# 实体标注说明书

这篇说明书按照安装、部署、配置、使用的过程中遇到的问题编写，介绍了BRAT标注工具的基本使用方法，还有很多没有用到的功能，如果遇到说明书之外的问题，大家可以在百度或者官网搜索相关内容，欢迎补充这篇说明书。

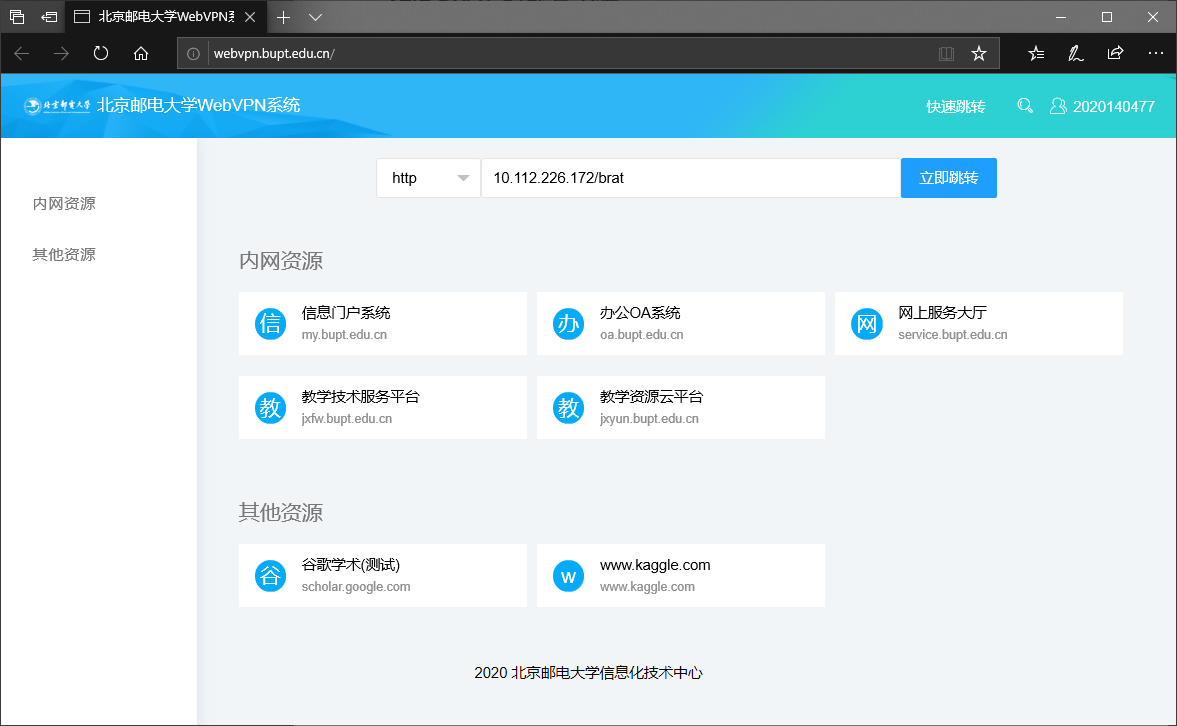
## 1 BRAT标注工具概述

BRAT是一个基于web的文本标注工具，主要用于对文本的结构化标注，用BRAT生成的标注结果能够把无结构化的原始文本结构化，供计算机处理。利用该工具可以方便的获得各项NLP任务需要的标注语料。

