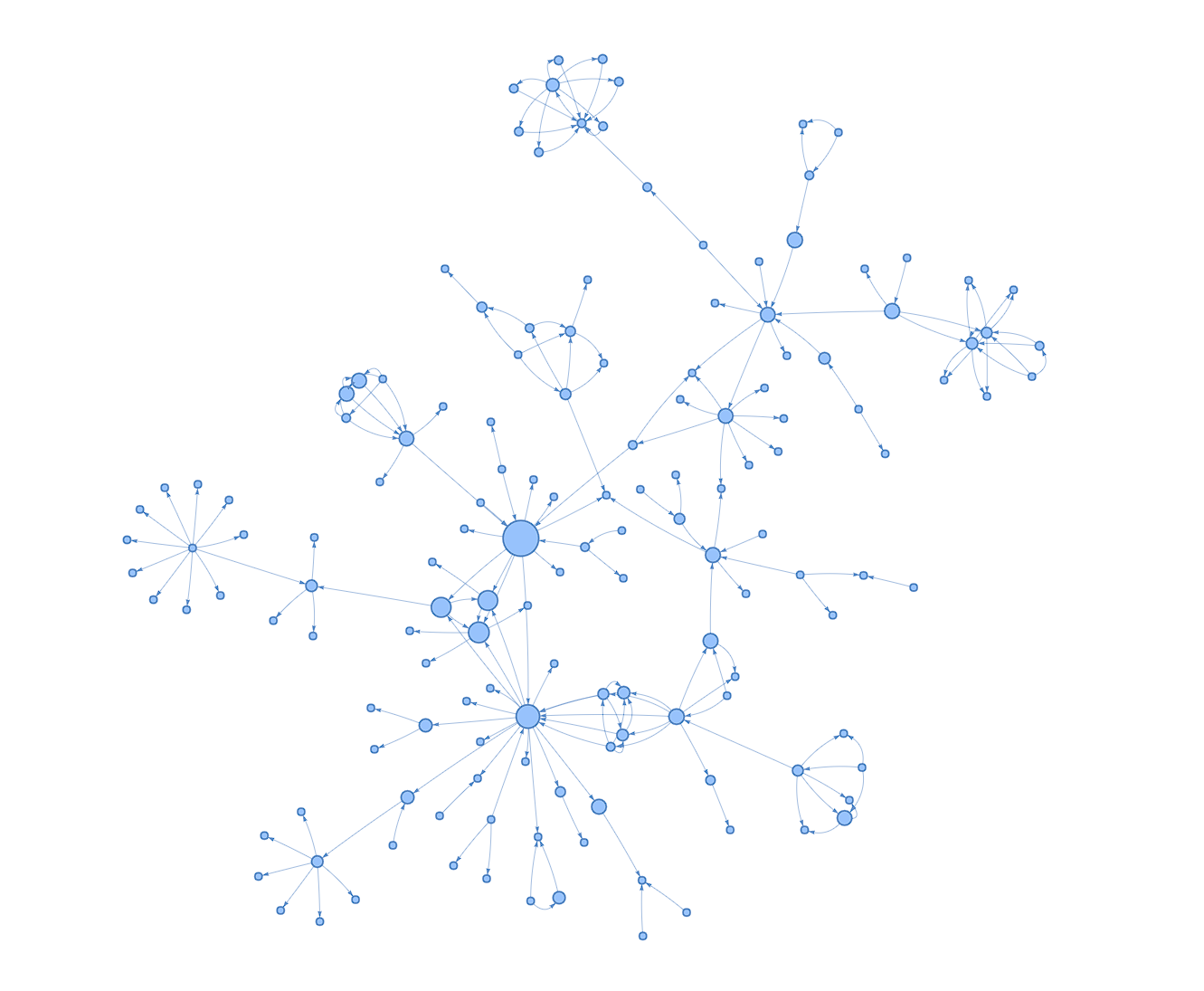
**Python大作业实验报告**

**题目1：人物图谱  
具体描述：以《西游记》文本作为输入数据，构建人物图谱。自动抽取人物及其关系，节点大小综合表示该人物出现次数，重要程度及法力大小等。自动生成人物名片，名片内容包含武艺、宝贝、外貌、性格、身世、口头禅等。用GUI显示人物关系谱图，当鼠标点击图谱中某人物时，显示其名片并以该人物为中心重点显示包含其相关人物的子图谱。名片中的关键词或句子都是超链接，鼠标点击后弹出小窗口，内容为对应原文中的片段。**

