

# Identify Experts In Baidu Baike

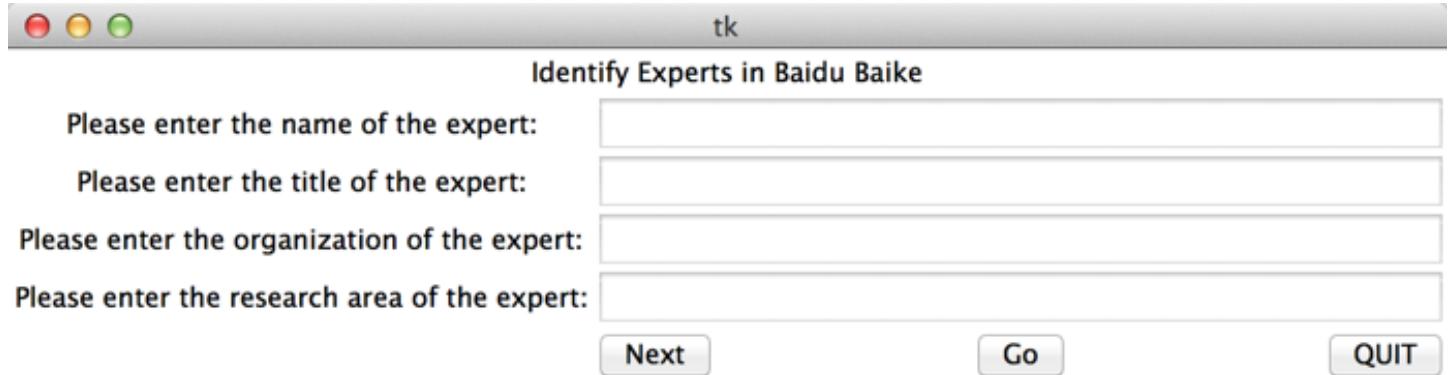
■ MachineLearning

## Language

Python

## GUI

利用python的Tkinter模块，编写了一个简单的GUI，外形如下。



关于将html嵌入GUI，我请教了清华的同学，似乎没有有效的方法，基本也都是控制浏览器。所以，这里我仍然沿用了我之前的方法。

## Function

### Normal Mode

运行后，显示GUI。

输入相应的信息，然后点击**Go**，程序即会利用Chrome Driver控制Chrome浏览器打开相应的web page.

为了方便，点击**Next**便可从数据集中通过解析XML获取下一个expert的信息。

效果如下：

Please enter the name of the expert: 秦樾

Please enter the title of the expert: 研究员

Please enter the organization of the expert: 中国科学院上海生命科学研究院

Please enter the research area of the expert:

Next Go QUIT

秦樾\_百度百科 - baike.baidu.com/view/236188.htm

新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

Baidu 百科 秦樾 进入词条 搜索词条 帮助

国家大剧院 管弦乐团

首页 分类频道 特色百科 玩转百科 百科用户 百科校园 百科合作 手机百科 个人中心

秦樾 编辑

本词条缺少名片图，补充相关内容使词条更完整，还能快速升级，赶紧来编辑吧！

秦樾[清]，字荫圃，江苏无锡人。善花卉翎毛，挥洒纵横，有天然生动之趣。《清朝书画家笔录》

中文名	秦樾	民族	汉
国籍	中国	出生地	江苏无锡

目录 1 秦樾简介 2 研究工作

1 秦樾简介 编辑

所系名称 健康科学研究所

性别 男

正在等待 c.baidu.com 的响应...

