

|  |  |
| --- | --- |
| 题目： | 课程相关信息密度系统分析设计报告 |

|  |  |  |  |
| --- | --- | --- | --- |
| 小组成员 | 学号 | 系统分工 | 报告分工 |
| 顾伟业 | 1800016625 | 检索匹配、可视化展示、txt文档转格式 | 检索匹配的设计与实现、可视化展示的检索匹配的设计与实现、试验数据分析 |
| 李雨佳 | 1800016630 | 音频切分、音频转文字、识别内容抽取 | 相关信息密度文献调研，系统整体结构，信息资源选择与采集的设计与实现 |
| 唐震怡 | 1800016602 | 视频获取、音频提取 | 系统目标、需求分析、线上教育课程内容检索文献调研 |
| 曾子欣 | 1800016623 | 标引模块、数据库维护、检索匹配 | 标引模块/数据库维护的设计与实现、检索匹配的设计与实现 |

目录

[一、 系统目标 1](#_Toc42329567)

[二、 需求分析 1](#_Toc42329568)

[三、 文献调研 1](#_Toc42329569)

[（一） 线上教育课程的内容检索 2](#_Toc42329570)

[（二） 相关信息密度 2](#_Toc42329571)

[四、 系统整体结构 3](#_Toc42329572)

[（一） 设计原理 3](#_Toc42329573)

[（二） 系统功能概况 4](#_Toc42329574)

[（三） 功能模块说明 4](#_Toc42329575)

[1、 信息资源选择与采集模块 4](#_Toc42329576)

[2、 标引处理模块 4](#_Toc42329577)

[3、 数据库创建与维护 4](#_Toc42329578)

[4、 知识组织模块 5](#_Toc42329579)

[5、 用户接口模块 5](#_Toc42329580)

[6、 提问处理/检索匹配模块 5](#_Toc42329581)

[（四） 模块联系与工作流程 6](#_Toc42329582)

[五、 主要功能模块的分析、设计与实现 6](#_Toc42329583)

[（一） 信息资源选择与采集模块 6](#_Toc42329584)

[1、 分析与设计 6](#_Toc42329585)

[2、 实现 7](#_Toc42329586)

[3、 试验结果展示 12](#_Toc42329587)

[（二） 标引处理/数据库创建与维护 13](#_Toc42329588)

[1、 分析与设计 13](#_Toc42329589)

[2、 实现 13](#_Toc42329590)

[3、 试验结果展示 16](#_Toc42329591)

[（三） 提问处理/检索匹配 16](#_Toc42329592)

[1、 分析与设计 16](#_Toc42329593)

[2、 实现 18](#_Toc42329594)

[3、 试验结果展示 21](#_Toc42329595)

[六、 开发环境与开发工具 23](#_Toc42329596)

[七、 试验数据分析 23](#_Toc42329597)

[八、 其他 24](#_Toc42329598)

[参考文献 25](#_Toc42329599)

1. 系统目标

由于新冠肺炎的影响，大部分学生通过在线教育的方式来学习。在线教育一般分为直播和录播两种形式，但不管是直播还是录播，学生都可以获取网课视频以便之后反复查看。但是，网课视频一般分为多段，且每段时长从一小时到两小时不等，虽然有倍速播放和时间轴调整功能，但在用户有明确复习目标（课程重点或具体知识点）的情况下，会在相关的课程内容片段的定位上花费过多的时间。

本系统的设计目标则是在在线教育的情境下为学生定位知识点、查看网课视频提供更多的便利，通过分析视频的相关信息密度的分布情况来帮助学生在数个网课录像中较为精确地定位到包含自己需要知识的片段，从而节省学生的复习时间，提高学习效率。

1. 需求分析

学生在上网课的过程中可能会遇到各种各样的状况，比如因为网络卡顿导致不能正常接收老师的直播讲解、因为走神错过了老师对某方面知识的讲授、对部分知识接受困难一时难以理解等等，在这些情况下学生就需要课后对直播视频进行查看。还有对直播讲授内容存疑、课后想要复习某方面课堂内容等情况，学生也需要对网课视频进行回放。在这些情况下学生是带着明确的查询需求来观看课程视频的，是需要尽可能快地定位到他们想要了解的知识的。所以在复看网课视频时，对于有明确查看目标（或复习目标）的学生来说，尽快定位到包含自己所需知识的视频片段上可以极大提高他们的查询效率。但是当前的视频回放定位只能凭借人工根据记忆去调整，如果记忆模糊或者错误就会耗费用户大量时间，不仅效率低下还会增加用户内心的焦虑，所以他们在此功能上也是存在较大需求的。

1. 文献调研

根据对用户需求的分析，我们将研究问题分为线上教育课程内容检索、相关信息密度两个方面进行文献调研。

1. 线上教育课程的内容检索

随着计算机和多媒体技术的发展，视频教学的不断普及，出现了基于内容的视频检索技术（Content based video retrieval technique, CBR），目前国际上也研发出多个基于内容的视频检索系统。比如MIT media lab研究的Photo Book系统，能够实现基于基本的视觉特征的视频检索；IBM公司的QBIC系统；美国哥伦比亚大学的 Visual seek系统——视觉特征搜索引擎[1]。在文献调研中发现，对于具体的检索过程主要可分为“描述模块”、“匹配模块”、“提取模块”、“认证模块”四个部分，同时需要建立多维度的索引[2]。在内容提取中既要注意视频特征又要注意语义概念，音频领域也要实现音频分割和分类[3]。具体实现上张强提出对多媒体课件建立松耦合多媒体数据库，使用倒排索引进行检索，并对于索引建立了一棵拼音检索树,以多媒体课件中的知识点为叶子,给出了树的创建、遍历、查找等算法,同时对这些算法进行理论分析论证，由此对倒排索引中索引项的冲突也给出了解决方法[4]。刘俊晓对视频结构进行了分析，自上而下分为视频序列、视频场景、镜头和帧，并提出从视频分类、概念学习和事件检测识别进行特定语义分析，对多个检索条件（用户检索词、颜色等）运用欧几里德距离公式进行相似度匹配，根据系统结果的检全率与检准率来进行评价[5]。