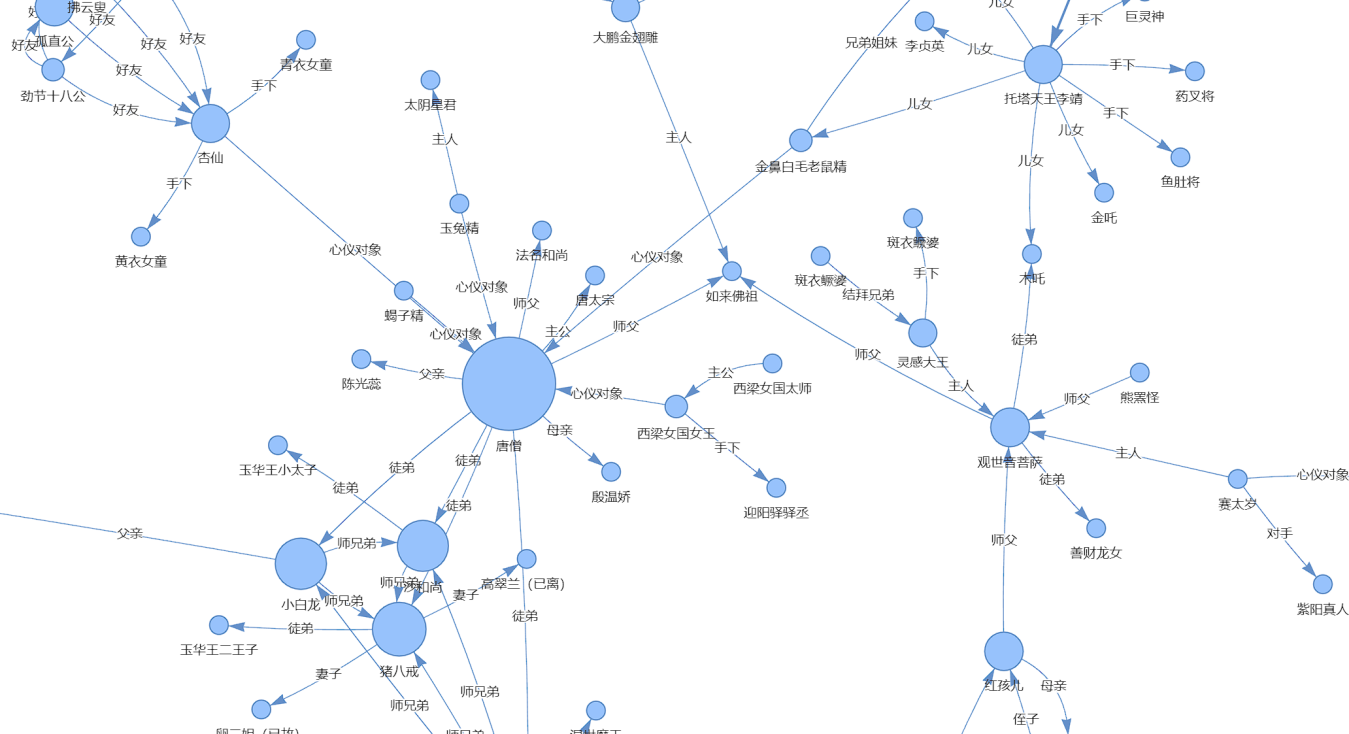
**代码运行结果：**

总览：



放大后：

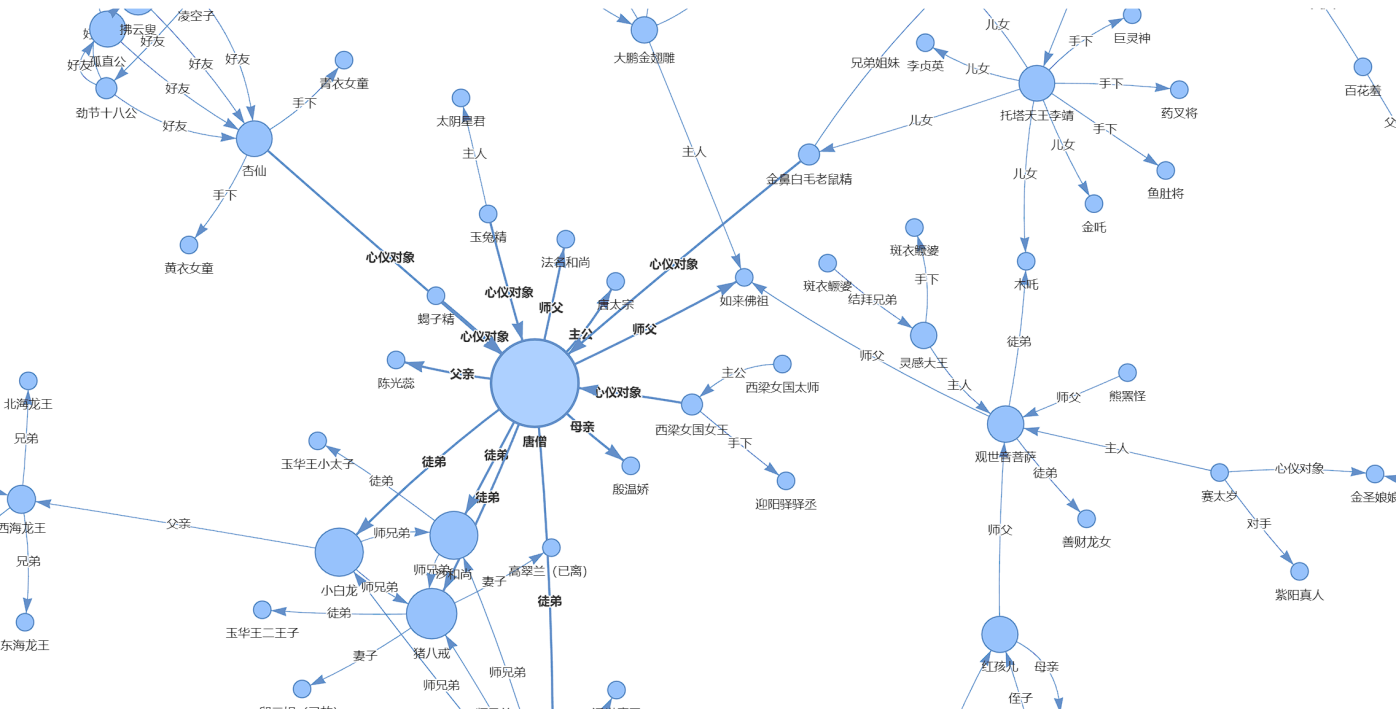
1.出现人物姓名和人物之间的关系。人物节点大小由出现频数+结点度数决定。



2.鼠标放在节点上可显示关系、名片、节点大小等相关信息。



3.当点击名片时，根据人物关系，会高亮显示与该人物相关的子图谱。

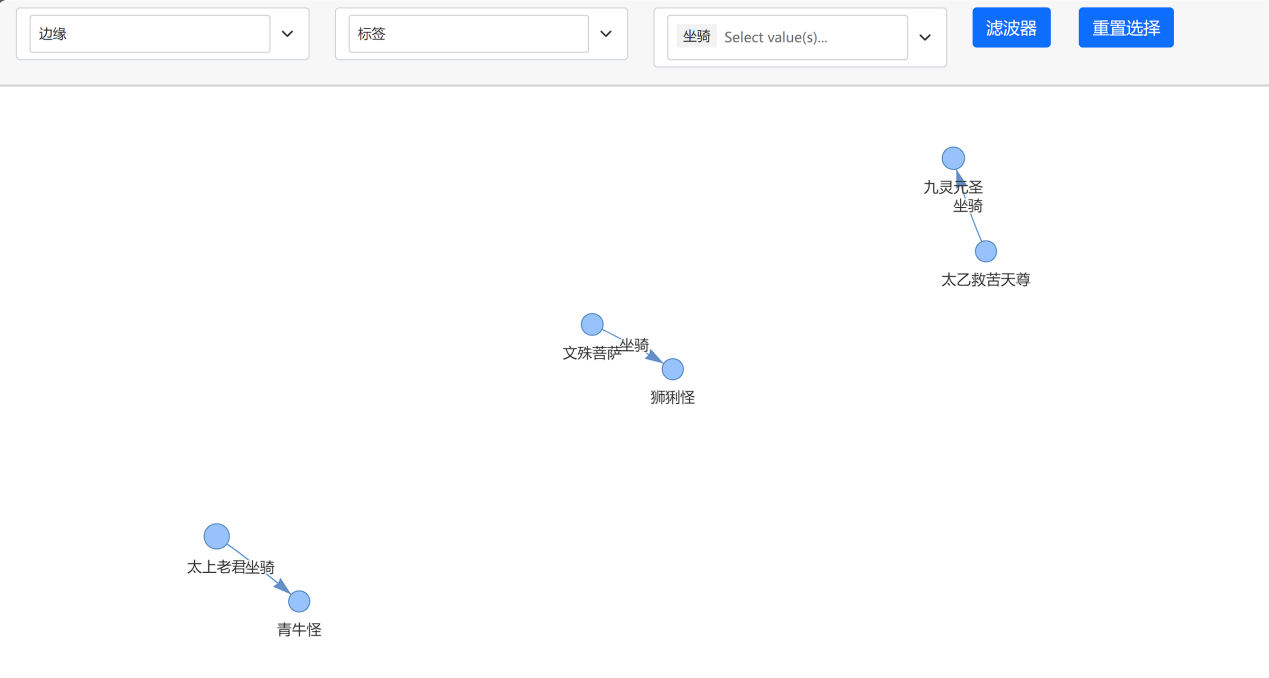


