In [1]:
```
import os

# Packages
import gensim
import jieba
import zhconv
from gensim.corpora import WikiCorpus
from datetime import datetime as dt
from typing import List


if not os.path.isfile('dict.txt.big'):
    !wget https://github.com/fxsjy/jieba/raw/master/extra_dict/dict.txt.big
jieba.set_dictionary('dict.txt.big')

print("gensim", gensim.__version__)
print("jieba", jieba.__version__)
```

```
gensim 4.3.1
jieba 0.42.1
```

In [5]:
```
WIKI_SEG_TXT = "wiki_seg.txt"
```

In [6]:
```
# %%time

from gensim.models import word2vec
import multiprocessing

max_cpu_counts = multiprocessing.cpu_count()
word_dim_size = 300  #  設定 word vector 維度
print(f"Use {max_cpu_counts} workers to train Word2Vec (dim={word_dim_size})")

WIKI_SEG_TXT = "wiki_seg.txt"

# 讀取訓練語句
sentences = word2vec.LineSentence(WIKI_SEG_TXT)

# 訓練模型
model = word2vec.Word2Vec(sentences, vector_size=word_dim_size, workers=max_cpu_counts)

# 儲存模型
output_model = f"word2vec.zh.{word_dim_size}.model"
model.save(output_model)
```

```
Use 12 workers to train Word2Vec (dim=300)
```

In [7]:
```
print(model.wv.vectors.shape)
model.wv.vectors
```

```
(1281108, 300)
```

Out[7]:
```
array([[ 1.3096823e+00,  1.6333671e-01, -1.7997390e+00, ...,
         6.3041550e-01,  3.9032009e-01, -2.7709281e+00],
       [ 1.5634978e+00, -1.0815852e+00, -2.7119610e+00, ...,
         1.4645466e+00,  8.0222178e-01, -3.7588122e+00],
       [ 8.0596614e-01, -1.8223846e-01, -2.8863567e-01, ...,
        -4.4410625e-01, -3.9572179e-01,  1.3776612e-01],
       ...,
       [ 1.5343905e-02,  7.4792691e-03,  5.1671248e-02, ...,
         9.6034782e-05, -1.4427365e-02, -1.1857649e-01],
       [-1.6592441e-02, -1.6240990e-02, -3.5799820e-02, ...,
        -3.8599610e-02, -5.4656074e-04, -2.9039632e-03],
       [-1.9124595e-02,  3.2623895e-02, -2.0442145e-02, ...,
        -8.3319256e-03, -3.1643976e-02, -7.3861657e-03]], dtype=float32)
```

In [22]:
```
# print(f"總共收錄了 {len(model.wv.vocab)} 個詞彙")

# print("印出 20 個收錄詞彙:")
# print(list(model.wv.vocab.keys())[:10])
from gensim.models import Word2Vec

# 加载并训练Word2Vec模型
model = Word2Vec.load('word2vec.zh.300.model')  # 替换成您的模型路径
vocab = model.wv.key_to_index

# 打印总共收录的词汇数
print(f"總共收錄了 {len(vocab)} 個詞彙")

# 打印前20个收录的词汇
print("印出 20 個收錄詞彙:")
print(list(vocab.keys())[:20])
```

```
總共收錄了 1281108 個詞彙
印出 20 個收錄詞彙:
['年', '月', '日', '中', '10', '12', '11', '小行星', '中國', '時', '-', '日本', '美國', '20', '香港', '臺灣', '15', '位於', '30', '站']
```

