Proj4-NB

Yifei Tang yt2547

2017/4/13/

In this file, we implement methods for Author Disambiguation problem in paper2 (Naive Bayes Model). We will first give an overall instruction about our process of Naive Bayes Method based on Nameset "AGupta", and show the results of all the Namesets at the end.

## Step 0: Load the packages, specify directories

setwd("C:/Users/yftang/Documents/GitHub/Spr2017-proj4-team-5/lib/")  
library(stringr)

## Step 1: Load and process the data

source("DataCleaningTZ.R")  
n=1 # Nameset "AGupta"  
num.paper<-length(data\_list[[n]])  
 auther.id<-NULL  
 coauther<-NULL  
 for(i in 1:num.paper){  
 auther.id[i]<-data\_list[[n]][[i]][[1]]  
 coauther<-unique(c(coauther,data\_list[[n]][[i]][[3]]))  
 }  
 a.i<-factor(auther.id)  
 uni.auther<-length(levels(a.i))  
 levels(a.i)<-seq(uni.auther)  
 auther.id<-a.i  
 df<-data.frame(auther.id)  
   
 num.coauther<-length(coauther)  
 get.coauther<-function(i){  
 coauthers<-rep(0,num.coauther)  
 for(j in 1:length(data\_list[[n]][[i]][[3]])){  
 coauthers[which(data\_list[[n]][[i]][[3]][j]==coauther)]<- 1  
 }  
 return(coauthers)   
 }  
 ##########################################################  
 #### build dataframe for auther coauther information######  
 ##########################################################  
 id\_co<-NULL  
 for(i in 1:num.paper){  
 id\_co<-rbind(id\_co,get.coauther(i))  
 }  
 colnames(id\_co)<-coauther  
 df<-cbind(auther.id,id\_co)   
 #df contains the auther coauther information in nameset 1: AGupta  
 id.num.co<-apply(df[,-1],1,sum)  
 df<-as.data.frame(cbind(id.num.co,df))  
 head.matrix(df)