网页功能：

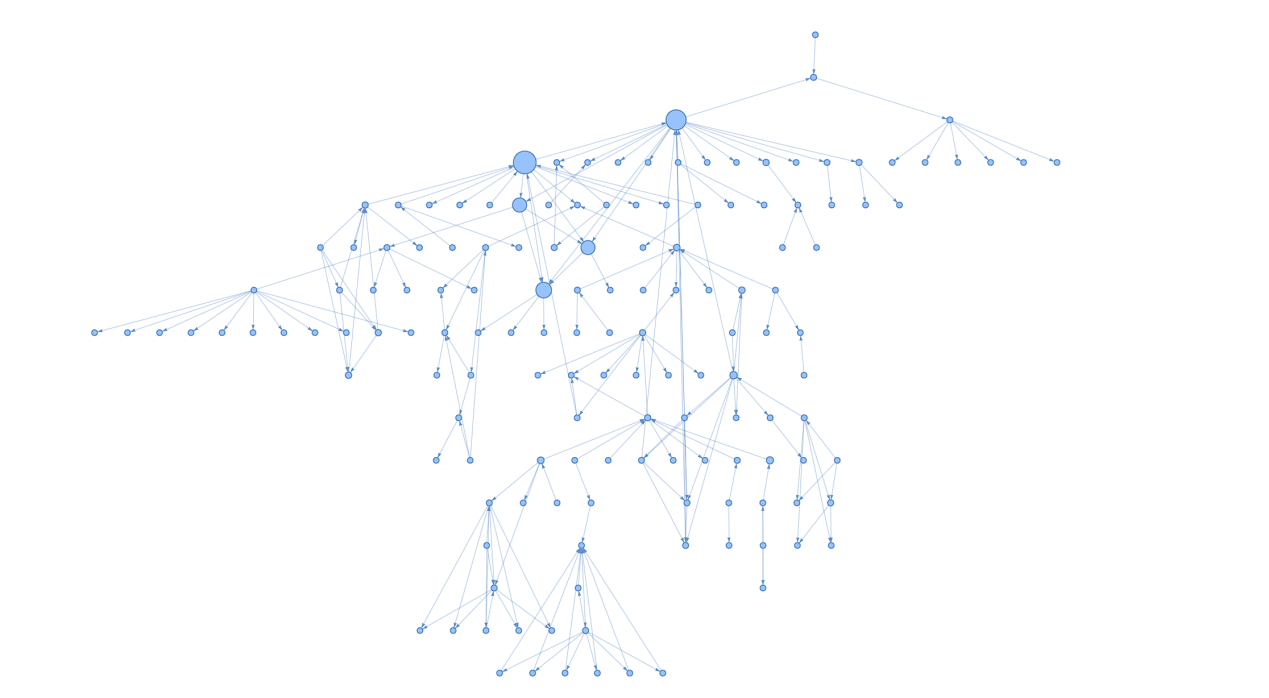
1.放大缩小，调整GUI显示（如颜色，边框），调整图谱重力，拖拽节点。

2.最上面一栏可以通过关键词搜索，得到特定节点和边：

如筛选“边”-“标签”-”坐骑”。



1. 调整图谱的形式，点击layout按钮可以切换成树状图。



**任务分析：**

构建人物图谱并实现相关功能，涉及文本处理、关系抽取、图谱构建、用户界面设计等多个领域。