由于我们小组主要根据已有网课视频进行研究，所以更加注重用户检索与呈现部分，而且由于网课视频的特殊性和用户的独特性，情境、颜色等视频中的特征内容并不属于我们的检索范围，因此我们研究的重点在于网课知识内容（文本）的检索过程与检索结果。

1. 相关信息密度

最早明确提出“相关信息密度”这一概念的中文文献出现于上世纪90年代，曹希平（1990）[6]提出，由于机检相较于手检文献量大，但有很大一部分文献仅是涉及到相关课题，并不是以相关课题为重点进行论述。而用户关心的不仅是文献中情报是否准确，而且还关心相关情报量是否充足（需要大于某一个固定阈值），这是相关信息密度提出的起因。同时曹希平提出，相关信息密度是单位文献中的相关信息量，提出用户所需最小相关信息密度公式，使用相关信息密度的概念修正了查全率和查准率的公式[[1]](#footnote-1)。之后一段时间，学者们大多聚焦于“信息密度”的探究，而此处的“信息密度”是期刊评价的指标之一，指的是每个印张所刊载结构概念完整的标题数量[7]。Chao Yang等人（2014）[8]通过可视化的方式展示Twitter推文每一天每一个时刻的推文数量与发出的位置，以此来呈现推文信息密度的分布情况。但是以上文献都没有对“相关信息密度”做出较确切的定义。王征，王林森（2015）[9]受到Chao Yang（2014）的启示，认为相关信息密度是指一定单位的信息中包含的有效信息量指标，对固定时间段内微博信息进行采集，计算单位时间内话题词汇数量来推算信息密度高的话题。与我们的课题“课程相关信息密度”最为契合的，是Benoit Favre等人于2007年发表的关于语音数据库相关信息密度可视化呈现的文献。文中指出，对于音频数据库来讲检索和定位都是比较困难的，为用户提供高频词汇又无法保证满足需求，因此采用用户输入检索式，系统将检索式出现的位置在时间轴上标出，从而用标示锚点的密集程度来显示相关信息密度，也就是直观地用某段时间内的关键词数量来表示相关信息密度。这与我们对相关信息密度的界定、可视化想法也是一致的[10]。

1. 系统整体结构
2. 设计原理

课程信息密度系统的设计，根本目的是为了节省学生复习时间，提高复习效率，而直接措施是返回用户关心的关键词在整个课程中的密度分布情况，帮助用户直接定位恰当地观看视频的时间节点。因此，最关键的是将词的密度与时间相对应。由此可见，我们对课程视频的处理，实际上是对课程视频中老师的话语进行分析，归根结底是将对视频的检索转化为了对文本的检索。因此首先需要将视频资源提取出音频，将音频转换为文字，然后进行根据用户输入的关键词进行密度的计算，最后为了更直观、更方便查看，使用可视化展示密度分布结果。

1. 系统功能概况

系统提供的主要功能为“定位”，用户首先输入想要查询的关键词（比如某个知识点词汇），然后该系统根据用户输入的关键词分别计算各个课程录像的相关信息密度，并返回根据相关程度排序后的结果，用户可以根据返回的结果找到他们需要的视频。

同时，对于选定的课程录像和关键词，用户可以选择将该录像的相关信息密度分布情况以可视化的结果展示，使结果的呈现更加的直观，方便用户更为精确的定位到具体的时间来进行观看。

具体的功能模块及工作流程如下。

1. 功能模块说明
2. 信息资源选择与采集模块

信息资源选择与采集模块主要是为后续模块的处理做准备。具体包括三个阶段的功能，分别是：（1）从视频提取音频（2）将音频切分为等时间片段（3）将音频片段转换为文字。

1. 标引处理模块

在标引处理模块我们主要进行切词处理，加入停用词表过滤掉无用字词。接着对词语所处时间建立倒排档索引，节省存储空间，提高检索效率。在标引模块我们没有进行词语加权处理，是因为考虑到一节完整的课程中出现的任何有实际含义的词语都有可能成为用户的检索词，同时每个词语的地位是不能预置的，它们地位是等同的，因此对每个切分出的词语权重没有加以区分。

1. 数据库创建与维护

数据库中存储的主要有三大部分内容。

一是切分并清洗后的课程每分钟中包含的词语组成的文档。二是课程内容相关的信息，即标引处理模块已经考虑到的（词语,时间）的倒排档文档。三是词语的密度信息，以（课程名，时间）为键，以时间段内的该词出现的次数为值。

在实际运用中，由于每次需要分析的课程内容是不同的，因此数据库内容是根据用户上传的视频内容变化而更新的。在试验中，我们选取了4节课程内容，因此数据库存储内容为这4次课程。

1. 知识组织模块

知识组织模块功能与其他模块联系紧密。考虑到用户需求，系统中的知识组织模块采用线性工具，主要是停用词典，在标引模块对切分后的课程内容词语进行清洗，同时在用户输入检索式后也对检索式进行了停用词的筛查。通过在两个阶段使用停用词表，筛选课程内容词语与用户检索词中有实际含义的词，方便匹配。

1. 用户接口模块

我们的系统中将用户接口做了淡化处理，主要有两个关键节点需要用户输入：（1）上传课程视频；（2）输入检索词以查看其在课程中的分布密度。

1. 提问处理/检索匹配模块

提问处理/检索匹配模块是系统中的重点功能模块。

主要功能是处理用户输入的关键词然后与数据库的索引文档匹配得到词语分布密度信息。对用户输入的检索式，进行分词、停用词表筛查，然后通过匹配得到其出现时间及频次。

在呈现匹配结果时，我们希望更直观地展示词语分块密度，方便用户查看与定位，因此将词语密度信息可视化，按时间线性排列、折线图形式展示。

1. 模块联系与工作流程

切词+停用词

建立倒排档索引

(词语,时间)索引文档

词语密度索引

(键值对)

