IR Assignment #3 **Social Network Clustering Analysis**

กำหนดส่ง : 30 เมษายน 2561 ภายในเวลา 23:45 น. ตามเวลาของระบบ

ข้อมูลที่ใช้ในการบ้านนี้จะเป็นข้อมูลที่ได้มาจากนักเรียนระดับ high school ของอเมริกาที่มี profiles บน Social Network ชื่อดัง ตั้งแต่ปี 2006-2009 จำนวน 30,000 คน โดยเก็บคำยอดนิยมจำนวน 500 คำที่ปรากฏใน pages ต่าง ๆ ของทุกคน จาก 500 คำนี้จะถูกเลือกมาเพียง 36 คำ ซึ่งเป็นคำที่สามารถแสดงความสนใจของบุคคล ใน 5 ด้านด้วยกันคือ กิจกรรมนอกเวลาเรียน (extracurricular activities), แฟชั่น (fashion), ศาสนา (religion), ความรัก (romance) และพฤติกรรมทางสังคม (social behavior) ตัวอย่างคำใน 36 คำที่เลือกว่า เช่น football, sexy, kissed, bible, shopping, death และ drugs

ข้อมูล dataset นี้อยู่ใน file ชื่อ snsdata.csv ซึ่งสามารถเปิดดูด้วย MS Excel โดยหนึ่งแถวแสดงข้อมูลของนักเรียน 1 คน ดังภาพ

	A	B	C	D	E	F	G	H	I	J	K
1	gradyear	gender	age	friends	basketball	football	soccer	softball	volleyball	swimming	cheerleading
2	2007	F	18.023	18	24	0	0	0	0	0	0
3	2006	F	18.932	2	22	2	0	1	4	0	0
4	2007	F	17.358	59	13	1	1	0	1	0	0
5	2006	F	18.261	20	12	3	0	0	0	0	0
6	2006	F	19.159	2	12	0	0	0	0	0	0
7	2008	F	16.537	47	12	0	0	0	2	1	0

ข้อมูลในแต่ละ column มีรายละเอียด ดังนี้

Column A-D แสดงข้อมูลของนักเรียน

gradyear ปีที่สร้าง profile บน social network (2006, 2007, 2008, 2009)

gender เพศ (M-Male, F-Female)

age อายุ

friends จำนวนเพื่อนบน social network

Column E-AN แสดงคำทั้งหมด 36 คำ ที่สามารถแสดงความสนใจใน 5 ด้านที่กล่าวมา

ตัวเลขในแต่ละช่อง แสดงความถี่ของคำ ๆ นั้นที่พบบน pages ของนักเรียน

(สมมติฐานคือ ถ้าใน pages ของนักเรียนคนใดถูกพบคำใดบ่อยๆ แสดงว่านักเรียนคนนั้นมีความสนใจในด้านนั้นเช่น พบความถี่ของคำว่า basketball จำนวนมากใน page ของนักเรียนคนหนึ่ง หมายถึงนักเรียนคนนี้มีกิจกรรมนอกเวลาเป็น กีฬา basketball เป็นต้น)

คำสั่ง

1. ติดตั้งโปรแกรม R studio ซึ่งเป็น IDE ของโปรแกรม R
2. ให้นักศึกษารันทีละคำสั่งผ่าน console แล้วดูผลลัพธ์ จากนั้นจึงรวบรวมคำสั่งและ capture ผลลัพธ์ที่ได้ในแต่ละคำสั่ง (ถ้ามี) ใส่ไฟล์ เพื่อนำมาส่งในรูปแบบ pdf โดยการ upload ไปที่กล่องการบ้านบน courseweb ของวิชา
3. รายการคำสั่งที่ต้องรัน มีดังนี้
   a. โหลดไฟล์ snsdata.csv เข้ามาใน working space ใส่ข้อมูลลงใน data frame ชื่อ teens

```
# Set working directory
setwd("/Users/Wirat/Desktop/R programs")
# read dataset file in csv format into teens data frame
teens <- read.csv("snsdata.csv")
```

