数据挖掘课程报告

二级标题

吕梦天 & 楚梦颖

# 项目概述

萧敬腾作为一名出色的歌手，同时被人冠宇“雨神”的称号。我们便通过萧敬腾的演唱会的行程与当地当天的天气进行对比，来验证“雨神”的称号是否合理。

# 数据收集与整理

## 2.1演唱会行程

项目开始之前，我们现在网上爬取了萧敬腾2011年至2016年间的国内演唱会的行程作为我们的初始数据集，然后通过每一条数据的日期和地点再网上爬取对应的天气，最终将两个数据集合并成一个数据集，如图1。

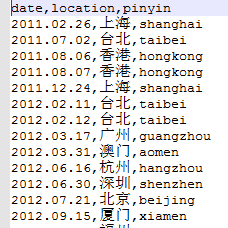


图1

合并数据集的代码如下：

library(readr)

mergeWeatherAndLocation <- function(){

#读取城市列表，调用爬虫函数，合并数据保存到一个文件

filepath = chartr("/","\\",getwd())

xiaojingteng<-read.csv(file=paste(filepath,"\\res\\xiaojingteng.csv",sep=""),header=TRUE,fileEncoding="utf-8",encoding="utf-8")

size = length(xiaojingteng$pinyin) #get row number

i = 1

result <- c()

for(i in 1:size){

wdata = getWeatherByCityNameAndDate(city=xiaojingteng$pinyin[i],date=xiaojingteng$date[i])

result <- append(result,c(wdata))}

wdataResult <- data.frame(matrix(result,nrow = 40,ncol = 6,byrow = TRUE))

names(wdataResult) <- c("high","low","weather","winddirec","windlevel","isRainy") #为数据集命名

w <- cbind(xiaojingteng,wdataResult)

write.csv(w,file=paste(filepath,"\\weather\\xiaojingtengwithweather.csv",sep=""),row.names=TRUE,fileEncoding = "utf-8")

}

library(RCurl)

library(XML)

#通过城市拼音和日期查询天气

getWeatherByCityNameAndDate <- function(city="beijing",date="2011.01.01"){

row <- substr(date,9,10)

row <- as.numeric(row) + 1 #get the rowth number

date <- paste(substr(date,1,4),substr(date,6,7),sep="")

ulchild <- paste("ul:nth-child(",row,")",sep="")

url <- paste("http://lishi.tianqi.com/",city,"/",date,".html",sep="");

freak <- html\_session(url)

table <- freak %>% html\_nodes("div[id='tool\_site']")

date <- table %>% html\_nodes(ulchild) %>% html\_nodes("li:nth-child(1)") %>% html\_text() #日期

high <- table %>% html\_nodes(ulchild) %>% html\_nodes("li:nth-child(2)") %>% html\_text() #最高气温

low <- table %>% html\_nodes(ulchild) %>% html\_nodes("li:nth-child(3)") %>% html\_text() #最低气温

weather <- table %>% html\_nodes(ulchild) %>% html\_nodes("li:nth-child(4)") %>% html\_text() #天气

windDirec <- table %>% html\_nodes(ulchild) %>% html\_nodes("li:nth-child(5)") %>% html\_text() #风向

windlevel <- table %>% html\_nodes(ulchild) %>% html\_nodes("li:nth-child(6)") %>% html\_text() #风力

isRainy <- grep("雨",weather)

if(length(isRainy) > 0){

isRainy = "有雨"

}else{

isRainy = "没雨"

}

cbind(high,low,weather,windDirec,windlevel,isRainy)

}