信息源

信息资源选择与采集

标引处理

数据库创建与维护

知识组织与管理工具

用户

用户接口

提问处理/检索匹配

数据库

视频提取音频

音频等时间分割

音频段转为文字

课程视频

课程相关检索词

检索式切词+停用词

与课程内容关键词匹配

可视化呈现

词语密度索引-(课程,时间)为键，词频为值

(词语,时间)索引

课程每分钟的词语文档

1. 主要功能模块的分析、设计与实现
2. 信息资源选择与采集模块
3. 分析与设计

信息资源选择与采集模块的工作主要是为后续模块进一步处理提供便利，因此也应以此为依据进行设计。

首先视频选取需要有代表性，并且视频之间需要形成对照来测试试验效果。视频中的音频质量要较高，避免因音频质量导致识别不准，影响后续分析。从视频中抽取音频时，因为要保证转换质量，因此全部采用.wav无损格式。

在音频分段上，考虑到用户的实际需要与效率问题，分段时长需要适中，因此将音频以1min为单位分成小段，这样处理也方便后期更长时间段（如3min、5min）的合成。

在音频转文字方面，最关键的是转换的准确率要高。目前很多公司提供语音转文字的开放式API，比如Google、IBM、百度。在此我们选用了IBM的Watson Speech to Text功能，主要原因有：Watson Speech to Text可以进行中文识别，这是最重要的一点；同时经过测试其转录的准确性较高；其对于音频的质量要求比较低，音质较差的音频也可转录；在识别前可以设置关键词，适用于一些产品名称、敏感主题、个人名称的情况，可以提高识别的准确性。

最后，将 ANSI格式的txt文档，转化成UTF-8格式以便于后续分析。

1. 实现
2. 视频选取

理想系统使用内置数据库，模拟存储用户上传的网课视频。在本次实验中，我们选取两门课程视频，分别信息服务学与太空探索。选取原因是两门课程直接提供完整课程视频，数据易得；因为是录播课程，音频质量较高，无杂音，方便后期识别；两门课程都是理论性较强的，有多个关键概念，便于后期试验。

在每门课程中又选取各选取两节课程，共四节课程，这样选取主要是方便在两门课程之间，课程内部具体的两节课之间进行横向和纵向的比较。命名及时长信息如下：

表格 2-选取课程信息

|  |  |  |
| --- | --- | --- |
| 课程 | 文件名 | 时长 |
| 信息服务学-第1节 | test1.mp4 | 00:57:00 |
| 信息服务学-第2节 | test2.mp4 | 01:38:55 |
| 太空探索-第1节 | test3.mov | 01:01:00 |
| 太空探索-第3节 | test4.mov | 00:45:05 |

1. 音频提取

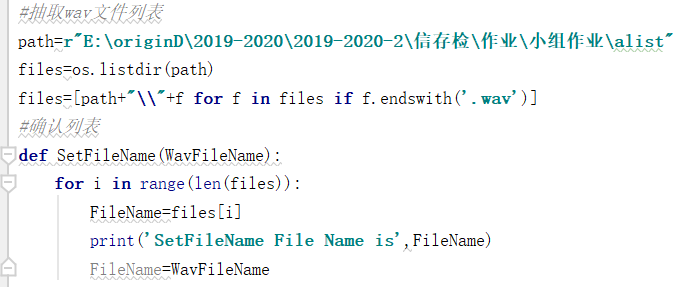
音频提取部分的功能为把视频文件（.mp4）转成音频文件（.wav），此过程一共调用三个模块来完成：

1. os模块：主要用于通过在目录树中游走输出在目录中的文件名；改变当前工作目录到指定的路径；分割路径，返回路径名和文件扩展名的元组。
2. glob模块：用它可以查找符合特定规则（或自己目的）的文件路径名，返回所有匹配的文件路径列表。
3. 从pydub模块中调用AudioSegment：pydub用于处理音频，有一个简单易用的高层接口，底层是依赖于ffmpeg工具。此部分使用该模块的作用主要为读取列表文件且将文件转为wav格式。



1. 音频切分

首先调用os模块抽取制定文件夹下的.wav文件列表，希望确认文件抽取情况因此定义输出文件名称函数：

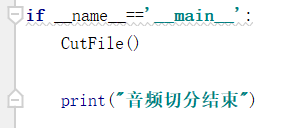


然后定义音频切割函数，主要运用数组运算的numpy模块以及处理音频文件的wave模块。基本思路是通过抽取音频参数，将二进制数据转化为数组，通过参数与循环计算每次切分的起止时间，然后配置基本参数后写入文件：