**设计思路与流程：**

****

步骤一：文本处理

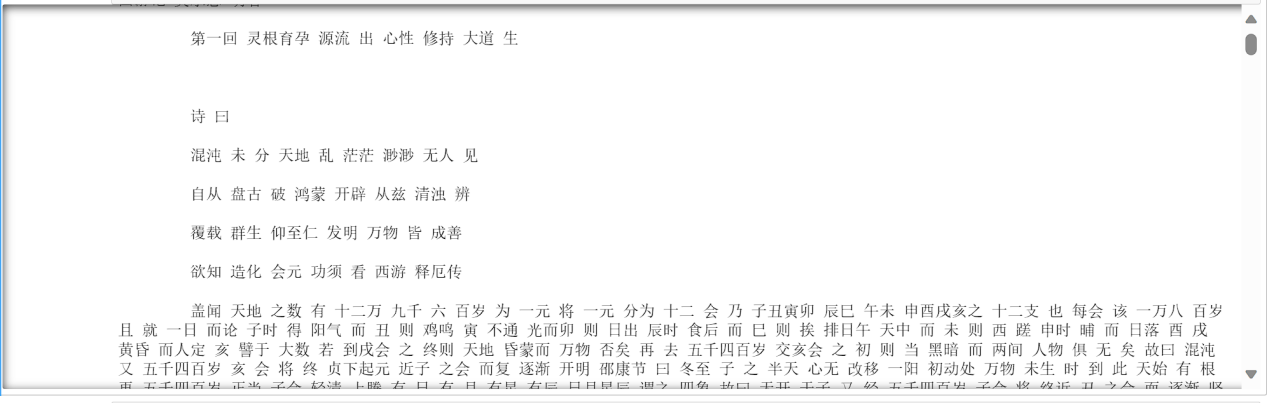
《西游记》相关文本数据，通过网络搜索或网络爬虫获得。

初期，我们尝试通过对《西游记》文本数据进行自然语言处理，进行实体识别和关系抽取。但在尝试过后，我们发现其中有许多工程问题难以解决，提取出来时实体和关系准确率低。因此我们退而求其次采用网络爬取和人工搜索相结合，获取了西游记中大部分的人物和他们之间的关系。

