信 息 内 容 安 全 实 验 报 告

实验名称： 人民的名义红楼梦人物关系分析

班级： SC011701

姓名： 李向娟（2017302235）

指导教师： 杨黎斌

实验时间： 2020.4.12

目录

《人民的名义》及《红楼梦》人物关系图谱分析

1、实验目的

①以《人民的名义》和《红楼梦》剧情梗概为材料，建立剧情人物关系图谱；

②掌握绘制关系图谱的一般方法

2、理论依据

2.1 jieba 分词

Jiaba 分词主要有三步：构建词典；构建有向无环图；计算最大概率路径。

1. 构建前缀词典：

基于统计词典构造前缀词典，统计词典有三列，第一列是词，第二列是词频，第三列是词性

2. 构建有向无环图：

根据前缀词典对输入文本进行切分，比如“北”，有北、北京、北京大学三种划分方式。因此，对于每个字，可以构建一个以位置为 key，相应划分的末尾位置构成的列表为 value 的映射。

3. 最大概率路径计算

在得到所有可能的切分方式构成的有向无环图后，我们发现从起点到终点存在多条路径，多条路径也就意味着存在多种分词结果。需要计算最大概率路径。

计算最大概率路径时，jieba 采用从后往前的方式，采用动态规划计算最大概率路径。每到达一个节点，它前面的节点到终点的最大路径概率就已经计算出来。

2.2 共现网络

实体间的共现是一种基于统计的信息提取。关系紧密的人物往往会在文本中多段内同时出现,可以通过识别文本中已确定的实体(人名),计算不同实体共同出现的次数和比率，当比率大于某一阈值，我们认为两个实体间存在某种联系。

2.3 Gephi 软件

Gephi 是一款开源免费跨平台基于 JVM 的复杂网络分析软件，其主要用于各种网络和复杂系统,动态和分层图的交互可视化与探测开源工具。可用作探索性数据分析、链接分析、社交网络分析、生物网络分析等。

Force Atlas:

基于力导向 (Force-directed) 的算法作为弹簧理论算法的一类典型，被广泛应用于描述社交网络等关系型信息图。它的原理其实非常易懂，我们可以把整张网络想象成一个虚拟的物理系统。系统 中的每个节点都可以看成是一个带有一定能量的放电粒子,粒子与粒子之间存在某种库仑斥力，使它们两两相互排斥。同时，有些粒子间被一些“边”所牵连，这些 边产生类似弹簧的胡克引力，又紧紧牵制着“边”两端的粒子。在粒子间斥力和引力的不断作用下，粒子们从随

机无序的初态不断发生位移，逐渐趋于平衡有序的终 态。同时整个物理系统的能量也在不断消耗，经过数次迭代后，粒子之间几乎不再发生相对位移，整个系统达到一种稳定平衡的状态，即能量趋于零。基本上绝大多数算法都遵循着这样的原则，即：

将网络看成一个顶点为钢环，边为弹簧的物理系统

不断迭代，使整个系统的总能量达到最小

3、代码实现过程

3.1 系统设计流程图

整个实验过程可以分为三步：

第一步：根据获取的剧情梗概文本，建立人物词典，包括姓名、频数、词性三方面内容，利用 jieba 分词对剧情文本进行处理，得到关系图中的节点信息，并保存在 csv 文件中；

第二步：创建角色关系即关系图中的边，这一步需要用到创建的

列表 lineName[i],它存储了每一段中出现过的人物，初始化情况下认为每一行人物两两相连，若人物之间未有边则权值置 1，否则权值+1,在输出结果时去掉冗余边，结果保存在 csv 文件中。

第三步：可视化网络，导入前两步生成的节点表和关系表，调整参数，生成人物关系网络。

3.2 使用 jieba 中文分词

中文分词主要使用的是 Python+Jieba 分词工具，同时导入自定义词典 dict.txt。

jieba 中文分词涉及到的算法包括：

(1) 基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；

(2) 采用了动态规划查找最大概率路径，找出基于词频的最大切分组合；

(3) 对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

jieba 中文分词支持的三种分词模式包括：

(1) 精确模式：试图将句子最精确地切开，适合文本分析；

(2) 全模式：把句子中所有的可以成词的词语都扫描出来，

速度非常快，但是不能解决歧义问题；

(3) 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

在本实验中，利用 jieba 精确模式进行分词，同时使用 jieba.possge 标记词性，建立姓名词典 Names、关系词典 relationships、lineNames,使用 Names 保存人物，该字典的键为人物名称，值为该人物在全文中出现的次数。Relationships 保存人物关系的有向边，该字典的键为有向边的起点，值为一个字典 edge，edge 的键是有向边终点，值为有向边的权值，代表两个人物之间联系的紧密程度。lineNames 是一个缓存变量，保存对每一段分词中当前段出现的人物名称，lineName[i]是一个列表，存储第 i 段中出现过的人物。

读入剧本每一行，对其分词，判断该词的词性是否为“人名”（词性编码：nr），如果该词的词性不是 nr,则认为该词不是人名，提取每一行中出现的人物集，存入 lineName 中，之后对出现的人物，更新他们在 names 中粗线的次数。

