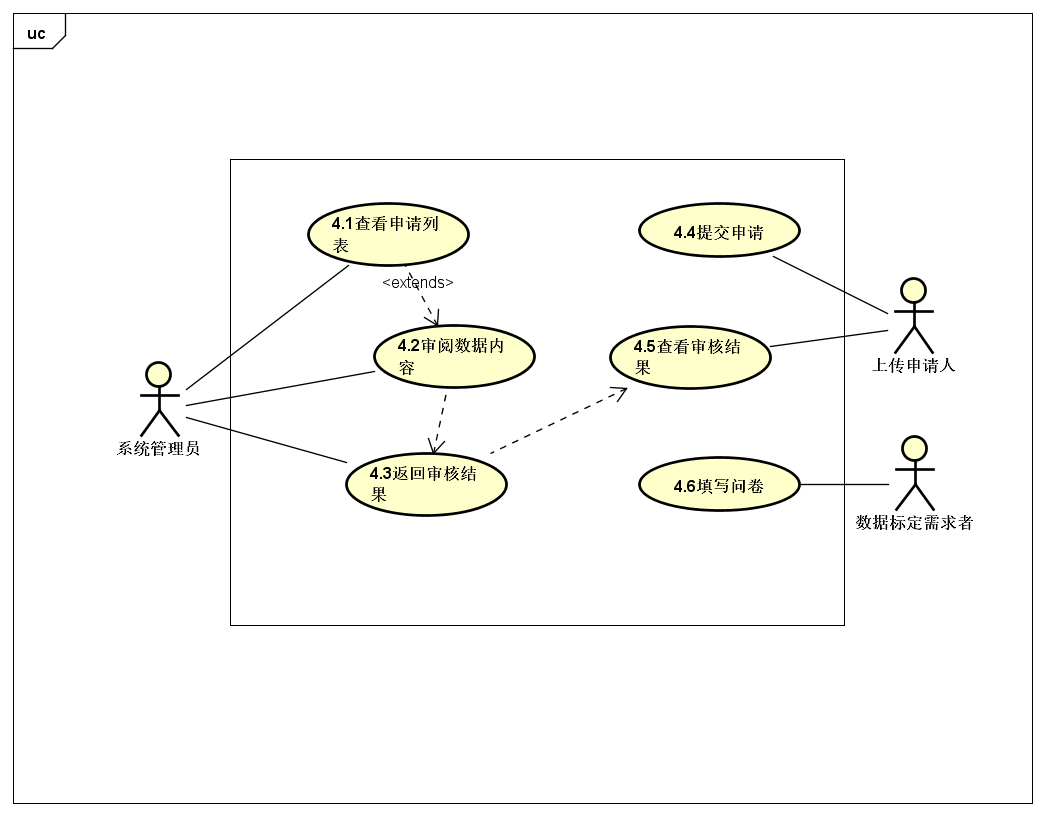
## 4.数据预处理

数据预处理主要有两个方面：1.上传预处理；2.标定完成后原始数据预处理。

上传预处理包含了数据分享用户上传数据的机器查重，人工审核，机器去冗余，数据有效化，和待标定数据的人工审核；标定完成后的原始数据的预处理主要是将标定完成的原始数据确定正确答案，计算准确率。

4.1 用例图



4.2 用例表

|  |
| --- |
| 用例：查看申请列表 |
| ID：4.1 |
| 简单描述：系统管理员查看数据上传申请列表 |
| 主参与者：系统管理员 |
| 副参与者：系统 |
| 前置条件：有未处理的上传申请 |
| 主流：   1. 系统管理员打开网页并以管理员身份登入 2. 系统管理员点击查看申请列表 3. 系统生成申请列表 4. 系统管理员查看系统生成的申请列表 |
| 后置条件：系统管理员完成申请列表的查看 |
| 附加流：无待处理申请时，显示给系统管理员：无待处理申请 |

|  |
| --- |
| 用例：审核数据内容 |
| ID：4.2 |
| 简单描述：系统管理员审核数据 |
| 主参与者：系统管理员 |
| 副参与者：系统 |
| 前置条件：系统管理员查看了申请列表 |
| 主流：   1. 系统管理员点击申请列表中的一个申请 2. 系统随机生成一个样本展示出来 3. 系统管理员审阅数据样本内容 |
| 后置条件：系统管理员完成查看某一申请上传的数据样本 |
|  |