词条统计 浏览次数：2457次 编辑次数：6次 [历史版本](#) 最近更新：2014-08-09 创建者：[jaful](#)

百度百科品牌视频 震撼发布

百科消息：

- 百度百科首页有奖调研
- 开学啦！百科商城为你护航
- 高校女神风云榜，谁是你的菜？
- 大学寝室必备神器，你值得拥有！
- 想成为百科大神，来蝌蚪导师计划！

开启百科新首页

点击**QUIT**，即可退出程序。

## Batch Mode

运行之后，从数据集中逐个读取数据，然后将结果存至`ans.txt`文件中。

效果如下：

```
5 张道宏 讲师 西北农林科技大学 No Found!
6 张新宇 教授 杭州师范大学 No Found!
7 谢云飞 副教授 江南大学 No Found!
8 秦樾 研究员 中国科学院上海生命科学研究院 http://baike.baidu.com/view/236188.htm
9 丁英涛 副教授 北京理工大学 No Found!
10 孔玲爽 副教授 湖南工业大学 No Found!
11 董志诚 研究员 中国科学院华南植物园 http://baike.baidu.com/view/3691757.htm
12 赵福平 助理研究员 中国农业科学院北京畜牧兽医研究所 No Found!
13 韦伟峰 研究员 中南大学 No Found!
14 吴骅 研究员 复旦大学 No Found!
15 叶丹 研究员 复旦大学 No Found!
16 蔡亮 研究员 复旦大学 http://baike.baidu.com/subview/1188672/14191199.htm
17 邓伟民 主任医师 广州军区广州总医院 No Found!
18 蔡天革 副教授 辽宁大学 http://baike.baidu.com/view/4102239.htm
19 刘俏 教授 北京大学 http://baike.baidu.com/view/9555916.htm
20 张影 教授 北京大学 http://baike.baidu.com/subview/2712157/13774271.htm
21 曹旗 副教授 新乡医学院 No Found!
22 侯兴亮 研究员 中国科学院华南植物园 http://baike.baidu.com/view/9667914.htm
23 周丰丰 研究员 中国科学院深圳先进技术研究院 No Found!
24 方敏 研究员 北京大学 http://baike.baidu.com/subview/473952/8497569.htm
```

## Strategy

### Available Data

- title
- organization
- research area

### Feature Definitions

- *title\_val*: If the corresponding web page contains the title name, this feature is 1; otherwise, 0.
- *organization\_val*: The times the corresponding web page contains the organization name.
- *keyword\_val*: The times the corresponding web page contains the research area keywords.

### Intuition

`title`的多次出现对结果的判定直观上没有太大的联系，其意义只在于其是否出现，所以`title_val`的值是binary的。

而`organization`出现的次数越多，则表明相应的expert与该organization的关系越密切，所以`organization_val`的值记录`organization`出现的总次数。

同理，`keyword_val`也记录的是`keywords`出现的总次数。

### Problems and Solvers

## 1. 在相应词条的网页上，同时也包含了同名词条的其他信息，如下图所示

陈辉是一个多义词，请在下列义项中选择浏览（共50个义项）

添加义项

陈辉 (清华大学美术学院教授)

锁 定

陈辉：1959年生于安徽合肥，1985年毕业于中央工艺美术学院，现为清华大学美术学院教授，博士生导师，学院学术委员会委员，学院学位分委员会委员，基础部主任，学院当代艺术研究所常务副所长，美术分部副主任，吴冠中艺术研究中心研究员，张仃艺术研究中心研究员，中国博士后基金评审专家，中国艺术研究院研究生导师，中国国家画院研究员，中国美术家协会会员，中国美术家协会中国画艺委会委员，北京市美术家协会中国画艺委会委员，中国画学会创会理事，北京市高等艺术教育协会理事，中国博士后基金评审专家，中国展览馆协会展示艺术专家委员会委员，第十一届全国美展评委。

中文名 陈辉 出生日期 1959年10月  
国籍 中国 职业 画家 教授 博士生导师

收藏 267 | 赞 43

这些同名词条的信息会干扰对页面的判定。为了滤除这些信息，我发现在百部百科词条页面的HTML固定有且只有一个的中文“拆分词条”可以用来分割同义词条信息与本词条的信息，所以我就应用re模块的split函数来将HTML分割，从而有效的滤除了相应的干扰信息。

