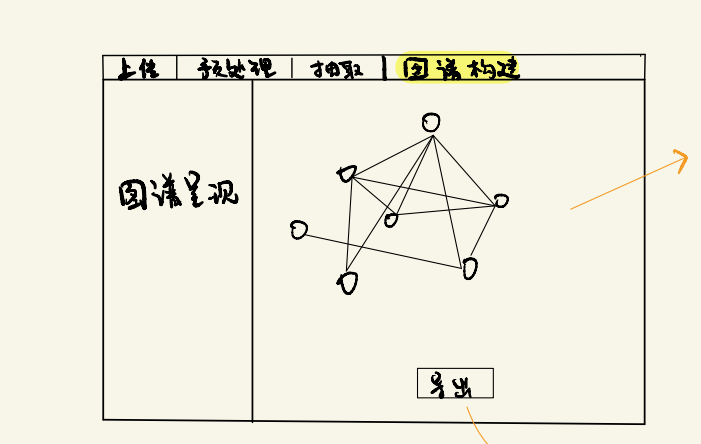


图4-1-1 图谱呈现

图谱呈现部分要呈现出图谱的内容也是交互式的，这个部分还可以添加搜索生成子图谱的样子。导出按钮时可以实现导出图谱。