Doccano标注工具使用说明

Doccano标注工具界面简洁清晰，适合多种文本标注任务，可用于文本分类标注、文本实体标注以及两文本标注（如翻译数据、问答数据等）等标注任务。基于管理员-用户系统，可以有效地对不同的标注任务进行多用户的分发，并且通过权限的设置和分配，保证不同用户之间的标注任务是相互独立的。权限的设置，也有效防止了用户的一些误操作，如删除数据、改变标签等。

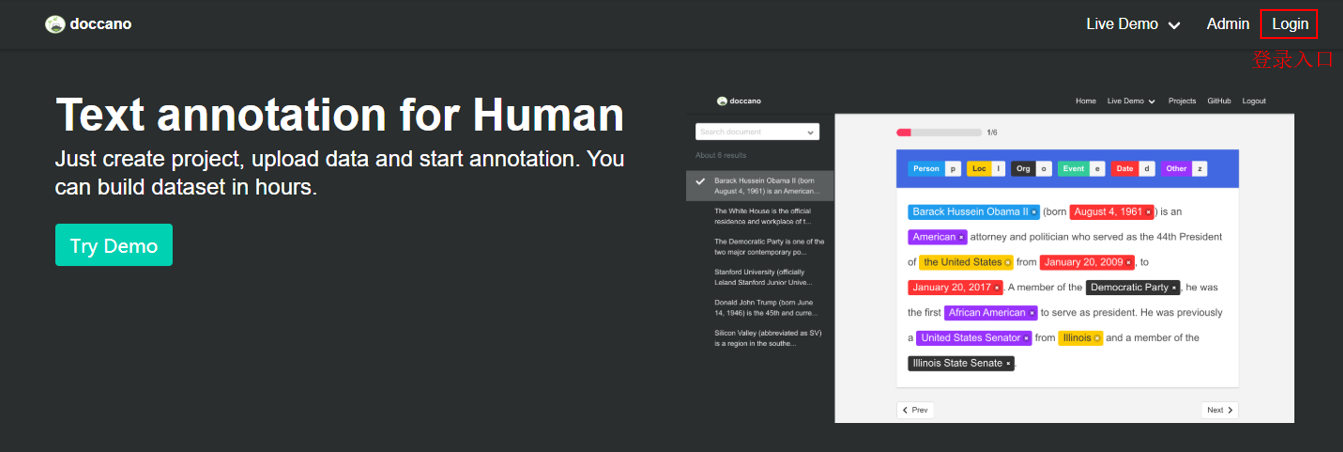
一、Doccano标注工具网址

<http://doccano.cc.163.com/>

二、用户登录

使用网易员工账号密码进行登录以及后续操作