## 2. 有时候，organization由两部分组成，如下图所示

Please enter the name of the expert: 秦俊

Please enter the title of the expert: 教授

Please enter the organization of the expert: 中国科学技术大学/火灾科学国家重点实验室

Please enter the research area of the expert:

Next Go QUIT

这时，在匹配organization时，可能会遇到这样的问题，在词条中只包含了中国科学技术大学而没有包含后面的部分，从而导致匹配的失败，这显然是不合理的。为了解决这个问题，不难发现在这种情况下，前半部分通常以“大学”，“院”等结尾。通过re的模块的split函数进行分割，然后对分割后的结果再进行匹配，这样的得到的organization\_val更加准确。

## Judging Method

因为大部分的样本3个feature均为0，或者只有title\_val为1，而且label均为0，形成了很大的干扰。如果利用相关的机器学习算法，如Logistic Regression或者Support Vector Machine，其效果并不理想。

所以我改而利用直观规则进行判断：先遍历所有子条目，并计算和每个条目的3个feature之和。如果，其中的最大值大于1，即将最大的子条目作为结果；反之，如果3个feature之和为0或者1，则预测没有找到相应的条目。

在实际应用中，由于匹配项与不匹配项往往有较大差距，实际测试效果，非常理想。

## Performance Analysis

### Definition

	<b>Positive</b>	<b>Negative</b>
<b>Positive</b>	A	B
<b>Negative</b>	C	D

这里我们定义：

- A为专家有对应的条目并准确预测的次数；
- B为专家有对应的条目但并未准确预测的次数；
- C为专家并没有对应的条目，但是却给出了预测的次数；
- D为专家并没有对应的条目且也预测该专家没有对应条目的次数。

从而  $Precision = \frac{A}{A+C}$ ,  $Recall = \frac{A}{A+B}$ ,

$F1 = \frac{2*Precision*Recall}{Precision+Recall}$ 。

## Experiment and Result

我人工标记了500组数据，结果如下：

	<b>Positive</b>	<b>Negative</b>
<b>Positive</b>	130	2
<b>Negative</b>	13	355

$$Precision = \frac{130}{130+13} = 90.9\%$$

$$Recall = \frac{130}{130+2} = 98.5\%$$

$$F1 = 94.5\%$$

## Analysis

因为人工筛选的基本原则也是看给出的信息有无匹配，所以实际测试的效果不错。

$Precision$ 降低的主要原因是当没有匹配条目时，会错误的将一些有部分匹配的条目作为结果预测，从而产生错误的预测。

对于未给出预测的专家，说明对应信息没有有效匹配。一般来说，是相当准确的，所以 $Recall$ 相当高。其中B类中的两个未给出有效预测的，是因为Data中给出的organization是英文给出的，所以未能有效的匹配。

总体来说，效果还是不错的。

## Further Work

- 单纯的匹配并不能完全说明问题，比如匹配到“清华大学”，并不意味这其在清华任职，也有可能只是就读过；研究领域也可能正好有交叉。但由于sample和feature都有限，所以采用Machine Learning的相关算法都难以解决这些问题，或许可以通过改进规则来提升 $Precision$ 。

- 有效的将英文organization名转化为中文，以利于更好的匹配，从而进一步提升Recall。
- Program的效率仍有待提升，可以采用多线程来提升Batch Mode下的效率。

## Appendix

实现代码可以在这里找到：<http://github.com/billy-inn/IdentifyExpertsInBaiduBaike>。代码环境为 MacOS X 10.9.4。

### 安装Selenium

命令行下输入：`sudo easy_install selenium` 即可。

### 安装Chrome Driver

下载地址：<http://code.google.com/p/chromedriver/downloads/list>

下载完解压后，将`chromedriver`拷至 `/usr/bin` 即可。