尝试代码（NLP.ipynb）：

1.使用jieba进行文本分词

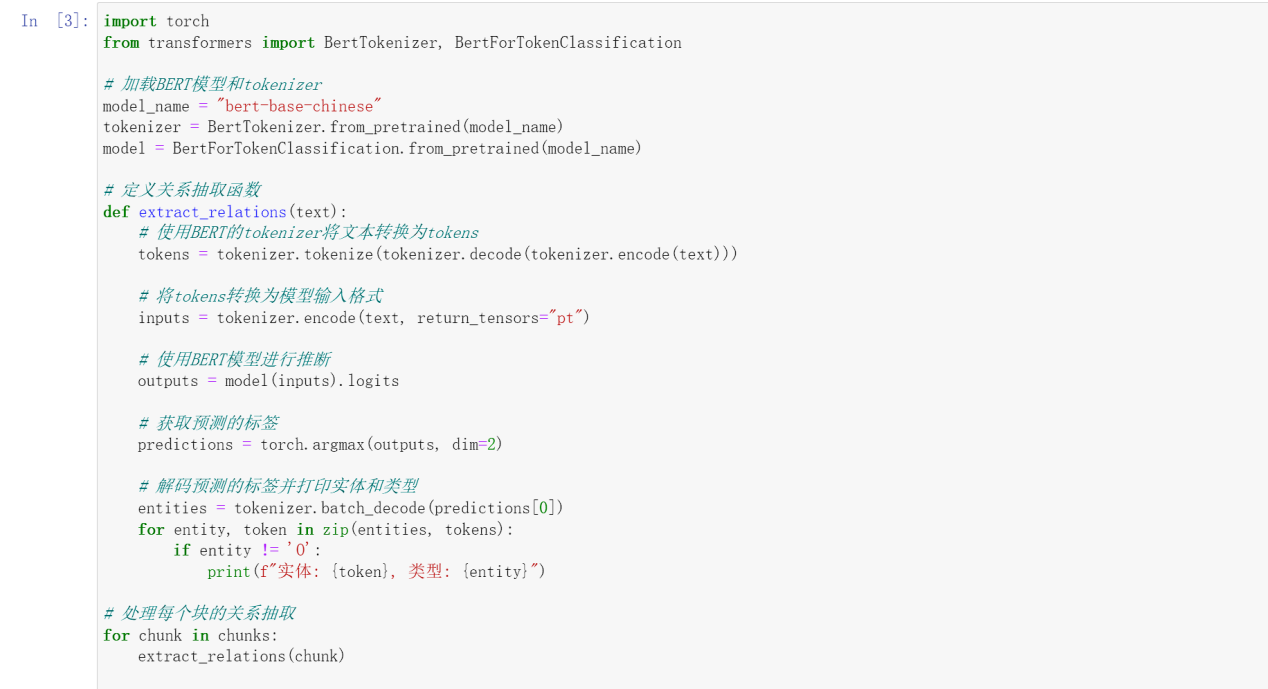




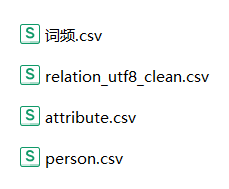
2.使用spacy进行命名实体识别



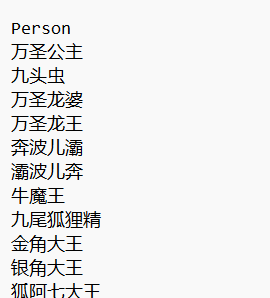
3.使用pytorch进行关系抽取



数据处理阶段获取的数据有《西游记》中的人物列表、人物之间的关系三元组、人物的属性信息、人物出现的词频。（文件在data文件夹中）

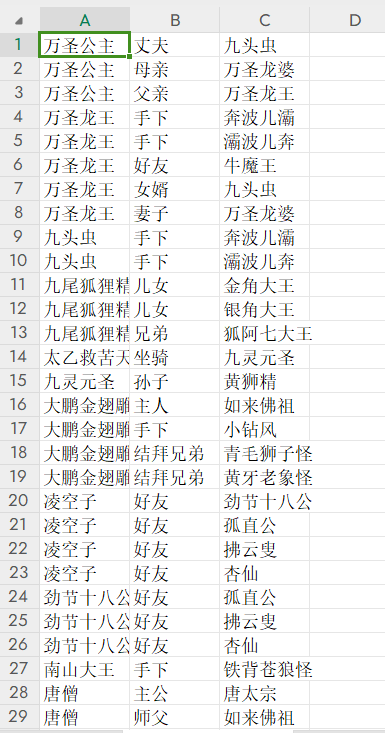


人物列表：

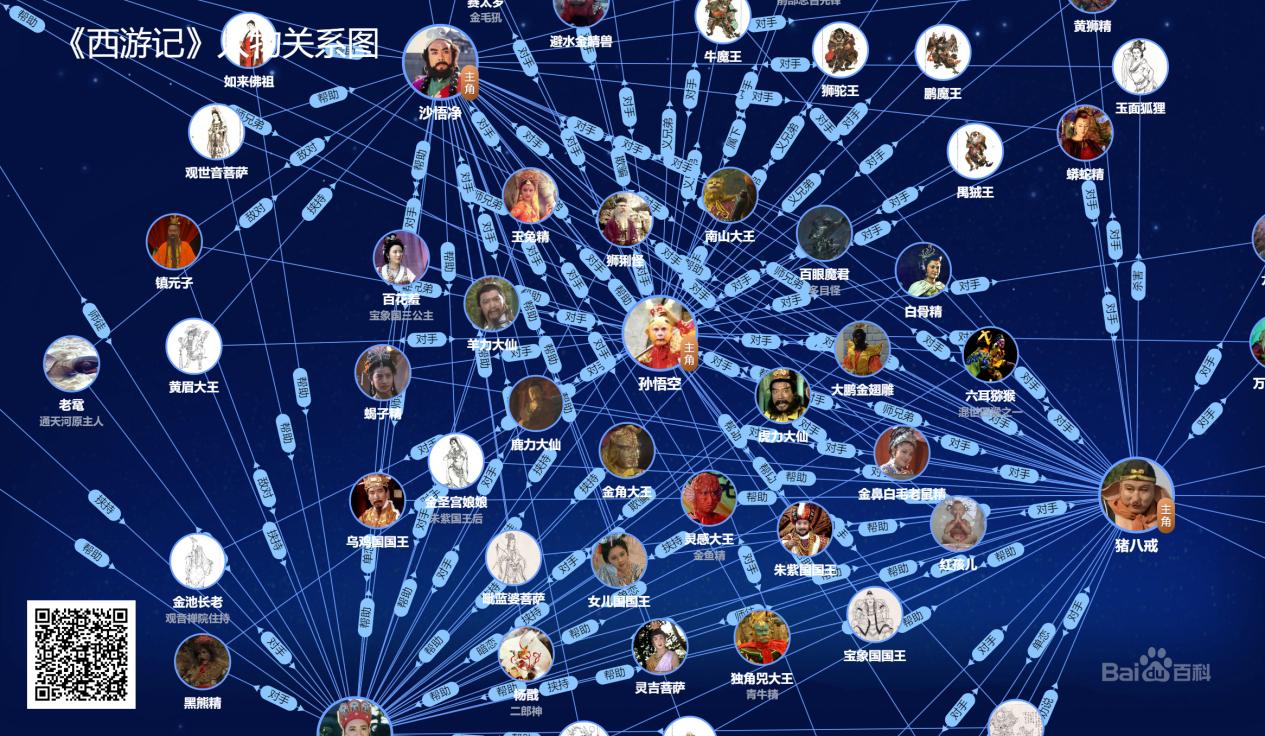


获取方式：通过spacy进行命名实体识别获取

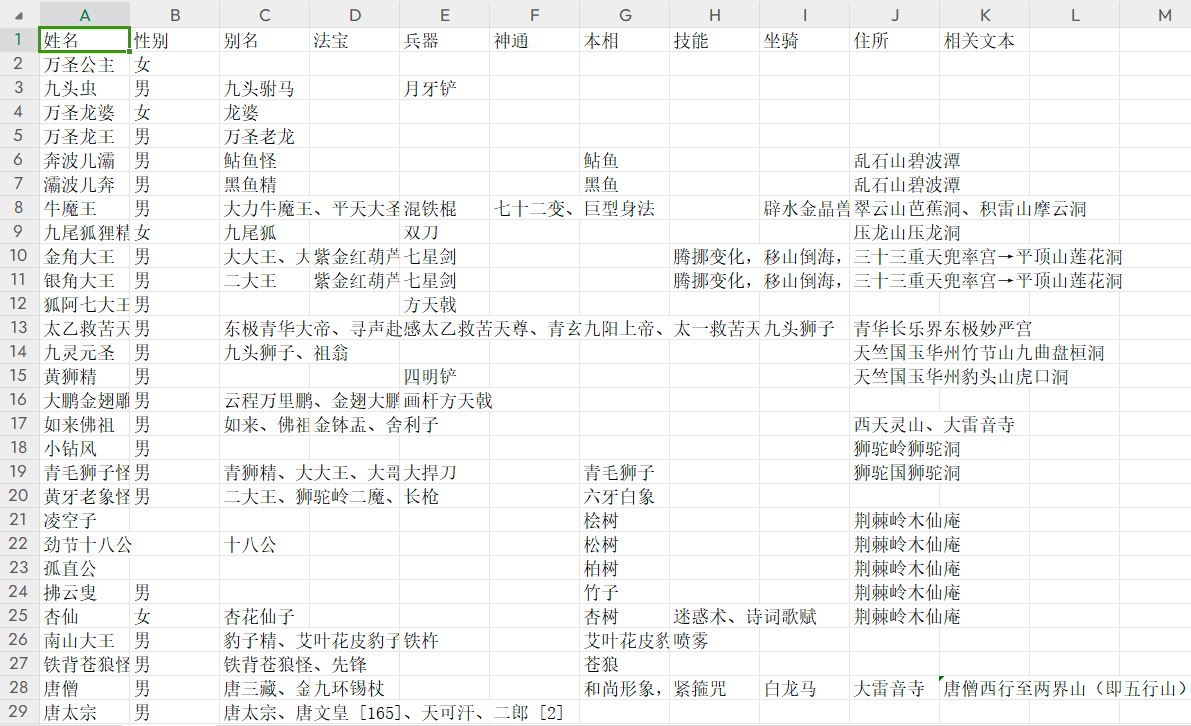
人物之间的关系三元组：



获取方式：参考百度百科西游记人物关系图进行构建（参考资料如下图）



人物的属性信息：

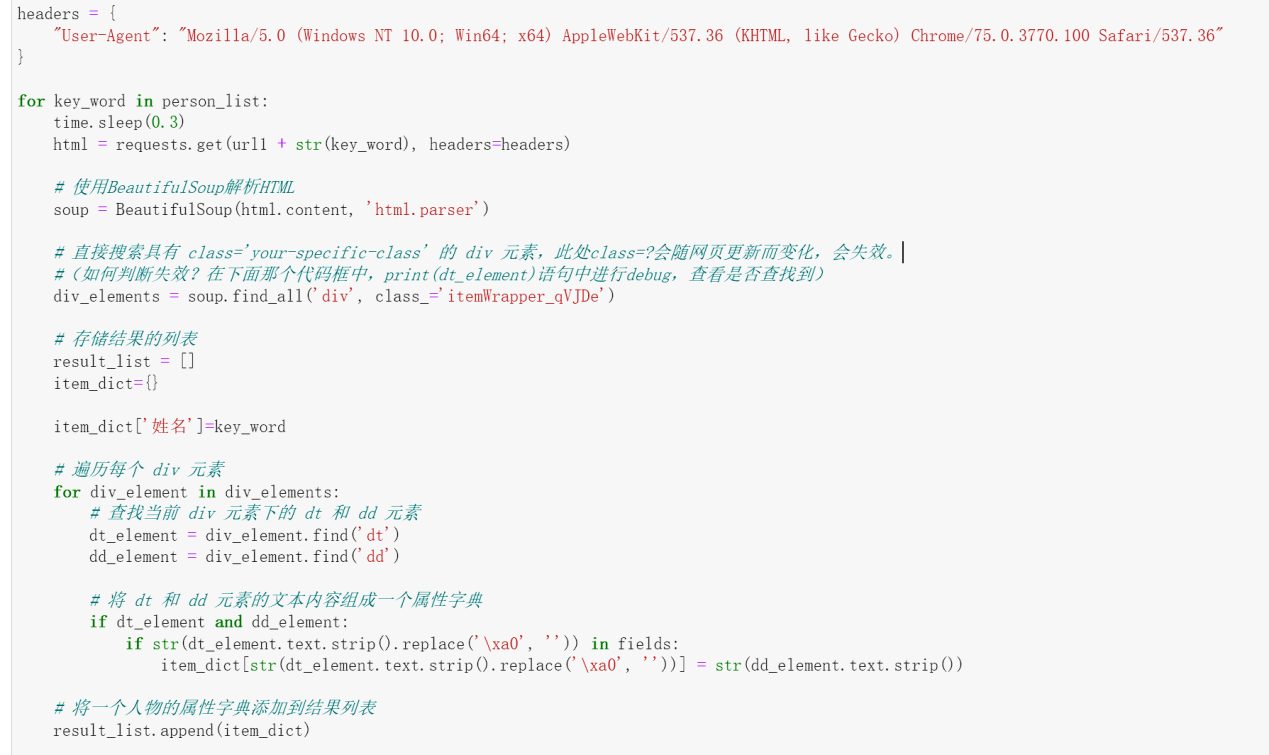