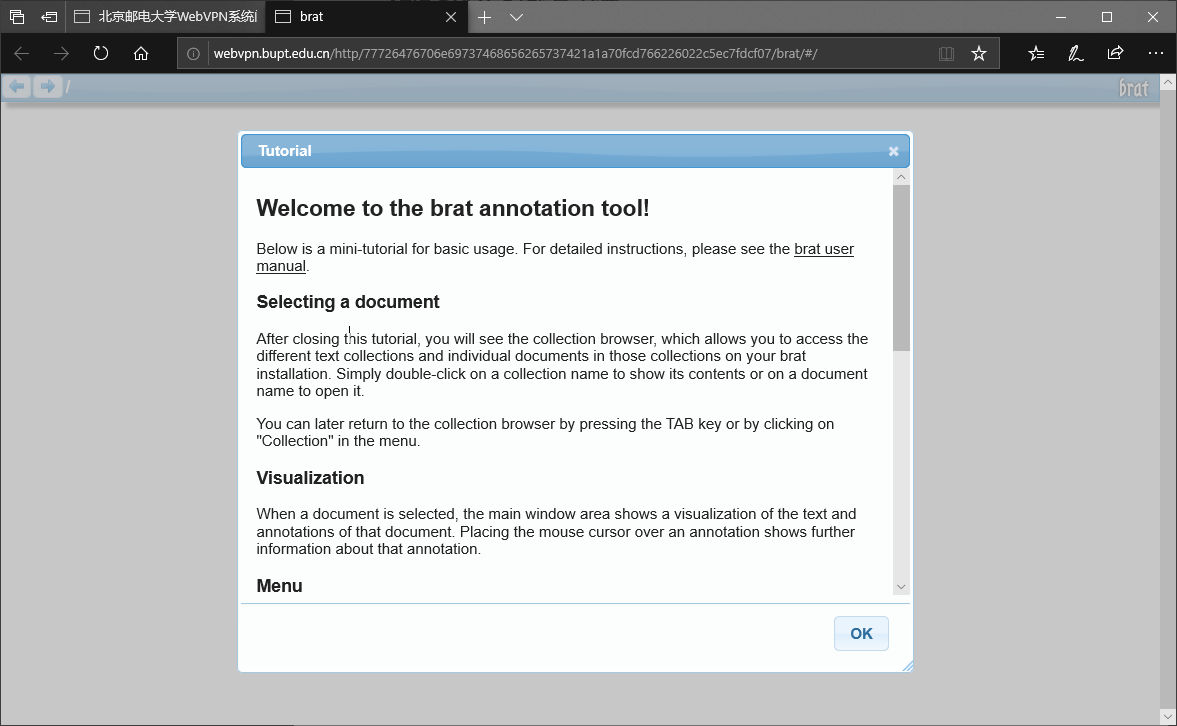
## 2 BRAT标注工具使用说明

### 2.1 BRAT的部署地址

校内网，或通过北京邮电大学WebVPN系统访问[10.112.226.172/brat](http://10.112.226.172/brat)（apple）：



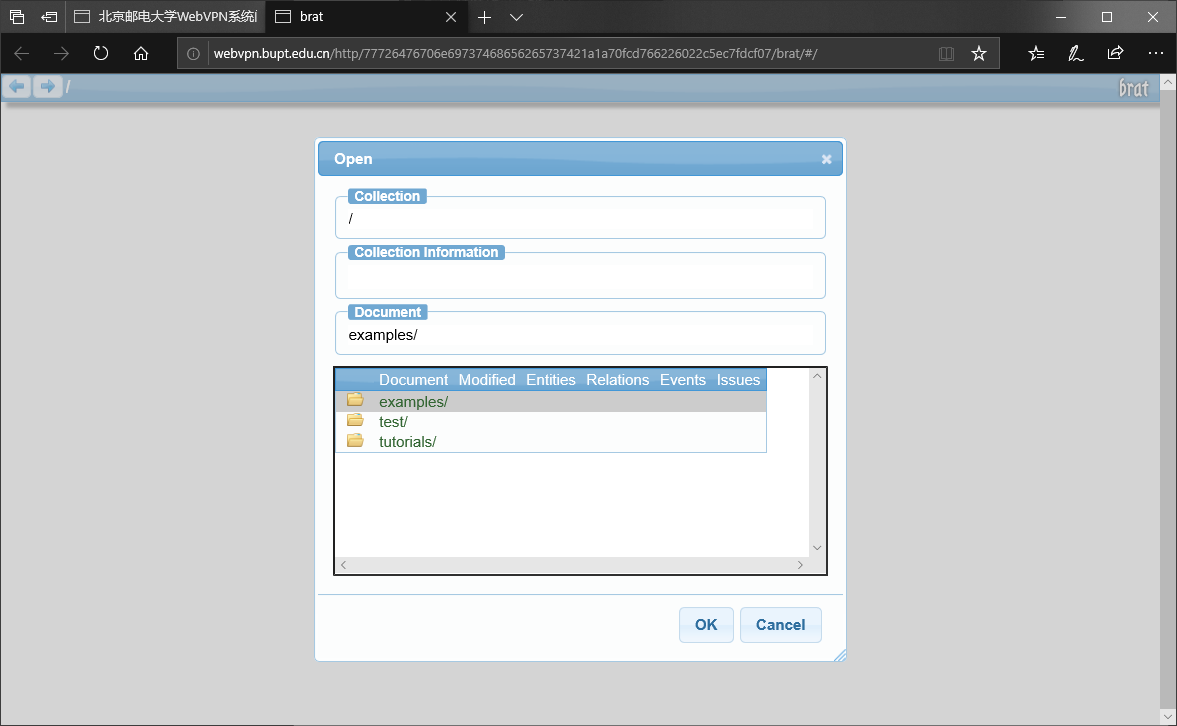
成功进入页面：



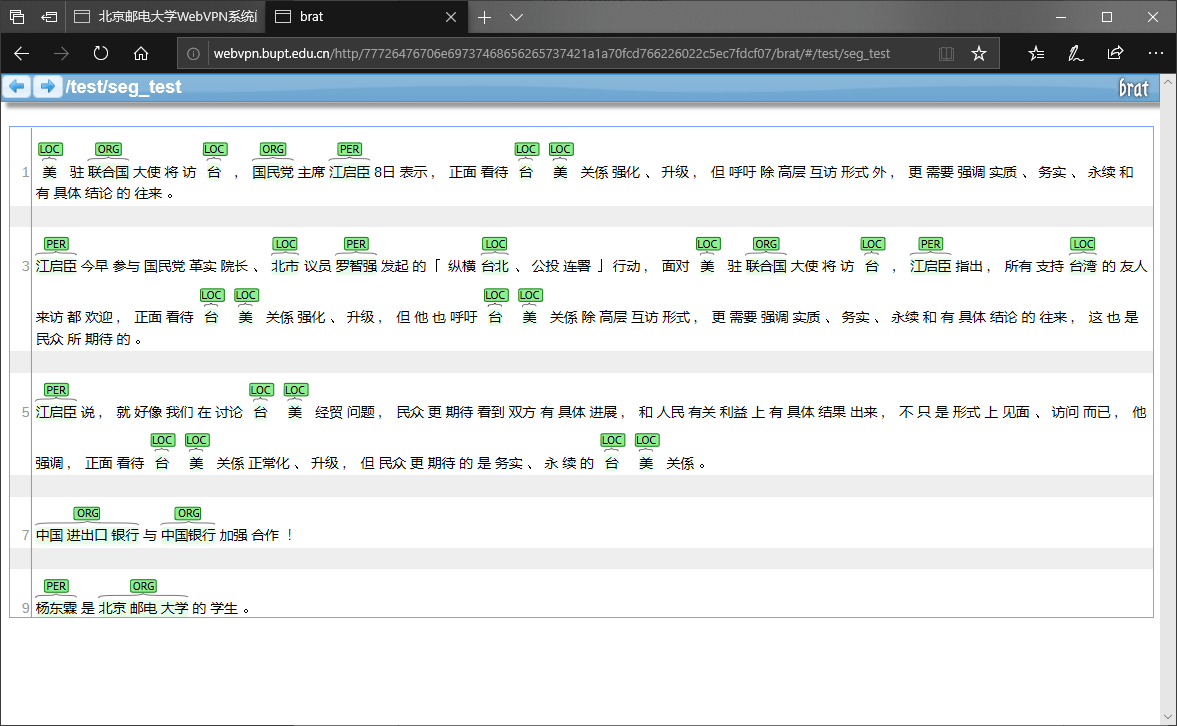
### 2.2 BRAT标注过程

#### 2.2.1 选择标注文件

主页面按Tab键或在菜单栏中点击collection，进入文件选择对话框，在collection bowser中选择要标注的文件：



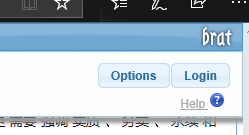
样例 ./test/seg\_test：



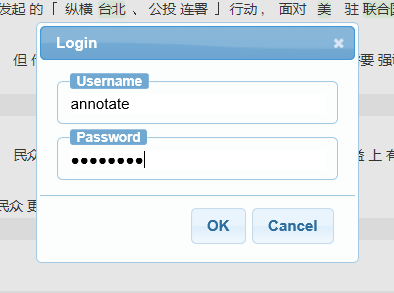
目前是只读模式还不能标注，登陆后可以进行标注工作。