## id.num.co auther.id B Kvande I Levinstein K Maly M Olson R Mukkamala  
## 1 8 10 1 1 1 1 1  
## 2 7 10 0 0 1 0 1  
## 3 8 10 1 1 1 1 1  
## 4 6 10 0 0 0 0 0  
## 5 6 10 0 0 1 0 1  
## 6 8 10 0 0 1 0 0  
## R Chambers R Whitney S Nanjangud C Vemuru H Syed H Abdel-Wahab M Kholief  
## 1 1 1 1 0 0 0 0  
## 2 0 0 0 1 1 1 1  
## 3 1 1 1 0 0 0 0  
## 4 0 0 0 0 0 1 0  
## 5 0 0 0 1 1 1 0  
## 6 0 0 0 0 0 1 0  
## S Dittakavi A Youssef C Overstreet E Stoica J Wild K Mary A Prabhu  
## 1 0 0 0 0 0 0 0  
## 2 1 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 1 1 1 1 1 0  
## 5 0 0 0 0 0 0 1  
## 6 0 0 1 0 1 0 0  
## A Gonzalez A Hamid S Ghanem X Zhu S Mynam E Al-Shaer M Kumar  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 1 1 1 1 0 0 0  
## R Srivastava M Moran R Bettati R Oppliger D Ferrari Q Nguyen R Yavatkar  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## W Heffner W Howe G Ventre K Rothermel S Ghosh S Pemmaraju C Mazumdar  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## N Gururaj M Kortright R Borie V Madhu P Singhal A Nandi H Paul  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## R Badrinath S Banerjee A Basu U Deshpande T Herman S Ranjan A Meka  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## A Chaturvedi S Shekhar M Karaata M Dubuisson-Jolly T O'Donnell C Liang  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## T Boult B Bhanu D Michael G Funka-Lea R White S Halliburton S Aharon  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## J Chew X Wang B Lee M Shamsuzzaman N Georganas L Orozco-Barbosa B ukic  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## P Konana D Stahl A Whinston B Jukic M Li M Parameswaran D Morrice  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## R Bapna P Goes Y Jin R Chellappa G Gupta A A A Chen J Marsden R Gopal  
## 1 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0 0  
## L Linden G Adomavicius G aruga G Karuga B Su Z Walter Y Tung A Biere  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## E Clarke P Williams J Kukula O Shrichman A Gupta Z Yang P Ashar  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## R Suryanarayan T Blalock J Hayes I Avila-Campillo T Green M Onizuka  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## D Raven D Suciu S Chaudhuri V Narasayya A Halevy K Chakraborty  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## P Mazumder M Bhattacharya S Kulkarni L Whitman R Agarwal B Verghese  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## M Rosenblum R Chandra S Devine J Chapin S Herrod W Weber D Cheriton  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## H Goosen P Boyle D Lenoski J Laudon J Hennessy K Gharachorloo P Gibbons  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## A Colburn I Tashev J Cadiz L He R Cutler S Silverberg Y Rui Z Zhang  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## Z Liu T Mowry D Nakahira L Stevens T Joe S Woo M Ohara E Torrie J Singh  
## 1 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0 0  
## M Horowitz M Lam M Martonosi T Anderson P Stenstr&ouml m m M Tambe  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## D Teodosiu T Lahiri A Tucker J Torrellas S Urushibara C Forgy A Newell  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## R Wedig V Soundararajan J Grudin Z Hakura D Ofelt J Kuskin J Baxter  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## J Heinlein M Heinrich R Simoni Q Liu P Nayak P Rosenbloom D Kalp  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## B Milnes E Rothberg E Sanocki R Bosch S Dresser A Acero A Acharya  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## E Rothburg A Agarwal A Agrawal N Chaddha T Meng A Balachandran G Jancke  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## F Li F Barachini H Mistelberger D Bargeron M Hill S Adve C Holt  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## T Totsuka G Venolia G Kimberly H Chesley S White N Omoigui M Levoy  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## B Smith D Douglas D Maier F Darema J Ratner O Lubeck P Messina  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## L Soul&eacute A Brush E Bugnion E Witchel A Kothari D Agrawal S Suri  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## A Abbadi M O'Malley D Siewiorek A Brennan P Edmond W Birmingham W Hwu  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## W Fuchs I Akyildiz R Fujimoto A Wong G Kim R Sriram S Gorti M Vemulapati  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## R Sekar A Tiwari H Yang J Frullo S Zhou T Shanbhag M Bhide M Mohania  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## S Pandey T Weymouth R Jain B Lud鋝cher M Martone X Qian J Tran  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## I Zaslavsky S Santini J Grethe A Hampapur B Horowitz C Fuller C Shu  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## J Bach M Gorkani R Humphrey M Wong G Sosinsky S Lamont M Ellisman  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## B Ludascher S Zhang R Moore C Baru A Rajasekar R Marciano M Wan  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## W Schroeder S Bhonsle A Katkere J Schlenzig P Kelly P Alarcon  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## A Pascual-Montano J Carazo B Lud&auml s scher P Alarc髇 I Altintas  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## S Bhagwanani D Buttler S Chandra M Coleman T Critchlow W Han L Liu C Pu  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## A Shoshani M Vouk T Wang Y Papakonstantinou P Velikhov V Chu  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## Y apakonstantinou Z Cheng M Worring A Smeulders M Trivedi G Burns  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## K Stephan M Chagoyen M Kurul P De-Alarcon P de-Alarcon M Goldbaum  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## V Kouznetsova D Nazafi E Ross H He S Moezzi A Taylor S Chatterjee  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## S Burgess Y Yang A Bagchi A Ray B Chanda S Tambawala H Hampapur R K&ouml  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## t tter J Ullman J Widom Y Sagiv D Quass V Harinarayan I Mumick  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## J Naughton P Deshpande R Ramakrishnan R Agrawal S Agarwal S Sarawagi  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## H Jagadish D Katiyar K Ross V Subrahmanian NA A Rajaraman L Haas J Rao  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## J Jos&eacute A Blakeley S Tiwari H Garcia-Molina A Swami H Young  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## N Nishimura D Kaller S Mahajan T Shermer A Proskurowski P Ragde  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## R Krishnamurti R Impagliazzo J Delgrande T Allen H Everett D Gaur  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## A Barranco-Mendoza S Buss S Cook V Ramachandran F Dehne J Sack  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## R Tamassia F Gustavson M Joshi S Toledo V Kumar A Karaivanov  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## J Wasniewski P Yalamov G Karypis D Moran G Ditlow R Williams T Wilkins  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## Y Muliadi P Rohatgi A Sameh Gupta A Kumar M Thorup R Rastogi C Chekuri  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## D Raz J Naor T Roughgarden M P&aacute l l B B&uuml l Yener J Kleinberg  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## J Lee R Krauthgamer & &Eacute v Tardos A Sinclair I Newman Y Rabinovich  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## F Zane A Andoni M Deza P Indyk S Raskhodnikova A Chakrabarti S Dasgupta  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## <U+FFFD> Tardos A Srinivasan K Dhamdhere R Ravi M Pal B Yener A Photiou  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## A Boals J Hashmi N Sherwani C Fan J Liu H Wang S Hambrusch E Doncker  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## K Rathbun P Ealy D Nelson A Genz J Ball S Bruell G Greenwood K McSweeney  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## W Maass D Motzkin S Malik A Fisher A Mukaiyama A Casavant K Wakabayashi  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## S Liu X Liu A Raghunathan S Bhattacharya C Wang M Ganai L Zhang M Reddy  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## M Siegel B Prasad P Reddy S Madnick F Pe&ntilde a a-Mora J Yeung  
## 1 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0  
## S Vadhavkar B Thuraisingham C Clifton E Ferrari E Bertino K Bansal  
## 1 0 0 0 0 0 0  
## 2 0 0 0 0 0 0  
## 3 0 0 0 0 0 0  
## 4 0 0 0 0 0 0  
## 5 0 0 0 0 0 0  
## 6 0 0 0 0 0 0  
## A Dhond J Kok P Putten M Butler M Feng P Huck K Hussein L Granowetter  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0  
## P Wang A Liu M Nagendraprasad S Ayyadurai L Mui  
## 1 0 0 0 0 0  
## 2 0 0 0 0 0  
## 3 0 0 0 0 0  
## 4 0 0 0 0 0  
## 5 0 0 0 0 0  
## 6 0 0 0 0 0

