Tf-idf

利用tf-idf值從book description中找出具有代表性的字將每本書轉換成向量，其中向量的元素表示出現過幾次對應的代表字。

例如：假設具有代表性的字是science, art, bag，統計每本書出現次數如下

|  |  |  |  |
| --- | --- | --- | --- |
|  | science | art | Bag |
| Book 1 | 5 | 2 | 7 |
| Book 2 | 1 | 8 | 3 |

則Book 1 = (5, 2, 7)而Book 2 = (1, 8, 3)

我們所找到具有代表性的字是['author', 'beauti', 'book', 'come', 'famili', 'help', 'includ', 'just', 'life', 'like', 'live', 'love', 'make', 'man', 'new', 'onli', 'power', 'stori', 'time', 'way', 'woman', 'work', 'world', 'year', 'young']，有些字像是不完整的單字是因為還經過stemming處理，將不同型態但是同一字源的字視為相同的字。

將每本書轉換成對應的向量後，再利用KMeans將全部的書分類，KMeans需要填入「期望的類別數」，經過幾次測試後選定7類作為最終選擇。而有些training data或testing data的書沒有出現在books.csv裡面，我們將這些書分在第8類，最後就得到所有書對應的分類了。