R05323012 曾院朋

文字探勘與其在健保資料的應用 投影片導讀

I. What is the question of the paper?
   A. 前半部先從介紹文字探勘開始，爾後導入實例，最後再分別介紹資訊抽取(Information Extraction)、情感分析(sentiment analysis)、主題模型(Topic Model)等三種技術。
   B. 後半段則是探討如何使用上述技術，利用既有的健保資料：診斷(diagnose)、藥物治療(Medication)、個人基本資料(Contextual information)，建立彼此之間關係預測，協助診斷。

II. What should we care about it?
   面對數據量暴增的現代，若能快速從大樣本中獲取想要的資訊，是非常有利的；而使用在醫療上，則可有效地協助醫生作出醫療決策。

III. What is your(or auther's) answer & How did you(or auther) get there?
   A. 前半部分作者分別針對三種目的提出多種工具，以下將針對各工具及其評量方法作粗略介紹。
   B. 針對健保資料庫，作者嘗試多種模型，並以 NDCG、MAP 等指標衡量後發現 LDA (Latent Dirichlet allocation)是在各指標中表現最佳的。

IV. Models and regressions.
   A. 資訊抽取(Information Extraction)：快速地將文章內的重要詞彙分類或取出
      1. Rule-Based Approaches:利用創建公式的方法，符合公式的項目及被擷取或分類。如 *Mr.Trump* 可以利用 *Mr\.\s+([A-Z][a-z]+\s?)+*找出，這方法相較於直接從字串尋找更有彈性
      2. Sequence Labeling Algorithms:目標是快速將文章內的文字快速自動分類，利用前後文及在整體文章中的排序為依準作計算，衡量方法可以由 Precision、Recall 等指標作依據。
   B. 情感分析(sentiment analysis)：自動判斷文章的意見傾向
      1. 為了達到快速判斷的目的，首先要先判定那些文章是有效的；確立使用的素材後，區分內容文字重要度(利用 TF-IDF 模型:該文章內越常出現的字彙影響越大；但若是大家都常用的字彙影響力則會下降)；最後再以不同的依據方法(線上字庫等來源)判斷該字詞的正負意涵，最後作出整篇文章的決定。

C. 主題模型(Topic Model):判斷文檔的主題

1. latent Dirichlet allocation (LDA):將一篇文章看成是許多主題的分配；同時各個主題又有其分別的字彙分配，而 LDA 模型則是要利用既有的結果(每個文字)去預測兩個分配的參數。在健保的應用上，則是分別將診斷藥物治療、個人基本資料各自計算 LDA。