# 1、任务概述

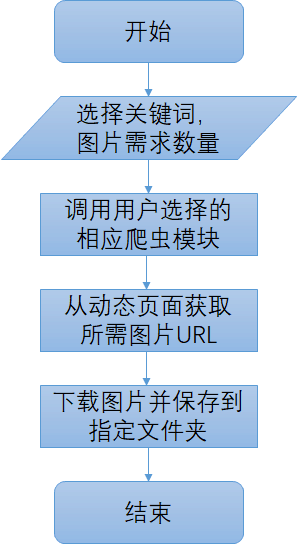
**1.1、目标**

软件系统的名称：网络爬虫软件

背景及目标：“爬虫”多用于从网页中爬取大量数据，为数据分析等操作提供了便利。但对于零编程基础的人来说，选择一款好用的爬虫软件，可以提高工作效率，达到事半功倍的效果。本软件实现功能就是如此。

本软件产品是一项独立的软件，而且全部内容自含。

**1.2、系统（或用户）的特点**



本软件面向不同领域、不同背景的广大群众，用户可以根据自己需要的数据集来使用本软件。本软件预期使用频度很高，或需配备素质过硬的软件维修人员，以便及时解决发生的问题。

**1.3、假定和约束**

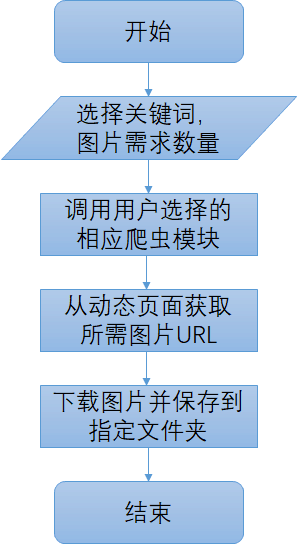
1. 开发期限为一个月。

2.模型数量受时间与人员数量限制，在管理方针、并行操作、安全与保密方面无约束。

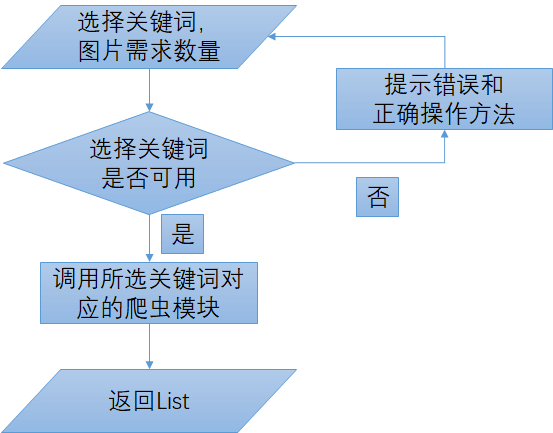
# 2、需求规定

**2.1、软件功能说明**

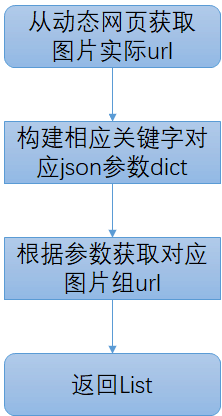
（1）爬虫主模块：



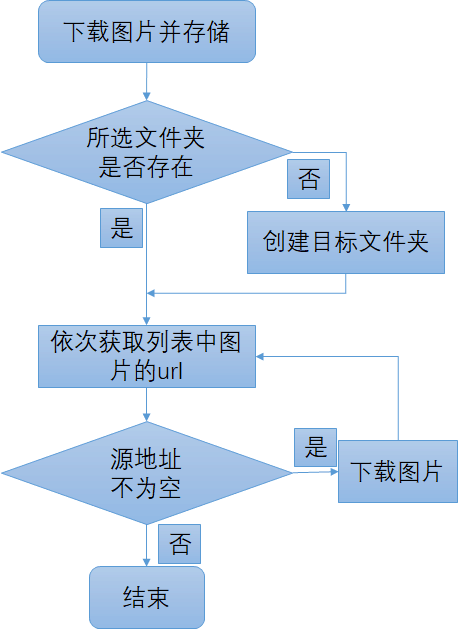
（2）选择调用模块：



（3）网页源码解析模块：



（4）图片下载解析模块：



（5）用户视觉模块：主页提供可选择爬取的关键词，后续可做成分类式，关键词不存在提示错误。输入下载图片数量，超出预定范围提示错误。

**2.2、对功能的一般性规定**

界面整体简单清晰易于操作，各输入模块都保证健壮性，对于不符合数据类型的输入进行模糊处理并进行提示，并提示用户如何正确使用。

**2.3、 用户界面**

用户界面使用Tkinter的grid网格来进行构建，界面图像简单清晰，通过点击选择需要爬取的图片，输入数量，按钮点击选择目录。点击进行爬取，操作简单。

**2.4、对性能的一般性规定**

**2.4.1、精度**

1.对于爬虫对象和数量要求保证精确，查询系统要求在对关键字查询的时候要精确及时。

2.在进行向数据库文件提取数据时，要求数据记录定位准确，在往数据库文件数组中添加数时，要求输入数准确。

**2.4.2、时间特性要求**

1.响应时间应在人的感觉和视觉事件范围内

2.更新处理时间：随着本系统的版本升级，系统将相应的进行更新

3.数据的转换和传送时间：1~2秒

**2.4.3、灵活性**

当需求发生某些变化时，本系统操作方式、数据结构、运行环境基本不会发生变化，变化只是将对应的数据库文件内的记录改变，或将过滤条件改变即可。

**2.4.4、输入输出要求**

输入方式是通过点选所需要爬取的数据库并输入所需的图片数量，因此输入的数据类型是整型。输出的是爬取的图片，图片主要格式有jpg,png等。本项目拟调用imageio库来讲图片处理成所需的特定格式，或者xml文件便于直接放入训练数据集。

**2.5、数据管理能力要求**

本项目进行的管理是，对于爬虫爬下来的图片，首先存到创建或者选择的文件夹中。要能实现把图片转换其他需要格式的功能，例如把图片转换成图片识别算法可用的RGB三通道xml文件，加入到Keras训练用数据集。

**2.6、故障处理要求**

故障处理要求能力要求很高，系统可以设置监控能力，一旦系统出现故障能能马上发出警告信号，并在必要时自动启动备用系统。后续在应用了数据库的情况下，可以设置数据库镜像，定时备份数据和日志，在出现严重故障时能及时恢复。

**2.7、其他专门要求**

当客户端需要下载大量图片时，可能出现图片占用空间过大的问题，于是可以进行图片压缩或磁盘管理，防止出现占用空间过大，剩余空间不足产生错误

# 3、运行环境规定

**3.1、设备**

本软件不需要特定的硬件或硬件接口进行支撑。

**3.2、支撑软件**

该软件可以支持操作系统：Windows，Linux，需要用到的编译程序有Spyder和PyCharm。

**3.3、接口**

客户端浏览器是搜索引擎和用户之间的接口。其首先获取用户搜索条件并加以分析，然后访问索引数据库进行匹配后获得检索结果，然后根据设定的处理后返回给用户。输出、输入的相对时间将由pc机本身处理速度来决定。对程序的维护，需进行必要的备份。

**3.4、控制**

本软件通过客户端浏览器，用户选择关键词及下载图片数量，从动态页面获取图片实际url，服务器接受信号执行操作响应。

# 4、尚需解决的问题

1. 构建的爬虫模型数量有限且有待优化；
2. 软件支持的客户端浏览器单一。