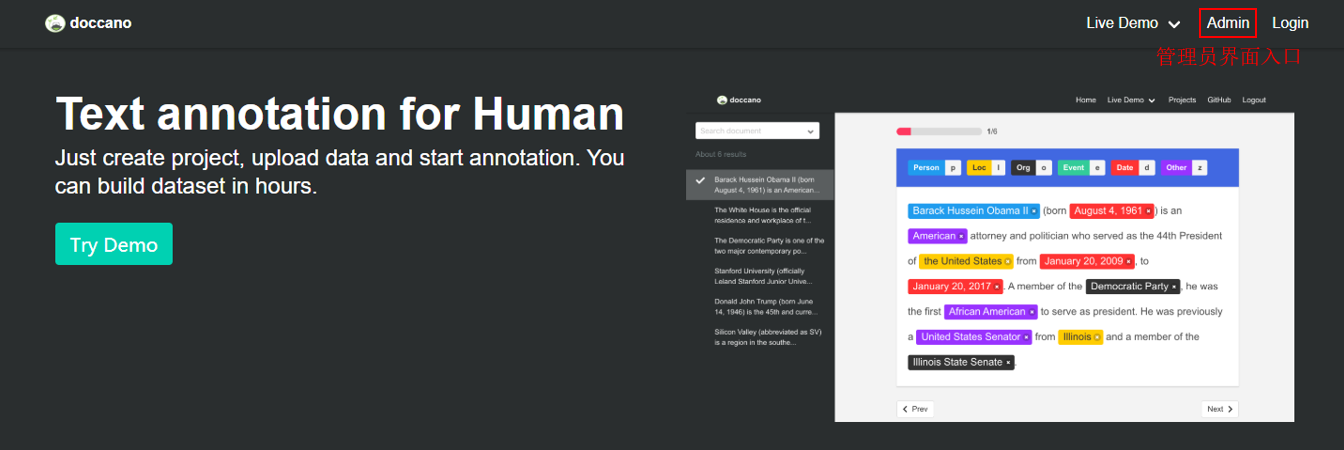
点击doccano首页头部右上角的Login按钮进行登录。



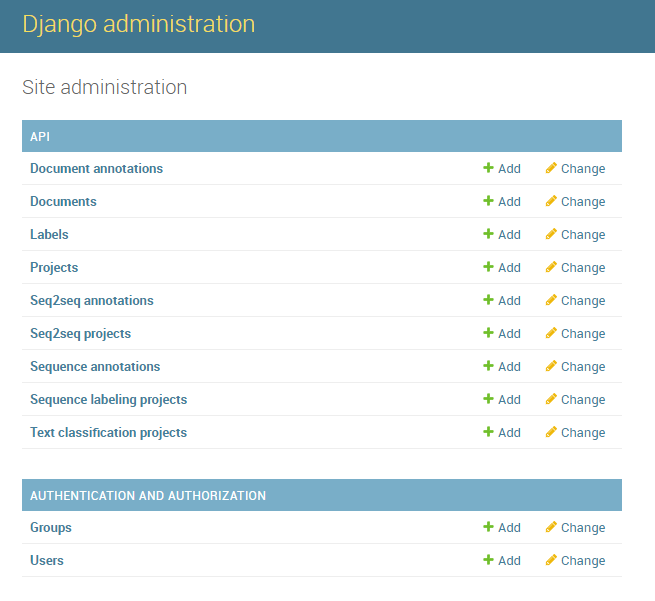
三、用户分组：管理员、普通用户

a.管理员页面、权限以及操作介绍

在管理员身份下，登录后拥有所有的权限，通过首页可以进入admin管理界面。



在进入admin管理界面后，管理员拥有里面的所有权限，如下图所示，具体分为两种权限：项目权限以及用户权限。



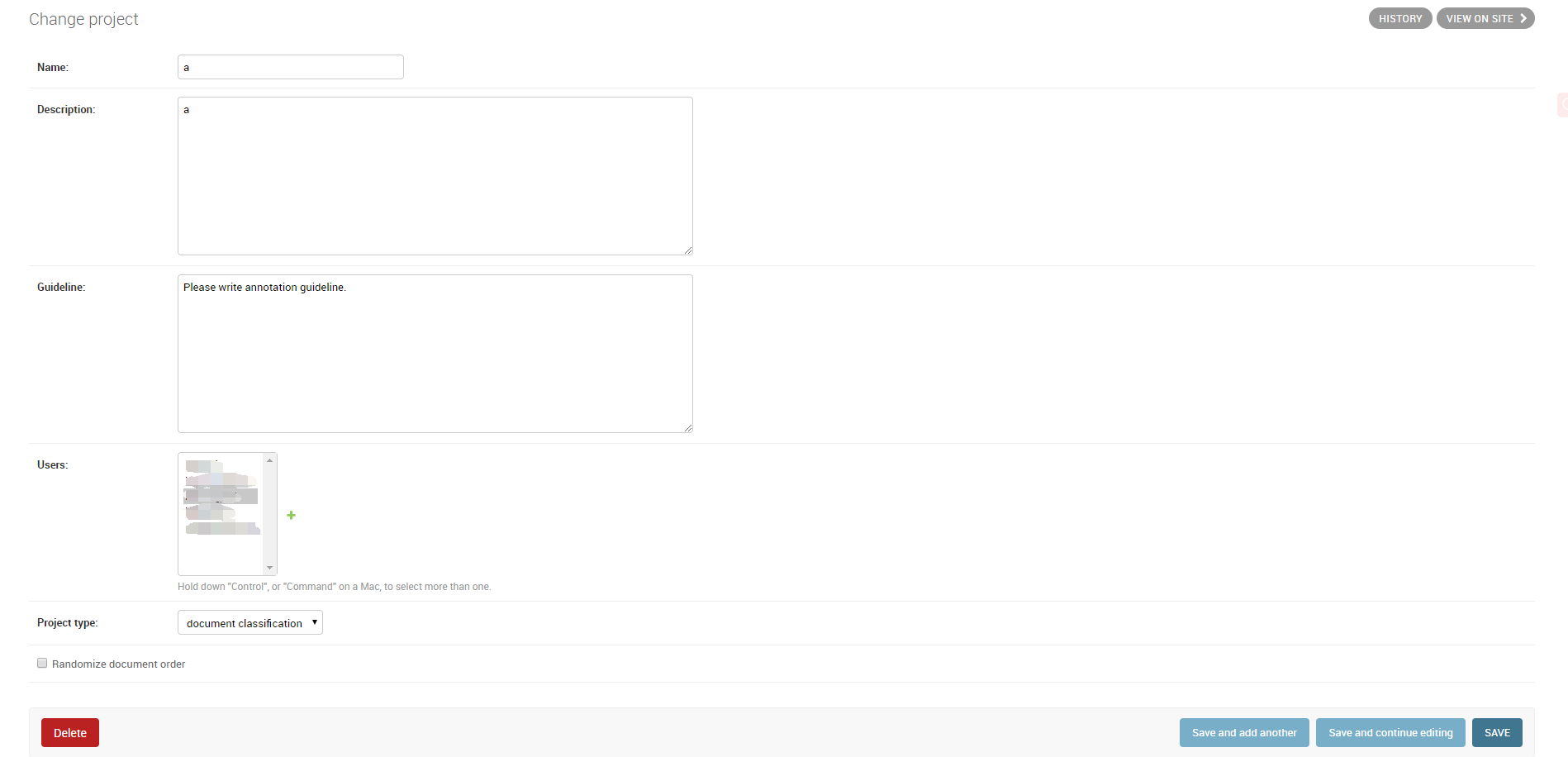
·项目权限

