# ResWander需求分析

小组：BUG小组

## 项目说明（杨帆）

### 1.1 项目背景：

在日常生活中，我们经常会遇到一下情况：当我们浏览网站，遇到喜欢的图片，文字，视频等资源的时候，想要将他们保留在本地，以便后期欣赏。然而在实际操作中往往会遇到许多问题，难以将资源下载。

* 耗时长久：一般的下载操作需要依次逐个手动进行游览、选中、点击、保存、确认等机械重复的操作，当下载量巨大时十分繁琐耗时。
* 无法下载：大多数网站并没有提供直接的资源下载按钮，用户往往不知道如何下载。
* 限制下载：许多网站会对下载做出限制，如需要用户注册登录甚至成为会员才能正常下载，导致资源在眼前却无法下载的状况。
* 禁止下载：少数网站甚至禁止用户的常规下载操作，如禁止右键点击、禁止常规下载菜单的弹出等等。

这充分说明了用户对自由下载资源的需求。

### 1.2项目目标：

在这种选题背景下，我们小组计划基于C#语言、运用网络爬虫和html文本原理、结合文件管理和文件筛选过滤技术，制作一个带有基本图形界面的网络资源爬虫程序。

程序具有以下功能：通过网络爬虫技术，用户可以使用本产品输入对应网页的网址以及需要下载的图片的格式或者其他的约束限制（图片，视频的大小，主题等等），从而自动批量下载获取该网页上的对应资源文件，而且可以预览资源，查看爬取情况等等。

### 1.3软硬件环境需求

操作系统：Windows10

数据库： Oracle 11g

开发工具：Visual Studio 2019

开发语言：C#

### 1.4使用的关键技术：

关键技术：C#语言，文件管理和文件筛选过滤技术，html文本原理

难点与亮点分析后台：

1.不同的网站对图片资源的存储方式具有差异性，同一网站对图片资源的处理也可能具有差异性。在爬取资源时可能会涉及到多种爬取资源的策略。  
2.保证资源的完整性。  
3.网站的反爬虫机制，以及图片资源被加密等无法下载的情况  
4.会涉及到大量网页的访问以及数据的下载，我们需要保证爬取的性能。  
5.根据用户的需求，我们需要对爬取下来的资源进行分类

6.获取资源的详细信息，为用户显示这些详细信息  
7.下载图片前，为用户提供图片预览功能  
8.为用户显示当前图片的下载进度以及预估剩余时间  
9.计算资源的实际下载时间，下载结束之后显示实际下载时间  
10.资源下载失败时，先再次尝试下载，若多次下载仍然未成功，需要提示用户并显示失败原因，用户点击可再次尝试进行下载。

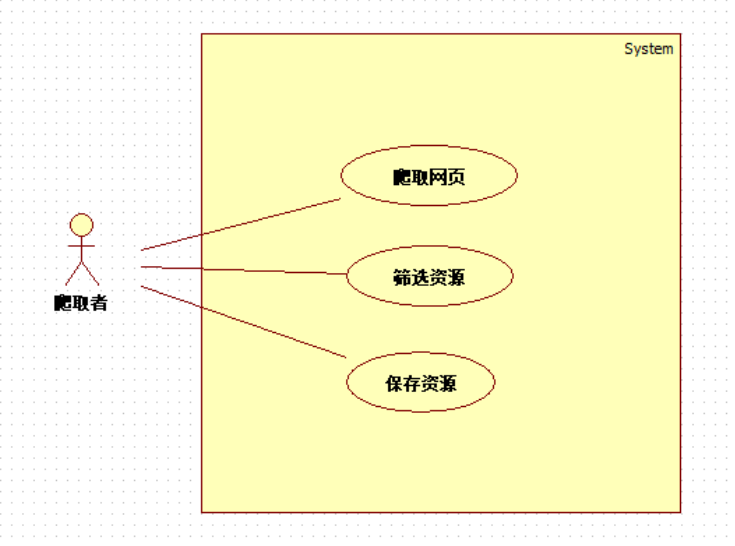
难点与亮点分析前端：

1.在原有系统中，窗口之间会有数据上的交互。  
2.在原有系统中，窗口之间会有行为上的交互。

3.如何使系统用所有窗口操作风格统一。  
4.当用户不按照窗口的打开顺序关闭窗口时，产生的数据混乱问题。  
5.控件分布合理与美感问题。  
6.界面美观问题。

