北京高校大学生创新创业训练项目

定级检查表

项目编号_____

项目性质：创新训练□ 创业训练□ 创业实践□

项目名称:基于机器学习的网课思维导图制作与内容解析平台

项目名称（英文）：Machine learning-based thinking map production and content analysis Platform for online courses

项目依托学院：计算机学院

项目负责人：胡鼎新

联系电话：15111400290

E-mail：2098507840@qq.com

指导教师：李蕾

E-mail：leili@bupt.edu.cn

起止年月：_____

填报时间 2020 年 9 月 15 日

填写说明

1、 本检查表所列各项内容均须实事求是，认真填写，表达明确严谨，简明扼要。

2、 检查表为大 16 开本（A4），根据内容需要可自行加页，但格式须与原件一致。

3、 检查表填写完毕后，须在“北京邮电大学大学生创新创业计划训练平台” 端口开放时间段内上传并提交，并请项目指导教师在网站上进行审核。

4、 提交检查表前，要确定所有团队成员和指导教师已经在“北京邮电大学大学生创新创业计划训练平台”网站上注册完成，否则无法正常提交。

5、 检查表填写时，“项目编号”“项目组成员签字”“指导教师签字”“指导教师综合评价”“评审意见”无须填写。

<table>
<tr><td>项目名称</td><td colspan="7">基于机器学习的网课思维导图制作与内容解析平台</td></tr>
<tr><td rowspan="2">项目负责人</td><td rowspan="2">胡 鼎新</td><td>学号</td><td>2018211932</td><td>所在学院</td><td>软件学院</td><td>手机号</td><td>15111400290</td></tr>
<tr><td>专业</td><td>软件工程</td><td>班级</td><td>2018211503</td><td>邮箱</td><td>2098507840@qq.com</td></tr>
<tr><td rowspan="2">指导教师</td><td rowspan="2">李蕾</td><td rowspan="2">职称</td><td rowspan="2">副教授</td><td rowspan="2">所在学院</td><td rowspan="2">计算机学院</td><td>手机号</td><td>13810927112</td></tr>
<tr><td>邮箱</td><td>leili@bupt.edu.cn</td></tr>
<tr><td>项目类别（类别说明见立项指南）</td><td colspan="7">☐智能硬件☐社交媒体☐数字娱乐☐通信网络☐医疗健康☐公共服务☐电子商务☐√教育文化☐房产家居☐理论研究☐机器人☐无人机☐智能制造☐智能交通
☐创意设计☐其他_____</td></tr>
<tr><td>检索关键词</td><td colspan="7">网课 思维导图 机器学习 自然语言处理</td></tr>
<tr><td rowspan="5">项目成员基本信息</td><td>姓名</td><td>学院</td><td>专业</td><td>班级</td><td>学号</td><td>电话</td><td>邮箱</td></tr>
<tr><td>张超伟</td><td>软件学院</td><td>软件工程</td><td>2018211503</td><td>2018211928</td><td>18210286788</td><td>zhangchaowei@bupt.edu.cn</td></tr>
<tr><td>余子羽</td><td>软件学院</td><td>软件工程</td><td>2018211503</td><td>2018211927</td><td>18860184155</td><td>790023173@qq.com</td></tr>
<tr><td>王俊博</td><td>软件学院</td><td>软件工程</td><td>2018211503</td><td>2018211925</td><td>13805192132</td><td>418866436@qq.com</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

一、 项目进展情况说明

1.1 项目计划要点（目标、内容、关键技术、创新点、商业模式）和调整情况

我们的项目是一个基于机器学习的网课思维导图制作与内容解析平台。

我们希望利用自然语言处理中文本摘要等技术，对课件进行分析，提取出课件中的文字信息，自动形成课程内容的思维导图。再依照思维导图，借助图像识别，语音处理等技术，智能地将网课内容进行解析，使得学习者可以有选择性地进行学习，并更加快速方便地将前后知识融会贯通，形成知识网络。

