2020 年大学生创新训练项目

申请书

项目名称　基于模板的微信公众号推送内容自动生成研究

项目负责人　邬语丝　　联系电话 18347943868

所在学院　计算机与网络空间安全学院

学　　号　201811113003

专业班级　2018 计算机科学与技术

指导教师　李春芳

E-mail　yusi@cuc.edu.cn

申请日期　2019 年 11 月 26 日

一、 基本情况

<table>
<tr><td>项目名称</td><td colspan="7">基于模板的微信公众号推送内容自动生成研究</td></tr>
<tr><td>选题来源</td><td colspan="7">❂规划项目 一般选题 5 新媒体前沿技术改造与实践</td></tr>
<tr><td>所属学科</td><td colspan="7">数据科学 计算机科学与技术 传播学 新闻学 网络与新媒体</td></tr>
<tr><td>负责人姓名</td><td colspan="2">邬语丝</td><td>性别</td><td>女</td><td>民族</td><td>汉</td><td>出生年月</td><td>2000 年 9 月</td></tr>
<tr><td>学号</td><td colspan="2">201811113003</td><td>联系电话</td><td colspan="4">18347943868</td></tr>
<tr><td colspan="2">项目负责人曾经参与科研的情况</td><td colspan="6">掌握 C，C++，Python，Java，MySQL 的基本操作，能运用 HTML+CSS+JS 进行网页制作。在小学期实践《日程管理系统》中负责产品设计和前端。对学科交叉感兴趣，在本专业计算机科学与技术的基础上学习了网络与新媒体双学位。</td></tr>
<tr><td>指导教师</td><td colspan="2">李春芳</td><td>联系电话</td><td colspan="4">13269636506</td></tr>
<tr><td colspan="2">指导教师介绍</td><td colspan="6">1. 职称/职务：博士（后），副教授，博士毕业于北京航空航天大学。
2. 研究方向：智能影视大数据、视频内容理解，空间大数据，复杂网络与软件网络、数据可视化、大型信息系统分析设计、IT 治理。
3. 所在学院：任教于中国传媒大学计算机与网络空间安全学院。
4. 相关经历：主持开发了中传艺考、如艺剧本系统，目前在研如艺智能影视平台、三维空间大数据平台。申请专利 3 项，授权 1 项，发表 SCI/EI 检索论文 30 多篇，申请软件著作权 4 项，主持及参与各类项目 10 余项。现为中国计算机学会和 ACM 会员，云安全联盟大中华区主席特别助理兼高校联络人，中国云体系产业创新战略联盟秘书长特别助理兼高校联络人，SCI 期刊 IJFCS 和 ETRI 复杂网络方向审稿人。
任课情况：数据可视化、Python 程序设计、数据科学导论（大数据技术导论）、社交网络分析、舆情分析与社会计算（研究生）。</td></tr>
</table>

	姓　名	学号	专业班级	所在学院	项目中的分工
项目组主要成员	王泽宇	201811153023	18 数据科学与大数据技术	计算机与网络空间安全学院	软件开发
	章颖	201811153012	2018 级数据科学与大数据技术	计算机与网络空间安全学院	软件开发-前端页面
	黄婧怡	201801213019	2018 级传播学(媒体市场调查与分析方向)	新闻学院	新闻模板设计
	谢清扬	201801113029	2018 级新闻学	新闻学院	数据新闻设计

项目组成员参与科研的情况	王泽宇： 曾搭建过一台服务器，至今仍在供自己及同伴使用；熟悉 C++、Java、Python 编程，曾参与设计过较大的数据管理系统，能够对大量数据进行储存、管理和分析；掌握新闻的基本知识，曾参与制作数据新闻《分析了 3220 个地名，我们寻到了中国地名的文化旧根……》。 章颖： 乐于学习新技术。有一定 Python 语法及爬虫基础，会使用 html/css/js 编写网页，租有一台云服务器(php+MySQL+Apache 架构)挂载自己的个人网站。 黄婧怡： 参与教育部哲学社会科学研究重大委托项目“高等教育大众化与媒介融合时代菁英女性培养与领导力提升研究”（项目批准号：15JZDW002），参与专著《连接力与粘合力：平台型媒体与全球菁英女性领导力自我呈现研究》的数据分析。 谢清扬： 调查研究《微信健康推送对特定群体传播影响》，对微信内容传播模板、传播方式有一定研究经验。