   b. แสดงตัวอย่างข้อมูลใน file ของ 3 คนแรก

```
head(teens,3)
```

   c. แสดงจำนวนข้อมูลทั้งหมด ได้แก่ จำนวนนักเรียน (rows) และจำนวน attributes (columns)

```
dim(teens)
```

   d. สรุปข้อมูล : จำนวน objects, variables และค่าของ variables แต่ละตัวที่ปรากฏใน file

```
str(teens)
```

   e. พิมพ์ค่า summary ของ age ได้แก่ ค่า Min, Median, Mean, Max และจำนวน NA (Not available)

```
summary(teens$age)
```

   f. คำสั่งให้เอาข้อมูลแถวที่มี missing values ออก

```
teens = na.omit(teens)
```

   g. แสดงจำนวนข้อมูลทั้งหมดที่เหลือ ได้แก่ จำนวนนักเรียน (rows) และจำนวน attributes (columns)

```
dim(teens)
```

   h. เลือกเฉพาะข้อมูล 36 attributes ในการทำ cluster analysis (เอาข้อมูล 4 columns แรกของบุคคลออก)

```
# Select data column 5-40 from teens to store in interests data frame
interests<-teens[5:40]
```

i. แปลงข้อมูลให้อยู่ในรูป standard score (z-score) เพื่อให้สามารถเปรียบเทียบ scores ที่มาจาก normal distributions ที่ต่างกันโดยใช้ scale() และ lapply()

```
# Apply z-score standardization to the interests data frame and store data in interest_z
interest_z <-as.data.frame(lapply(interests, scale))
```

j. สรุปข้อมูล : จำนวน objects, variables และค่าของ variables แต่ละตัวที่ปรากฏในชุดข้อมูลที่แปลงแล้ว

```
str(interest_z)
```

k. ทดลองแบ่ง clusters ของข้อมูลที่ได้ด้วย kmeans โดยทดลองแบ่งตั้งแต่ 2 ถึง 20 clusters โดยทดลอง 100 รอบต่อการแบ่งหนึ่งครั้ง เก็บค่า sum square errors ของทั้ง 100 รอบเพื่อหาค่าเฉลี่ย sum square errors ของแต่ละการทดลอง

```
range<-2:20 # k from 2 to 20
tries <- 100  # run k-means algorithm 100 times
avg.totw.ss <- integer(length(range)) # Set up an empty vector

for (v in range) {
        v.totw.ss <-integer(tries)  #Set up an empty vector to hold the 100 tries
        for(i in 1:tries) {
                k.temp <- kmeans(interest_z, centers=v)  #Run kmeans
                 v.totw.ss[i] <- k.temp$tot.withinss  #Store the total withinss
        }
        avg.totw.ss[v-1] <- mean(v.totw.ss)  #Average the 100 total withinss
}
```

ถ้าคำสั่งด้านบนนี้รันไม่จบให้กด Stop เพื่อ break เมื่อเห็นข้อความ warning ให้ทำคำสั่งถัดไปได้เลย

l. Plot ค่าเฉลี่ยของ within sum of squares errors ของแต่ละการทดลอง เพื่อหาจำนวน clusters (?) ที่เหมาะสมที่สุด (scree plot)

```
plot(range,avg.totw.ss,type="b", main="Total Within SS by Various K",
ylab = "Average Total Within Sum of Squares", xlab = "Value of K")
```

m. รัน kmeans อีกหนึ่งรอบด้วยจำนวน cluster ที่เหมาะสมที่สุด (?) ที่ได้จาก scree plat

```
teen_clusters <- kmeans(interest_z, ?)
```

n. แสดงผลจำนวน clusters และจำนวนสมาชิกทั้งหมดในแต่ละ clusters

```
teen_clusters$size
```

o. แสดงค่า centers ของแต่ละ clusters ที่ได้

```
teen_clusters$centers
```

p. แปลผล โดยสรุปลักษณะเด่นของแต่ละ clusters

ส่วนนี้ นักศึกษาดูจากผลที่ได้ว่าค่า attributes ใดของ cluster ที่มีความโดดเด่น และสรุปว่า cluster แต่ละอัน มีนักเรียนที่มีลักษณะคล้ายกันอย่างไร