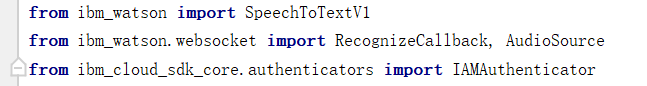


最后运行音频切分函数，得到结果：

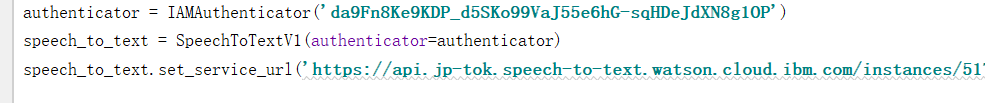


1. 音频转文字

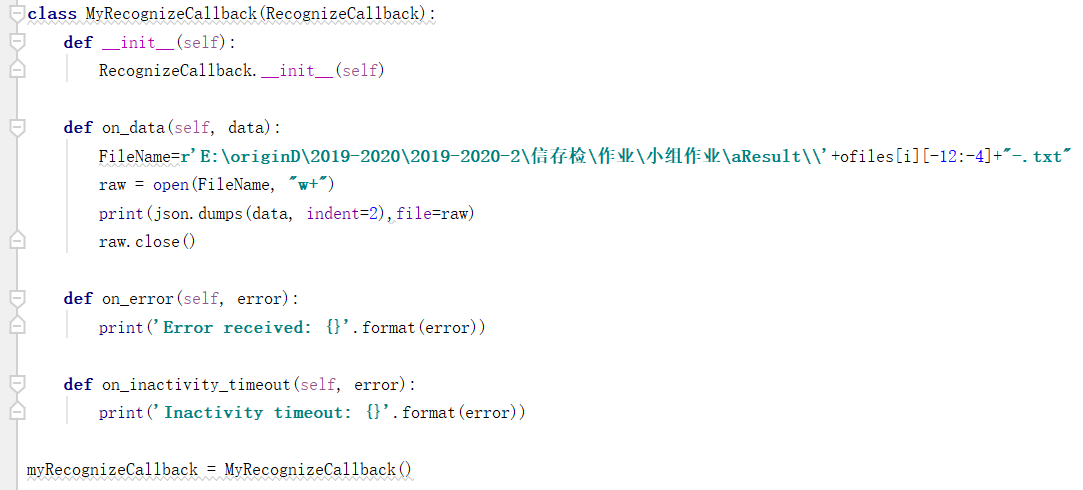
Python自带IBM Watson的模块，具体使用到的模块如下：



首先需要在IBM Watson的官网上注册账号，开通Speech to Text服务，复制服务凭证的密钥与url，然后在程序开始声名密钥和url信息：



定义识别反馈的类，包括写入路径、错误及超时反馈：

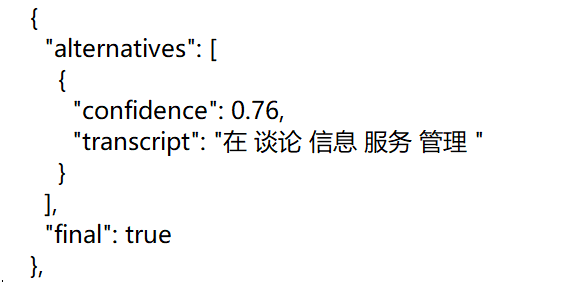


最后在指定文件夹内循环以识别所有文件。由于涉及到的专业概念较多，但同时是比较常见的，因此没有预设关键词及词频：



1. 识别内容抽取

通过Watson Speech to Text的转录结果格式如下图，它有一些转录信度等参数，真正的信息在“transcript”后，因此需要将实际有效的信息提取出来。



实现比较简单，只需要遍历制定文件夹文件，找出需要抽取信息的.txt文件，遍历每个文件中的每一行，找到包含“transcript”的行，进行切片然后抽取所需信息即可：



1. txt文档转格式

这一部分首先调用list\_folders\_files()函数，利用os库来读取文件夹中的所有文件，然后调用convert()函数，利用codecs函数来实现ANSI（这里等同于GBK）到UTF-8格式的转码



1. 试验结果展示

在对四个视频进行处理后，对各个视频得到对应时长的个数个UTF-8格式的txt文档，并存储在video文件夹的各个视频文件夹下，如下图所示（以时长57分钟的test1得到的txt文档列表为例）：



1. 标引处理/数据库创建与维护
2. 分析与设计

首先，对语料进行分词，并且通过停用词表去除掉标引价值较低的词语。我们选择的分词工具是常用的自然语言分析包jieba，根据检索的实际需要采用其中的精确分词模式。停用词表则是从Github上下载了中文停用词表、哈工大停用词表、百度停用词表、四川大学机器智能实验室停用词库四个停用词表后，在人工粗略浏览视频的语音文本后添加了一些词，最终合并成新的停用词表，本系统中运用的停用词表一共收录2318个停用词

然后，我们建立了倒排索引，采用数据结构是python中的字典，键选为关键词，值为一个由多个元组（a，b）构成的列表，其中a表示该词在哪个课程视频出现，b表示出现的具体时间（精确到分钟）。选用这一数据结构的原因是，不同的词项出现的次数差异可能较大，而列表中包含的元素个数可以调整，比较灵活，有利于节省存储空间。

1. 实现
2. 分词处理与建立倒排索引

在主函数中，先对停用词表进行处理，并载入先前建立的倒排索引和视频信息索引。其中视频信息索引记录的是文件名称和对应视频长度。然后，让用户输入新添加的课程文件名称和文件总数，并调用有关函数对倒排索引和视频信息索引进行更新。



下图中的程序将停用词表进行合并，再存储为新的停用词表。



下图显示了几个进行文件读取和存储的辅助函数。



BuildIndex（）函数是这一部分中最重要的函数，负责分词和建立倒排索引。



1. 试验结果展示

对前面得到的所有文本文档建立倒排挡索引后，得到的部分索引如下图截图所示：



如‘描写’:[(‘test1’,’5’)]即表示“描写”该词在四个视频中，在test1视频的第五分钟出现。

1. 提问处理/检索匹配
2. 分析与设计

当用户输入检索词后，系统首先对检索式进行分词，并使用停用词表进行清洗。之后，将检索词与倒排索引进行布尔匹配，并计算出每个视频的相关词数和视频中各时段的相关信息密度。此处对于相关信息密度的具体定义，结合之前文献调研的结果，我们选择以平均每分钟相关词数的数量，即相关词数/分钟，单位为 词/分钟。此处选择“相关词数/分钟”还考虑到如按照相关词数/总文本词量（或相关词数/总文本信息量）会出现因不同老师语速不一样而产生的相同时间文本总信息量不同的问题，导致不同老师的课程视频计算得到的相关信息密度无法比较。详细的信息密度计算公式如下：

为时间片段起点，为时间片段终点。表示视频d在时间段出现的词语的集合，在我们的系统中，将设置为，即计算的时间片段长度为1分钟。 为指示函数。表示视频的总长度。

每个视频的相关词数是以字典形式存储的，键为课程文件名称，值为相关词数。视频中各时段的相关信息密度是以字典形式存储的，键为一个元组（x，y），其中x为课程文件名称，y为出现时间，值为该时间段内相关词出现次数，如在该时间段内相关词没有出现，则记为0。

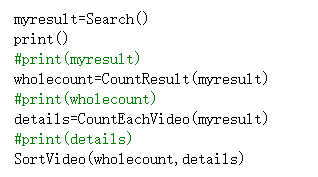
在计算各个视频的相关信息密度并排序的过程中，视频的相关信息密度即以储存的视频相关词数除以视频总时长得到。考虑到可能出现的视频本身信息密度不高，但是可能在部分时间集中出现的情况（即一堂课中，相关词语老师提及次数不多，但是在某几分钟内多次提及），加设了一个新的测量值，相关时长。如果在一分钟内相关词数出现超过一定次数，则这一分钟记为相关，统计被记为相关的时间即得到相关时长。经过对结果的多次人工试验后，将相关时间的判断标准定为该分钟内相关词数出现达到3次，即该分钟相关信息密度达到3词/分钟。