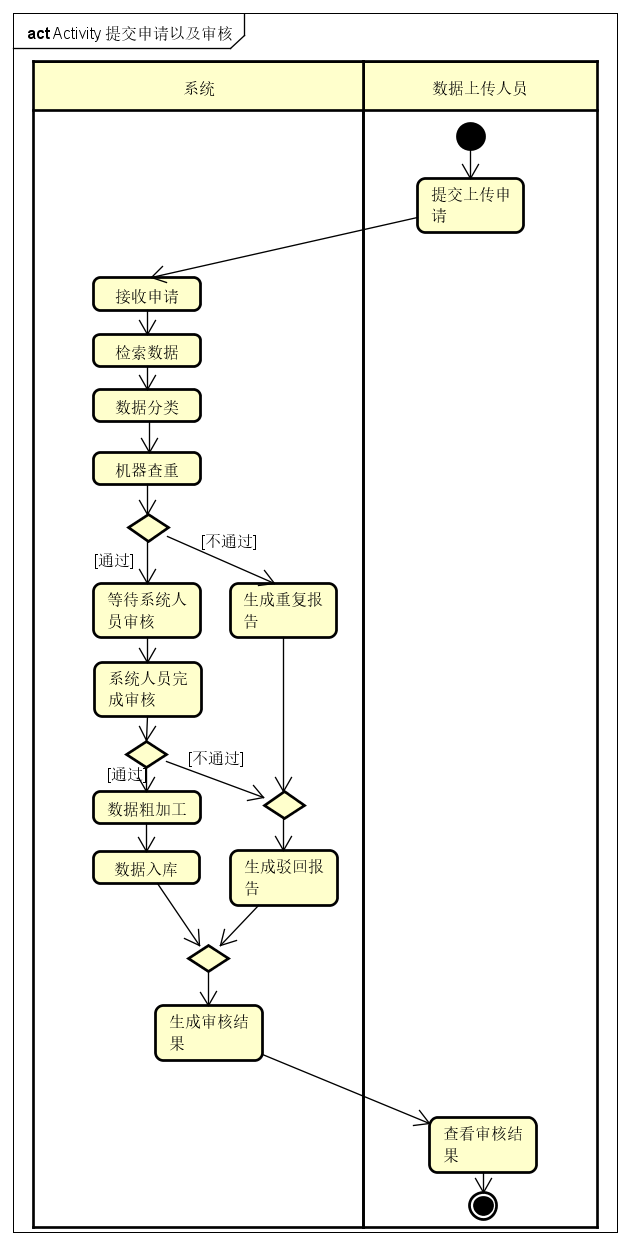
|  |
| --- |
| 用例：返回审核结果 |
| ID：4.3 |
| 简单描述：系统管理员返回审核结果 |
| 主参与者：系统管理员 |
| 副参与者：系统 |
| 前置条件：系统管理员查看数据样本 |
| 主流：   1. 系统管理员完成数据样本的审阅 2. 系统管理员点击通过，或者驳回 |
| 后置条件：系统管理员完成审核某一申请 |
| 附加流：点击驳回时，在旁边空白处生成文本输入框供系统管理员撰写驳回报告 |

|  |
| --- |
| 用例：提交申请 |
| ID：4.4 |
| 简单描述：上传用户提交审核申请 |
| 主参与者：上传用户 |
| 副参与者：系统 |
| 前置条件：上传用户拥有数据并已上传至服务器 |
| 主流：   1. 上传用户以用户身份登陆网页 2. 上传用户查看已有数据 3. 上传用户选择其中一个数据，填写申请表对数据进行描述并提交 |
| 后置条件：上传用户完成审核申请提交 |
| 附加流：系统未找到数据，显示：数据未找到或丢失！，返回 |

