**（一）进展情况**

一、内容部分

1.前期数据收集

小组以985、211高校作为对象，选取哲学、经济学、管理学、法学、心理学、社会学、教育学、政治学、文学、历史、新闻传播、艺术共计十二个人文社科类专业，进行讲座预告数据收集，涉及如“新传小馆”“中国传媒大学”“中央美术学院人文学院”等数十家公众账号。小组通过各高校及其下属学院、研究所等机构组织的官网及微信公众号，收集讲座基本信息。并最终按照时间地点、主持人、主讲人、简介、参与方式、主办承办等部分进行分类，最终形成了以1734条来自各大高校讲座的数据，作为最终成果实现的数据基础。

2.公众号策划

（1）公众号设计

为给讲座资讯提供内容载体，小组创立微信公众号“今日人文社科”，并设计相应头像、二维码logo以及头图等，形成相对完整的公众号运营体系。小组将“今日人文社科”定位设定为人文社科类讲座信息共享平台，“掌握讲座资讯，紧跟学科动态，聚焦顶尖高校，直击人文社科热点话题”是其创办理念。目前“今日人文社科”已发布34篇推送，拥有28名用户。





图1 公众号logo与二维码设计

（2）专题推送

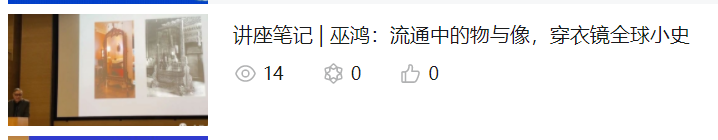
“今日人文社科”公众号开设“专题分享”模块，聚焦传媒领域前沿热点话题，为用户提供学界热点资讯、进行热门概念科普。目前该板块共有三篇内容，分别为《专题分享 | 什么是新基建？》《专题分享 | 什么是数字经济？》《专题分享 | 媒体融合的解构与建构》，对话题进行全方位、深层次的解读。如文章《专题分享 | 什么是新基建？》在对“新基建”给出严谨学界定义的同时，以直观图表形式介绍其“新”在何处。随后，推文分析了新基建如何被提出，并对其未来发展趋势进行了展望。整体多角度、多方位地呈现“新基建”带给社会的影响。



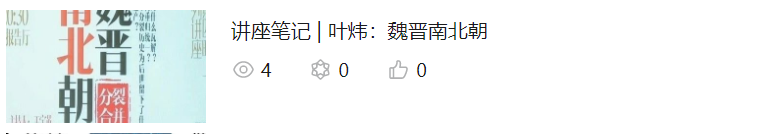
图2 公众号推送《专题分享 | 什么是新基建？》截图

（3）讲座推送

目前“今日人文社科”公众号共发出讲座相关内容推送31篇，选题涵盖计算机、新闻传播、艺术等领域，聚焦学术前沿，为读者提供相应领域顶级学者的讲座记录或名校公开课笔记。如北京大学数学科学学院教授耿直讲座记录《因果推断——数据驱动的因果作用评价》，世界著名美术史家、芝加哥大学教授巫鸿讲座笔记《流通中的物与像，穿衣镜全球小史》，斯坦福大学Coursera公开课CS228概率图模型学习笔记，中国传媒大学中国通史系列、北京大学著名中国史学家叶炜兼顾总《魏晋南北朝》听课记录等等。讲座笔记的发布为读者提供多领域学术前沿追踪、顶级学术资讯，可增强用户黏性，从而为讲座推送自动生成后功能的不断完善提供用户基础。







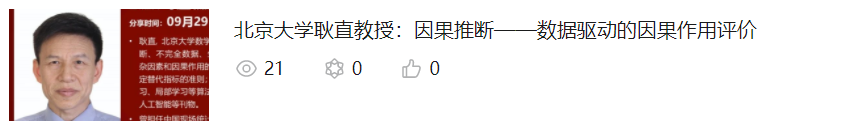


图3公众号后台推送截图

二、技术部分

## 1. 数据获取技术进展

目前主要以高校官网公布的讲座信息为主要信息来源，使用Python进行公开讲座信息的收集工作，辅以人工手段对文本数据进行进一步处理（包括提取关键信息）。收集的数据目前已上传至Github仓库：<https://github.com/JJYDXFS/little-innovation/tree/master/Text_Materials>。

