2016.01.11 Word2Vec 詞語分類測試：

步驟：
1.採用 TFIDF 對 100 個景點，各取 10 個結果出來，存成一語料檔
2.自定義類別
3.使用 Word2Vec 對語料進行相似度分析，以最相似的類別將該詞分類進去

結果：
https://github.com/spicyscap/AttrClassification/blob/master/Method_UseWord2Vec_Classification.ipynb

結論：
(1.)
在[藝術],[休閒],[懷舊]類別，分類準確率相當高
但'髒亂'的分類結果就不是很理想，
經查詢維基是百科型的語料，對於'髒亂'此字眼並沒有相關內容描述

→定義類別時：可先用小樣本，確定「類別間相似度低」及「類別分類結果佳」，再擴大樣本，並在過程中重複修正。

(2.)門檻值
與 "準確率" 成正比；
與 "成功率" 成反比。

→多次嘗試，去抓出最適合的門檻值

(3.)
辨識度：Word2Vec 約可辨識 85% 以上的詞語，尚可
成功率：即使門檻值已設 0.2，仍只有 50% 的詞語成功被分類，成功率略嫌過低

→從 'Word2Vec 模型修正' 或 '更改分群分類方式' 著手。
前者較難，少有能與 wiki 資料量相抗衡的中文資料，中國有網友用新聞語料去當模型，但還是可嘗試。