Gender Guesser

可根据姓名猜测性别的 PHP 类库

Wudi <wudi@wudilabs.org>

人看到一个姓名后是怎样猜性别的？

汪峰

王菲

刘欢

峰，叫这字儿的基本都是男的，这人应该是个男的。

菲，叫这字儿的基本都是女的，这人应该是个女的。

欢，这字儿男的女的都有叫的，不好猜这人是男的还是女的。

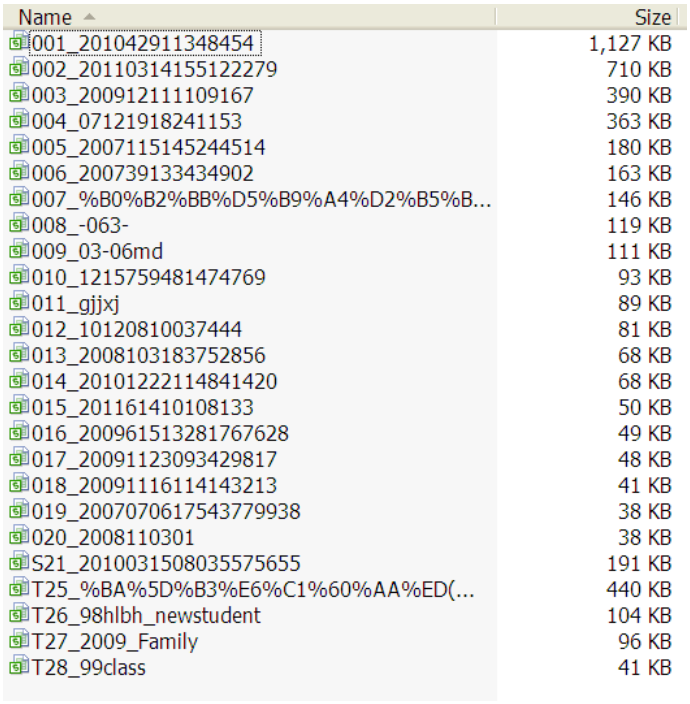
=> 根据已知的一些姓名和性别的对应关系来猜测

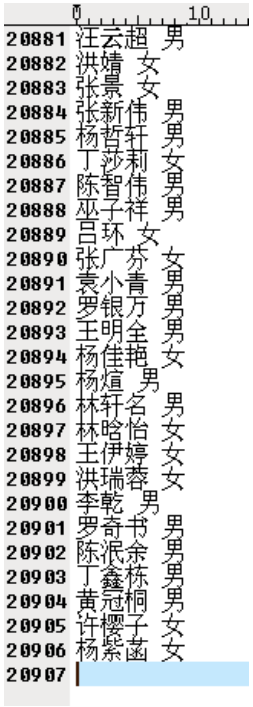
汉族人名的结构

- 姓+名
- 姓一般由一个汉字组成（单姓），也有少部分以两个或以上汉字组成（复姓）。
- 名一般习惯用一或两个字。
- 例如：
  - 高博　　（高+博）
  - 崔永元　（崔+永元）
  - 欧阳予　（欧阳+予）
  - 欧阳夏丹（欧阳+夏丹）
- 猜性别时只看名，而不考虑姓。
- 对于两个字的名，其第一个字和第二个字与性别的关联度不同。