二、 立项依据（可加页）

（一）研究目的

1.了解已有机器新闻和自动写作相关技术。

通过查询各大团队机器写稿发展情况，对该行业进行全面认知，并对用户市场需求进一步分析。

2.实现一个图文、视频融合的内容自动生成系统。

依据新闻稿件倒金字塔式结构、螺丝杆式结构、金字塔式结构，分别针对动态类、单一线条内容复杂类、趣味类会议消息进行自动生成。同时依据生成内容及用户自定义风格选择，广泛检索互联网图片与视频，最终完成个性化排版。同时，以“短、平、快”为特点，为内容生产者提供便利，为新媒体内容发展提供创新路径。

3.实现微信公众号内容自动推送。

通过自动填表单、爬虫及软件测试技术，在完成内容分析的基础上，达到生成内容与微信公众号发布平台的直接联动。用户可通过复制粘贴等方式，快速便捷地推送消息。

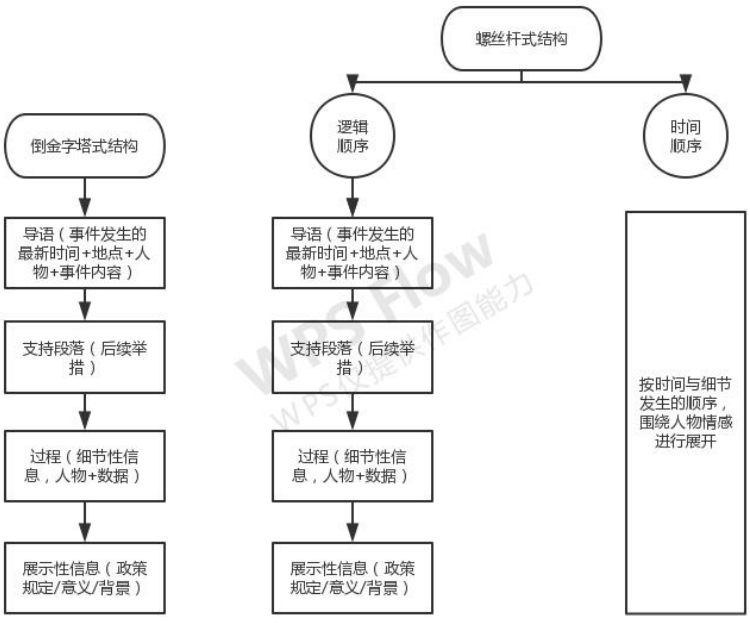
（一） 研究内容

用网站形式，基于模板搭建类似今日头条的根据海量数据自动生成内容的会议类微信公众号推送的应用。

一、 需求分析及信息来源

1.运用焦点小组访谈和问卷调查法研究用户对于即时发布内容推送的形式、排版的需求，获取动态需求信息。

2.通过内容分析法整合特定领域官方公众号平台的会议类推送，获取会议新闻各基本要素制作推送模板。列举两种新闻模板进行形式说明：



3. 将百度图片、必应图片、百度贴吧、豆瓣、花瓣、堆糖等图片网站作为图片检索信息源，以 bilibili、爱奇艺、腾讯视频等主流视频平台作为视频检索信息源，与生成内容结合形成推送整体内容。