我们的创新点在于将思维导图的生成过程自动化。借助统计自然语言处理的方法，我们能将课件中的关键信息摘要提取，再自动分析其中的层次结构后，生成对应思维导图。

不同于传统的思维导图，我们将思维导图与网课视频紧密结合。在获取网课视频内容资源的基础上，我们将网课内容的解析自动化。不再需要人工进行标记，借助图像对比，音频分析，我们将自动对网课的视频资源进行解析，并将裁剪好的小片段与思维导图一一对应，使学习者可以学得更细、更精。

项目的整体内容框架如下所示：

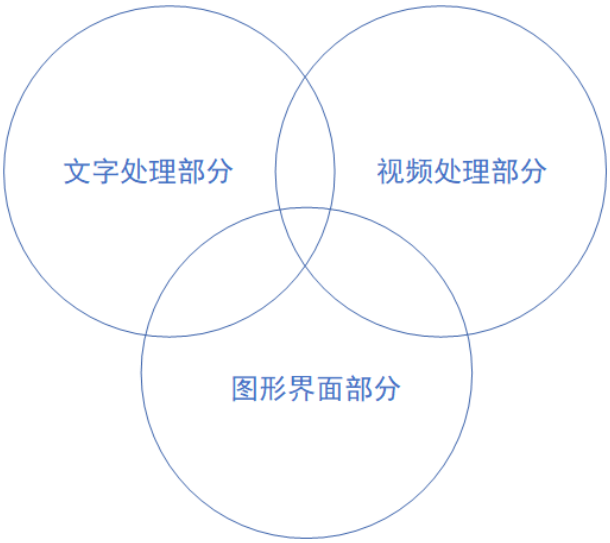

图 **1**　整体内容框架

经过需求的分析以及前期的学习，我们将整个项目主要分为三个方向，分别是：

1) 课件文档中文字信息的提取，内容的自动摘要

2) 课程视频素材裁剪对应

3) 思维导图图像界面的实现

1.2 目前工作主要进展

以下我们就项目实现过程中的各个部分目前的进展逐一汇总：

进展主要分为以下几个部分

1) 前期的专业知识的学习

2) 自然语言处理方面的进展

3) 视频素材识别对应方面的进展

4) 图像界面实现方面的进展

1.2.1 前期的专业知识的学习

时间：2020.5.17-2020.7.7

前期专业知识学习的方向主要是机器学习与自然语言处理。学习的主要途径是阅读相关教材以及学习相关公开课程。我们采取了白天学习相关内容，每天晚上在腾讯会议就相关内容进行分享交流的形式。我们发现这种方式一方面有利于保证进度的正常进行，另一方面在分享既有利于分享者加深对白天所学知识的理解认识，也有利于扩展聆听分享的人所接触的知识面。

以下是我们假期所阅读的一些书籍以及所学习的一些公开课。

《统计学习方法》李航　《统计自然语言处理》宗成庆　《python自然语言处理》

图 2　相关专业书籍

在《统计学习方法》中，我们对一些经典的机器学习模型有了基本的认识，并跟随其中数学公式的推导，对模型内部的运行机理有利更加清晰地了解，在《统计自然语言处理》一书中，我们初步接触了自然语言处理中的一些基本概念、理论方法和最新研究进展，也详细了解了一些简单的自然语言处理模型。而在《python 自然语言处理》一书中，我们利用 python 语言对之前所接触的一些模型进行了实践。

除去这些专业书籍的学习外，为了对机器学习有具体地了解，我们同时也学习了李宏毅老师地公开课。

图 **3**　相关公开课

在李宏毅老师地公开课中，我们对机器学习地工作原理，常见地机器学习模型，以及监督学习，半监督学习，无监督学习等机器学习方法有了更加生动地认识。

1.2.2 文字处理部分

文档中内容摘要任务分为两个部分，一是内容的提取，二是关键信息的摘要。

内容提取部分

我们选用一般课件通用的 pdf 格式作为信息来源，相较于其他文字储存格式，pdf 有着更强的通用性，但由于其本身是专为保持有固定的展示效果而设计，因此其中文字，图片等内容的编码方式繁琐，无法直接快速准确地读取。

在 pdf 内容读取方面，目前尚未有特别完善地解决方案，但索性已经有不少相关的 python 库可以进行解码操作，因此我们考虑以 pdfminer 库为基础，进行文字信息的提取。