第一步：收集姓名性别样本

- 通过搜索引擎查找各种带性别信息的人名单





- 经过整理，最终得到 20,906 个姓名性别信息

第二步：建立模型

- 对于任一汉字 C，其出现在男性姓名中的概率为
- $P_{男}(C)$ = C在男性姓名中出现的次数 / C出现的总次数

- 那么对于姓名中名为一个字 C_1 的人，其为男性的概率为
- $P_{男}(C_1)$

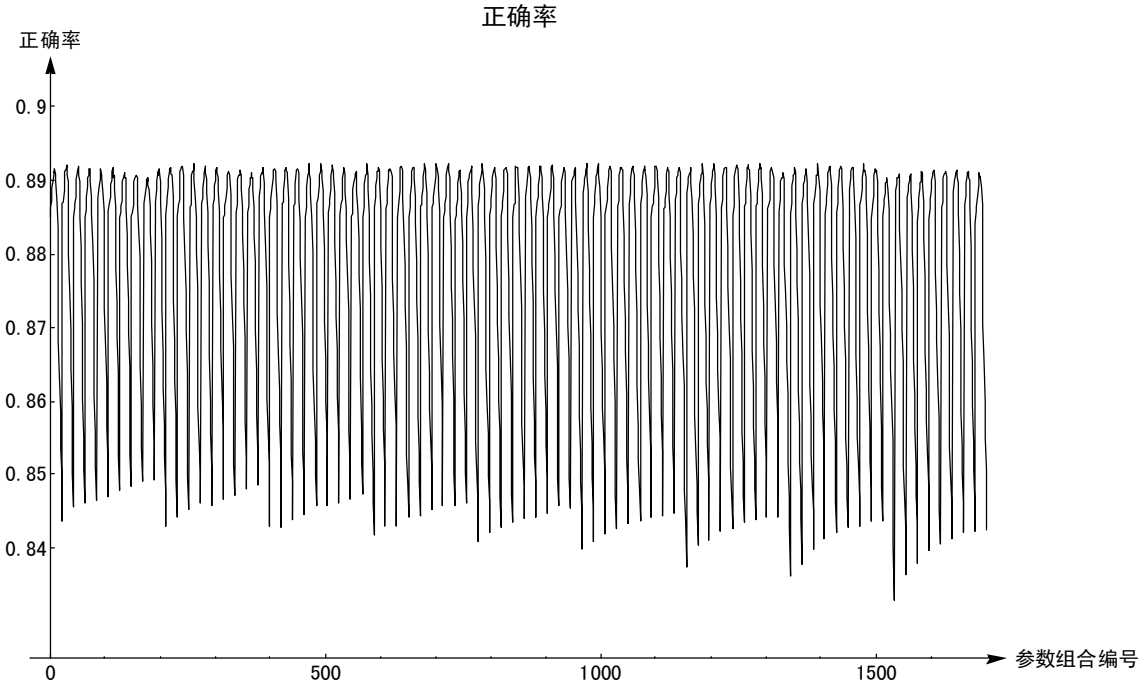
- 对于姓名中名为两个字 C_1C_2 的人，其为男性的概率为
- $w_1 * P_{男}(C_1) + (1 - w_1) * P_{男}(C_2)$
- 其中 w_1 为字 C_1 的权重，$(1 - w_1)$ 为字 C_2 的权重

第二步：建立模型

- 对两个字的名 C_1C_2，除了在猜测性别时对 C_1 和 C_2 给予不同的权重外，在由收集到的样本产生各个汉字 C 出现在男性姓名中的概率 $P_{男}(C)$ 的数据时，也对 C_1 和 C_2 给予不同的权重。
- 使用 v_1 和 v_2 参数调整 C_1 和 C_2 对 $P_{男}(C)$ 的贡献。
- 调整前，字 C_1 在男性姓名中出现一次记为 1 次，在女性姓名中出现一次记为 0 次。
- 调整后，字 C_1 在男性姓名中出现一次记为 $((1 - 0.5) * v_1) + 0.5$ 次，在女性姓名中出现一次记为 $((0 - 0.5) * v_1) + 0.5$ 次。

