文字探勘(Text mining)

Text Mining

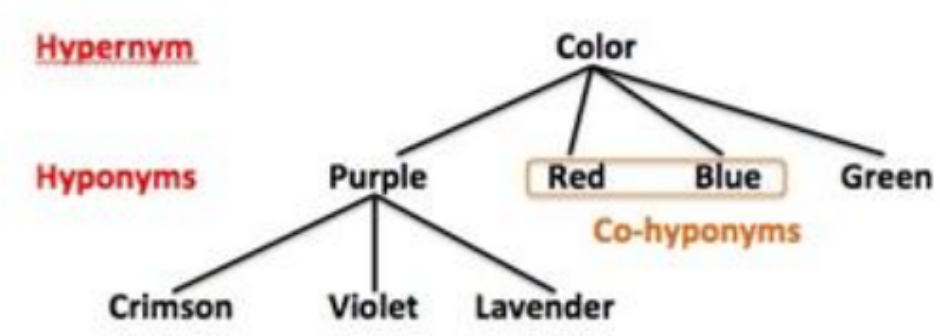
是以各種 Data Mining 方式來進行文件的文字資料分析，透過其分析來取得文字間的關聯性。與 Data Mining 不同之處，在於 Text Mining 是**針對文字進行分析，且文字多屬半結構化或非結構資料，因此要先對文字進行前處理**（Pre-Processing），**並透過某些統計方法與演算法**（例如：Term Frequency - Inverse Document Frequency，簡稱 TF-IDF），**對文字進行分析與運用**，進而取得必要的資訊，作為決策的參考依據。

前處理程序

1. Part-of-Speech Tagging：

   首**先進行詞性分析**，包括前後詞判斷，以及同義字（Synonym）、一字多義字（Polysemy）、反義字（Antonym）、泛稱（Hypernym）、具體名稱（Hyponym）...等；

   而單字可能與前後文字組成單詞（例如勞「作」、「作」業、工「作」、杵「作」、「作」文、磨杵「作」針等），因此 Text Mining 需要詞庫來進行標記（Tagging）處理。



圖一 泛稱與具體名稱之間的關聯（圖片取自維基百科）

2. Stemming：

   透過移除單字的後綴（例如：cats、catlike、catty，皆是以cat作為基礎，fishing、fished、fisher則以fish作為基礎），**將字根進行還原**。

3. Feature Selection：

   將擷取出來的單詞（Terms）進行過濾與篩選，首先**決定保留哪些詞性**的單詞（例如動詞或名詞），而後透過 TF-IDF 等統計方法或演算法，來分析單詞的頻率。

TF-IDF

是一種用於資訊檢索與文字探勘的常用**加權**技術，為一種統計方法，**用來評估單詞**對於文件的集合或詞庫中一份文件**的重要程度。**

1. TF（Term Frequency）：

   假設 j 是「某一特定文件」，i 是該文件中所使用單詞或單字的「其中一種」，n(i,j) 就是 i 在 j 當中的「出現次數」，那麼 tf(i,j) 的算法就是 **n(i,j) / (n(1,j)+n(2,j)+n(3,j)+...+n(i,j))**。例如第一篇文件中，被我們篩選出兩個重要名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 70 次，「富有」出現 30 次，那「健康」的 tf = 70 / (70+30) = 70/100 = 0.7，而「富有」的 tf = 30 / (70+30) = 30/100 = 0.3；在第二篇文件裡，同樣篩選出兩個名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 40 次，「富有」出現 60 次，那「健康」的 tf = 40 / (40+60) = 40/100 = 0.4，「富有」的 tf = 60 / (40+60) = 60/100 = 0.6，tf 值愈高，其單詞愈重要。 所以，「健康」對第一篇文件比較重要，「富有」對第二篇文件比較重要。若搜尋「健康」，那第一篇文件會在較前面的位置；而搜尋「富有」，則第二篇文章會出現在較前面的位置。

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

2. IDF（Inverse Document Frequency）：

   換個角度來看，假設 D 是「所有的文件總數」，i 是網頁中所使用的單詞，t(i) 是該單詞在所有文件總數中出現的「文件數」，那麼 idf(i) 的算法就是 **log (D/t(i)) = log D − log t(i)**。例如有 100 個網頁，「健康」出現在 10 個網頁當中，而「富有」出現在 100 個網頁當中，那麼「健康」的 idf = log (100/10) = log 100 − log 10 = 2 − 1 = 1，而「富有」的 idf = log (100/100) = log 100 − 1og 100 = 2 − 2 = 0。 所以，「健康」出現的機會小，與出現機會很大的「富有」比較起來，便顯得非常重要。

$$\mathrm{idf_i} = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

3. TF-IDF：

   最後，將 **tf(i,j) * idf(i)**（例如：i =「健康」一詞）來進行計算，以某一特定文件內的高單詞頻率，乘上該單詞在文件總數中的低文件頻率，便可以產生 TF-IDF 權重值，且 TF-IDF 傾向於**過濾掉常見的單詞，保留重要的單詞**，如此一來，「富有」便不重要了。

參考資料:

- 文字探勘之前處理與TF-IDF介紹
- 深入了解scikit Learn裡TFIDF計算方式