《网络群体与市场》2019 年课程设计

新闻数据中的社会网络挖掘

给定一批新闻数据（可从 obe-课件页面下载，文件名 gov_news.zip），从中挖掘出这些新闻中的社交网络关系。

数据格式：数据文件中包含了从人大新闻网采集的新闻内容（http://news.ruc.edu.cn/archives/category/important_news），包括新闻编号、发表时间、其他信息以及正文内容。 每一行为一条新闻，其中各个字段用\t 隔开。

基于该数据，完成下述内容：

一、数据预处理（0 分）

（1） 使用结巴分词，对其中的新闻标题和正文内容进行分词并抽取其中包含的人名、机构名和地名。结巴分词： https://github.com/fxsjy/jieba。注意，为了检测其中的实体，请使用 posseg 。Jieba 的词性标注列表请参考 https://blog.csdn.net/Yellow_python/article/details/83991967。相关的实体类型如下：

nr	人名
nrfg	人名
nrt	外国人名
ns	地名
nt	机构团体
nz	其他专名

（2） 统计其中包含的热门人物和机构，验证数据的正确性

（3） 建立社交网络图，两个人如果出现在同一篇新闻中，则假设这两个人有联系。两个人的联系强弱可以通过共同出现的文章的数目来表示。例如，假设 A 和 B 在 10 篇新闻中同时出现过，则 A-B 之间的边的权重为 10。

二、完成下面基础内容（30 分）：

图的验证： 提供界面，输入一个人 A 进行查询，可以输出和 A 关系最强的前 10 个人（邻居）。

图的统计：计算图的节点个数，边数，连通分量的个数，最大连通分量的大小。

影响力计算：使用 PageRank 算法计算每个人的影响力大小。并给出影响力最大的前 20 个人。

三、自选分析：在上面的基础上，完成下面内容几项内容中的任意 3 项（每项 15 分，共 45 分）：

（1） 小世界理论验证：计算该社交网络中，任意两个人之间的平均路径长度是多少？提供界面，输入A和B，找出A和B之间的前10条最优路径（路径越短越优，路径长度相同时，按照路径上权重总和由大到小排序）。
（2） 三元闭包验证：该数据上的社交网络关系是否符合三元闭包理论？请进行验证。
（3） 社区挖掘：挖掘该社交网络中的社区（Community Detection）。
（4） 中心性计算：计算每个节点的中介中心性，并输出中介中心性最大的10个人。
（5） 节点的聚集系数计算：计算每个节点的聚集系数，并输出聚集系数最大的10个人。
（6） 结构洞挖掘：挖掘该社交网络上的结构洞，输出结构洞数目和结构洞示例。

除上述内容外，也可以在图中融入机构名、地名等进行综合分析，分析出有价值有意义的现象。会根据内容适当加分。

三、使用计算机学报的论文模板和内容要求，撰写论文。要求格式规范，内容详实，有必要的思考，在摘要部分，清晰的写出完成了自选分析中的哪几项（未清晰写出的扣10分）。(15分)。

四、提交完整的代码、可执行程序、用户手册（10分）

五、按时提交，提交截止日期是2020年1月7日，晚一天扣2分。2020年1月14日后不再接受提交。

提示：可根据需要自学和使用开源工具，例如图数据库 NEO4J，图的可视化工具 vis.js, echarts, d3，图分析工具NetworkX，建议使用Python等。

论文格式可参考课件页面中的论文示例