鄭卜壬教授 Homework: Word2Vec

B00401062 羅文斌

Word2Vec 是一種 word embedding 的演算法，字面意思是將一個字對應到一個 vector，有了 vector 以後可以做各種運算，比如計算兩個 words 的 cosine distance，距離越近的字可能暗示彼此的關聯性越大。好的 word embedding 演算法可以提供許多資訊，比如上課所提到的，類似含義的字在 vector space 中彼此會 cluster 在一個區域，又或者可以透過 vector 的運算來完成 semantic 或是 syntactic tasks，比如：cry is to crying as dry is to ____，透過 crying - cry + dry 可能可以得到 drying。Word embedding 可以透過各類 machine learning 的方法，input 各式各樣的 documents 來 training 參數，最後得到一個 machine learning model。Word2Vec 有兩個重要的 training model，一個是 skip-gram model，另外一個是 CBOW （continuous bag of words）。兩個 model 都建立在 neural network（又稱為 deep learning）之上，neural network model 包含了 input layer、hidden layers、output layer，這兩者 model 的差別簡單來說如下：

- Skip-gram model：給定一個 word，經過 model 運算後預測 context。
- CBOW：給定一個字的 context，經過 model 運算後預測那個 word。

在 Google 所發表的 paper：Distributed Representations of Words and Phrases and their Compositionality，裡面特別探討了 skip-gram model，其 training data 為 window size 任意大小的 training context，目標為使正確 context 的 log probability 最大化，其中有三個重要的技術：hierarchical softmax、negative sampling、subsampling of frequent words。Hierarchical softmax 的 output layer 使用了一個特別的設計，就是 binary tree，其中 leaves 為所有可能的 words，而每一個 parent node 的值為其 child node 的 probability，透過 binary tree traversal，可以得到我們所要的 output vector，使用這個演算法的好處是效率高，因為要得到 output，只需要 O(log(n)) 的複雜度，對於 natural language processing 的問題來說，n 值（所有 word 的數量）通常非常龐大，結合 binary tree 可以使得 neural network 的輸出更有效率。針對 training context 龐大的問題，negative sampling 提供了一個另外的方法，其中一個重要的論點是，並非所有的 training data 都是品質良好，有些可能是所謂的 noise，如果可以透過一些手法，先針對 training data 做 sampling，也可以使 neural network model 的建造更有效率，logistic regression 為他們所使用的回歸分析來分辨哪些 training data 可能是 noise 而可以加以剔除。第三個技術為 subsampling of frequent words，這個技術可以刪除許多重複出現但是資訊量較少的 words，對應到老師上課提到的一個有趣現象就是，一個字出現的次數與它的排名相加起來幾乎是一個定值，越常出現的字往往是不具有特殊含義的字而一直被重複使用，在英文中的例子如 in、the、a 等等，中文的例子如個、這、此等等，老師也提到 Google search 在檢索時甚至會將這些字直接剔除。回到這篇論文上，作者也提到即使將這些字剔除，訓練完的 model 結果並沒有太大落差，又可以增加效率。

Word2Vec 演算法可以應用在醫療病歷的審查和研究上，醫學研究常常需要回溯過往病歷，如果可以先透過此 model 建立醫學病例常用的字彙即所對應的 vector，遇到一篇新的病歷可以較容易從中抽取出重要的資訊、或是病歷中異常的敘述，如此可以大幅提升醫學研究的方便性，以及健保審查上可以有更客觀的評判。