高性能计算课程报告

组员：

焦杰红 张萌 何毅晨 刘东 李宇宸

工作过程：

- 1.数据的收集和采集
- 高性能课程上我们总结了中国人才数据，在第一次课和第二次的课程我们收集数据库，首先这样的数据并不是现成的数据集，我们需要从不同的页面不同的表格去寻找数据，我们的数据集寻找方向是2016年的青年学者，每个人有120个人物情况的工作量。我们的寻找方向是网页寻找，很快我们找到了全部学者的名单，但是老师还是需要我们找到对应人的工作地点和工作时间以及每个人的学历所在地和获得学历等等，这样的信息就是复杂并且难以找到的，我们小组经过讨论决定经过不同的方法进行寻找，首先就是利用我们的优势，应用软件进行查找。

数据的收集和采集

- （1）python设计的网页爬虫程序
- 爬虫已经广泛应用到今天的互联网当中了，我们找到相应的网站，看到连接，然后应用爬虫程序下载表格这样大约600个人的信息我们就获取到了，然后偶们继续修改程序，找到每个人的相关属性，添加到表格当中，但是这样的信息并不是十分丰富，通知内容也不能够直接找到，而且很多的信息在网上是不能够找到的，所以我们在完成这样的工作后基本完成了数据库25%的查找工作。

数据的收集和采集

- （2）手动查找
- 对于找不到的部分内容，我们按照分工，把不能够找到的信息直接用手动搜索这样的信息在google和百度，找到部分内容，因为在信息库当中存在要寻找的关于获得者本科硕士博士的所在院校和对应时间的信息，但是这样的信息往往又不是很好找，简单的爬虫程序对于这样的深度检索体现的无能为力，我们进行手动检索，每个人分配为100人的工作量，完善表格，这样的寻找也很难完成所有人的对应的内容，完成后基本数据在50-60%这样的完成度上。

数据的收集和采集

- （3）完善数据
- 完成之后，如果我们要进行数据的可视化和分析这样的空余数据很难完成这样的工作，我们要对数据进行相关的处理，之后才能够只用，我们发现缺失严重的数据来自于最后的几部分数据，分析后我们发现对应于青年学者，他们在获得这个称号的年份，基本上学历相仿，我们可以采用类推得出没有详细数据的就读学历和年份，而学校往往对于他们的数据在最后的博士学校和相关的科研单位是有关系的，研究的专业也可以为我们分析一二，但是回顾数据集，我们可能要展示的重点是获得者的出生地和获得称号时候所在单位的重要性。

数据的收集和采集

- 对于缺失的数据往往又是不是绝对重要的因素，所以我们针对确实的数据进行了补充，当热这样的数据不会是正确的所以我们分成两个数据集，在可视化上如果数据集不完整，这样会难以识别，我们选择补充失真的数据集和对于不完整度很大的数据进行掩盖，无法显示他的可视化，同时两外的做法是只显示这一项上面数据完整的人的信息，这样集中可视化显示的结果是不一样的，尤其是对于我们收集到的并不是完全完整的数据集来说，而这样的结果我们呢也会继续的思考同时去想他们的解决方案。

克服数据收集的难度

- 2.克服数据收集的难度
- 在上面的构成中已经基本体现我们的处理方法就是根据不同的数据内容进行不同的收集方法，在应用大面积广泛搜索和细致范围的搜索。在手机过程中我们会采用手动弥补自动填充数据的不足，弥补数据的缺失，但是收据数据还是很难的过程，尤其是在广泛的查找还是没有收获，确实是信息确实严重。

增加数据维度

- 3.增加数据维度
- 数据的内容很多，那么我们对数据处理就需要技巧，每一条数据都包含了时间和空间的不同的数据，并不是在同一个空间同一个时间的数据，我们处理它们必须统一时间和空间，才能够继续处理，老师在上课的时候讲授了处理这样数据的方式和方法，我们拿出一组数据，对于不同人的相同组别是相同的内容含义，例如籍贯，这样的数据是没有时间属性的，因此不通用考虑时间不同，我们对照这个地点，找到对应的经纬度，这样就可以在地图上标标识出来，相似，对于后面的有空间信息的数据我们都可以这样进行处理，对于时间维度，我们进行时序分析，这样就可以得到相应的数据，同时可以在地图上显示出来。

数据处理

- 4.构架特征工程
- 在这样的数据当中我们构建特征，特征是指数据的处理提取，找到每个数据的典型特征有助于我们进行更好的分析和理解。
- 5数据分析方法：
- 采用上课老师说的方法，对于同一纬度，不同时间，相同时间不同维度进行数据的处理，数据会不同的表示方法，不同的处理。

Demo展示

人才所在区域的分布图