**方法2: find general tagset (陳力維)**

**目標 :**

我們想要利用神經網路預測出所有在文章中的字，哪些在是這個類別的tag而那些不是，此為輸入是word，output是0或1的regression task。

**輸入特徵 :**

接著我們將決定如何將一個word embed成向量，我們採取的作法是，看這個字在文章中出現幾次，如此一來就有一個維度為文章總數的向量，藉此來表示一個字。舉個例子，假如現在有三篇文章，分別為 :

1. What is spin as it relates to subatomic particles?
2. What is your simplest explanation of the string theory?
3. Lie theory, Representations and particle physics

依照上文，”theory”的向量表示法為，” string”的表示法則為。然而，因為文章數實在太多，無法直接將如此高維的向量輸入我們的神經網路，故先將此向量表示法做sparse SVD得到一個700維的向量表示法，在以此當作輸入的向量。

**模型架構 :**

D:\downloads\Untitled Diagram (5).png

此模型前面的Dense Layer都是用Relu作為Activation Function，最後一層的單神經元則是用sigmoid當作Activation Function，Loss採用binary crossentopy。

**Validation and Evaluation :**

我們利用biology和crypto類別的資料來當作training set，robotic的類別作為validation set，如此是考量到robotic與physics的主題較相近的關係。接著隨機初始化參數train 500次，並使用Early Stopping Method，然後將得出的模型去預測robotic類別的資料，取前五十個機率最高的字，將得到的tag set所有的字取其在答案中出現次數總和，最後保留其值最高的model。

最後我們將最好的模型去預測physics的tag set，得到一個基本的tag set，但因為這個方法沒辦法包含2-gram，故我們又做了一些處理。

**後處理 :**

我們將得到的tag set先刪掉一些在文章中出現次數過少的，在將這些單詞在經過前處理過的文本中，前後出現次數較多的字，將他們組成2gram放入tagset。

**成效與討論 :**

此方法只能作為輔助使用，但在主要方法加入此方法之後成效並不是很大，似乎沒有對分數有太大的影響。我認為是由於要找出整個類別的tag set這個task太難了，往往預測出的tag太過於general，對預測沒有太大的幫助。故先對文章進行分群之後，再利用此方法找出專屬於該群的tag set可能會比較有效。但是我發現的太晚了，由於SVD要求要有一定的文章數和字數才能實作，一開始設的700維似乎太高，之後沒有時間重新調整模型。且其他的方法成效似乎也不錯，故就先將此方法打住了。