word2vec

一、應用目標

生成特定文本中，詞彙的向量，相近的詞彙在向量中也會得呈現相近的關係。

二、環境要求

Python 3.9 (pkl檔要求，若使用其他格式則3.x即可)

pip install gensim

三、事前準備流程

* 文本爬蟲
  + 重大訊息
* 文字清理
* 斷詞
  + ckip tagger

四、事前準備的檔案:

* 斷詞結果
  + w2v\_full\_input.pkl
  + <https://drive.google.com/drive/folders/1nGFJEjWlKoR4dZq2NtsNpk3V97vuX6-u?usp=sharing>

五、介紹:

使用gensim library中的word2vec model進行word2vec訓練與實作。

* 教學與參數可以參考
* [Word2Vec的簡易教學與參數調整指南 | Kaggle](https://www.kaggle.com/code/jerrykuo7727/word2vec/notebook)
* [models.word2vec – Word2vec embeddings — gensim (radimrehurek.com)](https://radimrehurek.com/gensim/models/word2vec.html)

六、實作流程:

1.import library

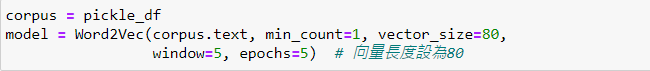


2.讀取資料(斷詞結果)



3.訓練word2vec model

調整參數可參考 [models.word2vec – Word2vec embeddings — gensim (radimrehurek.com)](https://radimrehurek.com/gensim/models/word2vec.html)



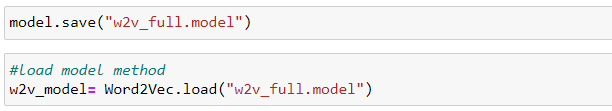
4.察看結果

觀察相似的結果是否和預期的有很大的出入，若成果不好再回到模型訓練調整參數。

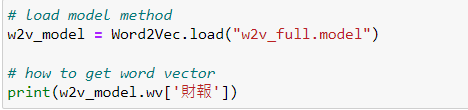


5.儲存模型

第4步ok的話，就可以儲存模型了。



6.模型的使用方法



7.模型轉換儲存的格式



七、Source code

[stock/word2vec.ipynb at main · samwang1228/stock (github.com)](https://github.com/samwang1228/stock/blob/main/%E6%B2%9B%E9%8C%A1/word2vec.ipynb)