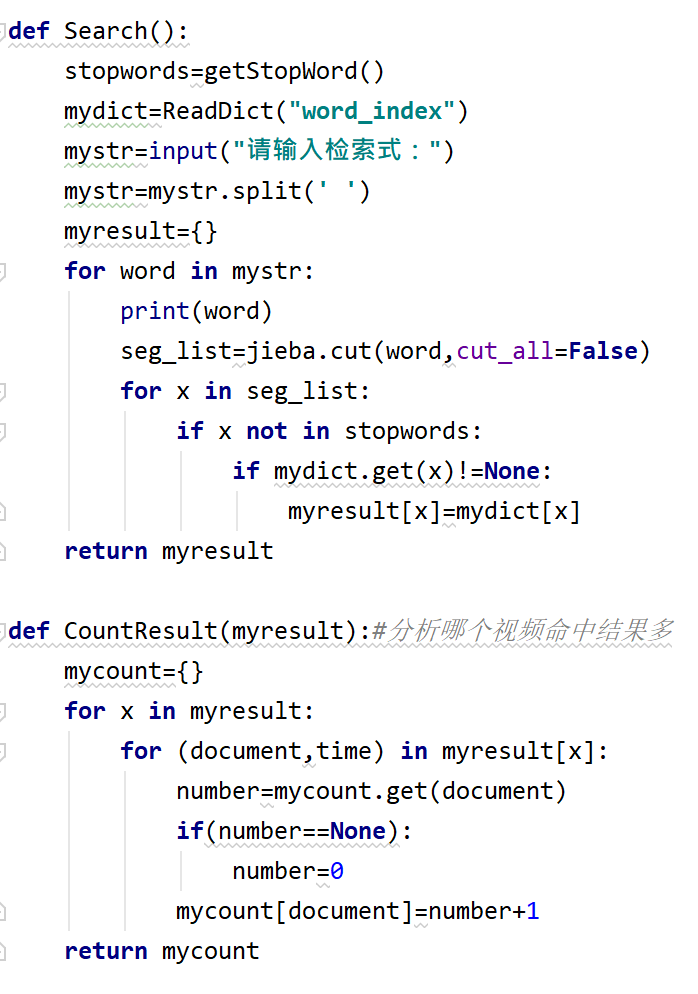
而在实际排序的操作中，一共分为3种情况。当存在视频的相关信息密度具有较高相关信息密度时，则将相关信息密度不为0的视频按照相关信息密度倒序排序，向用户呈现排在前三的结果；当视频的相关信息密度都较小，但存在视频有相关时长时，则将相关时长不为0的视频按照相关时长倒序排序，向用户呈现排在前三的结果；当视频的相关信息密度都较小，且视频都没有相关时长时，则返回“无相关结果”的检索结果。经过对结果的多次人工试验后，将视频具有较高信息密度的标准定为视频的相关信息密度达到0.3词/分钟。

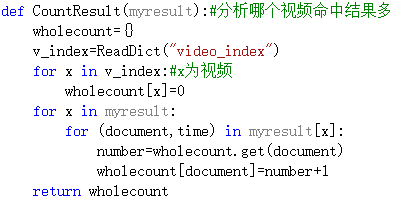
在可视化方面，根据前面储存的视频各分钟相关词数，也就是各分钟相关信息密度。使用折线图将选定视频的各分钟的相关信息密度进行可视化的输出，这样既可以展现视频各分钟的相关信息密度的分布与随时间的变化情况，也可以让用户直观的看到高相关信息密度的片段是在哪个时间，进行精准的定位。同时在图中加入一根与x轴平行的直线，值为视频的相关信息密度平均值，也就是视频的整体相关信息密度。也考虑到可能因为视频过长导致图中点过密的问题，在x轴加入可调节的控制条，用户可选择将自己关注的时间段放大查看具体的相关信息密度分布。

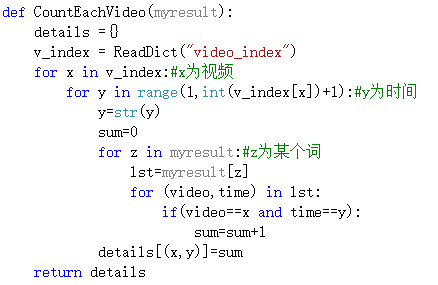
1. 实现
2. 计算文档相关词数与统计各时段相关密度

在主函数中，首先调用Search（）函数进行布尔匹配，该函数返回初步的匹配结果。然后在CountResult（）函数中对课程视频的总相关词数进行统计；在CountEachVideo（）函数中对每个课程各时段的信息密度进行统计。



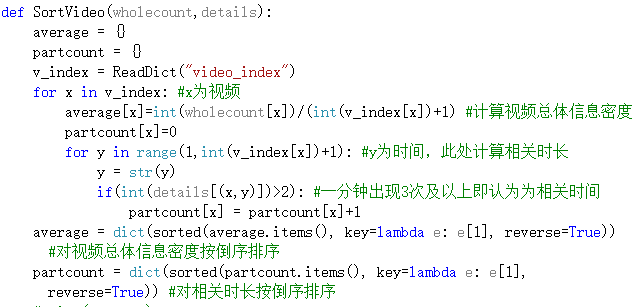






1. 视频排序和结果输出

在主函数中，调用SortVideo函数进行视频的排序，首先对视频总体的相关信息密度和视频的相关时长进行计算，并按倒序对这两组值进行排序。

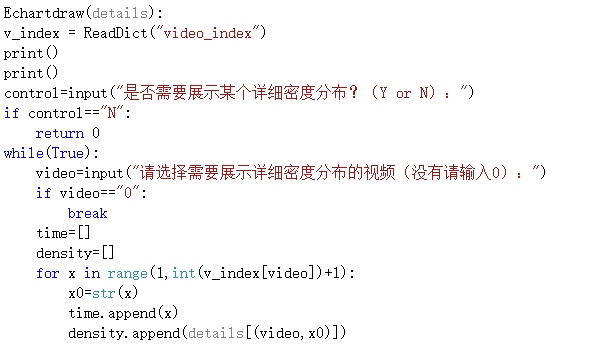


然后将结果确认为存在视频的相关信息密度具有较高相关信息密度；视频的相关信息密度都较小，但存在视频有相关时长；视频的相关信息密度都较小，且视频都没有相关时长三种情况中的一种，并将排序结果呈现给用户。



