中華大學生物資訊學系系統開發專題報告

以R語言實作外國人姓名英漢翻譯

Implementing English-Chinese Translation of Foreigner's Name using R Language

專題組員: 李蕙汝、高慈敏

專題編號：PROJ2018-BIOINFO-104005

指導老師:董其樺老師

1.摘要

現今國際化的世界，使用漢語的外國人日漸增多。為了讓外籍人士能在華語地區享有在地化的中文姓名，本專題構想出可幫助外籍人士取中文姓名。專題裡用 R 語言撰寫類神經網路的多層感知器模型，並使用 shiny 製作一個網頁，當使用者輸入他們的英文姓名後，藉由預先學習的類神經網路姓名翻譯模型，系統會輸出與使用者外語姓名相關聯的中文字。我們希望藉由專題，讓外籍人士能在使用漢語的地方可入境隨俗，減少隔閡感，更易融入當地文化。

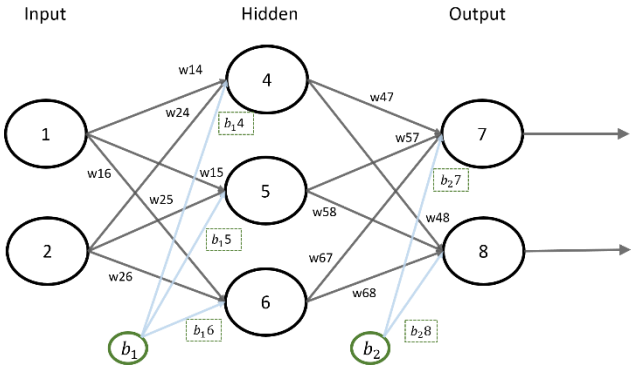
2.簡介

現今社會日漸發達，國與國之間的交流也越來越頻繁，再加上中國市場的崛起，吸引了不少外籍人士的注意。為了能在華人生活圈中溝通時能順暢無阻，所以現在來學習中文的外國人越來越多，讓自己取一個中文名字來代表自己，好方便與人交流。

根據「英語姓名的命名與知識」這本書中所述，我們了解到這些名字的涵義和特徵，有助於讓我們更深入探討姓名英漢翻譯之關係，以建立更精準的姓名含意。這本書是收錄英語民族中的不同的民族語言在這本書中有很明確地給我們對姓名解釋[1]。另外一本參考著作「取個有意思的英文名字」中，探討著各個不同國家取名的來源[2]。書中作者提及，取一個英文名字要道地，要符合世界習俗的道理就跟我們華人取中文名字一樣，從中也講述了各個國家在取名上的歷史與規則。此外，非洲人取名的習俗，他們取名的時候會依照父母行為或一種狀態來做取名的根據，所以我們用來打招呼的用語，像是「您好」或「再見吧」都可能變成他們的名字。在澳洲土著上，他們也有分小名與大名，分別代表著不同的意思，也因為他們的傳統，名字的涵義都代表著他們各自的圖騰。綜合以上幾點關於姓名取名的知識，或許皆可用於本專題想要開發的翻譯系統，讓各式各樣不同的英文字，能以意譯或音譯翻譯出有所呼應的中文字。

本專題使用 R 程式語言來進行，實作機器學習-類神經網路 (多層感知 Multilayer perceptron, MLP)[3-6]。

多層感知器的結構為:輸入層、隱藏層及輸出層三層的構造，其中隱藏層的層數，是根據資料的狀況調至 n 個隱藏層，而這些輸入的資料都是一個信號點，從輸入層到輸出層，共會經過多次的權重運算，運作原理如圖一。



圖一、類神經網路多層感知器之架構

$$h_k = \sum_{i=0} w_{ik}x_i + b$$

首先我們設

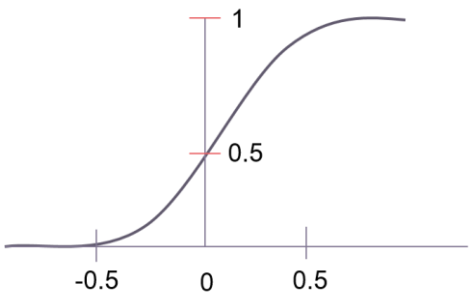
輸入層為 i；隱藏層為 h;輸出層為 o;