项目权限如图中API所示，分为文档标注、文档、标签、项目、seq2seq标注、seq2seq项目、实体标注、实体标注项目以及文档标注项目九项。

**项目**：管理所有项目（包括文档标注项目、seq2seq项目以及实体标注项目）的创建以及修改，其中可修改项包括名字（Name）、描述（Description）、参考（Guideline）、授权访问的用户（Users）、项目类型（Project type）以及是否打乱数据顺序（底部）。

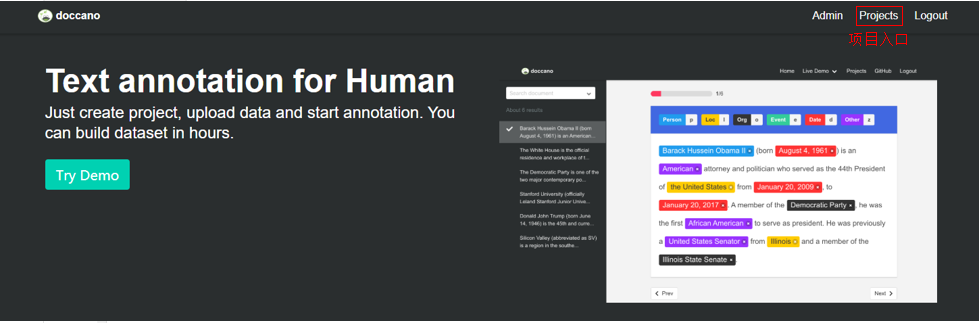
授权访问的用户可以控制当前可以访问该项目的用户，用于将不同的数据分配给不同的用户，如果需要添加新用户，请使用Ctrl+点击选中目标用户即可完成多用户的授权。