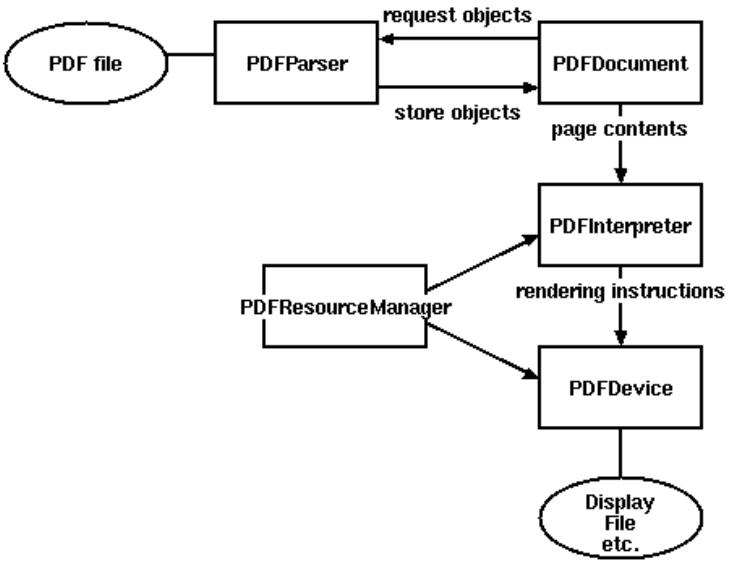

图 **4**　**pdfminer** 库结构图

pdfminer 是一个专为 pdf 中信息提取所设计的库，其各个组成部分分别实现了 pdf 内容提取，解码，储存等功能，除了从 pdf 的源文件中获取文本外，它还可以获取文本的确切位置，字体及颜色等相关信息。

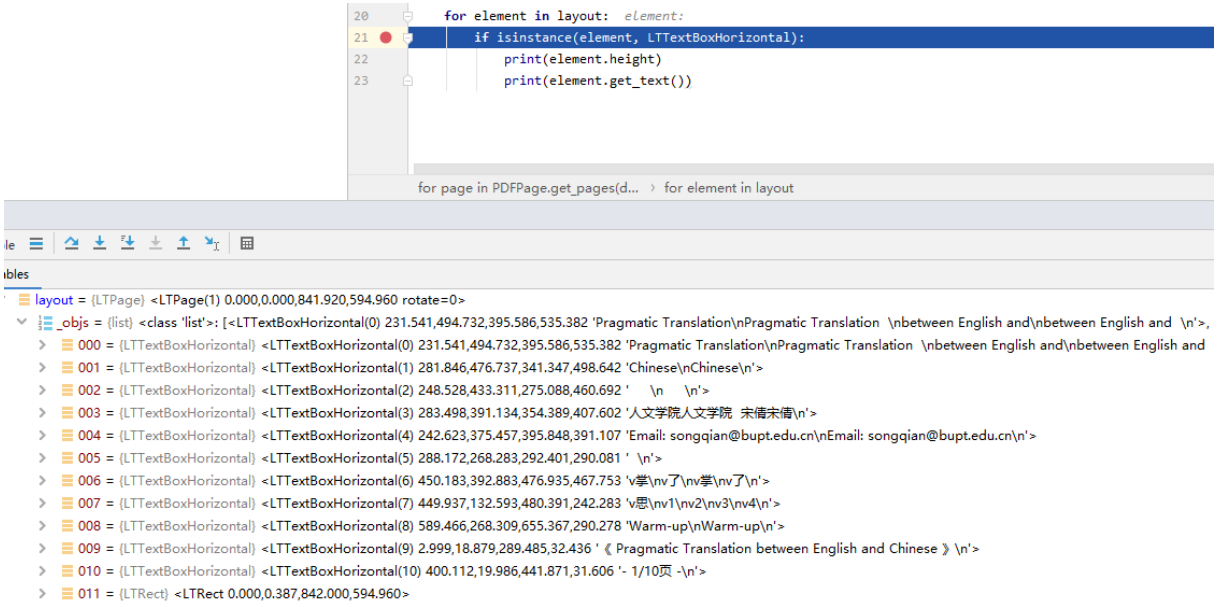

图 **5** 函数返回值

通过使用 pdfminer 库，我们可以有效地对 pdf 文件中的信息进行分析，由上述示意图可知，我们可以成功提取 pdf 格式中各个组成部分的信息，并用列表的形式进行储存，后续我们只需要对这些信息进行分析，便可以快速筛选出关键语句。