權重值為 w; 偏差值為 b

輸入層到隱藏層，值為h_k，k=1…n，是輸入訊號的加權值，而w_{ik}代表 i 個輸入層到第 k 個隱藏層的權重

得到 Hidden node 的輸出值 h_k

接著把的h_k的輸出結果設為 S_k，將上式所得到的h_k值帶入 sigmoid function 如圖二所示，所得之S_k值為h_k的輸出值，公式如下:



圖二、類神經網路中所用的 sigmoid function

$$S_k = \frac{1}{1 + e^{(-h_k)}}$$

隱藏層到輸出層的計算方法如同輸入層到隱藏層的概念，運用隱藏層的輸出值S_k，進行到最終 output 的權重運算，並算出整個 output 的最終結果。

$$O_j = \sum_{k=0} w_{kj}s_k + b$$

最後再經由反向傳遞(Backward propagation)，把最後的函數針對誤差進行權重及偏差的調整，誤差值越大則表示學學習狀況不好，並持續調整誤差值，使得各節點的參數收斂到最佳，這個過程就是我們所謂的「機器學習」。

當我們程式碼編輯完後，代入專為 R 打造的網路應用框架 shiny 來撰寫網頁的前後端，呈現出視覺化的互動窗口[7-9]，藉由 R 語言的程式和資料的集結，呈現出中文姓名翻譯網頁，提供外籍人士使用者可輸入英文姓名，獲得中文姓名的命名參考，從中也能增添取名字時的趣味性，藉此把這樣的方式分享給更多人使用，來促進國家之間的交流。

此英中姓名翻譯之查詢系統網頁，其畫面呈現，如圖三所示。使用者在輸入畫面裡分別輸入姓名之英文字。接著，系統會輸出最佳的中文姓名結果在畫面上，提供使用者參考。

圖三、英中姓名翻譯之查詢系統網頁

3. 專題進行方式

I. 資料的準備及整理

R 語言主要用於統計分析、繪圖、資料探勘，運用這些特質來進行本專題的製作。我們是根據書籍與網路上的內容[1, 2, 10, 11]，進行中英文姓名對照資料的收集和彙整。

	A	B
1	Ename	Cname
2	Abigail	艾碧蓋娥
3	Ada	艾妲
4	Adah	艾妲
5	Adela	阿德拉
6	Adelaide	阿德萊德
7	Adelina	艾德萊娜
8	Adeline	艾德琳
9	Adriana	艾德莉安娜
10	Afra	阿芙拉

圖四、資料的收集和彙整

II. 套件載入和讀檔

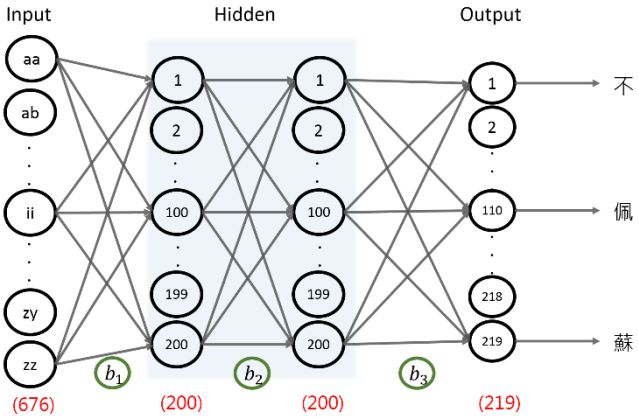
由於我們專題是運用類神經網路的概念去執行，所以在執行程式前必須先載入（neuralnet）及相關套件，載入完畢後才能做後續跑程式的動作。

```
#(1)載入套件及資料
library(stringr)
library(R.oo)
library(tidyverse)
library(NeuralNetTools)
library(DMwR)
library(nnet)
library(reshape)
library(devtools)
library(scales)
library(ggplot2)
library(readr)
library("neuralnet")
source('C:/Rdata/plotnnet.R')
names_input<- read.csv("C:/Rdata/firstname.csv",header=TRUE,sep=",")
```

