Gender Prediction In R

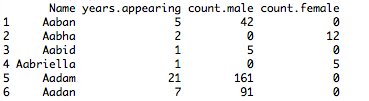
Pingyuan(Zoe) Zheng 7.2015

1. Base Project (GNB): <https://github.com/OpenGenderTracking/globalnamedata.git>

It is an R package containing statistical utility functions and demographic gender data. The functions include count males and females of every name every year. More reading on: <https://bocoup.com/weblog/global-name-data/>

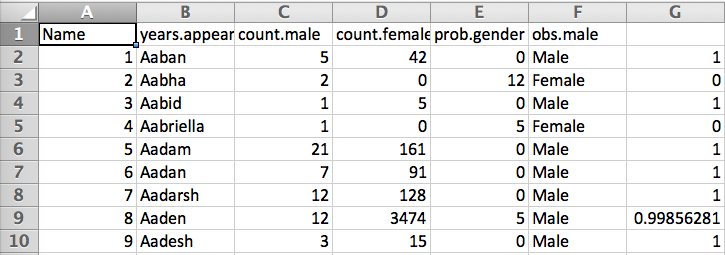
1. Major changes on the GNB project and logics:

I made changes on the classifier.R file, where the nameBinom() function can predict the gender of every name. The nameBinom function takes input data format as follows:



The binom.confint function will calculate probability of male and confidence interval. Here we are taking gender as a binomial event with random samples, so we can regard the frequency of male as the probability of male. If the probability of male is greater than threshold(default setting is 0.9), then the name will be predicted as male; if the probability of male is less than 1- threshold, then the name will be predicted as female; if the probability of male falls between threshold and 1-threshold, the function will perform a 10-time binomial trial with the given probability. If the result mean falls in the confidence interval, the name will be predicted as male, otherwise female.

The standard gender result in standard.csv file is as follows:

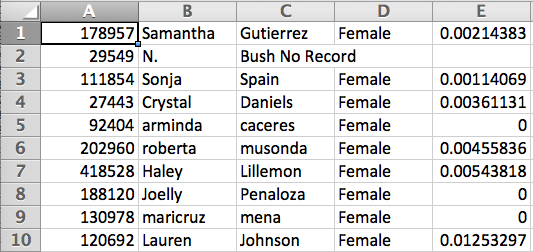


You can always use the standard.csv file.

1. Prediction

Basing on the standard list above, I compared every user’s first name to be predicted to the list, and retrieved their gender. Source code is in predict.py.

1. Result in predict\_result.csv file:



Summary:

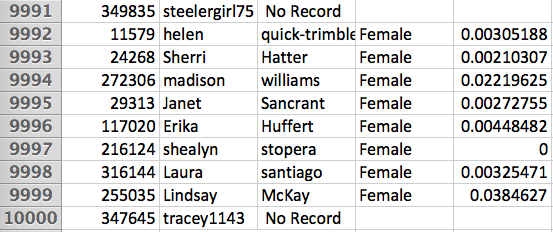
10000 users without gender information

Male: 1186, Percentage: 11.86%

Female: 6736, Percentage: 67.36%

No Record: 2078, Percentage: 20.78%

No Record is due to bad inputs of first names, like:



1. How to use this project

Github link: <https://github.com/LittlePenguinPenguin/globalnamedata.git>

The assets folder contains all the raw data.

The R folder contains all the R code.

The GenderPrediction folder contains all the result files, where you can see the standard gender file.

If you want to take a look at other utility functions, please refer to the blog of the author: <https://bocoup.com/weblog/global-name-data/>