二、软件与算法设计

1.使用爬虫技术获取推文素材，使用 NLP 技术进行词频分析，设计算法实现文案的自动摘要生成。

2.使用图像理解技术，设计算法实现自动选择配图。

3.借助指导老师实验室已有的视频识别技术并通过人工辅助，设计算法实现视频的自动摘要生成。

三、软件实现

1.使用 Python 和 Flask 搭建可运行系统。

2.使用 MySQL 进行数据操作。

3.使用 D3 进行网页的数据可视化操作。

四、基于软件测试的自动表单填写

使用自动表单填写技术，实现将已生成的推送自动上传到微信公众号后台，包括：文本、图、视频链接。

（二） 国、内外研究现状和发展动态

随着人工智能技术的不断发展进步，其应用也逐渐渗透到日常生活中的方方面面。其中，“机器新闻写作”就是其在新闻传播领域的重要应用。2016 年里约奥运会上，由北大计算机所万小军团队研发的“张小明”新闻机器人横空出世，13 天内，撰写了 457 篇关于乒乓球、羽毛球、网球等项目的消息简讯和赛事报道，一时间吸引了世人的眼球；而实际上，在“张小明”出现之前，新闻机器人就已经被不少世界上的主流媒体关注并加以使用——

国内外主要新闻机器人一览

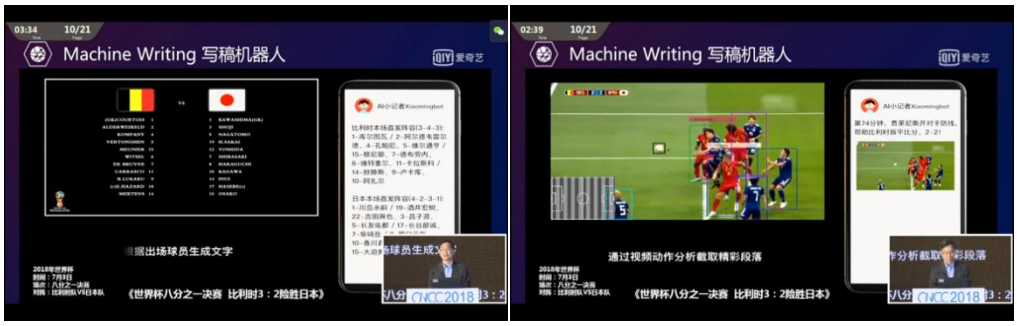
国内					国外				
名称	**所属机构**	**时间**	**领域**	**功能**	**名称**	**所属机构**	**时间**	**领域**	**功能**
Dream writer	腾讯	2015.09	财经	写稿	Quakebot	洛杉矶时报	2014.03	地震预报	写稿
快笔小新	新华社	2015.11	财经、体育	写稿	WordSmith	美联社	2014.07	财经、体育	写稿
DT 稿王	第一财经	2016.05	财经	写稿	Blossombot	纽约时报	2015.05	新媒体	编辑
张小明	今日头条	2016.08	体育	写稿	Heliograf	华盛顿邮报	2016.08	体育	写稿

其实，早在 2006 年，美国商业信息集团汤姆森金融就已经开始运用电脑程序自动生成部分财经新闻，可在上市公司公布业绩短短 0.3 秒内就发布出一篇盈利报道。这也被看作是“机器新闻写作”的初探，只可惜，在更加注重质量而非速

度的报纸时代，它并未引起太大关注。

在那之后，西北大学的科技精英们研制出了新一代智能写作软件 Narrative Science，这时的机器写作已经不仅仅局限于对数据的呈现，还已经具有了强大的运算能力，展示出一篇具有“人情味”的新闻报道——此时，已经具有了如今“机器写作”的雏形。

随着社会的飞速发展，传统的报纸、电视等传播媒介日渐衰微，数字化、碎片化的阅读成为人们最主要的信息来源。基于人们这样的需求，新一代“新闻机器人”应运而生，在不断的改进和充实中，逐渐满足了人们对新闻报道的各种需求——如今的“新闻机器人”，已经可以结合最新的自然语言处理、机器学习、大数据分析和视觉图像处理技术，通过语法合成与排序，基于现有的文本模板生成完整的新闻——此处所说的“完整”，不仅仅是指可以生成一篇完整的文本，更可以通过图片检索自己选择合适的配图，甚至模仿人类的语气，使新闻文本读起来更加亲切、抓人眼球。