## 网站搭建技术进展

项目最后的呈现方式为：微信公众号定期提供讲座分享+Web端网站提供讲座资源聚合服务及自然语言相关的应用，本部分主要侧重于说明WebApp的搭建进展。网站首页地址为<http://111.229.84.244:5000/>。目前使用html/css/js+Flask的前后端不分离形式进行网站demo的搭建，数据库使用MySQL8.0。前后端分工为由前端编写网页模板、后端Flask主要负责对模板进行渲染和对，动态路由也由Flask实现。

### （1）服务器环境搭建

服务器使用腾讯云服务器标准型S2，配置为1核 2GB 1Mbps，操作系统Ubuntu Server 18.04.1 LTS 64位，硬盘大小50GB。进行过基础LAMP（Ubuntu18+Apache2+MySQL+Php7）环境搭建，前期曾使用该方式进行过前后端分离的开发。目前在网站默认端口80也同步了网站主页。

### （2）前端

采用传统的html+css+js技术进行开发，基于os-templates的开源前端模板，根据项目需求对界面和功能进行改造。使用JQuery与服务器进行数据交换、实现动态加载；使用Vue.js简化DOM操作，后期考虑采用Vue组件化开发对网站前端结构进行升级。目前前端共编写了3个网页模板：首页模板、学科页模板、搜索结果页模板（其中首页模板编写进度较快，其余两个模板只有简单demo）。

#### 2.1 前端功能设计

<1>首页

* 首页预计实现如图所示功能，包括提供各类搜索功能和讲座预告信息，最终是否引入用户注册登录功能暂时未确定。
* 

图 4 主页设计

<2>学科页

学科页预计将提供学科词云和相关的预告、笔记。学科页功能设计如下。

* 

图 5 学科页设计

<3>搜索结果页

搜索结果页将根据关键词给出搜索的讲座信息，搜索结果将按匹配程度排序，并提供部分内容的预览。

* 

图 6 搜索结果页设计

#### 2.2 模板实现进展

<1>网站首页模板

如下图所示，使用GET请求初步实现了导航栏的动态加载和点击跳转以及搜索框的自动跳转功能。



图7 主页截图

<2>学科页模板

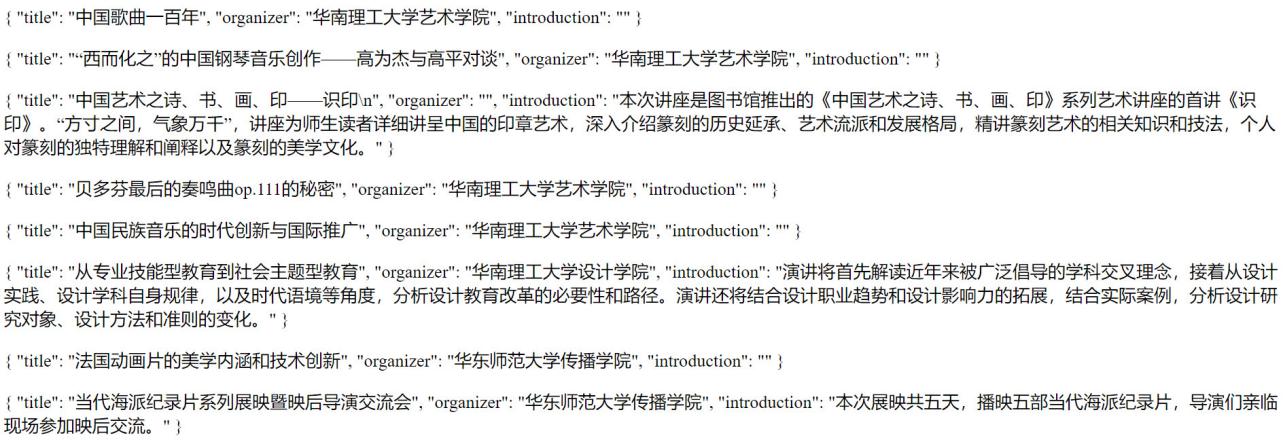
* 学科页暂未进一步细化模板，直接使用POST请求向后端调取对应数据显示在页面上。如图示结果为“艺术”学科页的当前效果图。
* 

