**定制爬虫技术文档**

**说明：**

**定制爬虫：**某些具有特殊链接结构和反常规dom模型（主要是为了反制爬虫）的网站，必须人工有针对性设置参数，调整参数，特殊编程才能爬取，就像给某些体型和嗜好特殊的人必须“定制”西服一样，而这个爬取过程一般也称之为“暗网”爬取。针对这一类网站的特殊定制爬虫策略及软件包，统称为“定制爬虫”。

# 主要功能

采集各大问答网站中有关癌症的问答数据。

## 功能价值

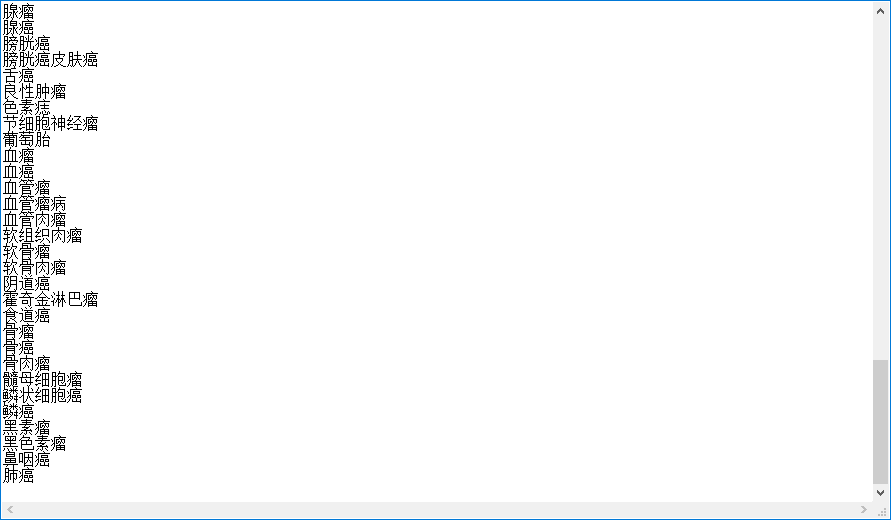
为整个癌症问答系统提供数据支持，为后续的数据分析、挖掘提供数据基础。

## 采集目标：

各大具有代表性的主流问答网站有关癌症的问答数据。

## 确定关键词：

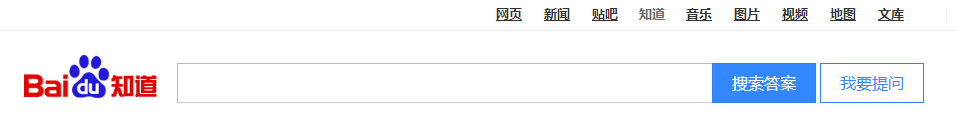
跟据各搜索引擎的对癌症的搜索结果，确定了116个热门的与癌症直接相关的名词作为下一步采集问答数据的搜索关键词。关键词示例如图：