同时，通过实际尝试，我们发现利用 pdfminer 所提取的文字存在如下几类问题

1) 被多余换行符分作数行

2) 字符之间存在英文乱码

3) 文本丢失

经过分析，我们发现问题一是由 pdf 的固有格式造成的，在 pdf 中，本身不存在换行符这一概念，文字以“行”而非“句”作为基本单位，因此在所有换行的位置都会存在这一误判的情况，而问题二与问题三则是由于 pdf 编码方式的复杂性以及 pdfminer 库本身的缺陷所致。

因此我们考虑对识别导出的文本进行二次处理，将错误的换行部分进行修正以解决问题一，并期望在后续进一步了解 pdf 格式的编码原理以解决问题二三。

对于转化出的文本格式错误的问题，我们针对一些样本提出了一些解决问题的方法。解决问题最关键的在于如何判断某一行是否被错误划分。为此我们所给定的判断标准如下：

1、以标点判断。如果该行除去换行符的最后一个字符为代表句子结束的标点，比如“。”、“?”以及“!”，那么该行被判定为正确分行。

2、以字符数（每一行的长度）判断。首先通过将所有行遍历一遍，可以得到每一行被填满时的最大字符数（最大长度）。当某一行达到最大长度，最后一个字符却不是

代表句子结束的标点的时候，那么该行判定为错误分行。

3、pdf 转换成文字之后可能已经丢失了文字大小与格式，仅剩文字，由于英文与中文的字符的占位长度并不相同，此时第 2 条无法得出每一行被填满时的最大字符数（最大长度）。在该情况下，如果文本基本为中文，在寻找最大字符数（最大长度）时仅将汉字作为字符。之后，英文也作为字符计数，当某一行达到最大长度，最后一个字符却不是代表句子结束的标点的时候，那么该行判定为错误分行。

关键信息的摘要部分

为实现自动摘要，我们阅读了大量相关论文以及在相关问题上比较成熟的解决方案，归纳出如下几个方向。

1) 文摘句的抽取（Extraction）

2) 相似点与关键词的总结归纳

文摘句的抽取方面

通过阅读相关论文，我们知道目前主流的自动文摘方法主要有 Extraction 和 Abstraction 两类方向，前者是抽取式的方法，通过提取文章中已有的关键词，句子生成摘要，后者则是通过生成的方式得到文章摘要，我们暂时考虑通过抽取的方式获取课件中的关键信息。

我们调研了实现自动摘要的相关算法，主要分为如下几类

图 **6**　自动文摘相关算法

通过前期《统计学习方法》等相关教材的学习，我们对这些方法已经有一定的认识

了解，后续我们将逐步实践，选取出效果最好的方法予以实践。

相似点与关键词的总结归纳方面

对前一步所提取的内容进行总结归纳，是思维导图区别于一般文摘的关键所在。需进行总结归纳，首先就得完成语法分析以及分词等基本任务。

为实现这一类基本的功能，我们上手并熟悉了几个在这方面功能完备的库。

1) Jieba


README.md

jieba

"结巴"中文分词：做最好的 Python 中文分词组件

"Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation module.

- *Scroll down for English documentation.*

特点

- 支持四种分词模式：
  - 精确模式，试图将句子最精确地切开，适合文本分析；
  - 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；
  - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。
  - paddle模式，利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。paddle模式使用需安装paddlepaddle-tiny，`pip install paddlepaddle-tiny==1.6.1`。目前paddle模式支持jieba v0.40及以上版本。jieba v0.40以下版本，请升级jieba，`pip install jieba --upgrade`。
    PaddlePaddle官网
- 支持繁体分词
- 支持自定义词典
- MIT 授权协议

图 7　**Jieba** 分词库

在 jieba 分词库中，可以通过不同的识别模式，对中文语句进行分词处理，并且可以自定义关键词，我们可以通过添加一些专业词汇，使其识别效率更高。

除开基本的分词功能外，jieba 库也提供了词性标注功能，这有利于我们进一步对句子结构进行分析，生成思维导图中的总结性语句，同时，jieba 库通过 TF-IDF 算法实现了关键词筛选的功能，对我们自动生成思维导图这一最终目的也有莫大的帮助。