## Step 2: Naive Bayes Model

We assume that each author’s citation data is generated by the naive Bayes model, and use his/her past citations as the training data to estimate the model parameters. Based on the parameter estimates, we use the Bayes rule to calculate the probability that each name entry (, where is the total number of candidate name entries in the citation database) would have generated the input citation.

Given an input test citation C with the omission of the query author, the target function is to find a name entry in the citation database with the maximal posterior probability of producing the citation , i.e., $ max\_iP(X\_i|C)$ (1) Using the Bayes rule, the problem becomes finding: $ max\_iP(C|X\_i)P(X\_i)/P(C) $ (2) Since does not depend on Then Function (2) becomes $ max\_iP(C|X\_i)P(X\_i) $ (3) Function (3) is the target of our method.

First, we use coauthor as the attribute to do the Naive Bayes Model.

### 2.1 Attribute - Coauthor

freq.id<-as.data.frame(table(df$auther.id)/length(df$auther.id))   
 # numbers of paper for each id  
 colnames(freq.id)<-c("auther.id","Freq")  
 num.id<-dim(freq.id)[1]

Split the Nameset into Train set and Test set

set.seed(29)  
sample\_split<-function(i){  
 Data<- df[df$auther.id==i,]  
 bound <- floor(nrow(Data)/2) #define 50 % of training and test set  
 Data <- Data[sample(nrow(Data)), ] #sample rows   
 df.train <- Data[1:bound, ] #get training set  
 df.test <- Data[(bound+1):nrow(Data), ] #get test set  
 return(list(df.train=df.train,df.test=df.test))  
}  
   
train<-NULL  
test<-NULL  
for(i in seq(num.id)){  
 train[i]<-list(sample\_split(i)[[1]])  
 test[i]<-list(sample\_split(i)[[2]])  
}   
   
TT<-NULL  
for(i in 1:num.id){  
TT <- rbind(TT,test[[i]])  
}  
test\_x<-TT[,-c(1,2)] # test set as data.frame  
test\_y<-TT[,2] # test lable