圖五、R 語言套件載入

III. 訓練模型資料階段

專題運用了類神經網路概念來進行資料的訓練[3-6]。我們最後所採用的多層感知器架構為，輸入層共676個節點(雙字母的組合數)、兩層隱藏層各200個節點、輸出層則為235/219/209個節點(依男性女性名字與姓氏的中文字而有所不同)如圖六。



圖六、女生多層感知器架構示意圖

在模型訓練之前必須把讀取進來的英文名字與對應之中文名字做逐一的拆解及統計，在過程中，因為英文名字普遍較長的關係，我們把英文26個英文字母拆成 aa 到 zz 的雙字母組合來進行統計的動作。例如，英文名 Adela 會分別統計成 ad 一次、de 一次、el 一次、la 一次，其餘組合為零次。處理完畢後，每組名字各可以獲得一組中英出現次數的統計結果。

```
      Ename       Cname aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax
1     Abigail     艾碧蓋娥  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   ay az ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz ca cb cc cd
1   0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz da db dc dd de df dg dh di dj
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   yg yh yi yj yk yl ym yn yo yp yq yr ys yt yu yv yw yx yy yz za zb zc zd ze zf zg zh zi zj zk zl
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   zm zn zo zp zq zr zs zt zu zv zw zx zy zz 不 洛 瑟 姆 德 雯 丹 妮 絲 內 娥 莉 莎 塔 奧 咪 黛 切
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
   諾 厄 曼 特 露 舒 拉 尤 雅 金 娜 西 菲 蜜 巴 卡 普 比 阿 麗 克 加 布 艾 兒 門 琳 珊 倫 古 葳 史
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
   雄 尼 珂 旭 姬 黛 琪 弗 吉 黛 勒 蕾 本 瓦 婷 萊 伊 凡 潔 希 多 蓓 芙 蛙 茉 琴 梁 爾 芝 貝 登 墨
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   亞 白 安 瑪 麗 托 婉 東 波 密 朱 思 葉 米 森 雪 凱 達 碧 黛 提 葛 亨 伯 佐 里 斯 佩 利 坎 佳 沃
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0  0  0  0  0
   肖 辛 芭 茵 邦 欣 帕 霞 梅 旺 昆 明 林 法 馬 南 哈 奎 威 蓮 敏 查 緹 邁 柯 珀 珍 科 業 約 芬 英
1   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

圖七、中英出現次數的統計結果

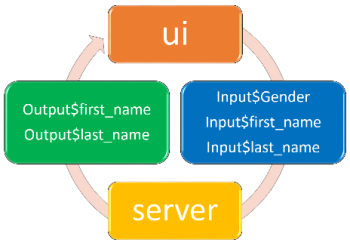
本專題預計可針對男性與女性個別給予英文姓名之翻譯，因此我們在訓練資料的過程是將男女生名字及姓氏分開做訓練，最終會獲得三組分別為男生名、女生名與不分性別之姓氏的訓練模型。一開始我們會從原始資料中(男生資料為1018筆；女生資料為949筆；姓氏資料為324筆)，經過隨機篩選平分的方式將這些資料分成獨立的八成以及另外兩成，其中取八成的資料（男生共807筆；女生共750筆；姓氏共273筆）當作多層感知器之訓練資料而另外兩成（男生共211筆；女生共199筆；姓氏共51筆）作為訓練過後之獨立測試資料。在資料訓練完畢後，便能用測試資料來進行模型評估，讓我們可以知道機器學習的狀況。