2) 哈工大 LTP

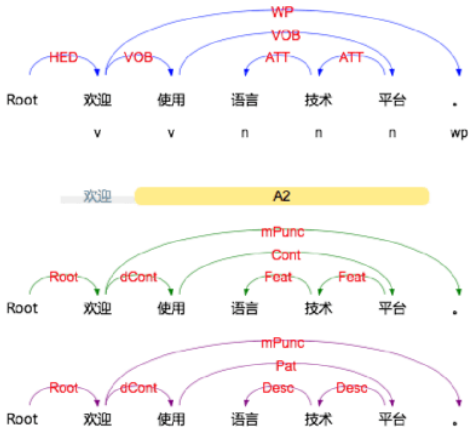

图 **8**　哈工大 **LTP**

LTP 提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等等工作。其中句法分析的功能，有利于我们进行词语依存关系的分析，使我们可以快速地生成文摘句。

1.2.3 课程视频素材裁剪对应部分

我们考虑对视频进行抽帧截图，通过图像比较以确定关键时间节点.

为了实现将图片集进行匹配对应的目的，我们尝试了一些图像匹配的方法，并通过对比从中选取效果较好的方法。我们首先尝试了均值哈希（aHash）、差值哈希 (dHash)、感知哈希 (pHash) 以及灰度直方图算法、三直方图算法和单通道直方图的图像比对方法，但这些算法通常需要将图片进行一定程度的压缩，而且 ppt 中的各张图片之间的差别并不大，匹配图像的准确程度并未达到预期的效果。

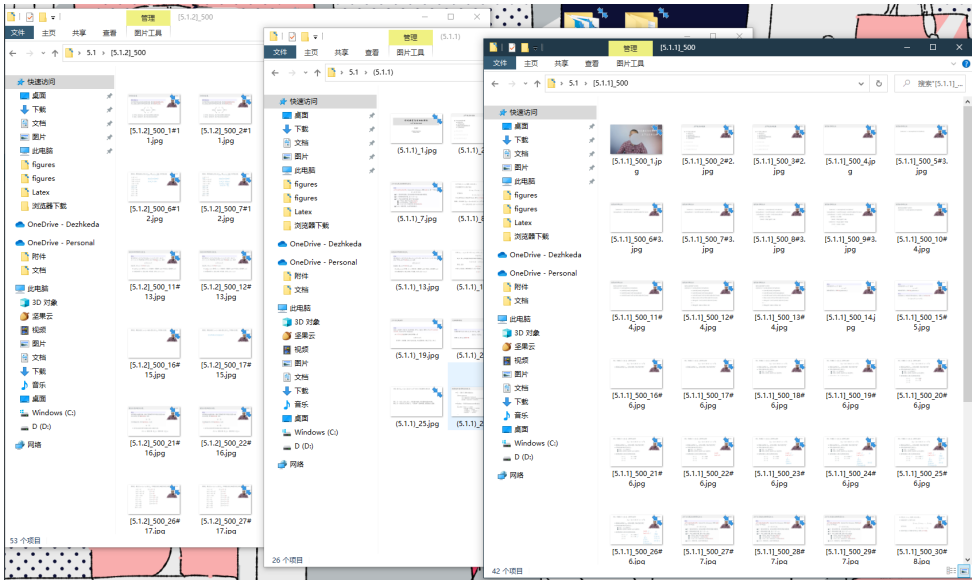

图 **9**　为测试而准备的图像集

我们考虑将图像中的文字进行提取，再通过文字进行图像匹配，但这样的步骤较为复杂，且需要耗费较多的时间，效率不高，而且 ppt 中往往还有一些不包含文字的图表，比如电路图、直方图等，仅仅使用文字进行图像匹配显然是不够准确的。