最后可以在页面的底部点击save按钮进行项目配置的保存，还有底部还有一个delete按钮，用于对整个项目的删除，包括该项目下所有数据以及标签。

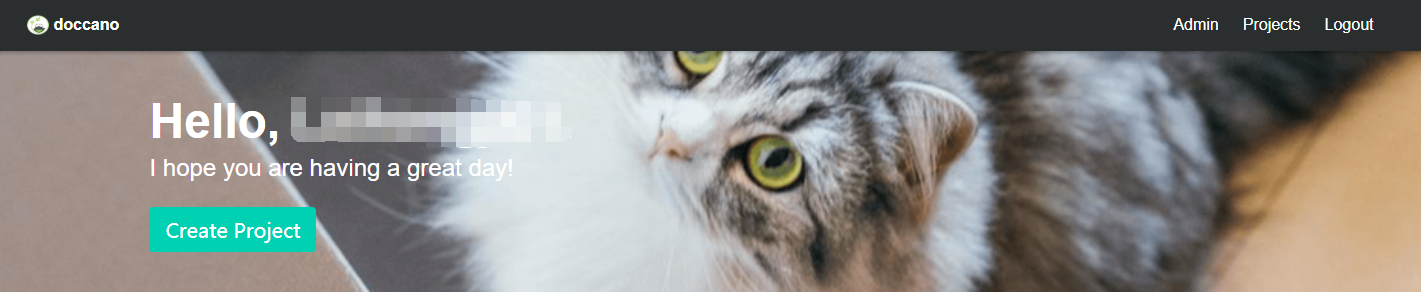


针对admin用户前端页面也可以直接创建项目。

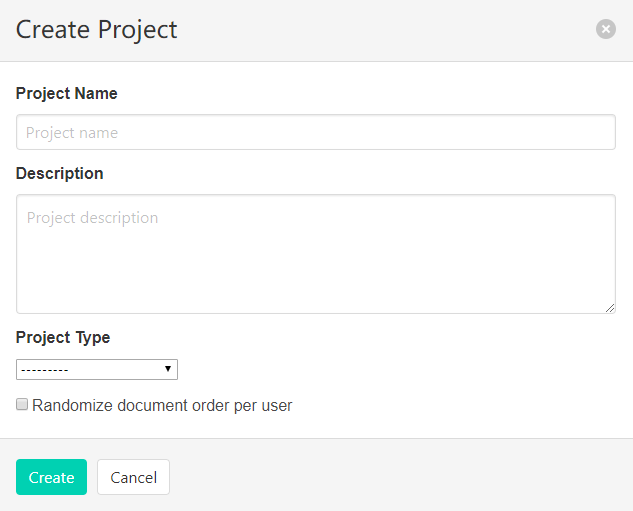
1.点击页面右上角的Projects的项目创建入口。



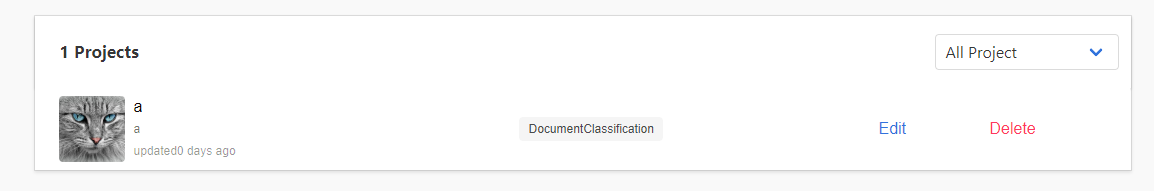
2.在进入项目入口后，点击create project按钮进行项目创建。



