Natural Language Processing

1. 詞袋模型(bag of words model)

i) 名詞解釋：

想像是放在袋子裡零散而獨立的物件，如此一個袋子代表一篇文件。詞袋模型的重點，不在於這個想像中的袋子，而在於其對待袋子中的詞彙方式，亦即每個詞彙都是獨立的單位，不考慮其相依性。

ii) 舉例：

文件 A 中的內容（如篇名）若為：「病人與醫生的糾紛研究」，以詞袋模型表示，則該文件可以表達成：「病人、糾紛、醫生、研究」這四個獨立的詞彙。又如文件 B 中的內容（如篇名）若為：「醫療缺失改善之探討」，以詞袋模型表示，則可表達成：「缺失、探討、改善、醫療」這四個獨立的詞彙。

iii) 分群：

文件中的詞彙代表空間中的一個維度，而維度與維度之間是獨立的，如此形成文件向量，便於後續的向量計算。如上例，文件 A 與文件 B 以（病人、醫生、糾紛、研究、醫療、缺失、改善、探討），8 個詞當維度，可以分別表示成（1, 1, 1, 1, 0, 0, 0, 0）與（0, 0, 0, 0, 1, 1, 1, 1）的向量。

自動文件分類中，也常以詞袋模型代表文件，將文件與類別的對應關係，如（文件 A，醫療類）、（文件 B，醫療類），分解成更小單元且會重複出現的詞彙與類別的對應關係，如（病人，醫療類）、（醫療，醫療類）等，以便於各種機器學習方法的運用。

iv) 缺點：

詞袋模型的缺點，則是其獨立性假設不太符合語言文字實際分布的狀況。例如，上述文件 A 與文件 B 的向量相似度為 0。但根據語言文字的出現機率，文件中談到病人、醫生的時候，醫療一詞出現的機率不應為 0； 若能考慮到詞彙的相依性，則文件 A 與文件 B 的相似度，就不會是 0。

2. TF-IDF（Term Frequency - Inverse Document Frequency）

i) 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。

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

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。

ii) IDF (Inverse Document Frequency)

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

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) = 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。

iii) 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。

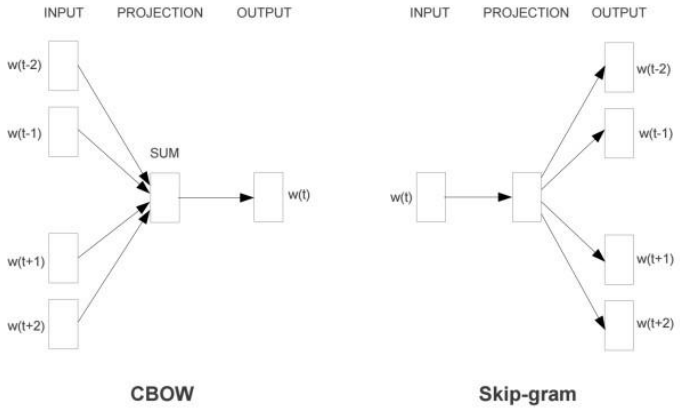
3. Word2Vector

i) Word2vec 是 Google 于 2013 年推出的开源的获取词向量 Word2vec 的工具包。它包括了一组用于 Word Embedding 的模型，这些模型通常都是用浅层 (两层) 神经网络训练词向量。

ii) Word2Vec 其實就是通過學習文本來用詞向量的方式表征詞的語義信息，即通過一個嵌入空間使得語義上相似的單詞在該空間內距離很近。

Embedding 其實就是一個映射，將單詞從原先所屬的空間映射到新的多維空間中，也就是把原先詞所在空間嵌入到一個新的空間中去。

iii) Word2Vec 模型中，主要有 Skip-Gram 和 CBOW 兩種模型，從直觀上理解，Skip-Gram 是給定 input word 來預測上下文。而 CBOW 是給定上下文，來預測 input word。



iv) 數學運算過程：https://kknews.cc/zh-tw/news/3j6yj2g.html