## 需求分析（张衡，邱壮）

### 用例图：



#### 爬取网页

参与者：爬取者

基本事件流：爬取者输入网页地址，点击开始按钮，系统尝试爬取网页中的资源。如果出现网页不存在、网页无法访问或网络无连接等错误，系统显示出错误信息；如果爬取成功，则呈现出网页中资源的预览图。

#### 筛选资源

参与者：爬取者

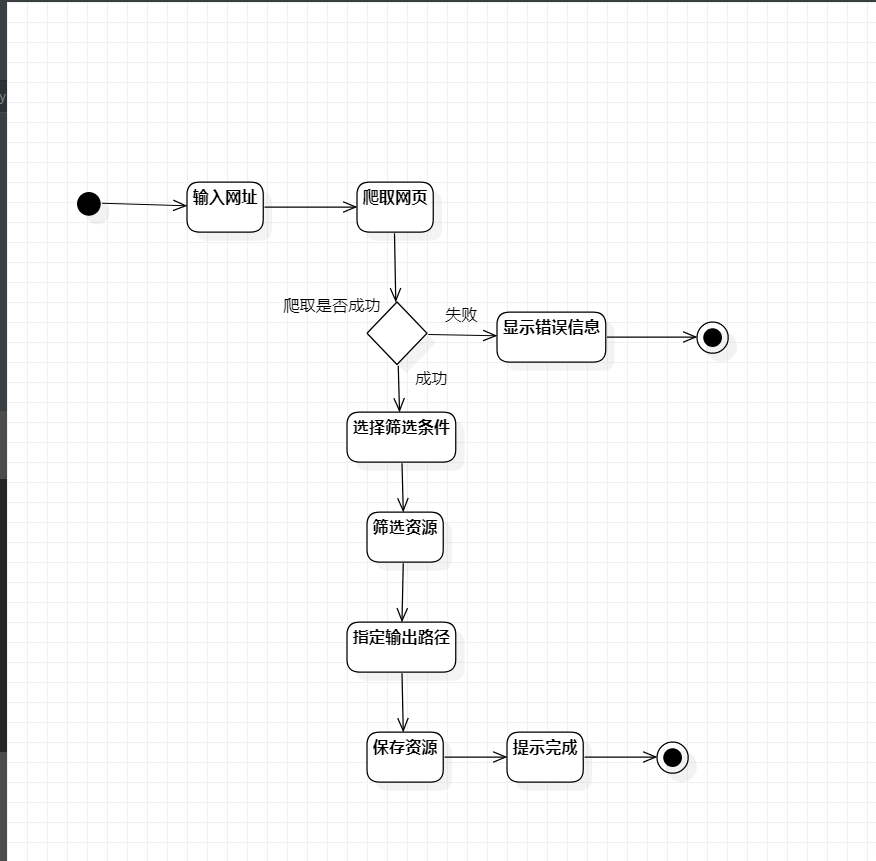
基本事件流：系统呈现出网页中爬取到的资源，用户改变筛选条件，系统呈现出经过筛选条件过滤后的资源。

