# 第三章

1. **阅读维基百科有关Soundex条目。使用Python实现这个算法。**

SoundEx 是一种拼音算法，用于按英语发音来索引姓名，它最初由美国人口调查局开发。 SoundEx 方法返回一个表示姓名的四字符代码，由一个英文字母后跟三个数字构成。 字母是姓名的首字母，数字对姓名中剩余的辅音字母编码。发音相近的姓名具有相同的 SoundEx 代码。可用Soundex做类似模糊匹配的效果。

规则：

1.保留名称的第一个字母，并删除a，e，i，o，u，y，h，w的所有其他出现。

2.用下面的数字替换辅音（在第一个字母之后）：

**b, f, p, v → 1**

**c, g, j, k, q, s, x, z → 2**

**d, t → 3**

**l → 4**

**m, n → 5**

**r → 6**

3.如果两个或多个相同编号的字母在原始名称中相邻（步骤1之前），则只保留第一个字母;同样由'h'或'w'分隔的相同数字的两个字母被编码为单个数字，而由元音分隔的这些字母被编码两次。这条规则也适用于第一个字母。

4.如果单词中的字母太少，以至于无法分配三个数字，则会附加零，直到有三个数字。如果有3个以上的字母，只保留前3个数字。

SOUNDEX 函数对英语单词效果最好。对于其它语言，它的用处不大。

**def** soundex(name, size=4):soundex\_digits = **'01230120022455012623010202'** sndx = **''** fc = **''  
 for** c **in** name.upper():  
 **if** c.isalpha():  
 **if not** fc:  
 fc = c  
 d = soundex\_digits[ord(c) - ord(**'A'**)]  
 **if not** sndx **or** d != sndx[-1]:  
 sndx += d  
 sndx = fc + sndx[1:]  
 sndx = sndx.replace(**'0'**, **''**)  
 **return** (sndx + size \* **'0'**)[:size]

**print** soundex(**'Robert'**),soundex(**'Rupert'**)  
**print** soundex(**'Catherine'**),soundex(**'Katherine'**), soundex(**'Katarina'**)  
**print** soundex(**'Rubin'**)  
**print** soundex(**'Tymczak'**)  
**print** soundex(**'Pfister'**)  
**print** soundex(**'Honeyman'**)

R163 R163

C365 K365 K365

R150

T522

P236

H555