（新一代的机器新闻成品，已经可以利用成熟的图片、视频抓取与识别技术自动生成语义明确、内容完整连贯的文章）

如今，新一代的机器写作的应用更加广泛。由最开始的集中于数据化、模板化的财经新闻和体育新闻报导为主，到现在，已经可以根据数据的宏观对比，得出变化、趋势等结论并加以呈现；同时，同一则新闻可以生产出多种个性化的版本。而国外由于机器写作研究的起步较早，拥有相对多的技术优势和实战经验，除了能够实现国内的应用外，还可以在房地产、商业等更多领域得到使用，对数据的追踪、复杂信息的消化、关系的挖掘等方面也要比国内更加完备与成熟，甚至还可以写出极具个性化的新闻报道。

首页 / 体育 / 正文

奥运羽球女子单打小组赛 奥运名将戴资颖(中华台北)2:0力克娜塔莉亚-佩米诺娃 轻松取分

AI小记者Xiaomingbot 2016-08-15 08:34

简讯：北京时间8月15日07:30时，现世界排名第8的戴资颖在奥运会羽毛球女子单打小组赛中胜出。戴资颖本轮的对手是现世界排名第52的娜塔莉亚-佩米诺娃，实力不俗。但经过28分钟的激烈较量，最终，戴资颖还是以总比分2:0战胜对手，笑到了最后，为中华台北延续了在这个系列赛事中最终夺冠的机会。这场比赛的各局比分分别是 21:12，21:9。

（“张小明”可以自己检索图片，还可以使用“笑到最后”“实力不俗”这种极具个性化的词语）

不过，由于相关技术仍然不够完备，当下的机器写作依然有一系列问题亟待解决，比如机器人对信息的理解深度还不够、新闻成品模板化明显，扁平而又千篇一律的文章依旧是绝大多数机器写作产品的常态，缺乏亮点；对信息的提炼和概括能力也还有所不足。这些也是研究人员们接下来将要致力于解决的方向。

奥运会羽毛球男子双打金牌赛 傅海峰/张楠组合2:1马来西亚组合 再添一金

AI小记者Xiaomingbot 2016-08-20 01:12

简讯：北京时间8月19日22:15时，中国的傅海峰/张楠组合在奥运会羽毛球男子双打金牌赛中以2-1的比分击败吴蔚昇/陈蔚强组合夺冠。比赛进行的精彩纷呈，高潮迭起。最终现世界排名第4的傅海峰/张楠组合发挥更加出色，以16:21，21:11，23:21的比分击败吴蔚昇/陈蔚强组合，获得金牌。马来西亚的吴蔚昇/陈蔚强组合遗憾获得银牌。

①

奥运会羽毛球男子单打金牌赛 名宿谌龙(中国)完胜李宗伟(马来西亚) 成就冠军荣耀

AI小记者Xiaomingbot 2016-08-20 22:15

简讯：北京时间8月20日20:20时，奥运会羽毛球男子单打金牌赛在里约会议中心-4号馆展开较量。现世界排名第2的谌龙迎战现世界排名第1的李宗伟，双方你来我往展开了激烈的较量。最后，耗时1时14分钟，谌龙率先在3局2胜制比赛中获得制胜分，以2:0拿下比赛，加冕桂冠，为中国夺得宝贵一金。双方各局小分分别为：21:18，21:18。

②

奥运羽球女子双打金牌赛 世界名将松友美佐纪/高桥礼华组合2:1取胜丹麦组合 成就冠军荣耀

AI小记者Xiaomingbot 2016-08-19 01:23

简讯：北京时间8月18日22:50时，奥运会羽毛球女子双打金牌赛，现世界排名第1的松友美佐纪/高桥礼华组合在里约会议中心-4号馆对战现年30和33岁的克里斯丁娜·彼德森/尤尔组合。比赛过程精彩纷呈，高潮不断。最后在3局2胜制比赛中，经过1时21分钟的等待，松友美佐纪/高桥礼华组合率先获得制胜分，以18:21，21:9，21:19，总比分2:1战胜对手，获得金牌。

