**《社交网络挖掘》实践项目说明**

1、项目简介

1.1 实践形式

本次项目实践通过小组合作方式完成，以组为单位提交结果，每个小组人数在3人以内，同学们请自行组队，并于4月2号(周一)之前把成员组成和选题在微信群里向助教汇报。

1.2 项目要求

1）项目选题：请在下文介绍的可选题目中任选一个完成，并按下述要求在规定的时间内进行实践成果的现场演示，同时提交项目报告、源代码等。

2）现场演示：每个小组对其实践项目成果在6月18日（第16周周一）的上机课上，向全班同学进行现场演示，做相关的介绍，并接受老师和同学们的提问。其他同学和老师、助教将共同为每组演示的成果进行现场评分。

3）项目报告：以书面报告（pdf格式）的形式详细阐述本小组完成项目课题的基本思路、数据处理方式、算法、实现工具/语言、实验效果、性能评价和小组成员的分工说明（必需）等内容。报告可用中文或英文书写，用英文书写且表述清晰明了的可酌情加分。若算法或实现代码是借鉴了他人工作的，请在论文中详细说明并附上参考文献。现场演示时展现的效果/结果与项目报告不符的，须做解释说明，否则会被扣分。

4）成果提交要求：项目报告以“学号1\_学号2\_学号3\_项目报告”命名并于6月15日（周五）前发送至助教邮箱；项目实践过程中用到的全部源代码（建议用Python或Java语言实现，数据库推荐MySQL）和其它相关资料请压缩成zip或tar.gz格式后以“学号1\_学号2\_学号3\_源代码”命名并于6月15日前发送至助教邮箱，文件过大的可以在现场演示时拷贝或通过网盘发送。教师和助教认为项目成果还需提供数据集进行检查的，必须提交数据集。以上成果的提交每超时1天，扣除该项目成绩最终得分的10%。

1.3项目评分标准

该项目实践评分共占期末总成绩的45%，以满分45分计算（完成加分项要求的最多可得55分），其中：

现场演示同学评分占15分+5分（加分项）

现场演示教师（包括助教）评分占20分+5分（加分项）

项目报告评分（教师和助教评判）占10分

重要提示：因为以下各项目选题的难度不同，因此教师（包括助教）对各组的评分会根据所做项目的实际难度做不同考虑，且能顺利完成加分项要求的可获额外加分（由教师、助教和同学共同打分，加分最多不超过10分）。

2、项目选题

2.1. 在线社交网络的社区发现与相关分析

2.1.1 数据集

斯坦福大学的社会网络分析实验室（SNAP）是国际领先的研究社交网络的科学实验室，其网站（<http://snap.stanford.edu/data>）上提供了很多著名在线社交网络的数据集，包括网络的用户结点、关系边、用户的评论/发言等等。

2.1.2项目要求

a） 选择SNAP中提供的一种社交网络数据，应用社区挖掘算法请将相应社交网络中的不同社区用可视化技术（具体展示工具不做硬性要求）展现出来。

b） 对所选的社交网络做进一步分析，例如应用各种结点中心性度量算法将最具影响力/权威性/中枢性的结点识别并凸显出来，通过图表展现网络的各种属性（如用户间平均距离、用户度分布、图（社区）的直径等）。

c） 加分项参考：利用社交网络数据集提供的信息，设计一种机器学习模型，预测好友关系的形成（即判断哪些结点间会产生新的边），同时说明预测模型性能评价的方法和结果。

2.2豆瓣电影的个性化推荐

2.2.1 数据集

在e-learning课程网站上提供下载（压缩包“豆瓣数据.rar”中的文件“douban.sql”），包含了豆瓣网上1000名用户和1000部电影的相关信息，对数据的具体描述请参见课程网站上的文件“豆瓣数据说明.txt”。

2.2.2 项目要求

a） 将1000名用户之间的社交关系用可视化工具展现出来。

b） 根据该数据集提供的信息，设计一种电影推荐的算法，从1000部电影中选出合适的电影推荐给这1000名用户，同时说明推荐效果的评价方法并给出推荐效果得分。

c） 加分项参考：自行设计爬虫上豆瓣网或其他网站（如百科或公开的知识图谱网站<http://kw.fudan.edu.cn/>）爬取更多的数据，从新数据中抽取新的特征加入推荐系统，并充分证明推荐系统性能的提升（课程网站上的豆瓣爬虫源代码可供参考，文件名“豆瓣爬虫.rar”）。

2.3基于学术文献的社交网络挖掘

2.3.1 数据集

DBLP（<http://dblp.org/>）是国际上最知名的收录计算机/IT领域学术文献的网址，其收录的数据包括论文的标题、作者列表、发表时间、发表刊物/会议等信息。清华大学的学术搜索网站ArnetMiner（<https://cn.aminer.org/>）基于DBLP和ACM等数据源收集了更丰富的文献数据，其中就有文献的引用数据集（<https://cn.aminer.org/citation>），其中包含了几十万篇计算机学术论文的相关数据，为txt格式，需自行解析。

2.3.2 项目要求

a） 设计聚类算法或社区挖掘算法，对数据集中的所有论文进行聚类，用可视化工具展现出各个领域（即社区，需注明对应的社区研究主题），并凸显各领域中最有影响力的学者。

b） 能对输入的任意一个学者，展现其自我中心网络（ego-network）的功能（参照ArnetMiner网站上的功能示例）。

c） 加分项参考：利用DBLP和ArnetMiner提供的其它数据，对更多的学者间社会关系进行分析建模和预测，例如预测两个学者间的合作或引用关系，预测一个学者将来会在哪个刊物/会议上发表论文。

2.4新浪微博社交网络挖掘

2.4.1 数据集

该项目不提供数据，需自行编写爬虫程序去新浪微博上爬取，课程网站上提供的豆瓣爬虫程序可供参考。

2.4.2 项目要求

a） 参考提供的豆瓣爬虫代码，自行设计爬虫程序，爬取新浪微博中与某主题（例如“大数据”）有关的一定数量（例如1000个）微博用户数据。主题有关的判定标准可自行确定，须在项目报告中对其合理性做详细说明。

b） 利用网络结点中心性度量等算法将最有影响力的前20位微博用户（意见领袖）找出来，并进行排序。

c） 加分项参考：针对爬取的微博用户画出标签云，或演示出意见领袖发表的信息在其社交网络中的信息扩散路径（树）。

2.5 自选题目

若小组想要做其它感兴趣的社交网络挖掘项目，可在4月2日之前向助教提出申请，要求详细阐述选择该项目的理由、项目的实践目标、数据集的获取方式等。助教会在一周内答复是否同意该自选题，或对所选题目提出修改要求。

如有任何疑问，可随时联系教师和助教。

附：参考资料

可视化工具：

Pajek:<http://pajek.imfm.si>

D3: <https://d3js.org>

NetworkX: <https://networkx.readthedocs.io/en/stable/>

<http://blog.sciencenet.cn/home.php?mod=space&uid=404069&do=blog&classid=141080&view=me&from=space>

Gephi: <https://gephi.org/users/>

<http://blog.csdn.net/zdw12242/article/details/8687644>

爬虫Cookie：<http://blog.csdn.net/zhyh1435589631/article/details/51307915>

<http://www.nowamagic.net/academy/detail/1302882>

爬虫防封技巧：<http://blog.csdn.net/zhanghaipeng1989/article/details/40828377>