**人工智能与社会学研究：在华外国人的生活和工作现状**

## 课题介绍

中国的经济高速发展，新鲜事物层出不穷，我们遭遇或接受着各种刻板印象，典型的有如来华外国人，我们是否真正了解他们的生活和工作现状？如果将外国人作为一种话语，那么官方媒体与民间小报分别如何呈现其形象？是否存在选择性报道？媒体形象与真实世界的差距是什么？而我们又该如何去研究这些事情的本质？带着这些问题我们将开始神奇的人工智能背景下的移民社会学研究的学习。



**图片来源：课题导师在田野调研时为华裔混血拍摄的纪录相片**

2017年5月，AlphaGo与柯洁在浙江乌镇举行围棋赛，这场“人机大战”以AlphaGo的绝对胜利昭示了人工智能的成功。AlphaGo体现了人工智能的前沿技术，其算法包括深度神经网络、监督/强化学习和蒙特卡罗树搜索等。如今人工智能（Artificial Intelligence）已经从科幻银幕进入现实，顾名思义，人工智能的目标是让机器通过模仿人的智力实现对真实世界的预测。

机器学习则是实现人工智能的途径，机器学习是用算法学习现有数据，从而进行预测。商业用途包括各类推荐系统，例如：淘宝的推荐商品模式、滴滴打车的派单系统和网易云音乐挖掘用户的音乐偏好等。机器学习的经典算法包括：聚类算法、决策树、随机森林、支持向量机等，而随着以神经网络为代表的深度学习算法出现，这种基于多个隐含层的非线就性方式的连接结构，大大提高了网络训练的效果和预测精度。

本课题主要关注机器学习的重要分支之一——自然语言处理（Natural language processing,NLP）。“自然”即意为与机器语言的区分，计算机无法直接理解自然语言，而自然语言处理技术则搭建起自然语言和机器语言的沟通桥梁。在互联网时代，我们可以轻易获得海量公开的文本资料，如：新闻、法律文书和虚拟社区讨论文本等。这些文本作为人类行为的无差别反映，如何利用这些宝贵而客观的研究材料，是本课题探讨的关键问题。

在新范式席卷各学科领域时，最重要的迅速把握新方法并解决该学科的经典问题。在社会学领域，福柯是后现代知识分子共同体中最具有传奇色彩的理论家之一，其思想颠覆性在于对现有知识体系的反思，并通过“知识考古学”方式追溯某种“真理体制”得以建立的社会和实践条件，而不是理所当然的态度。

福柯认为，知识是权力斗争的产物，也是权力斗争的工具。他将真理、权力和自我作为其研究的三个轴心。因此，我们首先要关注知识的创造者。例如，有学者感叹中华文化逐渐低俗化，其本质是因为历史是言辞和事迹的记忆，只能通过文字载体传递，在古代有机会学习文字的往往是社会高阶层，文化作为品位区隔，高阶层创造的多是高雅的“阳春白雪”；而随着教育和网络的普及，文化的营建者阶层扩大且下移，许多平台都可以供其发声，因此总体文化中“下里巴人”的部分有所增加。

其次，知识的本质与演化。福柯的话语分析关注真理产生的制度背景，尤其这种体制的历史变迁对于同一事物被建构的影响。例如，福柯分析惩罚方式的转变，从残暴的露天砍头的“剧场模式”转换为生产性的工厂模式，通过时间表和全景敞视监狱，对犯人身体进行规训，最终形成生产效用。这种个体化的技术，从监狱扩散到学校、工厂和兵营，因此福柯称现代社会是纪律社会，这就是惩罚向规训的转化，权力控制从宏观可见转变为微观隐蔽。

落地到本课题的机器学习方法论，作为话语分析的工具。我们主要将分析这两个方面：通过来华外国人的网络论坛文本数据，发现外国人在中国生活工作现状，他们如何融入到中国社会中？他们对中国的印象如何？另一方面，我们通过中文文本数据，去发现外国人在中国人眼中的形象，来自不同国家的外国人如何在中文文本中被建构。

## 二、学习安排

**第一部分 领域背景学习及研究方案的设计（2周）**

在了解社会学方法论的基础上，阅读福柯的经典著作，建立以文本分析事物的大体认知。进而了解机器学习的相关算法和研究流程，建立起“文本->数学表达->建模/预测”的思路；学生可选择导师提供移民社会学资料，设立研究目的并撰写研究计划，以研究为导向学习整个流程的方法论。

**第二部分 数据获取：学习通过Python抓取数据、建立语料库的方法（2周）**

学习基础的爬虫知识，以及如何针对选题（研究内容）获得文本数据（新闻、论坛和社交文本），使用正则表达式对文本进行初步清理，用分词工具对中文文本进行分词并建立语料库的方法。