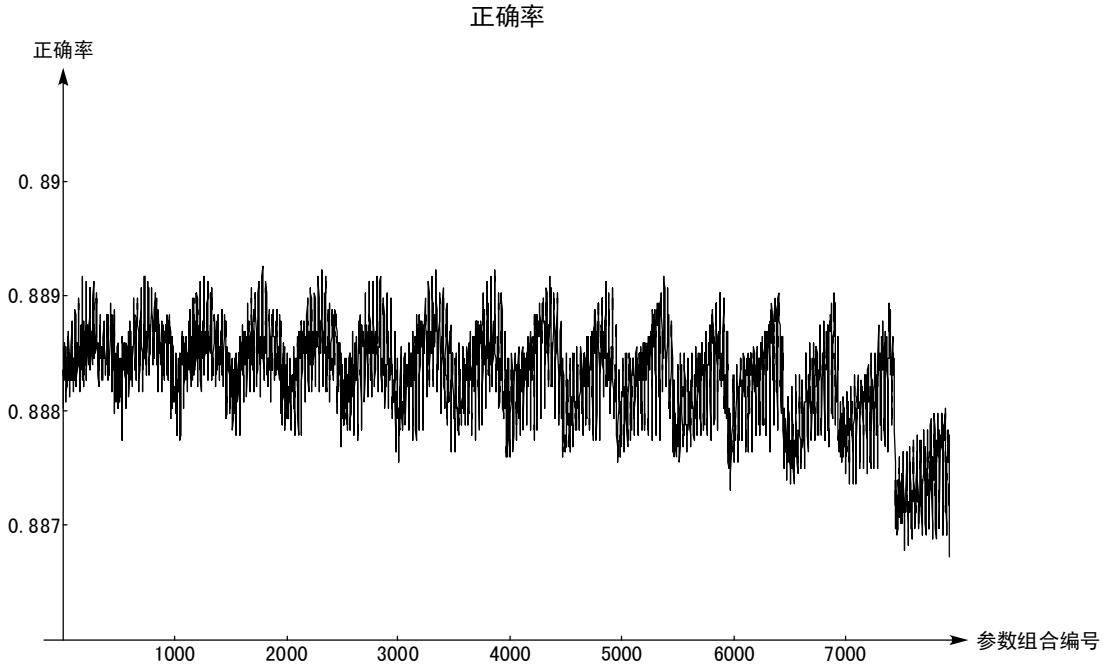
第三步：查找最优的权重参数

- 首先在如下比较大的范围内进行粗略的查找
- $0.6 \le v_1 \le 1.0$，$0.6 \le v_2 \le 1.0$，$0.3 \le w_1 \le 0.7$;
- 在该参数范围内，性别猜测结果的正确率分布在 0.830 到 0.889 之间



第三步：查找最优的权重参数

- 根据结果逐步缩小范围，增加分辨率，最终将范围确定为
- $0.85 \le v_1 \le 1.0$，$0.85 \le v_2 \le 1.0$，$0.43 \le w_1 \le 0.46$;
- 该参数范围内，性别猜测结果的正确率分布在 0.8867 到 0.8893 之间



