

“星火训练营” 003 |

“自然语言处理在数字人文中的应用及展望” 专题讲座

中国人民大学数字人文研究中心“星火训练营”旨在为学生研究员提供更为系统的数字人文学习指导和训练，借鉴国际上各数字人文中心普遍采用的“理论+方法+项目”的培养方式，立足RUC-DHC“数字人文新青年”的既有基础，邀请跨专业导师进行学术引领和指导，以参与实践项目推动数字人文理论学习与理解。

第三期学生研究员“星火训练营”专题讲座内容为“自然语言处理在数字人文中的应用及展望”，邀请到了中国人民大学信息资源管理学院讲师、师资博士后杨建梁担任主讲人。杨建梁博士是中国人民大学数字人文研究中心研究员、中国人民大学与麻省理工学院联合培养博士，现主持国家社科基金青年项目、中国博士后科学基金一等面上资助项目，参与多项国家社科基金重大、重点项目和国家自然科学基金重点项目。讲座由中国人民大学信息资源管理学院讲师、师资博士后祁天娇博士主持。



本次讲座内容主要分为四个部分：自然语言处理的概念；自然语言处理与数字人文的渊源；从数字资源到数据应用；对自然语言处理发展的展望。

杨建梁博士在讲座之初回顾了自然语言处理（Natural Language Processing, NLP）的概念，自然语言处理是人工智能和语言学领域的分支学科，主要探讨如何处理及运用自然语言。自然语言处理包括理解和生成两部分。“理解”是让电脑把输入的语言变成可计算的符号和关系，“生成”则是将计算机数据转化为自然语言。常见的自然语言处理技术包括文本分类、机器翻译、自动摘要和问答系统等。

随后，杨建梁博士从自然语言处理发展历程和该技术在数字人文领域中的诸多应用案例，说明了自然语言处理与数字人文的关联所在。在NLP技术的发展历程中，循环神经网络（Recurrent Neural Networks,RNN）是一个里程碑式的神经网络结构。RNN模型可以让机器初步理解人类的语言，并做出判断和预测，由此进一步实现文本校对、信息抽取和古籍句读等。但RNN也可能由于序列长度过长从而失去能力，研究人员进而提出了“神经网络的注意力机制”，该机制的提出让自然语言处理模型（基于深度学习）从“选择遗忘”向“选择关注”迈进。

接着，杨建梁博士分别从资源处理端、资源服务端和多模态融合等角度对自然语言处理技术进行了总结和展望，认为资源处理端在转录识别、自动抽取原数据和利用文本内容构建资源网络等方面还需改善，资源服务端在自动问答、自动翻译和摘要生成等方面的功能还需要进一步提升。此外，还应将图片、声音、视频与文字信息相结合，实现多模态融合的数字人文资源分析和挖掘。

讲座最后，杨建梁博士回答了师生们的热烈提问。中国人民大学数字人文研究中心学生研究员“星火训练营”专题讲座第三期圆满结束。

您是否希望重温本次讲座的精彩内容，是否为错过干货满满的训练营活动而遗憾？

中国人民大学数字人文研究中心贴心地为大家准备了视频回放，点击下方链接即可获取

什么是RUC-DHC学生研究员？

为进一步优化团队、培育人才、涵育创新，中国人民大学数字人文研究中心（RUC-DHC）面向各高校在校学生招聘学生研究员召集青年力量。中心于每年2月启动新一轮的学生研究员招聘，按照学术追踪组、学术传播组、学术期刊组划分组织，聘期为一年。首期学生研究员已于2021年2月完成集结，欢迎感兴趣的朋友持续关注公众号，敬请期待下次招聘~

撰稿：田清



公众号账号：rucdh2019

网址：<http://dh.ruc.edu.cn>

邮箱：rucdh@ruc.edu.cn



中心简介

中国人民大学数字人文研究中心集人民大学多学科优势，秉持融合文理、协同创新之理念，开展数字人文理论研究、实践探索、人才培养和学术交流。