```
set.seed(1117)
#取出樣本數的idx
t_idx <- sample(2,nrow(names_input),replace=TRUE,prob = c(0.8,0.2))
#訓練組樣本
traindata <- names_input[t_idx==1,]
#訓練模型
bpn <- neuralnet(formula = f, data = traindata, hidden = c(200,200),
                 learningrate = 0.01, threshold = 0.01,
                 lifesign.step = 1000, stepmax = 1e+07, lifesign = "full")
```

圖八、資料進行訓練之程式碼

IV. 查詢介面工作設計

R 語言中，有 shiny 套件可提供數值統計後的資料可視化。在 shiny 裡可分為兩個部分，一個是可以讓使用者在瀏覽器看到的網頁，稱之為 ui，另一個則是在背景執行 R 的程式碼，這部份稱為 server[7-9]。



圖九、R 語言 shiny 互動式介面架構

4. 主要成果

經過一連串的訓練後，我們把測試資料放入訓練模型進行預測，結果會顯示實際值及預測值之數據，並以 excel 檔的顯示，部分成果如下圖十所示。

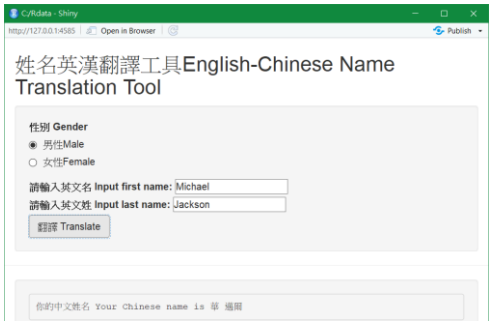
實際	翻譯結果
Adeline (艾德琳)	琳德
Clarice (克拉麗絲)	絲麗
Faustine (福絲婷)	福絲
Judy (朱蒂)	蒂朱
Juliana (朱莉安娜)	安娜
Juliet (茱麗葉)	朱麗
Letitia (莉蒂希雅)	蒂雅
Maria (瑪麗亞)	麗瑪
Marian (瑪麗安)	瑪麗
Marie (瑪莉琳)	瑪麗

圖十、女性名字英中翻譯結果

接著我們根據這樣的結果進行探討與分析。發現在這個測試資料集的結果中，可以統計出(男生名字約29%；女生名字約52%;姓氏約37%)的姓名翻譯，可正確翻譯出英文原名對應的中文字。但在這些準確率高的資料中，也有少數幾個名字輸出的結果中呈現前後顛倒的狀況，例如 Maria(瑪麗亞)在我們的系統中翻譯成了麗瑪。

然而除了準確翻譯的結果以外，也有部分是翻譯較不好的群體，原因是因為在訓練模型時是依照英文 aa~zz 之組合來進行的，我們認為在訓練資料不夠充足之下，可能導致某些英文雙字母的組合在訓練模型時沒有被訓練到，因而預測成果較差。例如，原名 Daryl(達麗兒)翻譯結果為「娜莉」，這是依據 da,ar,ry,yl 來決定中文名字，而根據分析，這四個雙字母中，所對應到的中文字剛好最多的正是「娜」字。因為在訓練資料「娜」的中文字與 da,ar,ry,yl 等雙字母的平均出現次數，是多於「達」、「麗」、「兒」等字。此外，平均次數第二多則為「莉」，因此 Daryl 才會被翻譯成為娜莉。

類神經網路訓練完後，再利用 shiny 套件將訓練後的模型帶入其中，建立出會呈現互動式姓名翻譯的視窗介面。在這個介面中，使用者可輸入其性別以及其姓名，按下翻譯後即可得到一組中文姓名。如圖十一所示，以 Michael Jackson 為例，依照網站上的性別點選男性、英文名與姓氏分別填入後，翻譯即可得到一組屬於他的中文名「華邁爾」。



圖十一、英中姓名翻譯查詢結果

5.評估與展望

藉由本次專題的探討，我們希望幫助外籍人士在取中文名字的時候，是有根據性的，而不是漫無目的。另一方面，我們希望向全世界推廣中國文化，協助大家取個有趣又好聽的名字時，了解漢字的美麗之處。

雖然本專題已實現輸入英文姓名顯示其翻譯的中文名字，但目前呈現出來的中文名與實際取名結果是有落差，且名字選擇性太少，網站的介面又太簡陋。

希望此研究可以把姓名呈現的資料做的更完善更精確，因為我們發現出利用雙字母組合來訓練且類神經網路太淺層，得出來的效果不是很理想，所以未來我們採取音韻或音節的分割和 RNN 遞迴式類神經網路，來作為訓練的依據，使翻譯出來的中文字可以更優美有意境。不但輸出應不只一組姓名可供選擇外，並且可顯示中文字的涵義，讓外國人可以自行選擇。

