Information Retrieval and Applications

HW01-Vector Space Model

資工所碩一M10915201陳牧凡

此次作業運用到的MODEL是vector space model，透過計算query以及document之間的cosine similarity來找出吻合度，並得到符合該query描述的document優先排序。而在計算相似度的過程中，將query和document轉換成向量，其由n個詞(term/word)組成，每個詞都有一個權重，不同的詞根據自己在文件中的權重來影響文件相關性的重要程度。程式中有用到的library包括:numpy、sklearn(後用numpy自己實作方法替代)、math等。

計算cosine similarity時不使用sklearn library而以numpy實作的原因在於效能問題，如下是兩者之間在計算速度上的比較(單位為秒):

numpy	1.867
sklearn	38.145

上傳kaggle之後也如預期輕鬆地超越了baseline:

不過在效能方面或許還有可改進空間，如下是各段工作所花費的時間(單位為秒):

loading file	0.488
making dictionary	0.143
calculating TF	4.555
calculating IDF	0.003
calculating TF-IDF	0.001
calculating cosine similarity	1.59
print out answer	0.127

可以看到計算TF時特別花時間，因為這邊只能拿每個文檔去count，雖然前面有用技巧讓dictionary維度少了非常多，但還是吃掉大部分的時間；其次是計算cosine similarity時因為也必須50個queriesX4191個documents一組一組做，所以也有點耗時。使用多執行緒之類的平行運算方法可能可以讓其計算更快。