In [10]:
```
vec = model.wv['數學家']
print(vec.shape)
vec
```

```
(300,)
```

Out[10]:
```
array([-7.29663312e-01, -2.04170108e+00, -2.08726811e+00, -2.00406528e+00,
        1.23695385e+00, -3.22162546e-02,  1.02139739e-02,  9.97790754e-01,
       -1.00164950e+00, -9.28038239e-01,  9.99963999e-01,  1.22361541e+00,
        1.07440972e+00, -1.01716733e+00, -7.80249894e-01,  4.82204668e-02,
        2.02229887e-01,  4.24782306e-01, -1.33372557e+00, -1.17479647e-02,
       -2.93124986e+00, -9.78583217e-01, -1.72899806e+00, -2.06405234e+00,
       -9.76091325e-01,  4.92442638e-01,  1.08806264e+00,  1.57044029e+00,
       -1.05221891e+00, -4.61767852e-01, -1.71692216e+00, -1.15216291e+00,
        7.27864742e-01,  5.68888849e-03,  2.44226810e-02,  1.88848341e+00,
       -1.30136621e+00,  2.09170961e+00,  4.51524079e-01, -1.44541895e+00,
       -7.35151231e-01, -3.17242789e+00, -1.84289038e+00, -4.77210999e-01,
       -5.06819598e-02, -6.38516605e-01,  2.39134097e+00, -1.15909004e+00,
        5.94701231e-01,  1.01881123e+00, -1.80936790e+00, -1.33932638e+00,
       -4.52207297e-01,  2.26115632e+00,  2.73495078e-01, -8.25931728e-01,
       -2.47877955e-01, -8.35634097e-02,  5.08708596e-01,  1.72053850e+00,
       -1.41429269e+00,  2.60252929e+00, -2.03497767e+00, -2.83284974e+00,
       -1.00708574e-01,  1.93940616e+00, -1.85330659e-01,  4.12466198e-01,
        6.30587637e-01,  2.06882167e+00,  3.59077722e-01, -2.22792172e+00,
       -1.77185702e+00,  2.02201128e-01, -3.33315581e-01,  5.41004360e-01,
       -2.75059676e+00,  1.26831782e+00, -5.00957549e-01, -8.06639254e-01,
       -1.95511091e+00, -1.82642603e+00,  1.13007855e-02, -1.42721009e+00,
       -2.07257584e-01, -1.02430415e+00,  7.68389046e-01, -1.75572753e+00,
        1.19209409e+00,  5.51771581e-01, -5.56405745e-02, -5.74681580e-01,
       -3.26856661e+00,  9.55411017e-01,  1.27508903e+00,  2.42995977e+00,
        2.62225151e+00,  3.55896056e-01, -1.43304002e+00, -1.30080771e+00,
        1.33091724e+00,  1.43545938e+00,  6.38025641e-01,  2.31637979e+00,
       -1.30736220e+00, -7.66243398e-01, -1.56304562e+00,  1.30084312e+00,
       -1.69872677e+00,  6.64506674e-01,  8.52019668e-01, -8.71260226e-01,
       -3.00369549e+00, -2.63318628e-01, -1.08460343e+00, -1.55742252e+00,
       -1.57011613e-01, -1.44376230e+00,  3.45521681e-02,  8.67060006e-01,
       -1.89181376e+00,  1.79826236e+00,  5.79327762e-01, -4.86085147e-01,
       -3.68616730e-01, -4.32814479e-01, -3.07655424e-01,  2.45765552e-01,
       -1.56716073e+00, -3.21549153e+00, -1.57656074e+00, -2.02823901e+00,
       -2.20462584e+00,  1.28402233e+00,  8.24476182e-01,  1.52029181e+00,
        7.69030452e-01, -2.07295108e+00,  1.48896229e+00,  1.05049801e+00,
        8.09115410e-01,  7.62654096e-02, -2.98700538e-02, -2.92015290e+00,
        1.38945508e+00,  3.32267976e+00,  9.82308447e-01,  3.06490898e-01,
       -2.13512444e+00,  3.41965646e-01,  1.12010145e+00, -1.29156566e+00,
        9.96370971e-01, -1.26623082e+00,  5.21973908e-01,  3.91560626e+00,
        2.27287436e+00,  4.45689440e-01,  1.83803177e+00, -4.52777147e-01,
       -1.26683021e+00, -2.16313195e+00, -1.22726989e+00,  1.65406978e+00,
```

```
 -4.16294903e-01, -1.20693362e+00,  1.56922317e+00,  9.00900483e-01,
 -2.63946950e-01, -2.24408793e+00,  1.36942491e-01, -3.08553457e+00,
  1.60712290e+00, -1.84919059e+00, -2.06452560e+00,  7.69858003e-01,
 -8.20989236e-02,  2.13722229e+00, -2.19921589e+00, -1.09491728e-01,
  9.94407296e-01, -1.67151821e+00, -5.20348549e-01, -1.21101117e+00,
  9.96180475e-01, -1.95859587e+00,  2.29596376e-01, -2.32872319e+00,
  1.48101854e+00, -2.74290442e+00,  5.61233044e-01,  2.10034919e+00,
  2.46877909e+00, -4.50825423e-01, -1.03584814e+00, -1.14000118e+00,
 -2.38541961e+00, -1.46805882e+00, -2.04078332e-01, -2.05508485e-01,
 -9.82613146e-01, -2.68874025e+00,  1.09272182e+00, -1.75672308e-01,
 -4.74937469e-01, -9.75526810e-01,  1.52819729e+00, -2.01562330e-01,
  2.54064798e+00, -1.19343758e+00,  1.00390255e+00,  2.75072336e+00,
 -3.18685830e-01, -2.54809737e+00,  1.48433638e+00,  1.33668089e+00,
  1.91030133e+00, -1.44425988e+00, -9.76709902e-01,  5.95131576e-01,
 -1.41748738e+00, -1.46174705e+00,  9.54195678e-01, -5.74497998e-01,
  7.55910456e-01, -2.30094886e+00,  1.32007515e+00, -6.11047328e-01,
 -6.95345163e-01, -7.09905863e-01,  9.54119444e-01,  1.91180241e+00,
 -1.82665467e+00,  4.49844360e-01,  1.69266784e+00,  2.28267998e-01,
 -1.31554937e+00, -1.99402118e+00,  5.34243405e-01,  7.86236227e-01,
  8.65973234e-01,  8.18367004e-02, -3.55294067e-03, -2.09629273e+00,
  7.65455484e-01,  1.00236225e+00, -1.34593749e+00, -1.22950315e+00,
 -6.84114099e-02, -1.70833063e+00, -4.31178242e-01,  2.01875591e+00,
  9.37773228e-01,  4.89351541e-01, -4.32685792e-01, -4.17008066e+00,
 -1.17980607e-01, -2.21484256e+00,  1.87447190e-01,  1.25197244e+00,
  1.35880911e+00,  1.17765605e+00,  3.66979331e-01,  1.77399188e-01,
  1.77892184e+00,  9.92396891e-01, -2.18532160e-01,  7.53022134e-01,
 -2.12726641e+00,  2.61834311e+00,  1.39605796e+00, -2.53855848e+00,
  1.45146370e-01,  1.11732972e+00,  1.09657741e+00,  2.35529900e+00,
  4.55185920e-01, -2.50180542e-01, -1.89471900e+00,  1.42322683e+00,
  2.25495672e+00, -2.34708846e-01,  7.97750354e-01, -1.38195440e-01,
  1.21414596e-02,  4.67886877e+00, -6.10259712e-01, -2.82048249e+00,
  4.63147610e-01,  1.62494898e+00, -5.98654509e-01,  9.72362101e-01,
  1.55744886e+00,  6.31877065e-01, -1.96382865e-01,  9.83198941e-01,
 -3.43257606e-01,  2.28850269e+00,  7.44235516e-01, -1.37282574e+00],
dtype=float32)
```