3.3 输出节点信息与人物关系信息

对于 lineNames 中的每一行,初始假定该行出现的所有人物两两相连，如果两个人物之间尚未有边建立，将新建的边权值设为 1，否则将已存在的边的权值加 1，在输出边的过程中假设共同出现次数少于 3 次的是冗余边，在输出时跳过这样的边，输出节点保存为包含

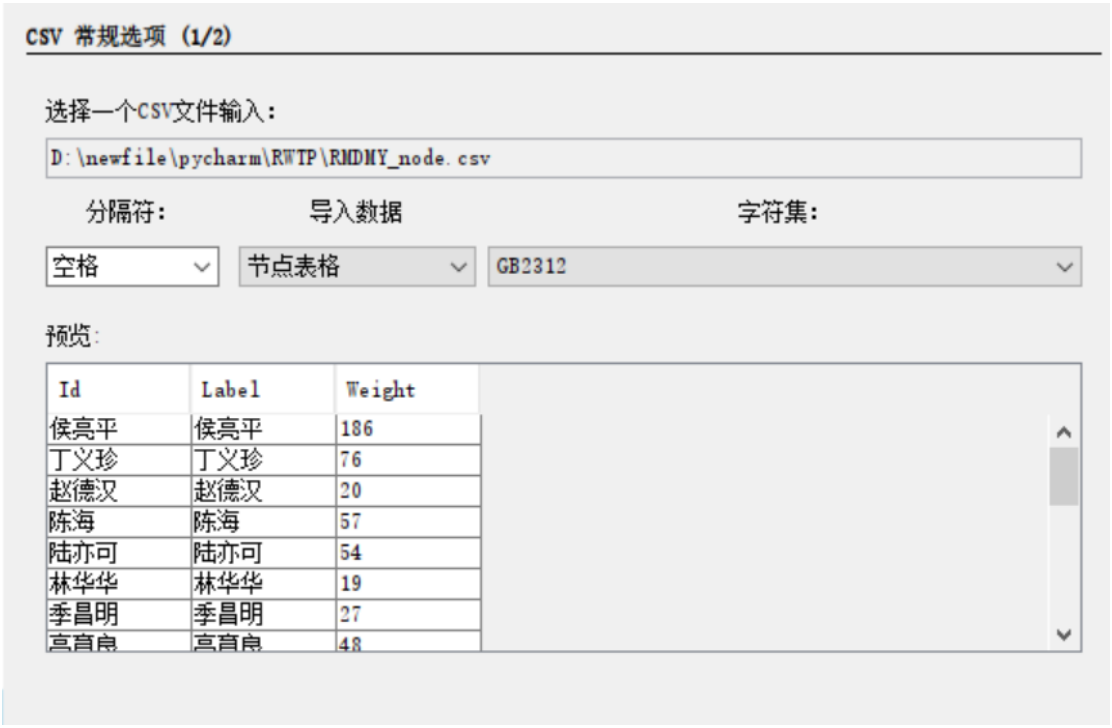
ID,Lable,Weight 信息的 RMDMY_node.csv,边保存为 RMDMY_edge.csv（包含 Source,Target,Weight),红楼梦中信息保存为 HLM_egde.csv 以及 HLM_node.csv。

3.4 Gephi 绘制关系图

安装 Geiph 软件并配置 JDK 环境,导入电子表格 RMDMY_node.csv 和 RMDMY_edge.csv,选择数值设定中的 Modularity Class,并调整连入度,布局选择 Force Atlas,再通过预览处理得到最终人物关系图谱。

4、实验结果及分析

《人民的名义》



CSV 常规选项 (1/2)