6.結語

我們為了讓外籍人士能在中國或台灣享有獨特的中文姓名，本專題運用 R 語言程式軟體，實作出類神經網路之多層感知器，並結合 shiny 套件將訓練所得的網路模型，呈現出一個互動式視窗網頁介面，使用者可以透過這個介面，當輸入他們的英文姓名送出時，經由後端的計算處理後，便會輸出一組該外語姓名相關之中文字。

藉由本次專題讓我們了解到外籍

人士對中文姓名的認知。同時，在專題製作過程中，我們也學習到 R 程式撰寫與機器學習概念。雖然這個網頁介面有點簡陋，且名字呈現太少，未來還有很多可持續研究的空間。另外，整個研究過程中所遇到最大的瓶頸就是，當我們在執行訓練模型程式時，由於資料所需要的記憶體空間過於龐大，電腦硬體記憶空間不足所以無法儲存暫存之資料，經常導致我們在訓練模型時，無法順利地執行完畢，在這過程中花費許多時間在重跑資料。

7.銘謝

謝謝董其樺老師盡心盡力的給予我們專題上的指導，並利用少數的閒暇時間，解決我們專題遇上的困惑，在我們電腦設備不足的情況下，及時幫助我們設置一台能夠執行專題的程式，才能讓我們的專題研究能順利進行，也謝謝組員們能在遭遇困難時，互相討論與扶持，讓專題能夠順利的完成。

8.參考文獻

[1] 准魯(2001)。英語姓名的命名與知識。笛藤出版圖書有限公司。

[2] 采詩(2017)。取個有意思的英文名字:中華文化名人英文名字三百六十家。台中市：白象文化事業有限公司。

[3] 尹相志(2009)。SQL Server 2008 Data Mining 資料採礦(第二篇 資料採礦演算法-第06章｜類神經網路)。悅之文化。取自 http://www.delightpress.com.tw/bookRead/skud00013_read.pdf

[4] skydome20(2016年5月23日) R 筆記 -(8)類神經網路(neuralnet)。取自 https://rpubs.com/skydome20/R-Note8-ANN

[5] StanleyJui(2016年12月27日)Day27 R 語言機器學習之類神經網路。取自 https://ithelp.ithome.com.tw/articles/10187683

[6] 類神經網路跟 Backpropagation 一點筆記(2016年11月23日)。取自 http://terrence.logdown.com/posts/1132631-neural-networks-with-backpropagation-one-notes

[7] 服務科學的分子廚房 MOLECULAR SERVICE SCIENCE (2015年8月1日)(統計 R 語言實作筆記系列 - 用 SHINY 套件極速打造你的商業智慧分析網站！)。取自 https://molecular-service-science.com/2015/08/01/r-shiny-business-intelligence-tutorial/

[8]Taiwan R User Group(2013年9月)。成人雜誌(R 講題分享-利用 R 和 Shiny 製作網頁應用)。取自 https://programmermagazine.github.io/201309/htm/article6.html

[9]Shiny from R Studio。取自 https://www.boca.gov.tw/sp-natr-singleform-1.html

[10]中國姓氏排名(2018年11月5日)。取自 https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%9B%BD%E5%A7%93%E6%B0%8F%E6%8E%92%E5%90%8D

[11]外交部領事事務局(外文姓名中譯英系統)。取自 http://shiny.rstudio.com/gallery/custom-input-bindings.html