#### 2.2.2 登录

点击菜单栏右边login：

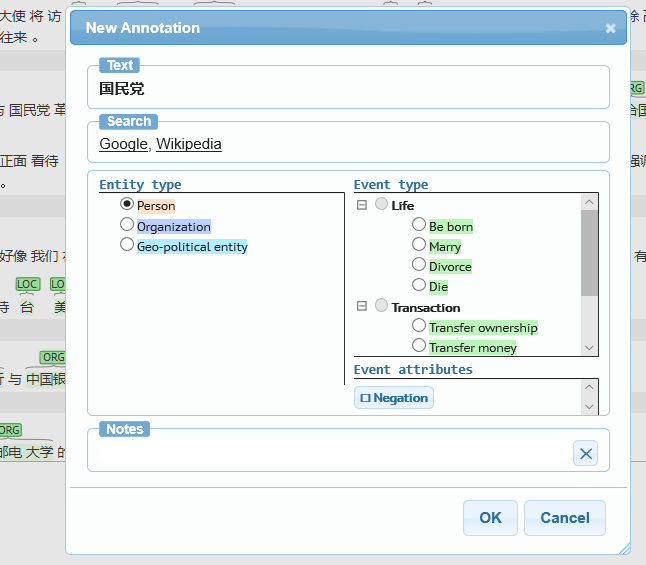


输入账号密码： Username：annotate Password：annotate



#### 2.2.3 标注

登陆后可以通过鼠标高亮拖拽等方式进行标注工作：



详细的标注方法不做进一步说明，请在官网查看使用手册，<http://brat.nlplab.org/manual.html>页面的Basic editing章节。

#### 2.2.4 原文件导入

回到collection bowser：



如何导入自己需要标注的文件？

BRAT搭建在apple上，工作目录为/var/www/html/brat，文件存放路径为/var/www/html/brat/data/，使用时在文件存放路径下创建自己的文件夹，可能需要sudo权限。例如test文件夹。将需要标注文件存放在，自己创建的文件夹目录下。例如myAnnotation.txt，内容片段如下：

**1964年9月10日，马云出生在杭州。**

**幼年的马云在人们的眼中是典型的坏孩子：叛逆、倔强、爱打架、逞强、顽皮淘气。**

**马云的父亲虽然是典型的江南人，但脾气却很火暴，马云从小在父亲拳脚下长大。**

**马云是看金庸的武侠小说长大的，行侠仗义、打抱不平的“侠义”情结在少年马云的内心深处早已生根。**

与txt文件对应，同一个目录下还需要同名的ann文件，初始可为空文件，系统才可以正常读取。使用BRAT对myAnnotation.txt的标注结果，会记录在myAnnotation.ann文件中。标注之后，再看看ann文件内容，如下：

Text

Description automatically generated with medium confidence

各字段说明如下：

* T1,T2所在的列，表示标注的类型和序号，比如如果是标注的实体间的关系会用R表示，这里因为只讨论命名实体，不涉及实体间的关系。
* 人名，公司名所在列表示标注词汇的实体类别；
* 第三、四列是标注词汇在整个txt中的起始索引和 (结束索引+1)；
* 最后一列是就是标注的词汇列。