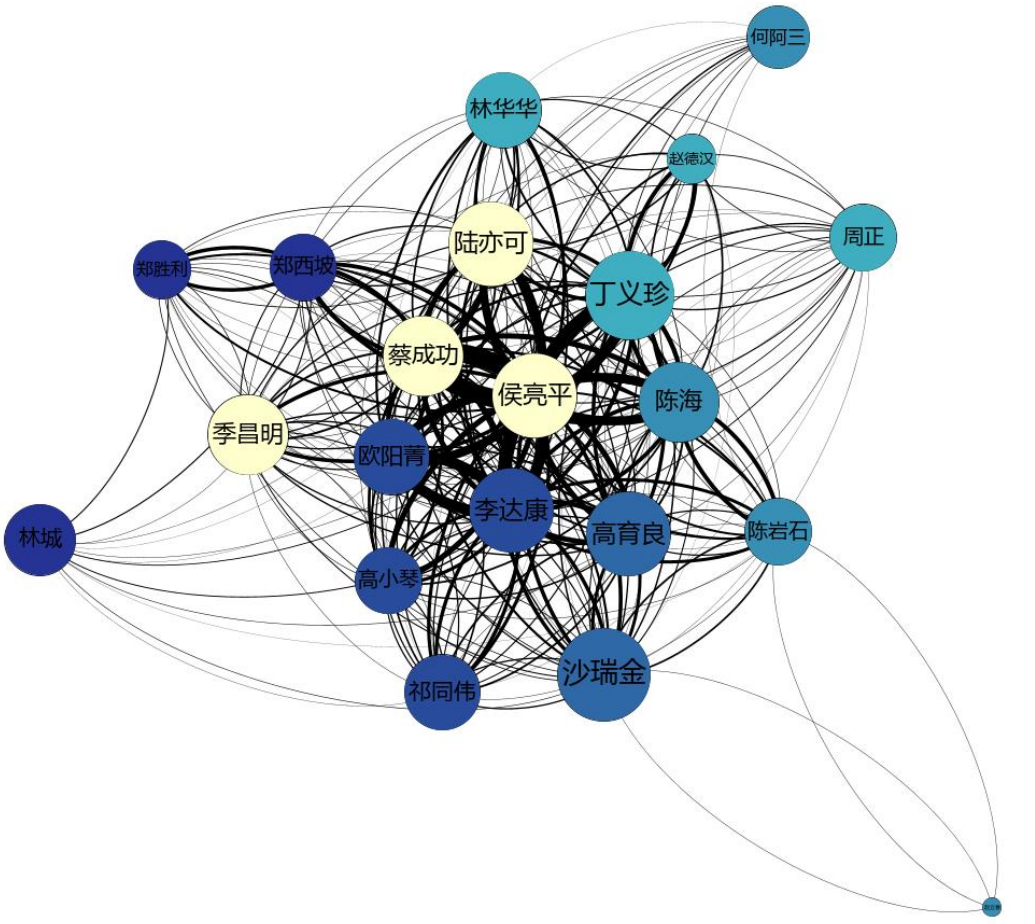
选择一个CSV文件输入：

D:\newfile\pycharm\RWTP\RMDMY_edge.csv

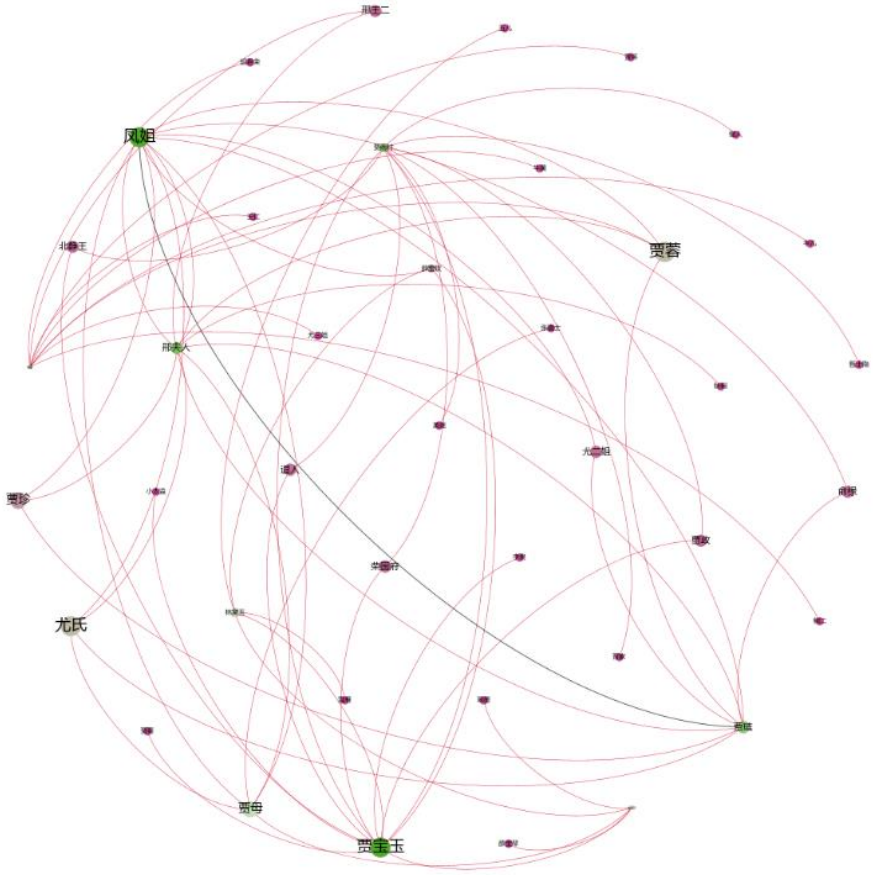
分隔符：	导入数据	字符集：
空格	边表格	GB2312

预览：

Source	Target	Weight
侯亮平	丁义珍	724
侯亮平	赵德汉	140
侯亮平	陈海	599
侯亮平	陆亦可	548
侯亮平	林华华	203
侯亮平	季昌明	250
侯亮平	高育良	357
侯亮平	银行卡	17



《红楼梦》



在本次实验中，通过对人民的名义中人物关系的分析，我进一步掌握了利用 jieba 进行分词并输出指定类型的结果,并初步了解和学会使用 gephi 软件进行关系网络图谱的绘制，在实验的最后，我尝试了用源码对《红楼梦》中人物关系进行分析并绘制关系网络图，根据结果来看仍存在需要改进的地方。