**第三部分 分析方法学习：主题模型、有监督算法、文本向量化（2周）**

自然语言处理作为最前沿的议题及研究方法之一，此模块将进行机器学习的算法模式（无监督和有监督）以及词向量化的方法学习。

（1）无监督算法模式是指研究者不提供分类指导，由机器自行判断分类，最有代表性的是LDA主题模型，其原理是将文本分解成：主题与关键词的矩阵╳关键词与文档的矩阵，以此将文本分为不同的主题类别。

（2）有监督算法模式则需要研究者提前给以机器分类指导，例如需要有手工标注的训练集，以此机器将按照研究者的理论进行分类，有监督的算法包括：支持向量机、随机森林等。

（3）文本向量化是通过词频-逆文本频率算法(term frequency–inverse document frequency，TF-IDF)对文本进行向量化，借此对比不同的语料库中词的位置，以及不同文本之间的差异，从而发现隐含在文本中的社会结构和社会关系。

**第四部分 数据分析与研究结论整理（2周）**

根据导师提供的外国人在华论坛文本（或学生自行抓取的网络文本），通过前三部分的学习，进行完整的实操练习，例如使用主题模型区分外国人讨论主题，同时使用有监督算法获得预设的分类模式，最终利用文本生成的变量进行回归分析，或者分析主题的年代变迁。同时可以利用词向量，分析不同词汇的空间距离，例如：中国与各种情绪词汇的距离，借此获得外国人对中国的整体印象。

## 三、预期目标及成果

本课题的重点探讨人工智能时代的社会科学研究法的转变，以方法论训练为主导。在完成对方法的学习后，会通过学生自选感兴趣的子课题进行实操，使学生能够体验完整的研究过程，打好规范做研究的基本功。

1、通过参与移民社会学研究，学生在博士导师的指导下收集到我国移民人群生活、工作的一手资料，了解我国移民人群的真实状况，在掌握先进的人工智能在社会科学领域的研究方法的同时丰富社会体验，并且在完成学习后，学生能够学会辩证的审视新事物，通过考察某种知识产生的过程，反思被社会建构蒙盖的真相。

2、在理论层面，理解福柯式的话语分析。

掌握福柯谱系学思想，通过历史分析或称为“当前史”分析：话语主体何以成为这样，以及社会何以成为这样。

3、在方法层面，掌握从数据收集到形成研究结论的方法流程。

掌握python爬虫的基础库：（1）Request：网页爬取；（2）BeautifulSoup：解析HTML，与正则表达式结合，获得数据。

掌握python机器学习的基础库：（1）NumPy：矩阵操作函数库；（2）Pandas：数据分析库；（3）Matplotlib：数据可视化库；（4）Scikit-learn：机器学习函数库。

掌握python自然语言处理的基础库：（1）Nltk：适用于英文文本的高效自然语言处理库，可用于词干提取等文本预处理；（2）Gensim：多功能的自然语言库，使用LDA主题模型（Topic Model）进行文本主题的挖掘；使用Word2Vec实现文本向量化，进而计算词语相似度。

4、培养科研思维，掌握科学研究的方法和科学论文的写作技巧，通过课题研究得到创新的、稳健的研究结论基础上，撰写研究论文。

## 四、推荐阅读材料

1、话语分析（必读）

福柯.规训与惩罚:监狱的诞生[M].出版社:生活·读书·新知三联书店, 2003.

福柯.词与物:人文科学考古学[M].出版社:上海三联书店,2001.

2、机器学习方法论

尤瓦尔·赫拉利. 《未来简史》[M].中信出版集团, 2017 （数据主义的未来：“信数据得永生”）

周志华.《机器学习》[M].清华大学出版社（通俗易懂的机器学习算法“西瓜书”）

3、前沿论文

Garg N , Schiebinger L , Jurafsky D , et al. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 115(16):E3635.（必读）

Börner K, Rouse W B, Trunfio P, et al. Forecasting innovations in science, technology, and education[J]. Proceedings of the National Academy of Sciences, 2018, 115(50): 12573-12581.（必读）

Zhou M, Chen W, Cai G. Chinese-language media and immigrant life in the United States and Canada[M]. na, 2006.（必读）

Kozlowski A C , Taddy M , Evans J A . The Geometry of Culture: Analyzing Meaning through Word Embeddings[J]. 2018.（选读）

**五、导师介绍**

李博士，中山大学社会学与人类学学院博士生；研究方向：移民社会学、社会学方法；发表数篇CSSCI论文，曾被“人大复印资料”转载。