获取方式：通过爬虫爬取百度百科相关资料进行获取（爬取的数据为下图部分）



代码：（data processing.ipynb）





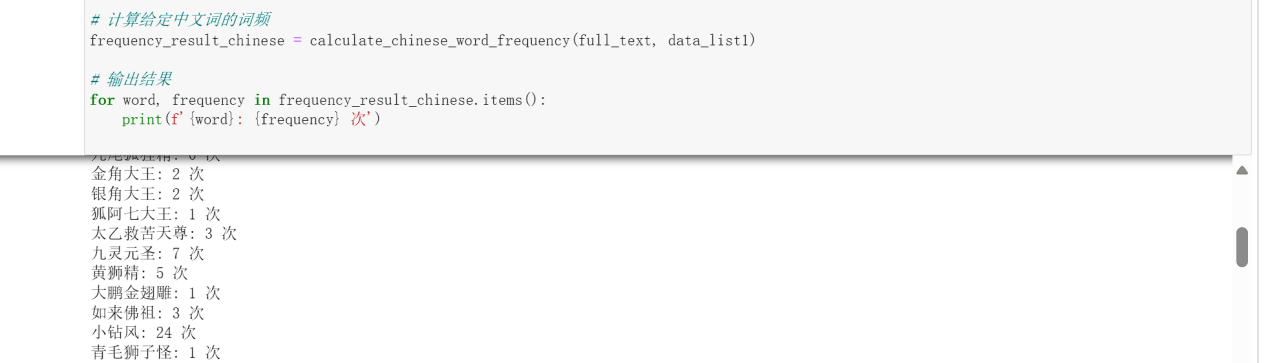


人物出现的词频：



获取方式：通过自然语言处理，依次统计人物列表中的人物在《西游记》中出现的次数

代码：（NLP.ipynb）



步骤二：读取生成的文本

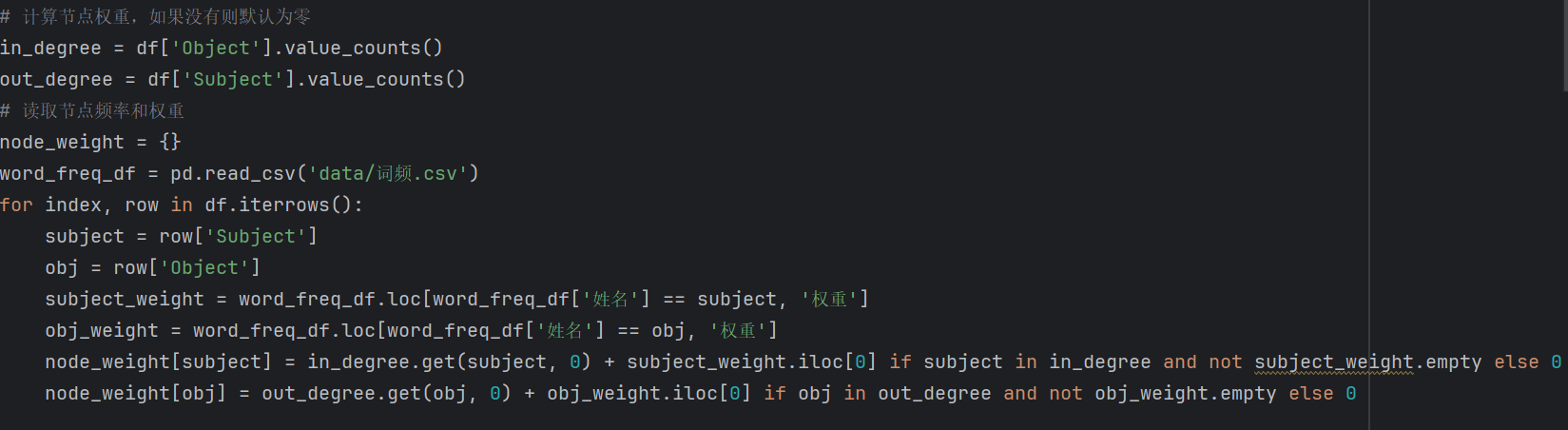
通过读取三元数组，获取每个人物的‘subject’，‘relation’，‘object’，即每个人物和与他相关的人物。

通过读取每个人物在原文中的相关文本，根据人物属性，构建人物名片，包含'性别、别名、法宝、兵器、神通、本相、技能、坐骑、住所、以及相关原文等信息。

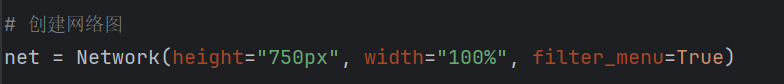


步骤三：修改节点属性

节点大小：通过读取词频.csv，获取每个人物在原文中出现的次数，人物节点的大小根据其在文本中的出现次数来确定。



步骤四：创建网络图