3.点击create project按钮后需要对各种参数进行定义。

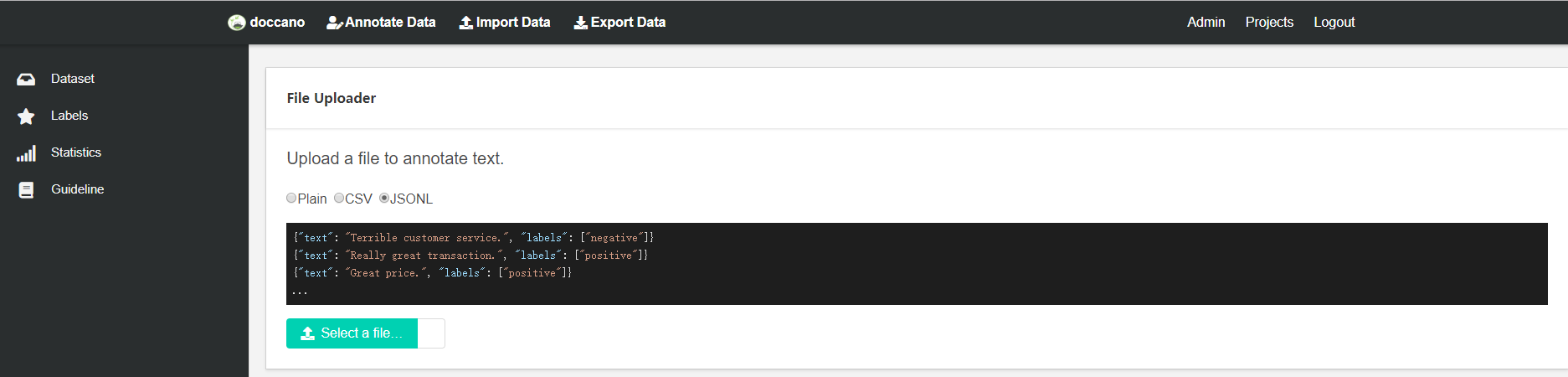


4.项目创建后会在Projects入口下得到以下的标签项。

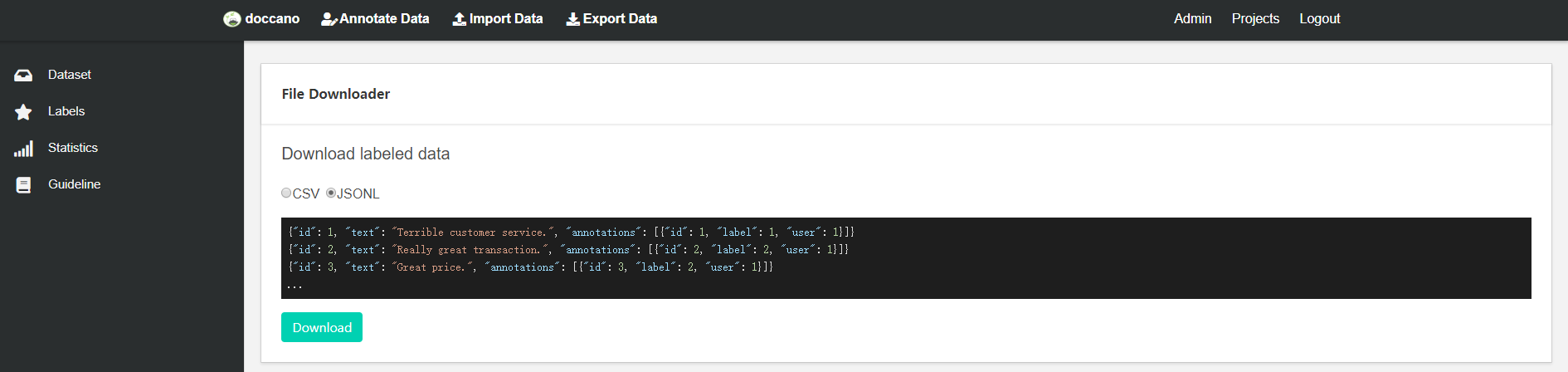


5.在项目内部，可以进行数据的导入、导出、标注以及标签的定义以及用户标注的情况。

数据导入：点击页面上方的Import Data按钮，数据格式有三种：plain、csv以及json格式数据，对应格式页面有示例数据，一般使用plain格式的文本数据。