#### 下载资源

参与者：爬取者

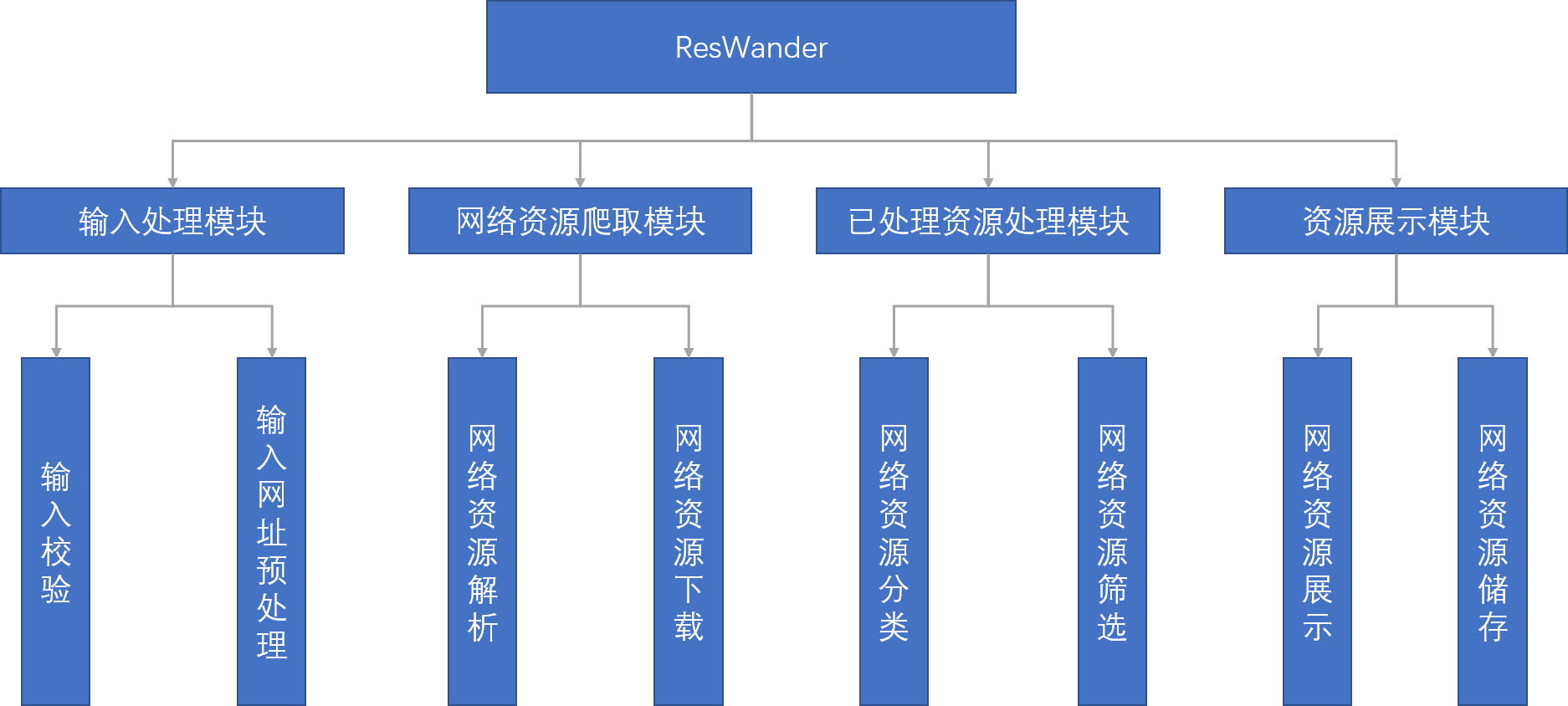
基本事件流：爬取者指定输出目录，爬取者点击保存按钮，系统将筛选后的资源保存在指定目录下，全部保存完毕后发出完成提示。

### 活动图：



## 3.概要设计（邢雄，杨翕然，李星鹏）

### 功能模块设计



#### 1.输入处理模块

输入：用户将爬取的网址、用户对爬取资源的要求(大小、类型)

