**分析京师大福问答数据**

1. 背景方向

针对北京师范大学出现的虚假搜索框现象，我们决定利用动态网络爬虫技术爬取京师大福所有问答数据。通过爬取这些数据，我们可以更好地了解该网站的问答情况。

在进行数据分析时，我们首先对爬取到的数据进行了清洗和整理。这一步骤包括去除重复数据、纠正错误数据、以及填充缺失数据。接下来，我们对数据进行了详细的分析，包括统计分析、可视化分析等。在统计分析中，我们计算了问答数据的数量、平均值、标准差等统计量，以了解数据的分布情况。相关性分析帮助我们探究了不同问答之间是否存在关联性，从而为虚假搜索框现象的成因提供线索。聚类分析则帮助我们识别出具有相似特征的问答，进一步揭示虚假搜索框现象的特点。

此外，我们还利用机器学习算法对问答数据进行了分类。通过训练模型，我们可以将虚假搜索框问答与正常问答区分开来。这一步骤有助于我们深入了解虚假搜索框问答的特征，为后续的整治工作提供支持。

综上所述，动态网络爬虫技术为我们提供了大量京师大福问答数据，通过深入的数据分析，我们揭示了虚假搜索框现象的规律性，为整治工作提供了有力支持。在未来的工作中，我们将继续优化爬虫和数据分析方法，以期更好地应对类似问题。同时，我们也呼吁相关部门加强监管，共同营造一个健康的网络环境。

1. 实现过程：关键步骤及创新实现

技术实现细节：

1. 断点续传功能：

此功能确保即使遇到断电、断网或封IP等意外情况，也不会导致工作前功尽弃。这对于代码的后续功能扩展至关重要，允许我们设定定期爬取间隔，以避免IP被封。

2. 保存首次登录信息：

由于统一身份认证平台的登录状态持续时间短，我们选择不保存cookies，而是加密保存上次使用的用户名和密码，以简化登录流程。

3. CLI美化：

使用theme模块为控制台输出添加分类着色，提升了命令行界面的美观度。

4. 代码结构：

通过Git进行版本管理，并将代码托管在Github上，以促进团队协作。代码目录结构清晰，多文件组织降低了代码耦合性，便于后续模块的复用和功能的添加。使用批处理运行Python代码，提高了运行效率。

5. 异常处理：

对统一身份认证平台登录密码错误、账号不存在等情况进行了异常处理。实测发现，BNU统一身份认证平台的登录具有一些特性，例如账号密码为空时不会产生错误信息，输入错误的用户名密码后，请求的服务会发生改变。断点续传功能会特别判断上次是否已爬取完成，以避免覆盖已爬取的数据。此外，使用UTF-8编码CSV，以处理京师大福上可能存在的GBK编码内容。

四．总结

在进行数据统计时，我们发现京师大福平台上的数据存在一些局限性，这使得统计结果的准确性受到了影响。

首先，京师大福平台上的问答数据只包含了相关部门已经回应的问题，而没有涵盖同学们所有的发帖内容。这种数据采集方式使得统计结果变得片面，无法全面反映同学们的真实需求和关注点。因此，单纯对这些数据进行统计分析，可能会得出与实际情况不符的结论。

其次，数据的时间戳存在明显的问题。从第17页到第51页的问答数据，时间戳都集中在2020年2月一个月内。这表明这些数据可能是统一放上去的，而不是实时更新的。这种情况下，数据的时间准确性受到了质疑，使得统计分析的结果失去了时效性。

此外，访问量的数据也存在问题。由于爬虫在抓取数据时会对访问量产生影响，因此我们无法准确获取实际的访问量数据。这使得我们无法根据访问量来评估问题的热度，进而影响了数据分析的准确性。

基于以上问题，我们认为爬取到的499条数据并不适合直接进行统计分析。相反，将这些数据保存为一个数据库，以便于后续的查询和分析，可能更为合适。

于是，我们将数据集接入了通义千问API，使用RAG实现了一个可以对话的京师大福智能体。使用京师大福AI我们可以方便地用自然语言去获取需要的信息。