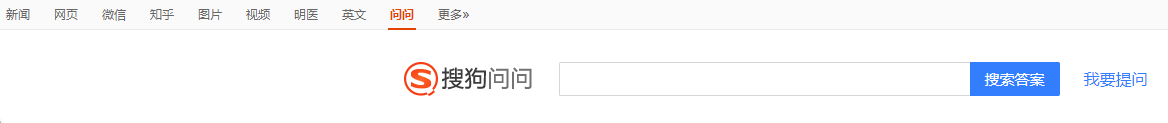
## 确定采集对象

采集对象为主要的问答网站。经过筛选最终确定采集网站为：

百度知道、360问答、搜狗问问等问答网站。







## 确定采集内容

采集内容为各问答网站中116个关键词的所有搜索结果，包括搜索关键词、问题、描述、回答。以百度知道为例

首先，搜索关键词，获取所有问题链接；



然后，进入问答详情页，采集相关字段；

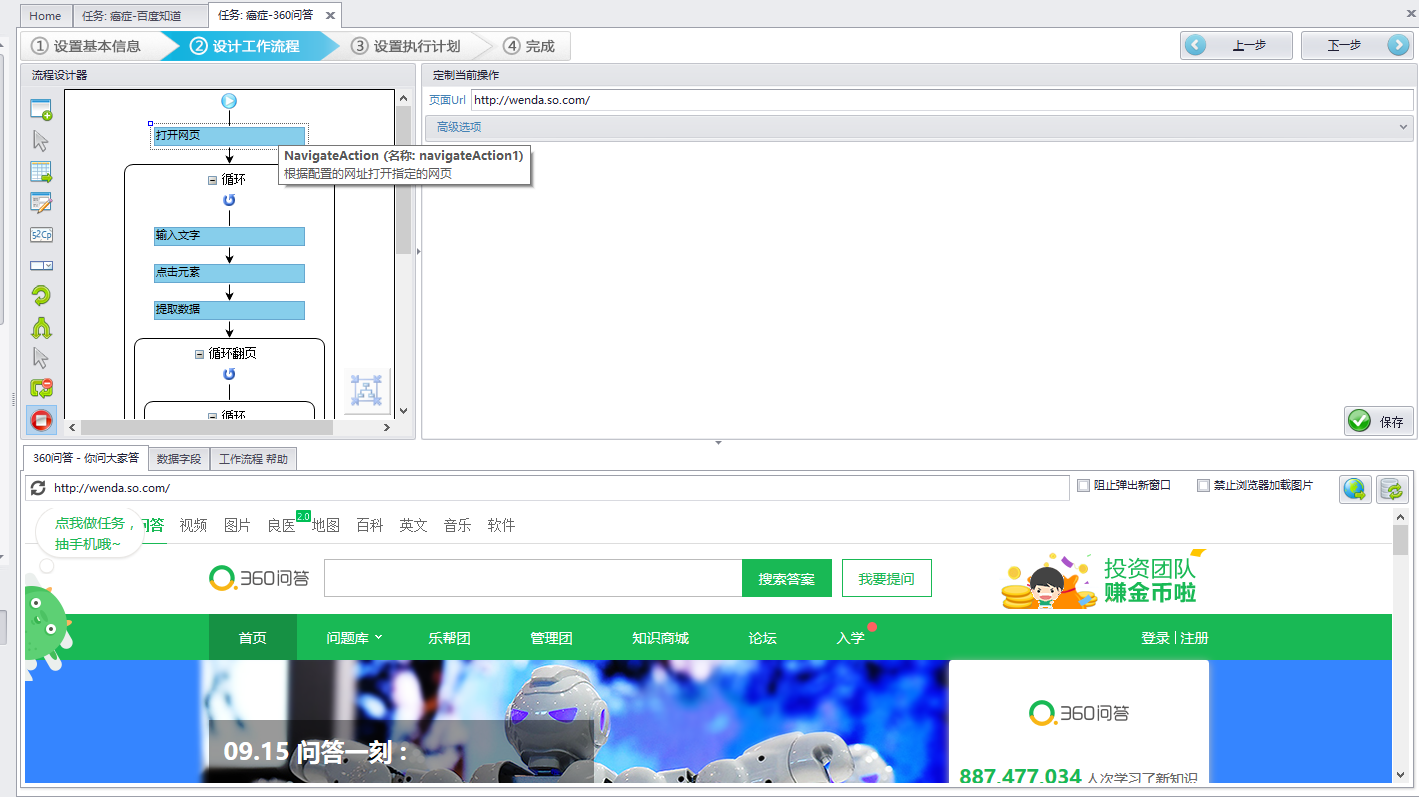




此处可以看出，并不是所有提问都有描述内容，所以在采集过程中需要注意“描述”字段可能为空

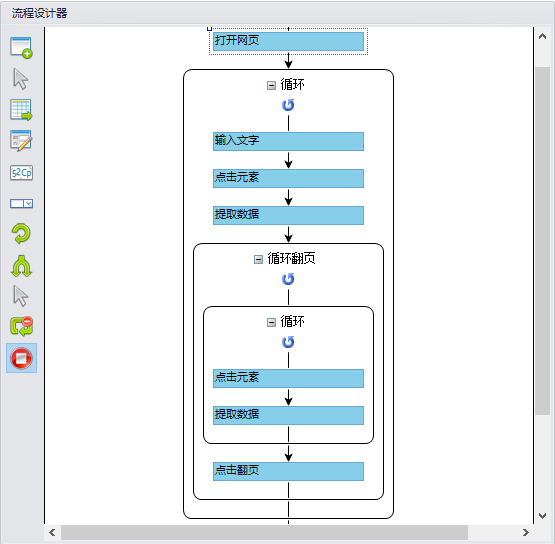
## 确定采集工具

基于采集的网站较多，对每个网站配置爬虫程序将花费较多时间的情况，选择使用八爪鱼采集软件进行数据采集。该工具模拟浏览器对网页进行捕捉，也方便对动态加载内容进行采集。软件爬虫任务设置界面如下：

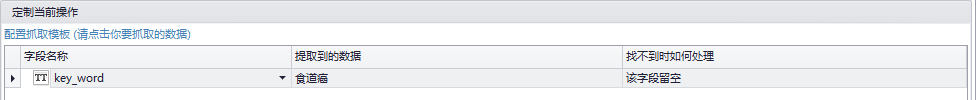


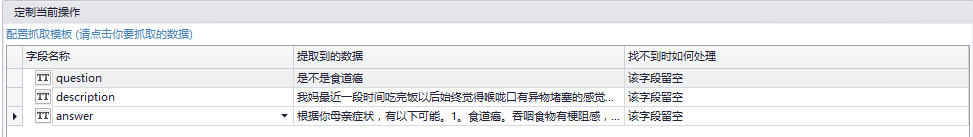
## 配置采集任务

采集任务的流程设计如图



采集的字段如图：





最终采集结果如图：