③

奥运会羽毛球女子单打金牌赛 球坛名将马琳获胜摘得金牌

AI小记者Xiaomingbot 2016-08-19 23:29

简讯：北京时间8月19日21:25时，奥运会羽毛球女子单打金牌赛，现世界排名第1的卡罗列娜-马琳在里约会议中心-4号馆对战现世界排名第10的V-辛德胡-普萨拉。比赛过程精彩纷呈，高潮不断。最后在3局2胜制比赛中，经过1时23分钟的等待，卡罗列娜-马琳率先获得制胜分，以19:21，21:12，21:15，总比分2:1战胜对手，获得金牌。

④

（新闻扁平化、千篇一律，是当今机器新闻的最大问题之一）

一言以蔽之，当前，“新闻机器人”是目前机器写作在我们日常生活中最为常见的应用，且已经证明了其高效性和可信赖性，但由于相关技术还不够成熟，其依旧有十分广阔的发展前景，有望在未来成为跨领域的多面手、人类记者编辑的好帮手，而本课题团队也想要以目前已经比较完备的基于模板生成完整新闻这一技术为基础，完成活动类微信公众号推文的实现，提高生产推文的效率，使机器写作这一技术在新媒体领域得到更加普遍的应用。

参考文献：

《新闻写作机器人的应用及前景展望——以今日头条新闻机器人张小明（xiaomingbot）为例》

《今日头条的算法生产新闻研究》，郝慧敏，《传媒论坛》，2018.08

《“Xiaomingbot”背后，写稿机器人的技术探寻——专访北京大学计算机科学技术研究所万小军博士》，刁毅刚、陈旭管，《中国传媒科技》，2016.09

相关网页链接：

http://media.people.com.cn/n1/2017/0111/c409691-29014245.html

爱奇艺视频：《字节跳动副总裁人工智能实验室主任马维英：AI 赋能内容创作和交流》

（四）创新点与项目特色

一、探索智能生成内容的新路径。

智能生成内容发展虽有相关路径可循，但针对会议类报道信息生产的研究较少，此次探索是一次合理尝试。基于传统新闻写作结构，从用户范围广泛的微信公众号入手，融合为新媒体用户所接受的阅读模式。并以会议类内容生成为核心，探索会议类新闻内容模板，开发“短平快”的信息写稿路径。

