Document Classification Workshop

Part I: Data preparation

1. ดาวน์โหลด corpus http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz
2. Unzip และตัดคำเอกสารทั้งหมดที่อยู่ใน folder ต่อไปนี้
   - sci.crypt โดยจะถือเป็น class Crypt
   - sci.electronics โดยจะถือเป็น class Electronics
   - sci.med โดยจะถือเป็น class Med
   - sci.space โดยจะถือเป็น class Space
3. สร้าง vocabulary จากคำทุกคำที่พบใน corpus
   ** นิสิตสามารถเลือกเฉพาะคำที่พบมากกว่า n ครั้งใน corpus ก็ได้ เช่น ต้องพบมากกว่า 1 ครั้งใน corpus จึงจะนำมารวมเป็น vocabulary
4. คำนวณค่า TF และ IDF ของแต่ละคำ ดังนี้

   TF (term frequency)

   tf_{ij} คือ ความถี่ของคำที่ j ใน document ที่ i

   IDF (Inverted document frequency)

   $$IDF = \log(\frac{N}{n_j})$$

   โดย N คือ จำนวนเอกสารทั้งหมดใน corpus

   n_j คือ จำนวนเอกสารที่มีคำที่ j ปรากฏอยู่
5. สร้าง document vector โดยให้เลือก weighting scheme LTC

$$a_{ij} = \frac{\log(tf_{ij}+1.0)*\log(\frac{N}{n_j})}{\sqrt{\sum_{p=1}^{M}[\log(tf_{ip}+1.0)*\log(\frac{N}{n_p})]^2}}$$

(ดูสูตรเพิ่มเติมได้จาก http://www.ipcsit.com/vol47/009-ICCTS2012-T049.pdf)

โดย a_{ij} คือค่า LTC ของ term ที่ j ใน document ที่ i

โดย vector ของแต่ละเอกสาร แต่ละตำแหน่งใน vector จะ associate กับคำใน vocabulary

ตัวอย่างเช่น สมมติ vocab = {a b c d e f}

DocA = {a b c} → vector of TF = {1 1 1 0 0 0}

DocB = {a a b d} → vector of TF = {2 1 0 1 0 0}

DocC = {b b e f} → vector of TF = {0 2 0 0 1 1}

Part II: Classification Model Learning

1. ดาวน์โหลดและติดตั้ง Machine Learning tool ที่มี support vector machine (SVM) เช่น weka, libsvm, tinysvm, svmlight, ect.
2. สร้าง training file ให้เป็นไปตาม format ที่ tool กำหนด
3. สั่งคำสั่ง train โดยในการ train ให้ทดลอง set ค่า kernel และพารามิเตอร์ต่างๆ และให้รันแบบ 10-fold cross validation