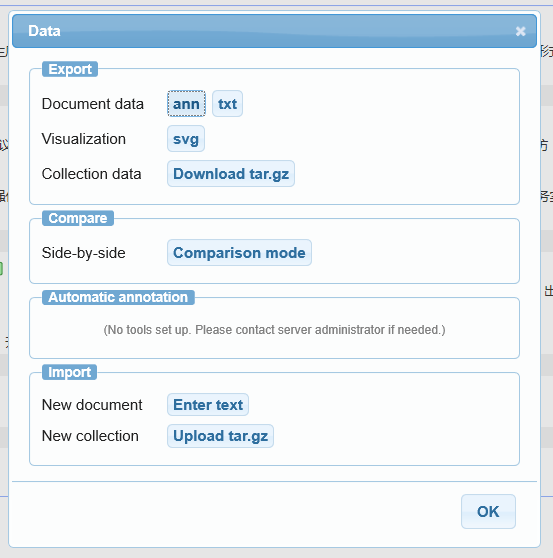
标注完成后，得到myAnnotation.txt和myAnnotation.ann两个对应的文件。

附一：

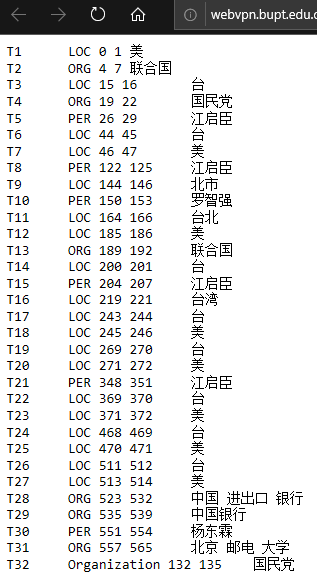
在web端菜单栏中点击data：



弹出对话框：



点击可以查看当前编辑的ann文件和txt文件，例如ann文件：



附二：

在命令行使用如下命令一次性批量生成目标文件夹下txt文件的同名ann文件（空文件）：

**find 文件夹名称 -name '\*.txt'|sed -e 's|\.txt|.ann|g'|xargs touch**

eg：find data/exa -name '\*.txt'|sed -e 's|\.txt|.ann|g'|xargs touch

附三：

Tips：可以使用python等工具将已标注文件（有一定格式）转为ann文件，使标注工具可读，进行可视化后，方便进行进一步标注。但需要十分清楚ann文件各个字段的含义及写法，可以在官网查看相关内容，<http://brat.nlplab.org/standoff.html>页面的Annotation files (.ann)章节。

#### 2.2.5 标注配置

成功导入文件后，BRAT工具可以在web端可视化的展示结果，使用者还需要根据自身需要修改标注配置。

BRAT通过配置文件来决定对语料的标注可以满足何种任务，包括四个文件

annotation.conf: annotation type configuration

visual.conf: annotation display configuration

tools.conf: annotation tool configuration

kb\_shortcuts.conf: keyboard shortcut tool configuration

一般只需要修改标记配置文件annotation.conf即可。

##### 2.2.5.1 配置annotation.conf

每个文件需要包含四类模块：entities、relations、events、attributes。各个模块都可以定义为空。各个模块详细介绍如下：

1）entities

用来定义标注的实体名称，其格式为每行一个实体类型，比如：人名、地名、英雄名、技能名等，可以采用tab来增加二级标注。

2）relations

用来定义实体间的关系，格式为每行定义一种关系，第一列为关系类型，随后是用逗号分隔的ArgNrelations用来定义实体间的关系，格式为每行定义一种关系，第一列为关系类型，随后是用逗号分隔的ArgN:实体名，用来表示关系的各个相关者。