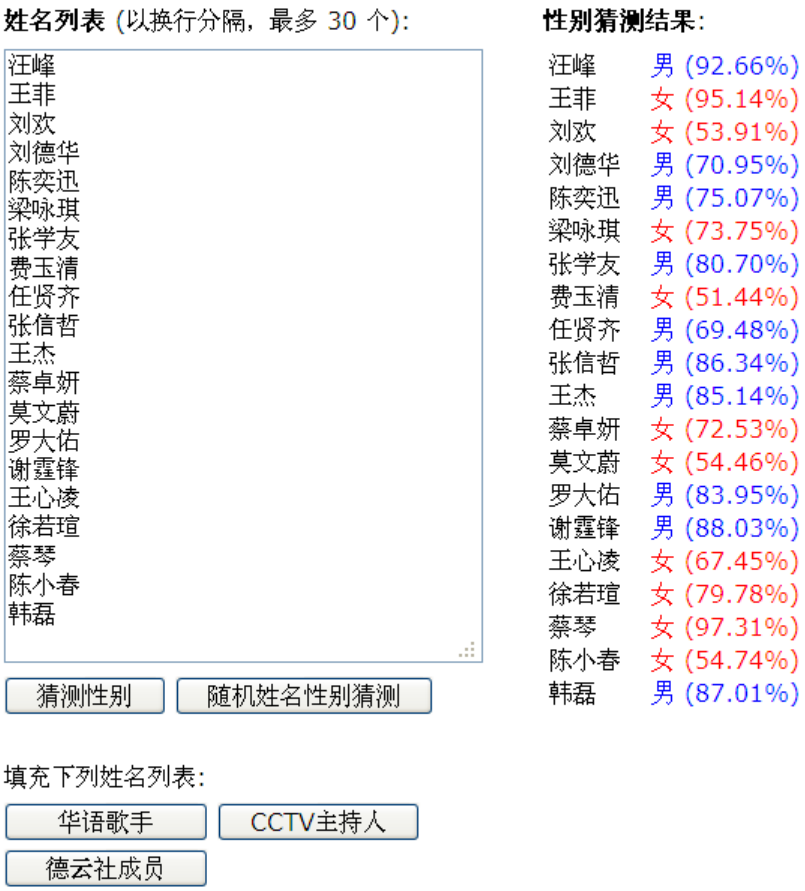
第三步：查找最优的权重参数

- 最高的正确率出现在 v_1=0.88, v_2=0.94, w_1=0.449 时，达到了 88.9266%
- 封闭测试结果中，猜测性别的概率很高但依然错误的样本，都是名字本身的性别倾向就不正常的。

```
共 20906 个，正确 18591 个，错误 2315 个（已显示的：41 个）。
正确率：88.93%，错误率：11.07%（已显示的，男：43.90%，女：56.10%）
高山：女 ---- 男（92.59%）
姚仪：男 ---- 女（90.10%）
黄翔：女 ---- 男（91.50%）
廖岚：男 ---- 女（90.93%）
肖瑶：男 ---- 女（90.61%）
林仪：男 ---- 女（90.10%）
王涛涛：女 ---- 男（93.35%）
吴豪：女 ---- 男（96.02%）
吴猛：女 ---- 男（91.14%）
李静：男 ---- 女（91.95%）
王斌斌：女 ---- 男（90.65%）
俞瀚：女 ---- 男（90.33%）
沈兰：男 ---- 女（92.60%）
王晴：男 ---- 女（94.28%）
```

v1	v2	w1	accuracy
0.85	0.9	0.449	88.9171%
0.86	0.92	0.448	88.9171%
0.86	0.92	0.449	88.9171%
0.87	0.93	0.448	88.9171%
0.87	0.93	0.449	88.9171%
0.87	0.93	0.45	88.9171%
0.88	0.94	0.448	88.9171%
0.89	0.94	0.447	88.9171%
0.89	0.95	0.448	88.9171%
0.89	0.96	0.449	88.9171%
0.9	0.95	0.446	88.9171%
0.9	0.95	0.447	88.9171%
0.9	0.96	0.45	88.9171%
0.91	0.95	0.447	88.9171%
0.91	0.96	0.447	88.9171%
0.91	0.96	0.448	88.9171%
0.92	0.96	0.446	88.9171%
0.92	0.96	0.447	88.9171%
0.92	0.97	0.446	88.9171%
0.92	0.97	0.447	88.9171%
0.92	0.97	0.45	88.9171%
0.93	0.97	0.446	88.9171%
0.93	0.97	0.447	88.9171%
0.95	0.98	0.444	88.9171%
0.89	0.95	0.449	88.9218%
0.91	0.96	0.446	88.9218%
0.92	0.97	0.448	88.9218%
0.88	0.94	0.449	88.9266%