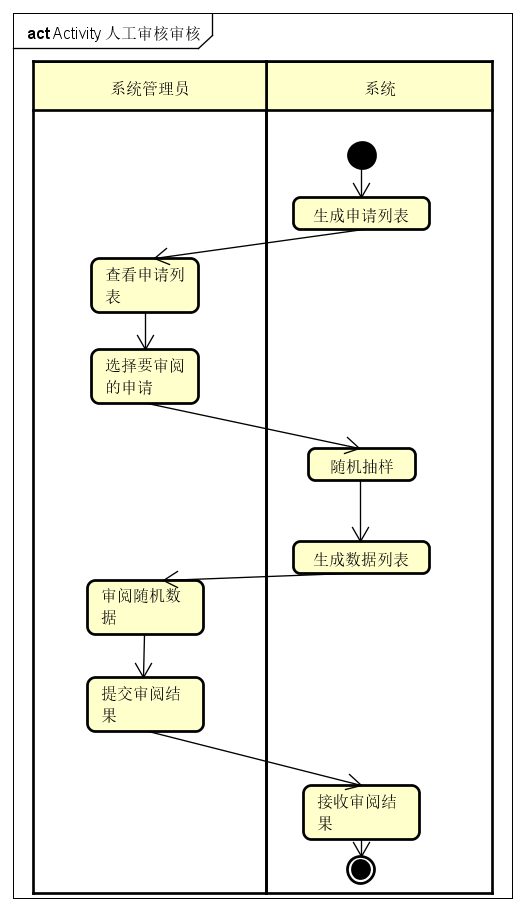
|  |
| --- |
| 用例：查看审核状态 |
| ID：4.5 |
| 简单描述：上传用户查看审核状态 |
| 主参与者：上传用户 |
| 副参与者：系统 |
| 前置条件：上传用户成功提交过上传申请 |
| 主流：   1. 上传用户以用户身份登陆网页 2. 上传用户查看系统生成的申请列表 3. 查看申请状态和详细信息 |
| 后置条件：上传用户完成查看审核状态 |
| 附加流：被驳回时，用户查看驳回报告 |
|  |

|  |
| --- |
| 用例：填写问卷 |
| ID：4.6 |
| 简单描述：数据标定需求者填写问卷以确定标定数据的准确性 |
| 主参与者：数据标定需求者 |
| 副参与者：系统 |
| 前置条件：数据标定需求者成功提交标定需求且系统已经初步完成标定任务 |
| 主流：   1. 数据标定需求者以用户身份登陆网页 2. 系统生成问卷 3. 数据标定需求者填写问卷 |
| 后置条件：数据标定需求者完成问卷填写 |
| 附加流：无 |

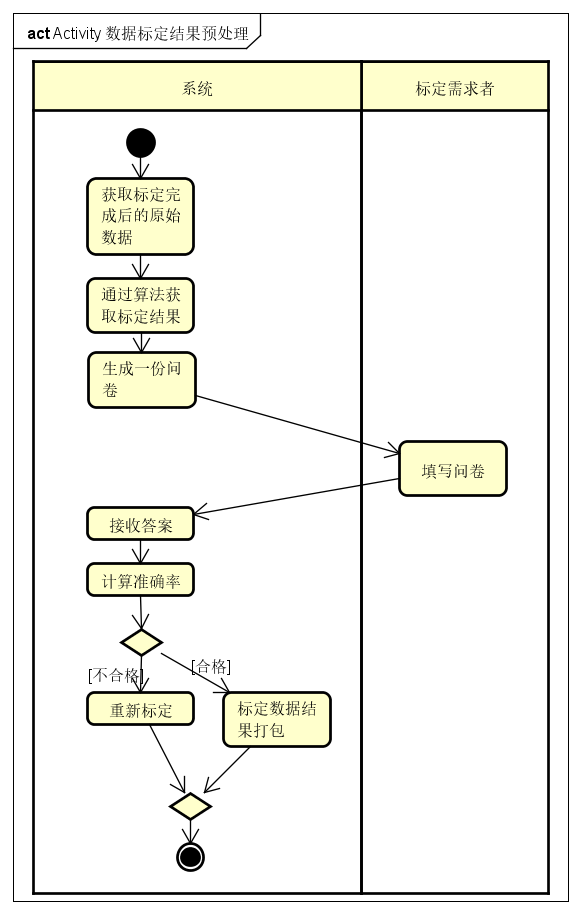
4.3 活动图



第一个活动图主要包含审核时，系统内部的工作。对于机器查重，我们可能会利用hashCode来查重。



第二个活动图是人工审核部分中人和系统的交互的活动图。



第三个活动图是标定完成后原始数据的预处理。

首先，我们可能会要求用户在填写的申请表中表明用户所要求标定数据答案的百分比，然后在后面确定答案时使用这个百分比。具体是这样来用，假如用户设定的是70%，并且我们系统设定一个待标定数据标定人数为10人。那么当某一个待标定数据同一答案的人数大于或等于7人时，我们就认为这个数据标定的答案是正确的。

标定完成后，为了保证数据标定需求用户的满意度，也为了保证数据的准确性，我们采用以问卷的形式，从用户那里获取部分标定数据的正确答案，与我们系统所获得的答案对比，得到这一组数据标定的准确率。如果准确率高于我们设定的一个数值，那么这一组数据标定任务就算是达成了。如果低了，就需要重新标定。