Assignment #1 **Corpus Analysis and Word Cloud using R**

กำหนดส่ง : 20 กุมภาพันธ์ 2561

เครื่องมือที่ใช้ R Studio และติดตั้ง package ดังต่อไปนี้

- tm สำหรับฟังก์ชันที่เกี่ยวกับการทำ text mining
- SnowballC สำหรับฟังก์ชันในการ stemming
- ggplot2 สำหรับการสร้างกราฟ
- wordcloud สำหรับการสร้าง wordcloud

คำสั่ง เขียนโค้ดของภาษา R เพื่อวิเคราะห์ข้อมูลใน corpus

1. สร้าง corpus ของตนเอง เพื่อใช้ในการวิเคราะห์ โดยเริ่มจากการเลือก topic ที่สนใจ แล้วสร้าง folder เพื่อรวบรวมไฟล์ .txt จำนวนอย่างน้อย 20 ไฟล์ สำหรับ topic นั้น ๆ
2. สร้าง corpus และตั้งชื่อด้วยคำสั่ง

   ชื่อ <- Corpus(Dirsource("C:/....."))
3. แสดงข้อมูลเกี่ยวกับ corpus ด้วยการพิมพ์ชื่อ corpus
   #ตัวอย่างคำสั่งในการแสดงรายละเอียดของเอกสารฉบับที่ 15

   writeLines(as.character(ชื่อ[[15]]))
4. ทำการ pre-processing ข้อมูลใน corpus
   - แทนที่เครื่องหมาย "-", ":", "'", "?","." ด้วยช่องว่าง
   - ถ้ามีสัญลักษณ์อื่นๆ ที่ไม่ต้องการให้ปรากฏในดรรชนีหลงเหลือให้แทนที่สัญลักษณ์นั้นด้วย space (ในขั้นตอนนี้อาจต้องใช้การ inspect ข้อมูลเพื่อตรวจสอบข้อมูลว่ายังมีสัญลักษณ์อื่นใดหลงเหลือ ก็ทำซ้ำ จนหมดหรือเหลือน้อยที่สุด
   - ลบเครื่องหมายวรรคตอน (removePunctuation)
   - แปลงข้อมูลทั้งหมดเป็นตัวอักษรเล็ก (lowercase)
   - ลบตัวเลขออกให้หมด
   - ตัด stopwords
   - ลบ whitespaces (stripWhitespace)
5. ทำ Stemming กับข้อมูลใน corpus
6. สร้าง Term-Document matrix
7. แสดงผลจาก matrix ที่ได้
   - จำนวน term ทั้งหมด
   - แสดง term ที่มีค่า frequency มากที่สุด และ term ที่มีค่า frequency น้อยที่สุด
   - แสดง term ที่ปรากฏใน corpus อย่างน้อย 50 ครั้ง
8. แสดง histogram ของ term ที่มีค่า frequency เกิน 50
9. แสดงผลในรูปของ wordcloud

<u>สิ่งที่ต้องส่ง</u> : zip file ตั้งชื่อเป็นรหัสนักศึกษา ในไฟล์ประกอบด้วย

- โค้ดของภาษา R ที่สามารถทำงานได้ตามคำสั่งข้างต้น
- ผลลัพธ์ที่ได้จากข้อ 3, 7, 8 และ 9 เป็น pdf
- โฟลเดอร์ที่เก็บ ไฟล์ข้อมูลใน corpus

ส่ง zip file ในกล่องการบ้านบน courseweb ของวิชา ภายใน 12:00 น. เที่ยงวันของวันกำหนดส่ง

แหล่งอ้างอิง :

https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/

https://rstudio-pubs-static.s3.amazonaws.com/265713_cbef910aee7642dc8b62996e38d2825d.html