Then, we use the train set to train the required probabilities

NB\_para<-function(data){  
 # Input: df.train   
 # Output: 6 probability used for NB model  
 paper.num<-dim(data)[1]  
 p\_0<-sum(data$id.num.co==0)/paper.num   
 #### P(N|X)  
 p\_1<-1-p\_0   
 #### P(Co|x)  
 colsum<-apply(data,2,sum)  
 data<-rbind(data,colsum)  
 p\_s\_cx<-sum(colsum>=2)/sum(colsum>=1)  
 #### P(Seen|Co,X)  
 p\_u\_cx<-1-p\_s\_cx  
 #### P(Unseen|Co,X)  
 total.num.co<-colsum[1]  
 p\_a\_scx<-colsum/total.num.co   
 p\_a\_scx<-p\_a\_scx[-c(1,2)]   
 #### P(A|Seen,Co,X)  
 p\_a\_ucx<-1/(num.coauther -sum(colsum>=1))   
 #### P(A|Unseen,Co,X)  
 outcome<-list(P.N.X = p\_0,  
 P.Co.X = p\_1,  
 P.S.Co.X = p\_s\_cx,  
 P.U.Co.X = p\_u\_cx,  
 P.Ak.S.Co.X = p\_a\_scx,  
 P.Ak.U.Co.X = p\_a\_ucx)  
 return(outcome)  
}  
   
 P.N.X<-NULL  
 P.Co.X <-NULL  
 P.S.Co.X<-NULL  
 P.U.Co.X<-NULL  
 P.Ak.S.Co.X<-NULL  
 P.Ak.U.Co.X<-NULL  
 for(i in seq(num.id)){  
 P.N.X[i]<-NB\_para(train[[i]])$P.N.X  
 P.Co.X[i]<-NB\_para(train[[i]])$P.Co.X  
 P.S.Co.X[i]<-NB\_para(train[[i]])$P.S.Co.X  
 P.U.Co.X[i]<-NB\_para(train[[i]])$P.U.Co.X  
 P.Ak.U.Co.X[i]<-NB\_para(train[[i]])$P.Ak.U.Co.X  
 P.Ak.S.Co.X[i]<-list(NB\_para(train[[i]])$P.Ak.S.Co.X)  
 }  
 freq.x<-NULL  
 for(i in seq(num.id)){  
 freq.x[i]<-dim(train[[i]])[1]   
 }  
 px<-freq.x/sum(freq.x) # px is the prior for the nameset

Build the Naive Bayes Model and use test set to do the prediction and calculate the accuracy rate.

NB\_model<-function(data){  
 # input: Test datafram  
 # data.df=test\_x  
 # out<-NULL  
 # for(m in seq(dim(data.df)[1])){  
 # data = data.df[m,]  
 if(sum(data)==0){  
 out<-which.max(P.N.X)  
 }  
 else{  
 condition<-function(id){  
 i=id  
 # i=id, k = k-th coauther  
 P\_Ak\_X<-NULL  
   
 significant<-length(P.Ak.S.Co.X[[i]][data>0])  
   
 for(k in seq(significant)){  
 P\_Ak\_X[k]<-P.Co.X[i] \* ( P.S.Co.X[i] \*   
 P.Ak.S.Co.X[[i]][data>0][k] +   
 P.Ak.U.Co.X[i] \* P.U.Co.X[i])  
 }  
 return(prod(P\_Ak\_X))  
 }  
 target<-NULL  
 for(i in seq(num.id)){   
 target[i]<-condition(i)\*px[i]  
 }  
   
 out <-which.max(target)  
 }  
 return(out)  
 }  
 est\_y<-apply(test\_x,1,NB\_model) ## test estimate  
   
 accuracy<-sum(est\_y==test\_y)/length(test\_y)  
 accuracy

## [1] 0.902027

## Step 3 Outcomes for all nameset

source("NB\_Co.R")  
  
summary.all

## Nameset mean.accuracy StdDev.accuracy  
## 1 A Gupta 0.9077703 0.01788024  
## 2 A Kumar 0.7919355 0.02242651  
## 3 C Chen 0.8081928 0.02125709  
## 4 D Johnson 0.8053476 0.04941182  
## 5 J Lee 0.7986486 0.01655061  
## 6 J Martin 0.7844828 0.06514818  
## 7 J Robinson 0.8534091 0.03278212  
## 8 J Smith 0.7951064 0.01085130  
## 9 K Tanaka 0.9062937 0.01899445  
## 10 M Brown 0.8337500 0.03283481  
## 11 M Jones 0.7526316 0.02309303  
## 12 M Miller 0.9290476 0.01317556  
## 13 S Lee 0.7934840 0.01377755  
## 14 Y Chen 0.8664625 0.01350176