输出：合法的规范化的网址，用户对爬取资源的要求

功能概述：提示用户输入待爬取的网页，以及对于爬取资源的要求，检验用户输入网址的合法性并作一定的加工处理，输入合法且规范化的网址，以及用户对爬取资源的要求。

##### 输入校验

输入：网站网址字符串

输出：合法的网址字符串

功能概述：根据用户输入，判断网址的字符串的合法性，若网址字符串合法，则输出网址输出，否则提示用户再次输入网址，直到输入合法。

##### 网址预处理

输入：合法的网址字符串

输出：规范化的网址字符串

功能概述：通过合法的网址字符串进行一定的处理，例如：加上：“http：//”等等，输出一个标准化、规范化的网址字符串，方便爬取函数的使用。

#### **2.网络资源爬取模块**

输入：网址，爬取的资源类型

输出：当前网站上指定类型的资源

功能概述：利用规范化的网址在特定网站上进行资源爬取，并所有指定类型的资源下载下来。

##### 网络资源解析

输入：网址、爬取的资源类型

输入：目标资源的URL

功能概述：利用规范化的网址输入，对网页的html代码进行解析，找到目标资源的url，并将其输出。

##### 网络资源下载

输入：目标资源url

输出：网络资源实体

功能概述：利用每个资源对应的url，将目标资源实体下载到本地。

#### 3.已爬取资源处理模块

输入：下载至本地的网络资源实体，用户对资源的要求

输出：分类清晰、满足用户需求的网络资源实体

功能概述：对原始的，刚下载完毕的资源进行分类整理，并根据根据用户的要求，对资源进行一定程度的筛选，最后将分类、加工完成的资源展示给用户

##### 网络资源分类

输入：网络资源实体

输出：不同类别的网络资源实体

功能概述：对输入的网络资源实体进加工，根据网络资源的属性(文件类型、文件大小等)对网络资源分类，并输入分类好的网络资源

##### 网络资源筛选

输入：网络资源实体，用户对资源的要求

输出：进行过筛选的网络资源实体

功能概述：根据输入的用户要求，对资源进行筛选，过滤掉低质资源（例如：文件大小过小）以及用户不需要的资源，将精简过的资源输出。

#### 4.资源展示模块

##### 网络资源展示

输入：网络资源

输出：图形化界面

功能概述：生成图形化界面，并将输入的网络资源展示在图形化界面上，与用户进行交互，供用户查看，并下载图片。

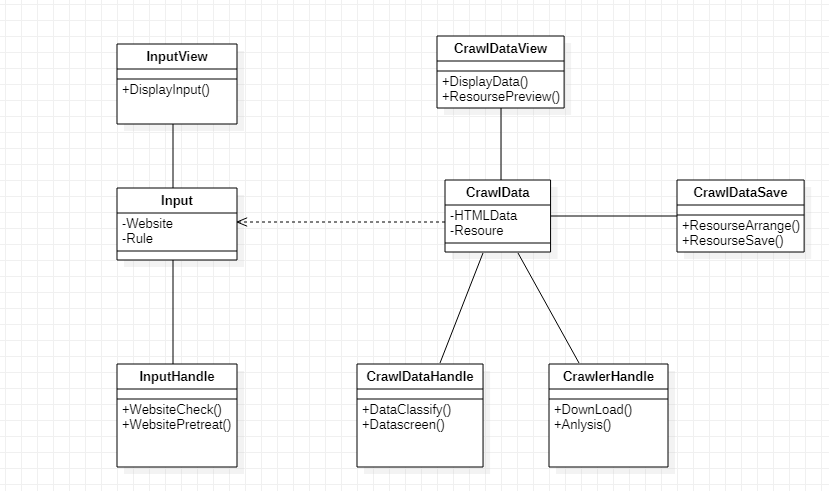
##### 网络资源储存

输入：待存储的网络资源实体，存储路径

输出：存储至本地的网络资源实体

功能概述：用户可以通过图形化界面选择需要进行存储的资源以及资源保存位置，系统可以将指定网络资源存放到操作系统的指定位置。

#### 5.核心类图



## 4.UI设计简图（邢雄）