除此之外，我们也尝试了模板匹配的算法，OpenCV 提供了 6 种计算两幅图像相似度的方法：

差值平方和匹配 CV_TM_SQDIFF

标准化差值平方和匹配 CV_TM_SQDIFF_NORMED

相关匹配 CV_TM_CCORR

标准相关匹配 CV_TM_CCORR_NORMED

相关匹配 CV_TM_CCOEFF

标准相关匹配 CV_TM_CCOEFF_NORMED

经过使用各种算法进行图像匹配操作并进行横向比较，最终我们通过对模板匹配中的标准相关匹配（CV_TM_CCOEFF_NORMED）算法进行一定程度的利用，得到了匹配程度较高的方法。我们对样本图片进行验证，证实了这样的图像匹配方法达到了预期的效果。

1.2.4 图像界面方面

考虑使用 pyqt 实现桌面软件应用程序：

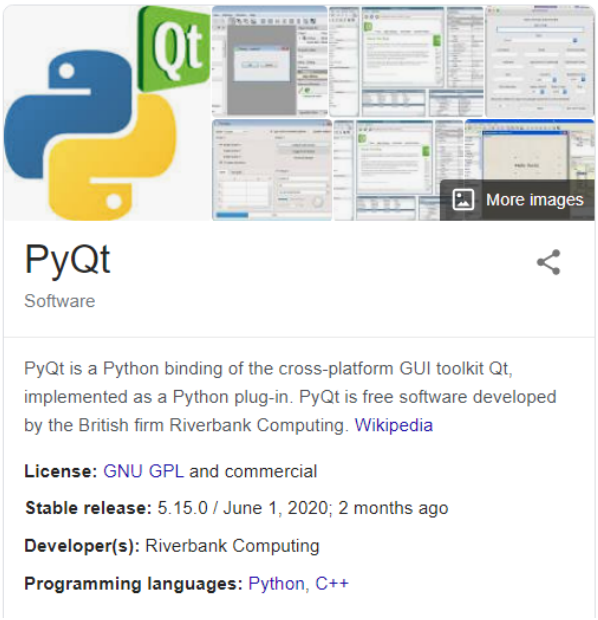

图 **10** **pyqt**

基本的界面设计：创建继承 QWidget 的界面类和继承 QPushbutton、QAction 的组件类。用信号和槽的方式实现包括点击增加节点一类的功能（on_button_clicked()），用鼠

标事件实现包括节点的拖曳、双击展开等功能 (QMouseEvent)。使用 setstylesheet 内置写好的 QSS 设计样式。使用 QMultiMedia 组件控制实现包括视频播放器的功能。结合 python 的 SQLite 调用数据库中的存储数据，用 Veiw 的方式进行展示（QSqlTableModel）和修改（QSqlQueryModel）。下载分析 Xmind 源码进行分析，考虑使用 Graphics Veiw 实现思维导图的树状链接格式。