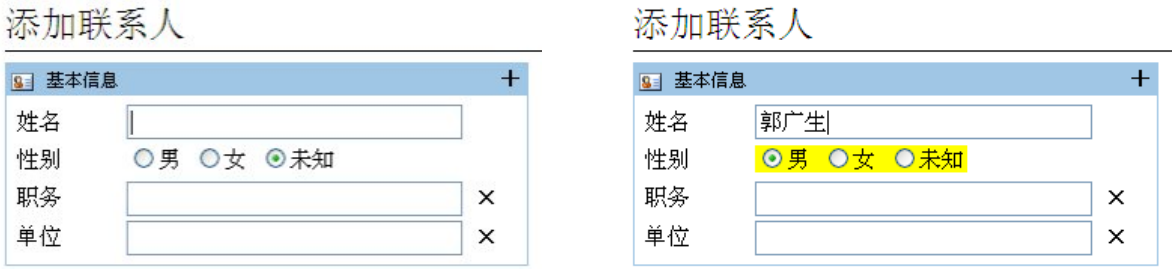
完成



http://demo.wudilabs.org/lab/gender_guesser/

该应用的意义

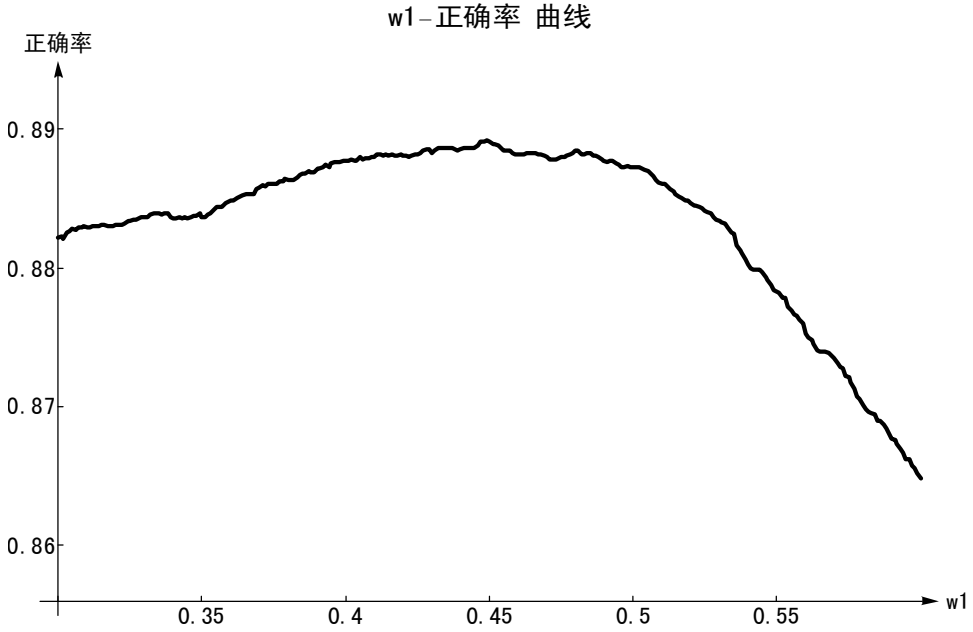
- 在向通讯录中添加联系人时，由程序自动填写性别，多数情况可免去一次点击。



- 在批量导入不含性别信息的联系人时，由程序自动猜测性别，再人工纠错，可节省时间。
- 可对人名单进行批量性别猜测和统计，得到大致的性别比例。
- 外国人起中文名时，可用该程序判断所起的名字性别倾向有没有问题。

一些发现

- 以下是 v_1=0.88, v_2=0.94 条件下的 w_1-正确率 曲线。在 w_1 = 0.5 的右侧，正确率降低较左侧快很多。而 (1 - w_1) 为双字名中第二个字的权重，这和一般人们的经验是相同的，即双字名中第二个字和性别的关系更大。所以 (1 - w_1) 应该大于 0.5，w_1 应该小于 0.5。



一些发现

- 以下是封闭测试中性别猜测错误的 2315 个姓名的概率-错误数量曲线。可以看出猜测出是男性的概率大于 50% 的错误的比概率小于 50% 的明显要多，也就是说女性起男性名的现象相对于男性起女性名的现象要常见的多。

The End

http://blog.wudilabs.org/tag/genderguesser/