In [11]:
```
word = "這肯定沒見過 "

# 若強行取值會報錯
try:
    vec = model.wv[word]
except KeyError as e:
    print(e)
```

```
"Key '這肯定沒見過 ' not present"
```

In [12]:
```
model.wv.most_similar("飲料", topn=10)
```

Out[12]:
```
[('飲品', 0.8034288287162781),
 ('軟飲料', 0.7063937187194824),
 ('酒精類', 0.694480836391449),
 ('酒類', 0.6846030950546265),
 ('果汁', 0.679166853427887),
 ('含酒精', 0.6671655774116516),
 ('蘇打水', 0.6448999047279358),
 ('提神', 0.6378534436225891),
 ('瓶裝', 0.636431097984314),
 ('罐裝', 0.6351532936096191)]
```

In [13]:
```
model.wv.most_similar("car")
```

Out[13]:
```
[('truck', 0.6898175477981567),
 ('tikita', 0.6747397780418396),
 ('seat', 0.658551812171936),
 ('motorcycle', 0.6424771547317505),
 ('limousine', 0.6411415338516235),
 ('cab', 0.632193386554718),
 ('wagon', 0.6217625737190247),
 ('driving', 0.6124157905578613),
 ('motor', 0.6108798384666443),
 ('sedan', 0.6072883605957031)]
```

In [14]:
```
model.wv.most_similar("facebook")
```

Out[14]:
```
[('臉書', 0.8054246306419373),
 ('面書', 0.742925226688385),
 ('專頁', 0.740338921546936),
 ('instagram', 0.7256566882133484),
 ('貼文', 0.7056513428688049),
 ('twitter', 0.6894757151603699),
 ('推特', 0.6717672348022461),
 ('粉專', 0.6632859706878662),
 ('粉絲團', 0.6570253968238831),
 ('網誌', 0.6474297642707825)]
```

In [15]:
```
model.wv.most_similar("詐欺")
```

Out[15]:
```
[('欺詐', 0.7272929549217224),
 ('詐騙', 0.5991770625114441),
 ('竊盜', 0.5756206512451172),
 ('慣犯', 0.5518917441368103),
 ('詐欺罪', 0.550362229347229),
 ('敲詐', 0.5362935066223145),
 ('金光黨', 0.5334990620613098),
 ('詐騙犯', 0.5274579524993896),
 ('師篇', 0.5268533825874329),
 ('背信', 0.5230058431625366)]
```

In [16]:
```
model.wv.most_similar("合約")
```

Out[16]:
```
[('合同', 0.7789182662963867),
 ('簽約', 0.7078366279602051),
 ('續約', 0.6748090386390686),
 ('續簽', 0.6076886653900146),
 ('簽下', 0.5910363793373108),
 ('租約', 0.5828015208244324),
 ('短約', 0.5813875198364258),
 ('買斷', 0.5810321569442749),
 ('選擇權', 0.5714226365089417),
 ('解約', 0.5621798038482666)]
```

In [17]:
```
model.wv.similarity("連結", "鍵接")
```

Out[17]:
```
0.71518385
```

In [18]:
```
model.wv.similarity("連結", "陰天")
```

Out[18]:
```
0.0019686818
```

In [19]:
```
print(f"Loading {output_model}...")
new_model = word2vec.Word2Vec.load(output_model)
```

```
Loading word2vec.zh.300.model...
```

In [20]:
```
model.wv.similarity("連結", "陰天") == new_model.wv.similarity("連結", "陰天")
```

Out[20]:
```
True
```

In [ ]:
```

```