图 8“艺术”学科页截图

<3>搜索结果页模板

搜索结果页暂未进一步细化模板，直接使用POST请求向后端调取搜索结果（标题中含有关键词的讲座信息）显示在页面上。如图示结果为关键词“世界”的搜索结果。

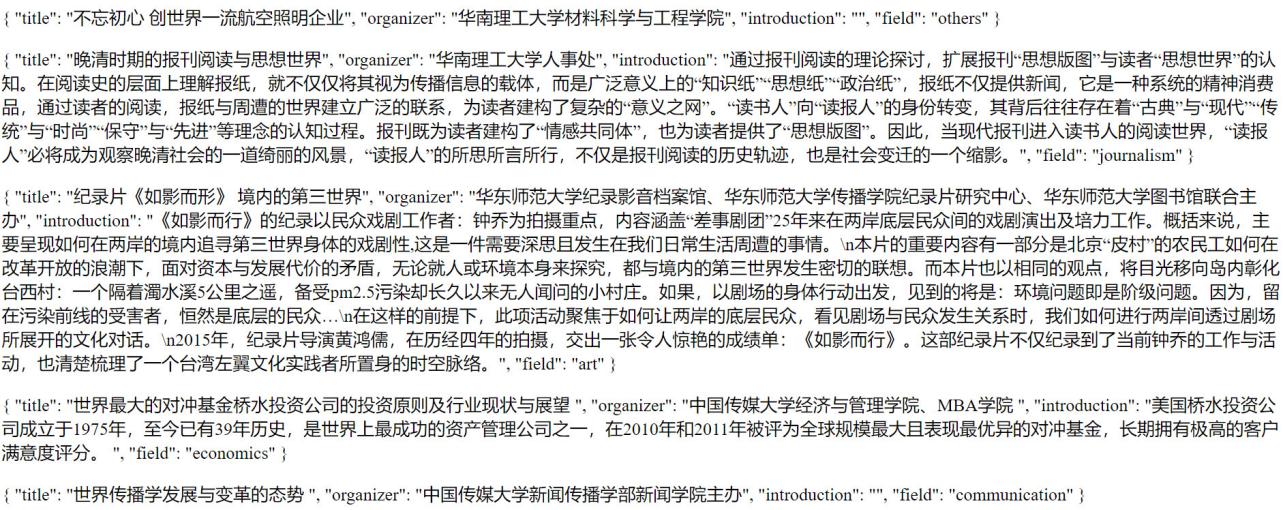
* 

图 9 关键词“世界”搜索结果

### （3）后端

使用Python Flask框架，目前实现了学科页和搜索结果页的动态路由、上述三类模板页的渲染以及三个接口：

* 1. GET：获取学科列表
  2. POST：根据学科返回该学科包含的讲座信息
  3. POST：根据关键词返回标题中含关键词的讲座信息

接口具体实现：

<https://github.com/xiaochuang-JRRWSK/JRRWSK_web/blob/main/API/JRRWSK-2020-11-29.png>