二、媒介融合实践，打造图文视频融合媒体。

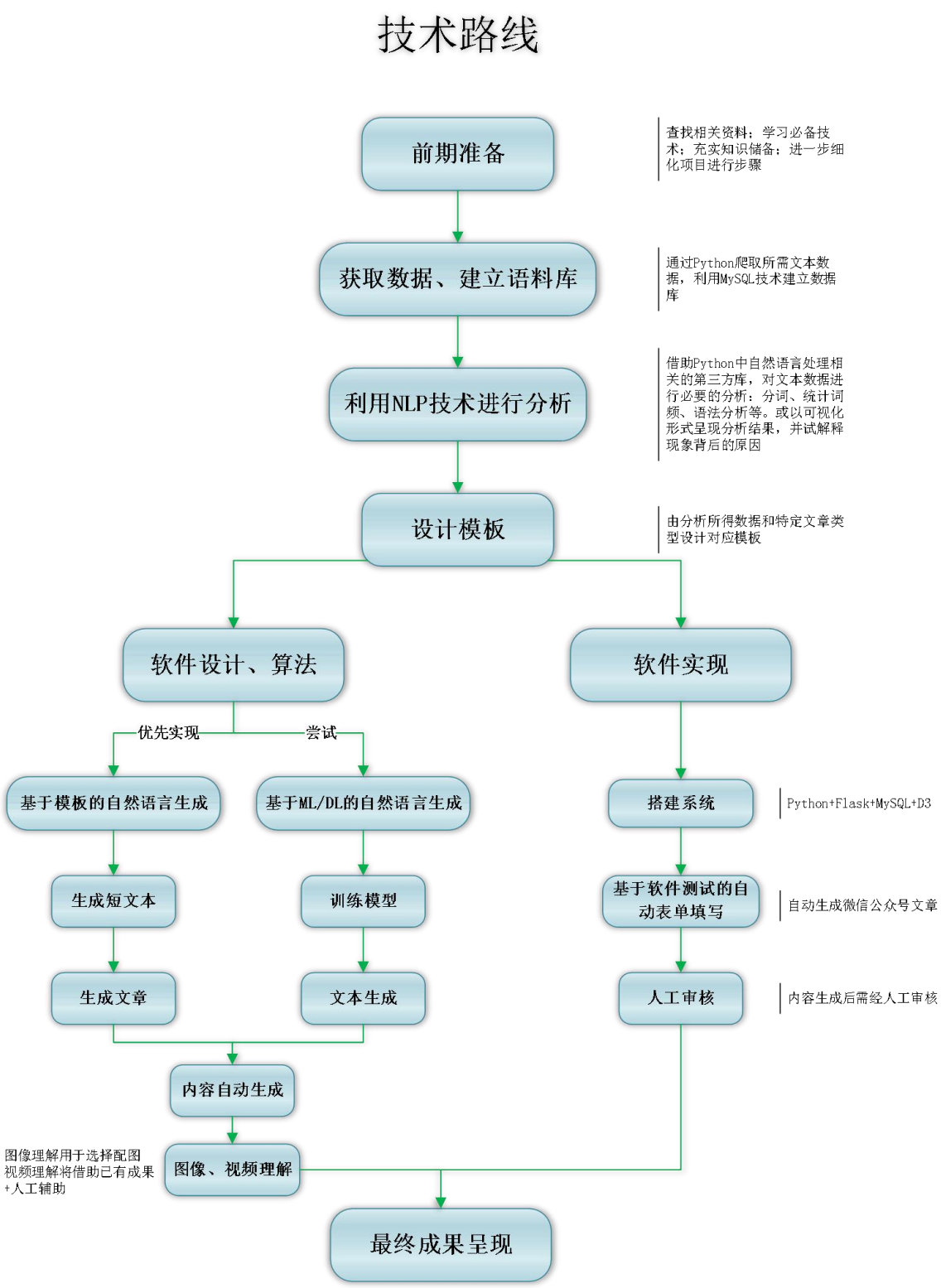
此次研究计划在自动形成文案的基础上，依据生成内容进行全网智能检索，为推文匹配合适的图片与视频，实现推文文字、图片、视频三合一，为受众带来更优越的视听体验。

三、操作便捷，自动填充。

依据新闻稿结构，形成内容模板。用户只需选择适当的结构并输入新闻要素，系统自动生成新闻稿件并填写到微信公众号后台，经由人工审核后即可发出。提高内容生成效率，降低人力成本。

（五）技术路线、拟解决的问题及预期成果

一、技术路线



前期进行必要技能的学习；初期利用 Python 爬取文本数据，结合 MySQL 建立数据库；再通过 Python 的 NLTK 库和处理中文的第三方类库（例如 SnowNLP 等）自然语言处理技术对文本数据进行分析；由分析的结果设计相应类型文章的模板。软件设计和算法方面：基于设计出模板选择“模板填空模式”或“机器学习/深度学习”的方式实现特定类型推送文章的自然语言生成。在上述步骤已实现的基础上，将进一步对图像理解和视频摘要生成进行研究，使系统具有自动配图等功能。软件实现方面：利用 Python+Flask+MySQL+D3 技术搭建系统，并实现基于软件测试的自动表单填写，辅以后期的人工审核，保证软件的顺利实现。

二、拟解决的问题

1.将获取的数据分类建立语料库。

2.通过自然语言处理的现有手段对中文文本进行合理的分析并得出有价值的结论。

3.设计可行的文本模板。

4.实现基于模板的自然语言生成。

5.实现可应用系统的搭建。

三、预期成果

1.通过 NLP 技术对文本进行分析后设计出特定文本模板。

2.实现基于模板的文本内容生成。

3.实现可以自动生成内容的程序。

（六）项目研究进度安排

立项－2020.1

研读与微信活动类推送有关的文献并搜集至少三百条活动类推送。

2020.2

组建焦点小组，根据搜集的活动类推送编制访谈内容，进行开放、深入的访谈，对所得访谈结果进行梳理。根据访谈结果制作问卷，问卷经修改通过信度效度后，进行问卷的发放、回收及数据分析。