1. 可视化呈现

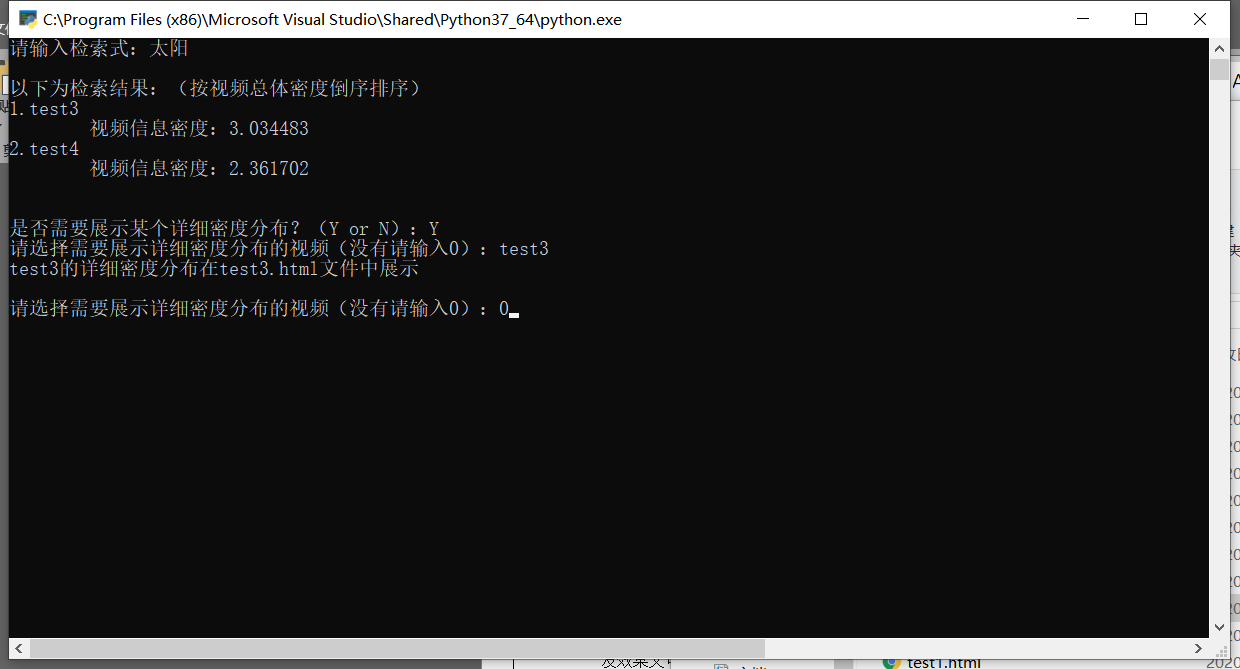
在SortVideo中，如返回的不是“无相关结果”，则会调用Echatsdraw函数，来根据用户选择的视频可视化输出，这一部分调用echarts实现，并在可视化图像成功生成后提示用户输出的可视化图像的存储名字。

