**《社交网络挖掘》实践项目说明**

**1、项目简介**

1.1 实践形式

项目实践通过小组合作方式完成，以组为单位提交成果，每个小组人数在3人以内，同学们请自行组队，并于10月8日之前把成员组成和选题通过群内腾讯文档形式提交给助教。

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第18周之前（具体时间待定）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。其他同学和教师、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（word或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，要求不超过8页A4的篇幅。若算法或实现代码是借鉴了他人成果的，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：2023年 1月5 日（第18周周五）晚24点前将以下文件发送至助教邮箱sjwlwj2023@163.com

a） 项目报告，以“学号1\_学号2\_学号3\_项目报告”命名；

b） 项目的全部源代码和其它相关资料请压缩成zip或tar.gz格式，并以“学号1\_学号2\_学号3\_源代码”命名。文件过大的可以在现场演示时拷贝或通过网盘发送；

c） 教师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47%（另50%为期末闭卷考试成绩，3%为评教成绩），以满分100分计算，则其中：

* 现场演示的同学互评评分占50%
* 现场演示教师和助教的评分占30%
* 项目报告评分（教师和助教评判）占20%

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评分时做不同考虑。

**2、项目选题介绍**

课题1：社交网络节点分析与社区挖掘

任务说明参考：

a） 应用某种社区挖掘算法划分网络中的不同社区，并用可视化技术展现出（具体展示工具不做硬性要求）。

b） 对网络做进一步分析，例如应用各种节点的中心性度量算法将最具影响力/权威性/中枢性的节点识别并凸显出来，通过图、表等形式展现网络的各种属性度量结果（如节点间平均距离、度分布、图/社区的直径、网络结构演化等）。

c） 附加任务（非必做，可任选其一）：

（1）链接预测：利用网络数据集提供的信息，设计一种机器学习模型预测网络中边的形成，同时说明预测模型性能评价的方法和结果。针对不同网络数据集，链接预测可对应不同的应用任务，如好友预测、科研合作预测、购买行为预测等。

（2）节点分类：设计一种模型判别网络中各节点的类别，同时说明模型性能评价的方法和结果。根据实际数据集可以实现二分类或多分类。

数据集1：社交网络ego-Facebook <http://snap.stanford.edu/data/ego-Facebook.html>

数据集2：投票网络Wiki-Vote network <http://snap.stanford.edu/data/wiki-Vote.html>

数据集3：学术引用网络Citation-network <https://www.aminer.cn/citation>

更多可选数据集：<http://snap.stanford.edu/data/>

课题2：社交网络中的用户行为挖掘

任务说明参考：

1. 多维度分析社交网络中用户之间，或者用户与资源（对象）之间交互的特性。例如，学术合作网络中，学者间的论文合作、引用等行为预测；商品、帖子、图片、视频的热度（购买率、点击率、点赞数等）增长与变化模式如何？分析其原因。
2. 用户行为标签（多）分类。类似节点分类，利用社交网络结构特征和其他用户的行为数据等预测目标用户的不同行为类型。

参考文献见elearning上的参考资料。

课题3：异常用户（僵尸、水军、垃圾用户）检测

任务说明参考：

a）可视化用户之间的社交关系。

b）从数据集中抽取用户社交关系网、发言行为等特征，设计一种机器学习模型来识别社交网络中的虚假、恶意用户，并说明评价方法和结果。

c）展现虚假用户和正常用户的（局部）网络，基于课程所学知识分析两个网络的异同。

将一个邮箱地址对应为一个用户，则该任务变成垃圾邮件/邮箱检测。

参考数据集：

Tagged <http://github.com/shobeir/fakhraei_kdd2015>

MuMiN <https://paperswithcode.com/dataset/mumin>

课题4：基于地理位置的社区挖掘与预测

任务说明参考：

a） 根据用户的兴趣地点（POI）对用户进行社团划分并做可视化展现。

b） 可视化展现用户（可选取代表性用户）POI的时间和空间轨迹，并基于分析结果从社交角度做出深入解读。

c） 根据用户历史上的POI，设计模型预测/推荐用户新的POI。

数据集1: Foursquare

<https://www.kaggle.com/chetanism/foursquare-nyc-and-tokyo-checkin-dataset>

数据集2: Brightkite

<http://snap.stanford.edu/data/loc-Brightkite.html>

参考文献：

<https://www.cse.cuhk.edu.hk/irwin.king/_media/presentations/p325.pdf>

课题5：用户评论观点挖掘/情感分析

任务说明参考：

1. 利用文本主题模型、语言模型等机器学习模型（传统模型、深度模型、大模型都可）分析社交网络中用户评论的观点或者情感倾向，也可以预测用户对给定对象（新闻、商品、电影等）的情感倾向类别。模型可利用的特征来源视数据集实际情况而定，不强求使用社交关系数据。
2. 可视化展现用户观点/情感分析的结果，如主题关键词、情感特征词的数量、分布等。