（4）数据库

数据库ER图如图所示，目前数据库中一共收录讲座数据1048条，涵盖学科12类。

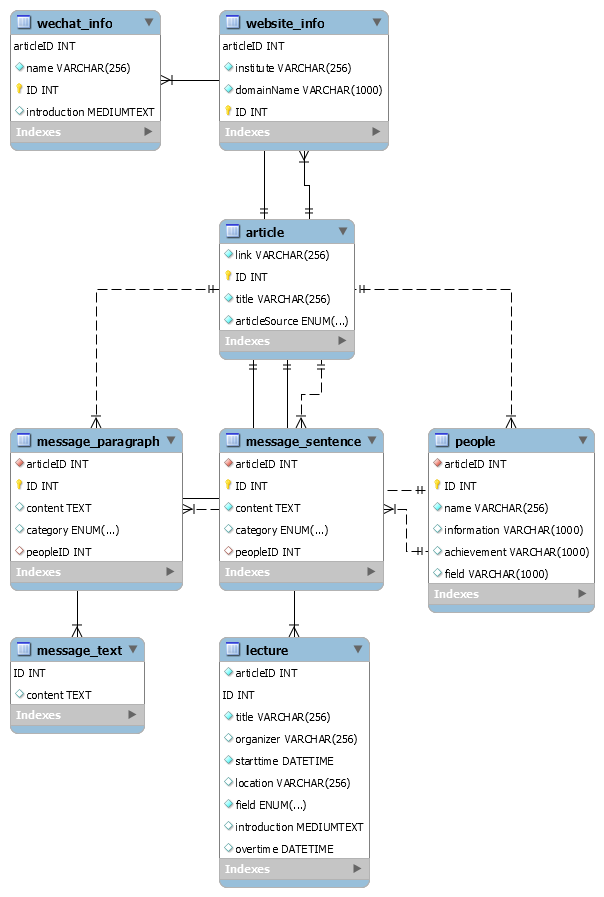


图 10 数据库ER图

1. **开支情况**

小组购置了一台腾讯云云服务器，用来搭建项目网站。有效期从2020年2月25日至2021年2月25日，支出99元。到期后仍会续费。

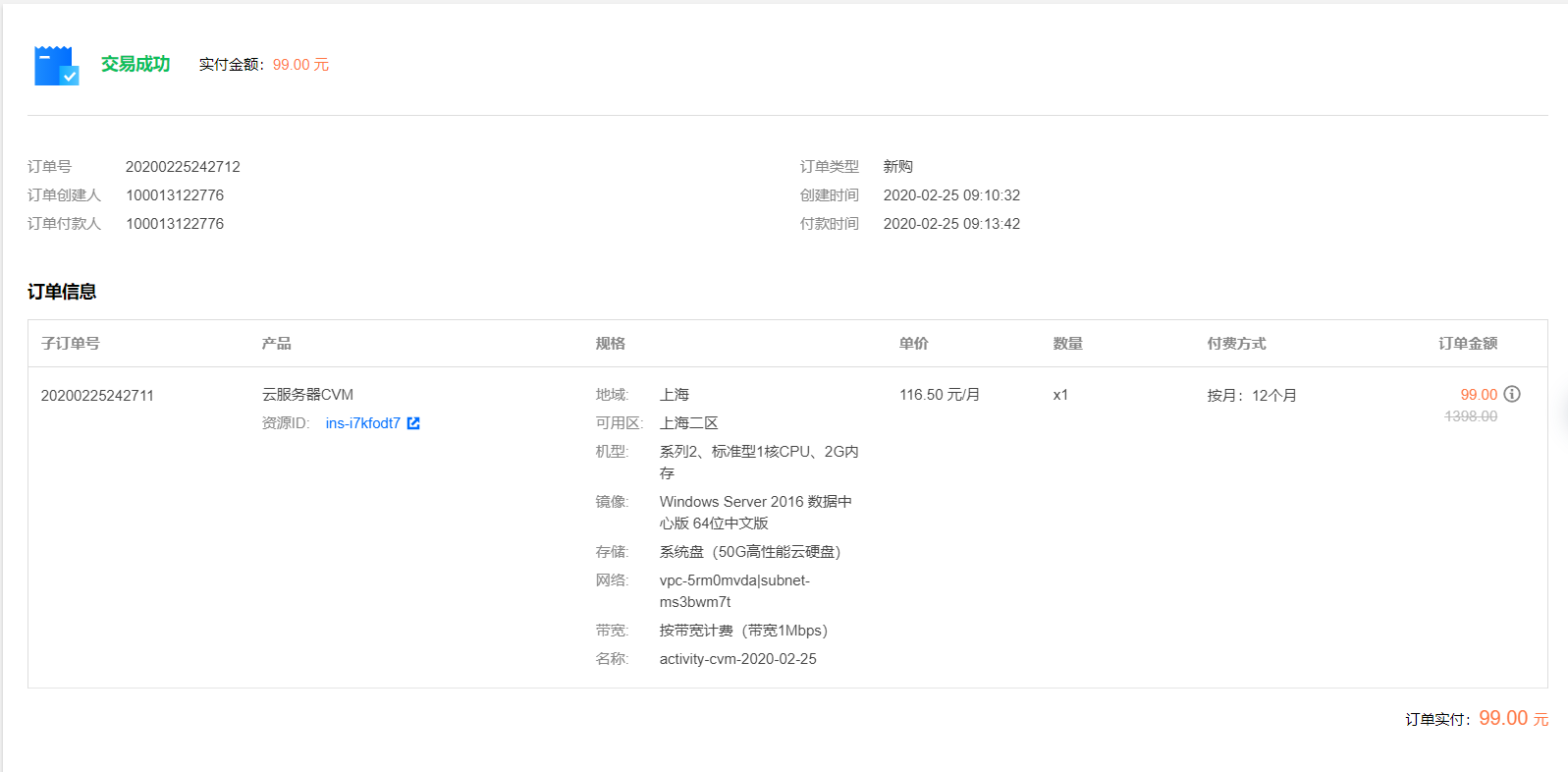
****

图11 腾讯云服务器订单截图

**（三）取得成果**

1. 整理模板

考虑到985、211高校的讲座资源较为丰富，根据全国985、211高校名单，小组成员寻找了各高校汇集讲座信息比较多的一两个公众号或官方网站，并收集这些网站和公众号2015