# “知乎数据爬虫”软件系统的需求构思及描述

## 背景介绍

随着互联网的普及和发展，现在网民数量已经达到了9.89亿人，其中相当一部分人对知识是十分渴求的，所以知乎这类问答平台变得火爆起来了。知乎作为一个问答软件，许多人将其当做阅读、学习、和业内人士沟通的工具。知乎的月活跃用户高达9430万。虽然知乎的用途广、用户多，但是缺点同样明显。

## 欲解决问题

知乎的缺点如下：缺少垃圾回答屏蔽系统；答主排序、检索系统；信息搜集、可视化系统。

在知乎查资料的时候经常会遇到垃圾回答，比如广告、引战、辱骂、搞笑段子。这些问题可以通过数据清洗来去除，因为垃圾信息往往同质化严重，当重复度高于某一数值时，就可以列为垃圾回答。

答主排序检索系统是可以将所有答主的人气信息搜集起来的系统，人气包括：赞数、评论数、粉丝数、发表数、收藏数、转发数、引用数。通过这些数据综合排序就可以或者相对高质量的作者。

现在知乎想要知道某板块的热度或者某些热词的出现频率，只能通过购买数据搜集网站的服务来统计热度。而我们可以解决它，用项目来给用户带来便利。

## 软件创意

在爬虫部分，我们使用了requests，json，parsel，re，threading，time，random，pyaml，proxy\_pool，os，pandas，matplotlib，queue，contextlib框架，后端主要两个大框架：mybatis，springboot，前端：vue，element，echarts，检索库:elasticsearch,通过这些部分的合作，我组才能实现如此庞大的项目。我们的软件创意点在可视化数据，通过云图、动态柱状图等多种形式的可视化，增加了该软件的趣味性和直观性。

## 可行性及潜在风险

该项目可行性是非常高的，第一，有前人完成的例子可以进行学习，各网站也有足够多的相关教程。上述框架已经有相当大量的应用于此方面的先例，在兼容性方面是不用担心的。组员分工明确，各个组员工作量适中，不用担心猝死，而且可以高效率工作。

潜在风险：1.隐私泄露。我们的项目会将数以百万计的答主的数据收集来，数据量会达到TB级别，这些数据一旦泄露，我组按照法律会被罚款。信息是有归属权的，我们大量爬取信息的行为有可能面临知乎公司的起诉。2.性能骚扰。过于大量的信息涌入校园网服务器会导致网络拥挤、瘫痪。造成其他同学的使用不便。3.封号风险。知乎网站会在用户登录后对cookie的调用进行检测，如果调用次数过于频繁，账号有可能会被封禁。