参考数据集：

GoEmotions https://github.com/google-research/google-research/tree/master/goemotions

Stanford Sentiment Treebank https://nlp.stanford.edu/sentiment/index.html

IMDB dataset https://ai.stanford.edu/~amaas/data/sentiment/

课题6：信息传播预测

任务说明参考：

1. 去社交网络上爬取信息传播的数据（如转发微博的用户列表）和社交网络结构，展现信息级联的相关情况，包括信息热度分布、时间变化曲线、转发树等。
2. 设计一个级联预测模型，即根据已知的参与级联（传播信息）的用户列表（序列），预测下一时刻会参与级联传播的用户（微观预测）、整条信息传播的热度（宏观预测）等，说明模型评价方法和结果。

参考数据集：

<https://github.com/zhitao-wang/Hierarchical-Diffusion-Attention-Network/tree/master/data>

参考文献见elearning上的参考资料。

课题7：用户标签画像

任务说明参考：

1. 设计算法模型，利用社交关系、标签间的语义关联等特征为标签缺失的用户打上标签，或对已有标签的用户进行标签补全、优化。
2. 根据用户的标签对用户进行分组（可结合社交关系），并对每组用户特点进行分析，可视化分组情况。

参考数据集：

微博用户标签数据集（待补充）

课题8：个性化推荐

任务说明参考：

a） 将用户之间的社交关系用可视化工具展现出来。如用户太多，可视化工具展现不清或计算耗时，可以设计一种抽样方法只展现部分用户。

b） 利用数据集设计一种推荐模型并验证其推荐效果，对模型的推荐结果进行分析说明。推荐模型既可以选择传统的浅层学习模型（如协同过滤、矩阵分解）或深度学习模型、大模型。

c） 一般而言，加入模型的用户/物品特征越多，模型推荐效果越好。因此可自行设计爬虫或调用API获取相关网站的数据，从中抽取更多关于用户和物品的属性信息，并针对这些信息设计有效的算法提取特征，融入推荐模型。

d） 附加任务：为推荐结果设计可解释算法提供可解释说明（模型生成）。

参考数据集：

可用eLearning上的豆瓣数据和网易云音乐数据，亦可从参考论文获其他渠道自己寻找相关推荐数据集。

亚马逊数据集：<http://jmcauley.ucsd.edu/data/amazon/>

课题9：领域知识图谱构建及应用

任务说明参考：

a) 基于某特定领域数据自动化地构建领域知识图谱（KG），可以考虑基于文本语料库实现实体识别、实体链接或者关系抽取等功能。

b) 可视化展现所构建KG的结构、数据等。

c) 附加任务：基于构建好的KG实现一个简单的知识问答系统，即对于用户提出的一个关于KG中的知识问题，系统从KG中选择若干知识（实体、概念、属性或者文本描述等）作为该问题的答案，对问答系统效果进行评估并进行分析；或者融入知识图谱的推荐系统。

参考数据集：

<https://wiki.dbpedia.org>

<https://github.com/villmow/datasets_knowledge_embedding>

课题10：自选题目

若想要做其它感兴趣的社交网络挖掘项目，可在9月29日之前向助教提出申请，要求详细阐述选择该项目的理由、实践完成的目标、数据集的获取方式等。教师或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。

\*以上所有课题任务皆可自行爬取数据，或从eLearning上参考文献（后续更新）中提供的数下载链接下载数据集，相关模型、算法亦可借鉴参考文献。

**附：**其他参考资料

可视化工具：

包括Gephi, Echarts, Highcharts, D3, Pajek, NetworkX等。

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

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

D3: <https://d3js.org>

Neo4j: <https://neo4j.com>

NetworkX: <https://networkx.org/documentation/stable/index.html>