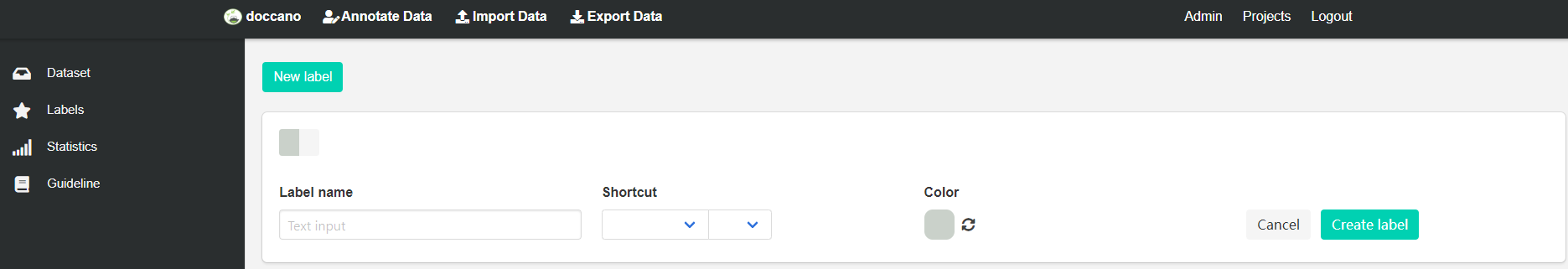


数据导出：点击页面上方的Export Data按钮，数据格式有两种：csv以及jsonl格式数据，一般使用jsonl格式导出。



数据标注：点击页面上方的Annotate Data按钮，然后根据自定义的label进行标注。

标签定义：点击页面右方的Labels按钮进行标签定义，New label按钮可以创建新标签，包括标签名、标签快捷键以及标签颜色。



用户标注情况：可以查看用户标注的各个标签的数量分布情况以及用户已标注的数据量。

**文档标注项目、seq2seq项目以及实体标注项目**：设置与项目一致，只是里面只存放了相应的项目。

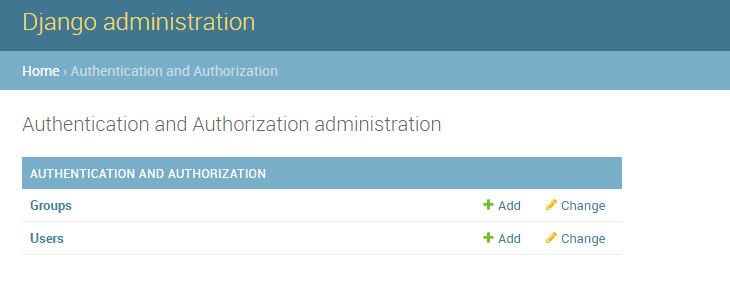
**文档**：存放了管理员上传的所有文本信息，管理员可以进行逐条的删除以及数据的添加，一般不建议这样操作，数据的删除应该在admin页面下的删除项目时附加执行，而数据的添加应该在非admin的前端页面进行项目创建时进行项目的导入。

**标签**：存放各个项目下定义的标签的详细信息，包括标签名、所属项目、文字颜色以及背景颜色等。

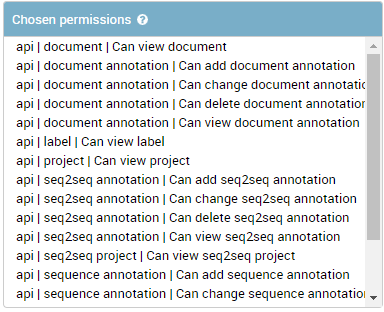
**seq2seq标注、文档标注、实体标注**：存放了对应项目下所有的文本数据以及其标注结果。

·用户权限

用户权限如图中AUTHENTICATION AND AUTHORIZATION所示，分为Groups以及Users两项。



**Groups**：用于定义用户权限的分组，针对标注任务，可以对普通用户赋予以下的权限分组，后面的文本标注项目的相关权限也类似设置即可。



**Users**：用于用户的管理，包括权限的定义以及增加和删除。针对用户的邮箱地址需要填写用户的网易邮箱。

b.普通用户权限以及操作介绍

普通用户可以通过页面右上角的Projects进入项目页面，然后点击Annotate Data按钮进行数据的标注即可。

四、工具即将开发需求以及现有解决方案

问题一：用户在标注数据的过程中，若由于某些突发状况或者自身原因关闭了当前的标注页面，在重新进入数据标注项目页面进行标注数据时会返回第一页的页面。

解决方案：后续会开发用户标注数据定位功能，使得用户在重新进行标注页面时自动到达标注过的最后一个词所在的页面。在功能未完善前，当前需要用户记住之前标注的大概位置并自主跳转页面，跳转方法有两种：（1）通过修改网址的参数实现跳转，在当前标注页面的网页地址的尾部添加#offset=[要跳转的页码]后再回车并按F5刷新页面即可；（2）通过标注页面中的标注框底部的页面跳转功能实现跳转，在底部的页码左边部分的显示框内输入要跳转的页码然后F5刷新即可，但是有些时候标注网址在#标志符后无offset参数时该功能失效，此时可以先参考第一种方法修改网页后即可再次正常使用页面跳转功能。

问题二：用户在标注数据的过程中，若标注数据过快，会导致部分数据不能被标注，即按标注快捷键后不能打上标签。

解决方案：（1）标注数据的速度控制在5条/3秒，过快的速度也会增加错判的几率；（2）若真的出现这种情况后，可以通过刷新页面进行解决。

问题三：用户在标注数据的过程中，出现漏标数据。

解决方案：后续会开发漏标数据提示功能，实现方式未确定。当前需要管理员进行漏标数据的线下检测再把漏标数据的索引号反馈给对应的用户进行数据的重新标注，用户通过管理员反馈的索引号进行对应的页面跳转操作进行重标即可。

问题四：有些在Windows下进行操作后的文本数据在写出后原数据会在头部增加一个不可见的’ \ufeff’编码字符，影响后续的数据提取。

解决方案：该问题的出现是由于Windows下使用Notepad++查看或者编辑数据使得文件自动被编码为utf-8-bom的格式，而原始文件是utf-8的。因此为了避免该问题，每次修改或者查看或者传递数据的时候，需要确保当前文本文件基于utf-8进行编码。