3）events

用来定义事件，每行定义一类事件，第一列为事件名，随后是用逗号分隔的Participant:实体名，用来表示事件的各个参与者。

4）attributes

用来定义属性，每行一个属性，第一列为属性名，随后是用逗号分隔的Arg:<模块类型>, Value:属性值。注意属性值可以有多个。

例一：

Graphical user interface, text

Description automatically generated

例二，如果涉及events模块可以看一下：



##### 2.2.5.2 配置visual.conf

需要配置可视化颜色、标签显示内容时查看此章节，不需要可以跳过。

1）配置不同的类别用不同的颜色

visual.conf可以配置不同的类别用不同的颜色显示，找到如下段落，更新内容：

**[drawing]**

**时间 bgColor:yellow**

**地点 bgColor:blue, fgColor:white**

**人名 bgColor:deepskyblue**

**组织名 bgColor:green, fgColor:white**

**公司名 bgColor:purple, fgColor:white**

**产品名 bgColor:pink**

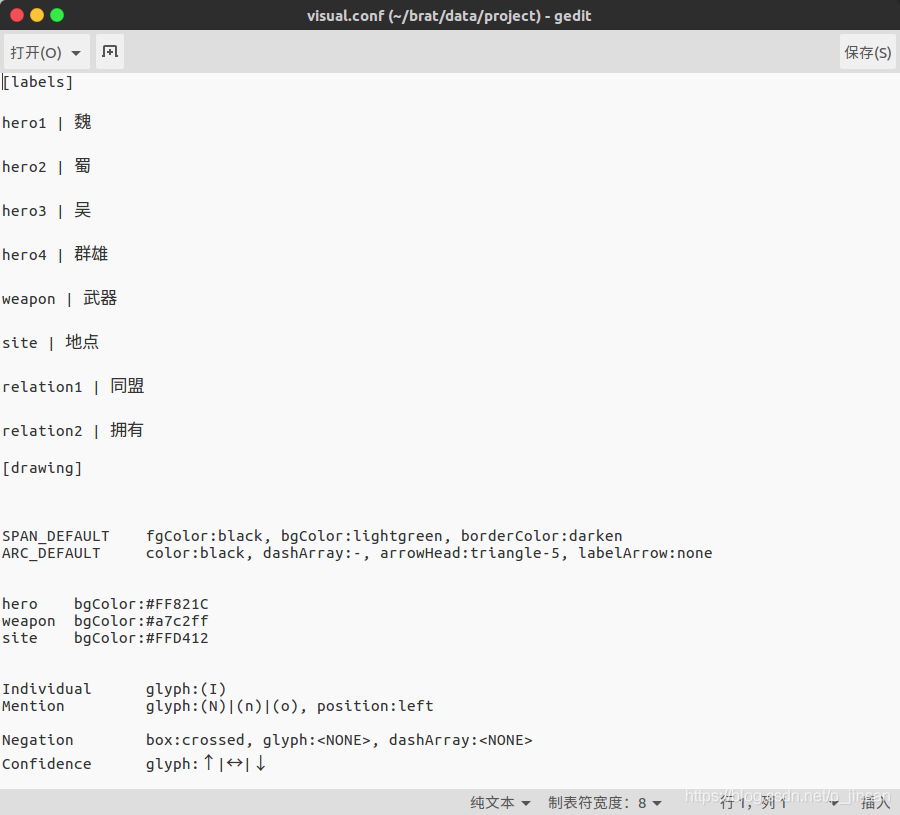
2）名称替换

visual.conf可以配置对annotation.conf中的名称进行替换，格式如下：

**[labels]**

**hero1 | 魏**

例：



## 3 数据标注样例

针对文档级关系抽取举例。

使用哈工大开源的NLP工具ltp进行命名实体识别。

使用python处理命名实体识别结果，处理为ann文件。

按照上述方法可以批量处理现有语料。

将标注好的文件放到BRAT工作目录下，实现命名实体识别的可视化。

使用BRAT进行关系抽取结果标注：预计使用远程监督等方法初步获取关系抽取结果，同样使用python进行处理，这样标注工作只用筛查远程监督关系抽取结果，极大减少了工作量。

## 参考

[Brat标注工具（本地）安装及使用](https://blog.csdn.net/p_jinsan/article/details/96152078)

[用深度学习做命名实体识别(三)：文本数据标注过程](https://zhuanlan.zhihu.com/p/92335448)