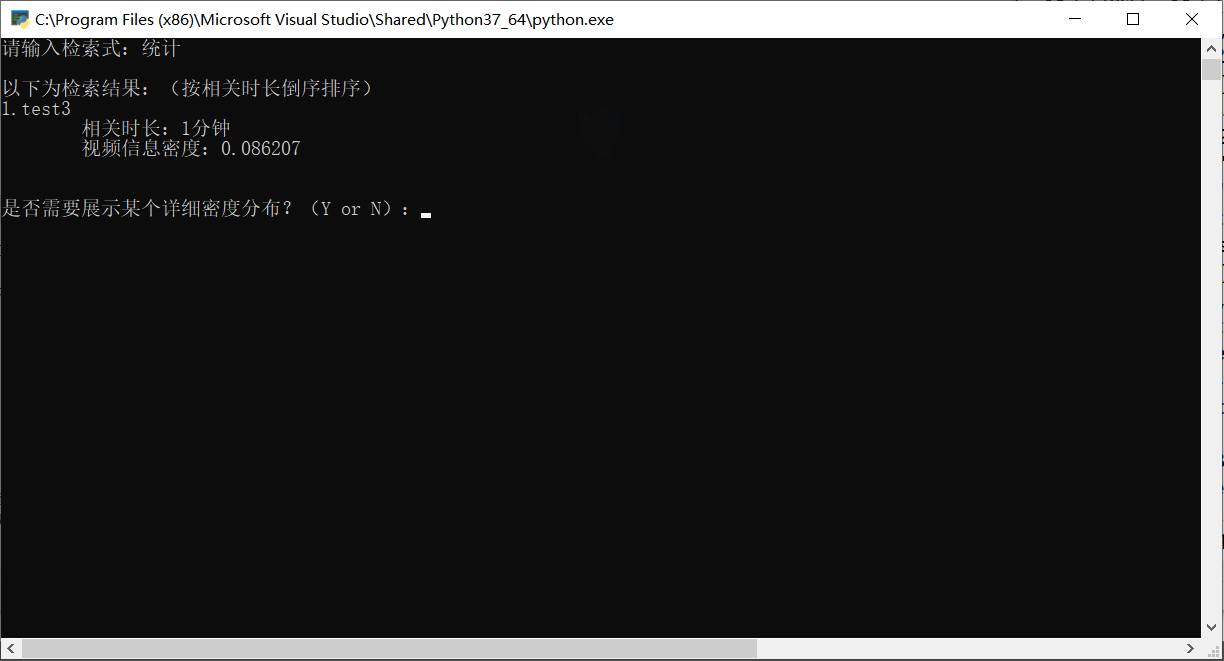


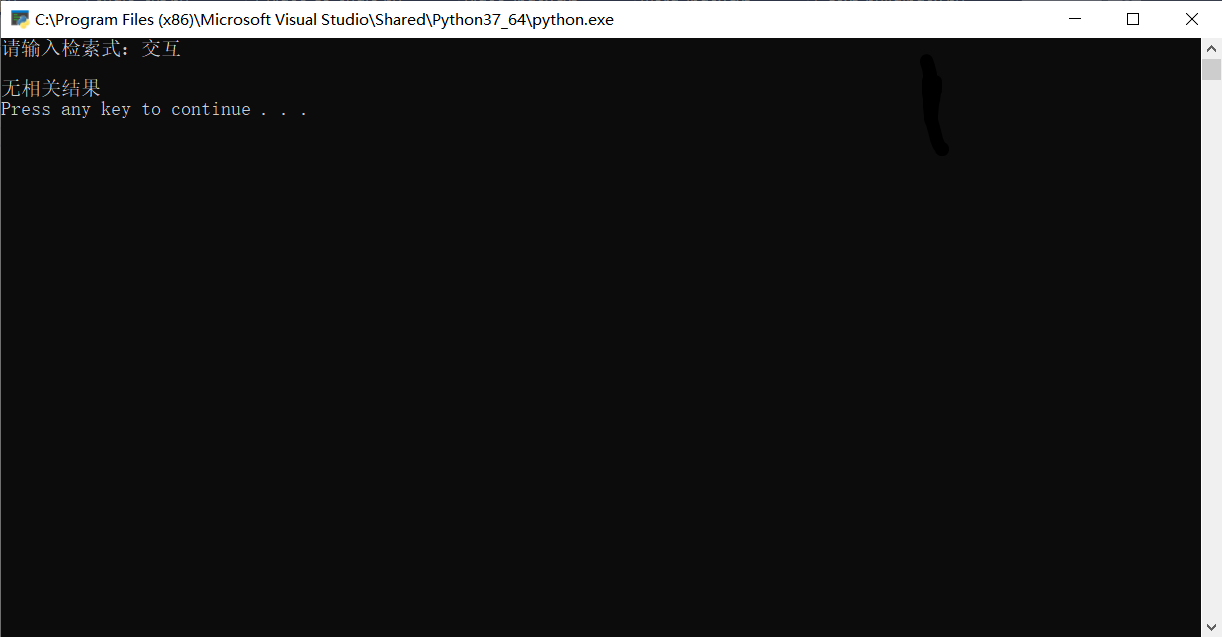


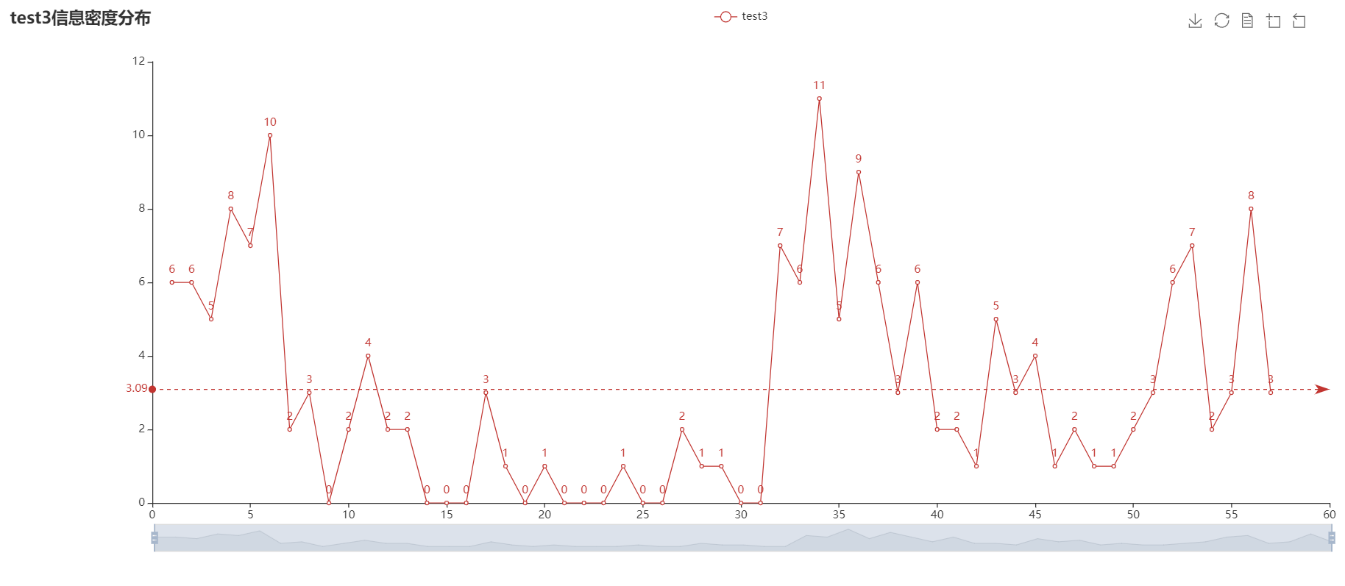
1. 试验结果展示

这一部分即用户的使用界面，用户输入检索词后系统返回检索结果，如非“无相关检索”情况，用户可选择视频进行可视化输出，具体运行截图如下（三幅图分别对应视频的相关信息密度都较小，但存在视频有相关时长；视频的相关信息密度都较小，且视频都没有相关时长三种情况）：







以检索式为“太阳”时的test3为例，其此时的相关信息密度分布的可视化图像如下：

1. 开发环境与开发工具

Windows10环境下实现，采用python语言开发。选择python的原因是，编写简单快捷，同时提供大量功能强大的包，为功能实现提供便利，既能保证开发效果又可以节省时间。比如数据预处理过程中使用到的ibm\_watson相关包，可以直接调用Watson Speech to Text，使系统的连贯性和集成性更好。在信息密度计算模块使用jieba库进行分词，分词准确，效率高。最终的可视化展示使用echarts库作为输出工具，echarts是百度开源的数据可视化JS库，可以生成多种类型图标，使用灵活、高效。