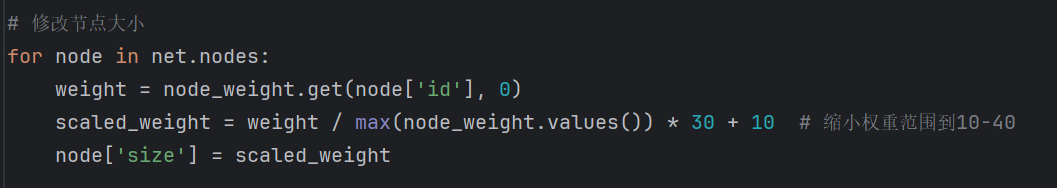
步骤五：添加节点和边

节点和边的构建：遍历Pandas数据框（DataFrame）中的每一行，并返回每一行的索引和数据，根据当前行的 'Subject' 列值，在 profile\_data 数据框中筛选出对应的个人资料信息，并将结果保存在 profile 变量中，然后添加节点，标签，以及其他相关信息，构建有向的人物关系图谱。



步骤六：修改节点大小

遍历网络中的每一个节点，根据节点的权重值，计算并设置节点的大小属性，以便在后续的数据可视化中使用。

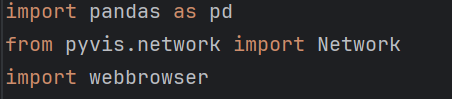


步骤七：创建控制面板并打开生成的HTML文件

将网络对象进行可视化，并将可视化图形保存为HTML文件，并使用浏览器打开指定的URL或本地文件



代码所用到的包或模块：



pandas用于数据处理和分析。

从pyvis库中导入Network类，用于创建交互式网络可视化图形和操作网络图。

webbrowser模块用于控制和自动化浏览器的操作。