考虑在网页上进行实现的进度：HTML5+CSS 设计界面，学习包括 jsp 函数控制和 ajax 动态生成实现功能。

1.3 阶段性成果

PDF 文件中文字内容的读取与修正，关键字提取

图 **11**　**pdf** 中文字提取

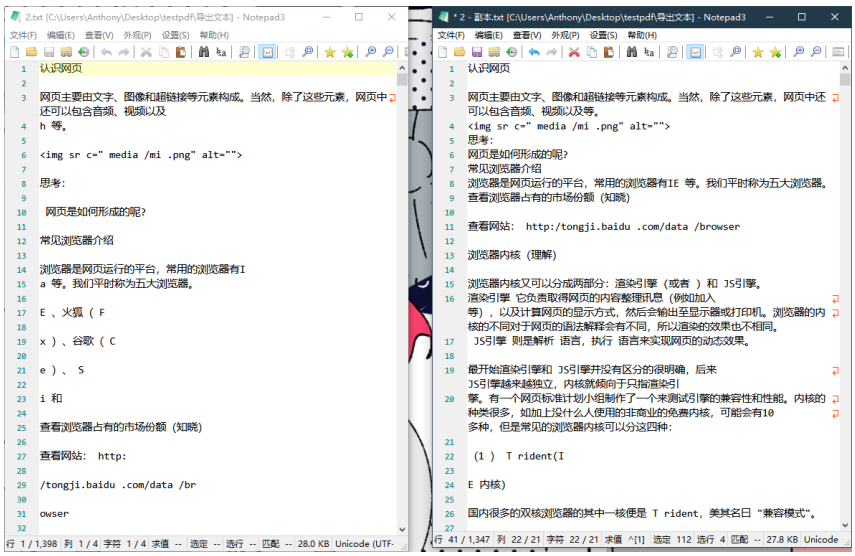

图 **12**　提取文字进行修正

```
C:\Users\Anthony\PycharmProjects\untitled18\venv\Scripts\python.exe C:/Users/Anthony/PycharmProjects/untitled18/pdf.py
Paddle enabled successfully......
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\Anthony\AppData\Local\Temp\jieba.cache
Paddle Mode: 网页/主要/由/文字/、/图像/和/超链接/等/元素/构成/。/当然/，/除了/这些/元素/，/网页/中/还/可以/包含/音频/、/视频/以及/等/。
 浏览器/。
Loading model cost 0.799 seconds.
Prefix dict has been built successfully.

Process finished with exit code 0
```

图 **13**　分词功能的实现

关键时间点处视频图像的确定

图 **14**　视频时间点与课件的对应

思维导图图形界面的框架

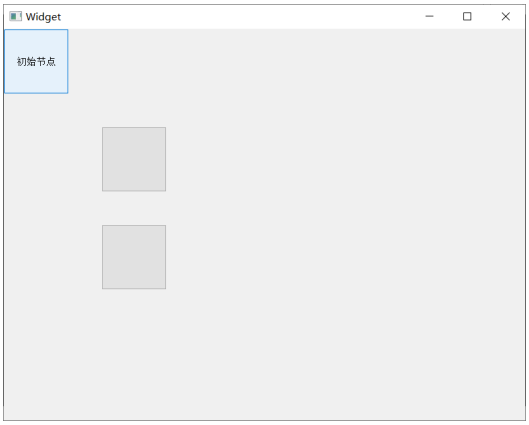

图 **15**　可拖拽的节点的 **UI** 实现

图 **16**　可播放视频的节点的 **UI** 实现

1.4 经费使用情况

无经费开销

二、 项目成员分工及完成情况

时间	胡鼎新	王俊博	余子羽	张超伟
-7.19	阅读《统计学习方法》《统计自然语言处理》	阅读《统计学习方法》就其中公式了解推导过程	前端方面相关内容 阅读《统计自然语言处理》	学习李宏毅相关课程 阅读《统计学习方法》
7.20-7.25	完成《统计学习方法》	阅读《统计学习方法》	阅读《统计自然语言处理》	学习李宏毅相关课程
7.26-8.3	完成阅读《Python自然语言处理》	阅读《统计学习方法》	实践Qt相关功能	学习李宏毅相关课程
8.4-8.9	学习scrapy爬虫使用 熟悉jieba,pypdf,pdfminer库	python爬虫学习 文字处理库学习	实践Qt相关功能	python爬虫学习 python视频处理练习
8.10-8.16	熟悉jieba,pypdf,pdfminer库	使用python进行pdf处理	实践Qt相关功能	制作视频训练数据集
8.17-8.23	阅读关键词提取，自动摘要相关论文	pdf提取后文字的二次处理	实践Qt相关功能	制作视频训练数据集
8.24-8.30	了解LTP中文句法分析	总结pdf文字提取流程	实践Qt相关功能	制作视频训练数据集

图 **17**　成员分工

三、 项目下一阶段工作计划及预期成果

3.1 下一阶段主要研究内容和工作计划

自然语言处理方面。

下一步我们考虑利用库函数，通过收集并整理相关专业词汇，提高关键词识别的精准度，在识别关键词的基础上逐步实现文摘功能，对关键词逐步总结归纳，信息分类以至最终实现思维导图的自动生成等相关功能。

关键时间点处视频图像的确定方面

下一步我们将尝试更多图像相似度比照的方法，提高相似度识别的精准度。并设定视频抽帧的算法，减少性能的开销，提高识别的速度。

图形界面实现方面。下一步我们将把基本功能一一实现，并对界面进行美化处理，并考虑网页版本的实现：使用 HTML5+CSS 设计界面，学习包括 jsp 函数控制和 ajax 动态生成实现功能。