年至2020年2月的讲座预告，再依据据表格对这些预告内容进行编码，编码内容包括标题、时间地点主题基本信息、主持人、主讲人、简介、推荐书目、参与方式、主办方、承办方八个个部分，由此整理出了严肃类讲座预告的模板。

C:\Users\H黄婧怡\AppData\Local\Temp\1607609556(1).png

图12 编码内容

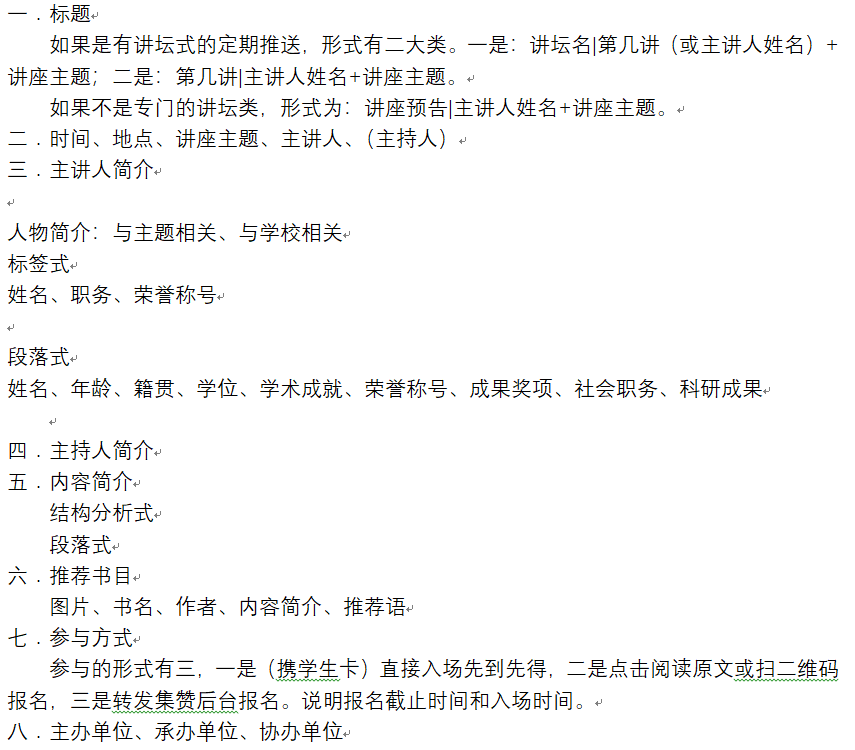


图13 模板内容

讲座类推送的内容一般分为讲座开始前的预告和结束后的总结两类，主要有严肃和活泼两种风格。考虑到讲座后的总结已经没有太大意义，受众更多地是想获取预告信息以便自己能参与讲座，所以我们选取了预告类；又因为严肃类较活泼类更容易整理出模板内容，小组成员先尝试从严肃类开始。