**设计中遇到的问题及解决方案：**

1. 问题：在使用NetworkX+Matplotlib方法可视化时出现了点击没有高亮，或者高亮后二次点击任意处无法取消高亮，可视化呈现的图形过于凌乱拥挤，不够美观的问题。

解决：尝试了多个不同的宏包，包括NetworkX+Matplotlib、Graph-tool，iGraph，Plotly，Pyvis，Neo4j图数据库，最终选择了能较好满足题目要求，可视化美观，GUI代码简单的Pyvis包。

1. 问题：自然语言进行关系抽取难以实现怎么办？

解决：参考网络上的资料进行人工记录。

1. 问题：自然语言进行属性获取难以实现怎么办？

解决：使用爬虫获取百度百科中的属性数据，再选择需要的特定数据。

**总结：**

1. 通过完成《西游记》人物图谱设计项目，我们小组学习到了自然语言处理的常用方法和一些相关包及模块的使用方法。
2. 学习到了GUI设计的一些方法和相关设计代码，比较了不同宏包中所呈现的不同效果，拓宽了知识面。
3. 掌握了利用爬虫在互联网上爬取所需内容的方法。
4. 接触了不同类型的文件，以及了解到了它们的具体用法。
5. 锻炼了小组成员之间的协作能力。
6. 提高了查阅资料的能力和自主学习能力。