3.1.1 预期研究成果

上线软件，并根据研究成果完成相关文档材料或论文

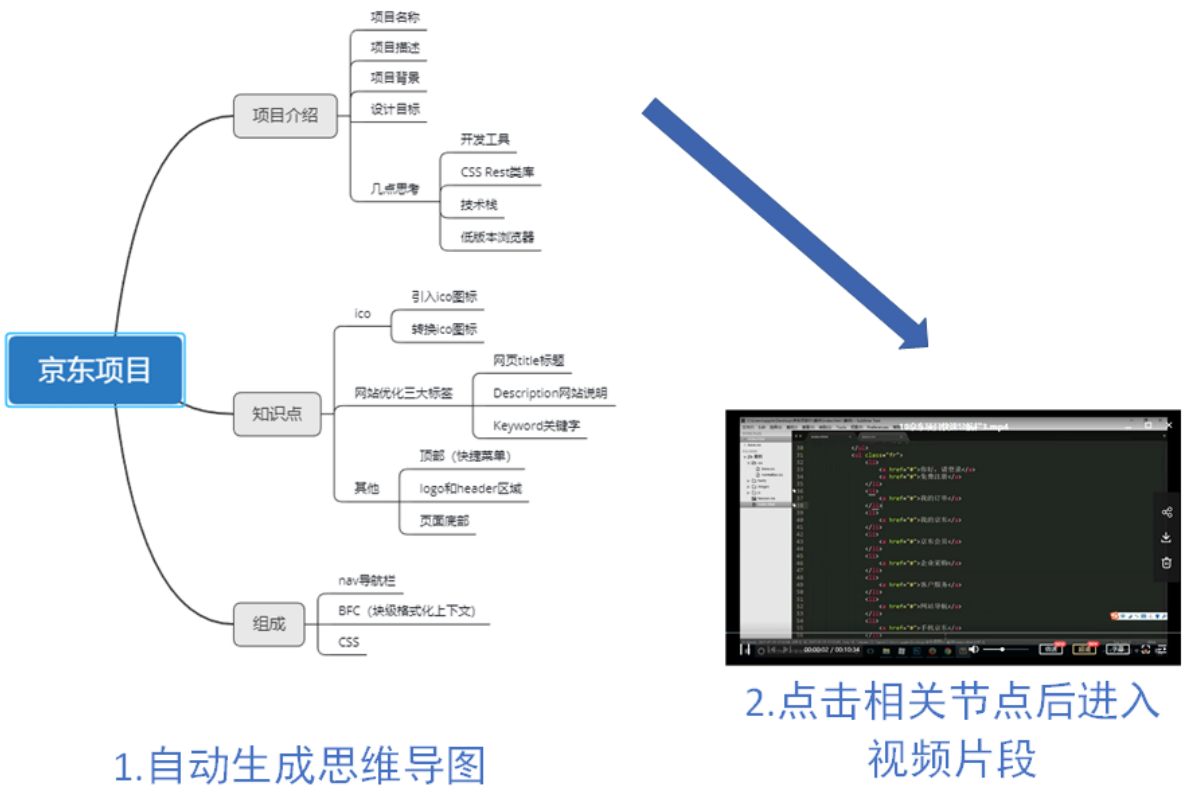

图 **18**　预期实现的功能

3、 经费使用预算

支出项目类别	支出项目说明	支出金额（元）	测算依据
1、业务费	打印费、复印费、装订费、书费、资料费等费用	500	专业书籍购买,资料打印
2、仪器设备购置费	购置或试制专用仪器设备，对现有仪器设备进行升级改造等费用	2000	训练模型所需服务器
3、材料费	芯片、模块、元器件、电路板等低值易耗品费用	0	无
4、外协费	支付给外单位的检验、测试、化验、维修、租赁和加工制作等费用	1000	训练数据集
5、差旅费	开展科学实验（试验）、科学考察、项目调研、学术交流等所发生的外埠差旅费	500	软件需求分析调研
6、会议费	学术研讨、咨询、培训等费用	0	无
7、专项业务费	版面费、专利申请及其他知识产权事务等费用	3000	论文版面费，专利申请
合计（元）	8000		

四、指导教师综合评价		
项目组成员 签字		年　月　日
指导教师签字		年　月　日
五、评审意见		
专家组评价意见： 负责人签字： 年　月　日		