1. 公众号

公众号“今日人文社科”创立于2020年3月2日，前期已经设计好logo、二维码等品牌信息。公众号旨在成为一个人文社科类讲座信息共享平台，方便读者掌握讲座资讯，紧跟学科动态，聚焦顶尖高校，直击人文社科热点话题。目前公众号仍是人力写作与发布，最后将实现依据模板的自动生成内容。开创至今，公众号已发出34篇推文，内容包括话题和讲座总结两方面，话题关注新基建、数字经济、媒体融合；讲座总结则是团队成员依据自己关注的学科，每周听完学科内讲座后总结而成，领域包括计算机、新闻传播学和艺术学。



图14 公众号二维码

1. 数据库

小组已将爬取下来的讲座信息进行过数据清洗和人工预处理，并且已经将处理好的文本结构化，存储到服务器上搭建的数据库之中，数据库设计思路及数据之间的相互关系已在之前的ER图中给出。目前，数据库内存储了2015-2020年来1048条、12类不同学科的讲座信息，包括讲座题目、主办方、开始时间、结束时间、举办地点、简要介绍等；同时，对每场讲座的主讲人也进行了信息整理和存储，记录了每名主讲人的姓名、基本信息、主要成就、研究领域等。对非结构化的文本信息进行结构化存储，为后续的文本分类处理、自动讲座推送生成奠定了基础。

1. 网站搭建

目前，网站的搭建已经初步完成。网站首页上方的工具栏提供了登录界面和搜索界面，在登录界面上，用户可以登录自己的账户或注册进行操作；搜索方面，提供了两种搜索方式，一种是在搜索栏中通过输入进行关键词匹配搜索，进行搜索后，将会从数据库中返回存在相应关键词的对应讲座的题目、主办方、简要介绍等信息；另一种方式是通过学科下拉菜单选择对应学科的讲座信息，点击后将返回此类学科的所有讲座信息。未来将继续优化完善搜索界面，实现搜索结果的跳转功能，为每一条讲座信息设置单独的浏览界面，使用户能够直接通过点击讲座标题获取讲座具体内容信息。主页头图下方存有已整理好的讲座笔记和未来讲座预告信息。



图15 网站首页

**（四）存在的问题和困难**

一、问题和困难

1. 复杂内容无法整理出模板，可整理成型的模板单一

目前已经整理出的模板只有严肃类讲座预告，除此之外的多种讲座推送形式，由于风格多变、内容不一，无法整理出统一模板。也就是说，现有的成型模板形式单一。

1. 部分文本数据存在空缺，为后续处理带来了一定的困难

当前已经构建好的数据库中，不少数据项存在空缺，在使用模板进行文本的自动生成时会造成一定的困难，因此下一步需要考虑如何处理存在部分空缺信息的文本信息。

1. 未来讲座预告的整理

当前数据库中的数据均为已经结束的讲座预告信息，且对文本信息的处理都基于人工操作，面对未来大批量的新数据时处理仍有难度，需要寻找更为高效的方法；且目前仍缺少未来讲座预告的获取途径，需要进一步掌握并及时爬取数据新的讲座预告信息、进行处理后添加到数据库当中。

二、拟解决的方案：

1. 设计和整理更多的模板，并设计多种模板之间的组合方式，利用自然语言处理技术，对不同粒度的文本进行组织和排列。
2. 设计默认填充的算法方案，解决数据缺失的问题。优先使用完整数据，并在数据有缺失的情况下争取填充，使每一条数据物尽其用。
3. 扩大数据来源，保证数据源的时效性和稳定性，争取使用爬虫的方式，实现半自动定期更新数据。

**（五）建议和要求**

1. 当前工作重点：
2. 完成网站的开发
3. 完善公众号的运营细节
4. 工作组织方面：
5. 保持沟通。根据工作相关人的数目，灵活召开工作会。并保证工作内容每周跟进一次。
6. 有效分工。根据每个人的工作特点，合理分配任务，保证小组工作的有序推进。
7. 后期工作规划：
8. 首先把网站搭建好，再对后端包括但不限于文本生成算法进行优化。
9. 扩大数据库，找到更稳定的数据来源，减少人工成本。
10. 丰富公众号内容，传递有价值的信息，吸引更多用户。
11. 尽可能创造项目经费来源。