初步需求调查分析

1. **背景介绍**

知乎是中文互联网高质量的问答社区和创作者聚集的原创内容平台。知乎以问答业务为基础，经过近十年的发展，已经承载为综合性内容平台。

知乎也面临着一些问题，一旦有人从中灌水，就会引发一连串的连锁反应，这就是[破窗效应](https://baike.baidu.com/item/%E7%A0%B4%E7%AA%97%E6%95%88%E5%BA%94" \t "https://baike.baidu.com/item/%E7%9F%A5%E4%B9%8E/_blank)。知乎也意识到这种威胁，一方面继续采用严格的邀请制度，另一方面采取了许多措施来抑制破窗效应。比如对提问类型进行了限制，允许用户举报重复低质量的问题和隐藏“没有帮助”的答案等。但是，破窗效应的威胁依然存在，随着不断开放和用户数增长，如何保持回答的质量和社区的氛围依旧是不小的威胁。

1. **欲解决的问题**

解决数据质量驳杂冗余的问题

解决有效数据的获取困难问题

1. **软件功能描述**

对数据进行爬取并分析，并进行可视化分类，定期更新保证信息的有效性

通过算法分析数据的真实性和有效性，并进行分类排序和标注

提供更便捷的搜索和下载功能

1. **系统的组成与部署**

初步设计为爬虫、后端、前端三部分

可选用：python+scrapy+redis+mongo

数据采集工具：分布式 python 爬虫

分析工具：ElasticSearch + Kibana