**基于社交网络的爬虫与数据分析**

**一、 背景及原理介绍**

**1.1 实验背景**

知乎（www.zhihu.com）是一个中文问答社区，通过使用机器学习技术，挖掘平台上海量数据的价值，帮助构建优质内容生产和消费闭环，辅助进行更高效的平台运营。于2011年1月正式上线，以“让人们更好地分享知识、经验和见解，找到自己的解答”为品牌使命。其社区式的管理环境，带来了很强的社交网络属性，通过对其社群用户的数据特点进行爬取、分析，可以获得有关中文社区网络结构、用户关联以及社区内容检测的粗略一瞥。

因此，本次实验中将以爬虫的方式，通过设定一些种子用户——“知乎大v”作为锚点，对其社区中的用户关联尝试进行刻画，并获取、分析用户的基本信息（如：id、粉丝数、城市、关注数量、文章数量等），通过一些简单web应用实现可视化界面以进行筛选、展示，并尝试通过NLP处理模型或调用GPT4、kimi等大语言模型API，实现社区检测等有意义内容，如：热点话题、作者归类等。

**1.2 爬虫技术**

获取网络内容有两种方式，其一是通过正常浏览网页获取完整的HTML/SHTML数据，再进行网页解析和异步请求分析来获取所需要的数据；第二种是通过API（Application Programming Interface）的方式获取数据。前者的优点是可以避开平台权限限制，但是当前新型前端架构导致数据较难获取，且解析网页也较为复杂；后者可以帮助我们精确获取无冗余的高质量数据。网络爬虫即为通过上述两种技术手段，模拟浏览器发送网络请求、接收请求响应，自动地抓取互联网信息的程序。

本次实验中利用python实现网络爬虫，通过调用requests库发出请求、处理并接收响应以抓取知乎信息，并通过广度优先的算法递归实现对于用户关系的刻画；此外，利用pandas库与numpy库实现数据清洗处理，为后续分析、展示做准备。

**1.3 简单Web应用**

在对互联网爬虫爬取到的数据进行清洗处理后，还需要根据标签、特点进行筛选与分类，方便对数据进行可视化分析与处理。因此，我们利用简单的Web应用，将爬取到的用户数据、社群特点集成到网页前端，利用networkx库可实现网络结构的分析；通过pywebio、pyecharts等库实现简单的Web应用构建；这些库大大简化了编写Web网页所需的HTML文件、CSS文件和JavaScript脚本的过程，将这些内容以及浏览器与HTTP服务器交互的逻辑转化为python代码的编写，为快速构建对UI要求不高的Web应用提供了极大便利。这些库主要实现简单的用户表格上传与页面图片展示，并可通过jieba等库实现中文分词，为后续社群检测做简单准备与尝试。

**1.4 社群检测与分析处理（选做）**

在获取了网络内容、进行分类筛选处理后，通过自然语言处理模型（NLP）或调用现有大模型GPT4、kimi等的api接口，可实现进一步的数据解构分析与归类，分析文本情感、内容归类以及内容总结等，也可对用户进行划分归类，实现更进一步的社群检测与数据归类总结。

**二、方案设计**

本次实验主要基于python实现，用到如下python库：

* requests（构建HTTP请求并接收响应）
* pandas、numpy库（实现数据清洗处理）
* networkx库（对网络结构进行刻画分析）
* pywebio、pyecharts库（构建简单的Web应用，实现数据的筛选与可视化）

（选做）此外，尝试接触NLP模型，对GPT4、kimi等现有大语言模型api接口进行一定了解，尝试将其融入社区检测、内容分析中，进一步对数据进行处理与刻画。

具体的实现步骤如下：

首先，根据所需要的数据编写python代码，搭建HTTP请求客户端，将所需数据按照一定格式爬取为csv文件并储存，爬取过程中需要注意随机sleep一段时间，防止由于反爬频率过高导致封号；

第二步，根据爬取到的csv文件以及数据存储的格式，对数据进行清洗处理，根据一定标签、特征进行分类与统计；

第三步，构建web应用，将第二步中处理完毕的数据以可视化方式呈现在web应用中，并通过pywebio库实现可选表格对数据进行筛选；

（选做）最后，通过外接NLP或现有大预言模型，实现对筛选的内容进行进一步归类、分析。

**三、分析及验证方案**

实验完成后，应得到能够根据筛选给出相应数据的可视化界面的web应用以及爬取到的csv文件，通过对csv文件的简单观察、结合互联网检索即可判断爬取到的信息是否正确，若随机抽取的数条csv文件内容与知乎社区内检索情况符合良好，即可认为第一第二步数据的爬取与处理基本实现。

随后通过对Web应用的调试使用，结合csv文件原始数据与知乎社区内容检索，可以判断第三步是否成功搭建Web应用，并实际得出Web应用的完成情况与可用度。

（选做）最后，通过对原内容分析与由外挂模型分类、总结或分析的结果对比，即可简单得出社区检测是否成功。

需要注意的是，以上几种分析验证方案均需随机抽取数条数据进行检测，才可得出一定结果，对试验成功与否得出结论。

**四、实验分工及时间安排**

首先共同完成HTTP请求客户端的搭建、数据处理代码的编写。（周五完成）

分别对不同种子用户进行爬取。（周六、日）

然后分别对不同数据进行不同的可视化处理构建，详细分工后续讨论得出（周一、二）

对Web应用进行一定优化，并尝试外接大语言模型或NLP。（选做）