1. 试验数据分析

经过人工输入检索式对系统进行测试，系统根据检索式正常返回对应的视频排序结果，也都实现了指定视频的相对信息密度分布的可视化呈现。将结果与视频对照比对后，检索得到的结果与视频实际都大致相符，因此可以认为系统基本上实现了最初的系统设计目标，即帮助用户定位到自己需要的知识的视频与片段。

但是系统仍然存在诸多不足和可改进的地方。首先为了试验的方便，一共只选取了四个平均长度65.5分钟的视频，分属两个课程，试验的数据量并不大，同时在试验的过程中主要是人工对多个检索式进行测试，并没有进行全面且系统的测试。虽然系统基本上满足了设计目标，但是具体的准确性仍有待测试。

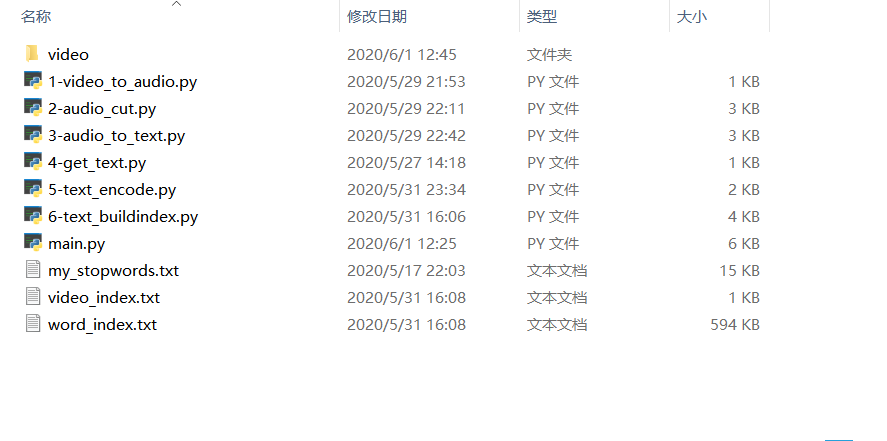
其次系统中用于视频排序的两个指标（视频具有较高信息密度的标准为视频的相关信息密度达到0.3词/分钟；相关时间的判断标准为该分钟内相关词数出现达到3次，即该分钟相关信息密度达到3词/分钟）为人工进行数次测试后得到的经验指标，并没有大量的数据或理论进行支撑，在系统的具体运用中这两个指标仍需要进行调整或重新确认。

另外系统中的停用词表虽然结合了多个中文停用词表，并人工添加了一些停用词，但是该停用词表没有综合考虑课程录像的特点而设置专门的停用词，也没有对类课名中出现的“服务”（“信息服务学”）这样的词进行限制。停用词表还需要进行完善，同时也缺少近义词表或同义词表来辅助检索，达到更好的检全率与检准率。

最后在对视频的语音文本进行分析的过程中，可以看到，因为老师习惯问题，会让语音文本中包含大量的语气词，这一点在“信息服务学”对应视频test1和test2中十分明显。语气词会导致分词建立索引的过程中，部分词语或词组被错误切割，导致检索结果出现偏差。

1. 其他

系统由以下文件组成：



其中 “1-video\_to\_audio.py” “2-audio\_cut.py” “3-audio\_to\_text.py” “4-get\_text.py” “5-text\_encode”五个python程序为系统的信息资源选择与采集模块，依次实现视频的语音转换、语音文件的切割、语音转文字、文本获取、txt文件转换格式；“6-text\_buildindex.py” python程序为系统的标引处理模块，实现对video文件夹中的语音文本建立倒排档索引；“main.py” python程序为系统的提问处理和检索匹配模块，是用户使用该系统的程序，用户打开后可输入检索式进行检索，并选择视频进行可视化输出，输出图像保存在该文件夹下。

另外，video文件夹中存储的是从视频中提取的语音文本,“my\_stopwords.txt”为系统停用词表，“video\_index.txt”为视频索引，“word\_index.txt”为建立好的倒排挡索引。

参考文献

[1] 路程. 视频内容检索技术概述[J]. 山西科技, 2018, 33(02): 56-58.

[2] FANG H G, WEN Z Y, LIU J, et al. Research on the retrieval process of teaching video resources based on content[C]// Research on the retrieval process of teaching video resources based on content. 2009 2nd IEEE International Conference on Broadband Network & Multimedia Technology. IEEE: 472-476.

[3] DIMITROVA N, ZHANG H J, SHAHRARAY B, et al. Applications of video-content analysis and retrieval[J]. IEEE multimedia, 2002, 9(3): 42-55.

[4] 张强. 多媒体数据库检索技术研究与应用[D]. 大连交通大学, 2008.

[5] 刘俊晓. 视频的内容分析与检索技术及其教育应用研究[D]. 山东师范大学, 2007.

[6] 曹希平. “相关信息密度”的概念、意义和应用[J]. 情报科学, 1990(01): 33-36.

[7] 国家科委科技信息司, 中共中央宣传部出版局, 新闻出版署期刊管理司. 关于颁布五大类科技期刊质量要求及评估标准的通知[Z]. 1994-08-02.

[8] YANG C, JENSEN I, ROSEN P. A multiscale approach to network event identification using geolocated twitter data [J]. Computing, 2014(96): 3-13.

[9] 王征, 王林森, 赵磊. 基于信息密度的微博突发话题检测模型研究[J]. 情报理论与实践, 2016, 39(03): 125-129.

[10] BENOIT F, JEAN-FRANCOIS B, PATRICE B. An Interactive Timeline for Speech Database Browsing[C]. 8TH Annual Conference of the International Speech, 2007: 2720-2723.

1. 用户所需最小相关信息密度=广义相关文献量/(阅读能力×激励能力)，修正后的查全率R=A(x)/B(x)，查准率P=A(x)/C（A(x)是相关信息密度大于x的检出文献量，B(x)是文献库中相关信息密度大于x的文献量，C是检出文献总量） [↑](#footnote-ref-1)