学习自然语言处理相关知识，使用 Python 的 NLTK 和 SnowNLP 等类库，建立短文本处理简单模型。

2020.3－2020.5

在阅读相关文献的基础上，参考已有的严谨研究，编制内容分析法中使用的编码表，并对编码表进行试编码及信效分析，得出最终的编码表；培训编码员，检测编码员间的信度，再对几百条活动类推送进行编码，整理编码所得数据，运用内容分析法得出构成活动类推送的关键指标和建立文本模板所需的语料集。

学习图像理解技术与视频理解技术，使用 Python 爬虫技术与 MySQL，建立配图数据库与视频数据库，并设计活动类推送关键指标与配图数据库的匹配算法。

2020.6-2020.9

根据调查得出的活动类推送关键指标与语料集，使用 MySQL 建立文本模板数据库，

并设计关键指标与文本模板数据库的匹配算法。
后端设计基本完成。

2020.10-2020.12
搭建网站，设计网页，并实现前后端联调

（七）已有基础

一、与本项目有关的研究积累和已取得的成绩

1.本课题团队已研究今日头条、新华社万小军等机器新闻写稿前例，对基本流程有一定掌握；同时，在新媒体传播形式、内容及影响中有相关的研究经验，例如通过对微信公众号平台内容在一定群体中的传播及影响的研究。

2.小组成员可以熟练掌握并应用定量分析、采访访谈技巧等研究方法。

3.能够使用爬虫爬取网页，并对爬取下来的内容进行分词处理、词频分析等，即已经可以抓取文本并且对文本进行最基本的分析操作。

4.数理统计知识准备充足，已经可以对大量数据进行结构化的归类和统计分析，整理出想要的有用数据信息。

二、已具备的条件，尚缺少的条件及解决方法

1.已具备条件：

（1）掌握MySQL基础知识与操作，能够熟练使用数据库并进行数据库的管理等，比如建立图片库，为文章生成合适的配图；

（2）具备Python编程基础，为对NLTK类库和SnowNLP类库的学习和实际应用做好了准备；

（3）有扎实的数学、概率、数理统计基础，对人工智能领域的基本算法已有初步了解并能够合理应用，为更进一步的研究打好了坚实的理论基础；

（4）初步学习过机器学习相关知识，为日后生成模板文本和完整连贯文章等研究成果提供理论基础；

2.缺少条件：

（1）课题团队目前缺少对市场需求细化的把握、对活动类微信推文大数据资源的掌握，尤指论文资料、课程书籍等；对用户的具体需求了解还有所不足。

（2）目前尚缺乏对NLP技术的了解，暂时难以通过自然语言处理的现有手段对文本进行分析；

3.解决方法：

（1）针对目前问题，团队将通过市场调查分析及模式探索创新解决已有问题，采用问卷、当面采访等方式，了解用户需求，以实时更新团队的具体研究方向，呈现出符合用户需求的研究成果。

（2）学习NLP自然语言处理技术、进一步深化机器学习领域的探索和研究，以解决生成文本模板、自动生成连贯的文章、完成完整排版的技术。

三、 经费预算

开支科目	预算金费（元）	主要用途
云服务器购买	2000 元	计算、网络、存储
网站域名购买	1000 元	搭建网站
存储硬件购买	1000 元	本地存储，如移动硬盘、U 盘
资料费	2000 元	材料打印、文献期刊书籍购买
参与费	2000 元	支付编码员工资
场地租用费	1000 元	租用访谈所需场地
预算经费总额：	9000 元	

四、 项目组成员承诺

我们承诺所填报的内容真实有效。如果获得立项资助，我与本项目组成员将严格遵守学校有关规定，认真开展项目，按时报送材料。如违反承诺，将自行承担后果。

承诺人（全体成员）签名：

年 月 日

五、 指导教师意见

导师（签章）：

年 月 日

六、 大学生创新训练计划专家组意见

负责人（签章）：

年 月 日