In [1]:

```
%load_ext memory_profiler
# !pip install -q zhconv
```

In [2]:

```
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 [3]:

```
ZhWiki = "zhwiki-20230501-pages-articles-multistream.xml.bz2"

!du -sh $ZhWiki
!md5 $ZhWiki
!file $ZhWiki
```

```
2.6G	zhwiki-20230501-pages-articles-multistream.xml.bz2
MD5 (zhwiki-20230501-pages-articles-multistream.xml.bz2) = 27e78ff901b
cd3803955d9373537dd3f
zhwiki-20230501-pages-articles-multistream.xml.bz2: bzip2 compressed d
ata, block size = 900k
```

In [4]:

```
import spacy
nlp_zh = spacy.load("zh_core_web_sm")
nlp_en = spacy.load("en_core_web_sm")
# spacy.cli.download("zh_core_web_sm")  # 下載 spacy 中文模組
# spacy.cli.download("en_core_web_sm")  # 下載 spacy 英文模組
```

In [7]:

```
from f import preprocess_and_tokenize
```

In [8]:

```
%%time
%%memit

print(f"Parsing {ZhWiki}...")
wiki_corpus = WikiCorpus(ZhWiki, tokenizer_func=preprocess_and_tokenize, token_min_l
```

```
000gn/T/jieba.cache
Dumping model to file cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0
000gn/T/jieba.cache
Dumping model to file cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0
000gn/T/jieba.cache
Dumping model to file cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0
000gn/T/jieba.cache
Dumping model to file cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0
000gn/T/jieba.cache
Dumping model to file cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0
000gn/T/jieba.cache
Dumping model to file cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0
000gn/T/jieba.cache
Loading model cost 1.629 seconds.
Prefix dict has been built successfully.
Loading model cost 1.593 seconds.
Prefix dict has been built successfully.
Loading model cost 1.618 seconds.
Prefix dict has been built successfully.
Loading model cost 1.537 seconds.
```

In [9]:

```
g = wiki_corpus.get_texts()
print(next(g)[:10])
print(next(g)[:10])
print(next(g)[:10])
```

```
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Loading model cost 1.252 seconds.
Prefix dict has been built successfully.
Loading model cost 1.272 seconds.
Prefix dict has been built successfully.
Loading model cost 1.285 seconds.
Prefix dict has been built successfully.
Loading model cost 1.294 seconds.
Prefix dict has been built successfully.
Loading model cost 1.266 seconds.
Prefix dict has been built successfully.
Loading model cost 1.287 seconds.
Prefix dict has been built successfully.
Loading model cost 1.275 seconds.
Prefix dict has been built successfully.
```

```
['歐幾裡', '西元前', '三世', '紀的', '古希臘', '數學家', '幾何', '之父', '此
畫', '為拉斐爾']
['蘇', '格拉', '底', '死', '雅克', '路易', '大衛', '所繪', '1787', '年']
['文學', '狹義上', '一種', '語言藝術', '語言', '文字', '手段', '形象化', '客
觀', '社會']
```

In [11]:

```
WIKI_SEG_TXT = "wiki_seg.txt"

generator = wiki_corpus.get_texts()

with open(WIKI_SEG_TXT, "w", encoding='utf-8') as output:
    for texts_num, tokens in enumerate(generator):
        output.write(" ".join(tokens) + "\n")

        if (texts_num + 1) % 100000 == 0:
            print(f"[{str(dt.now()):.19}] 已寫入 {texts_num} 篇斷詞文章")
```

```
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.cache
Loading model cost 0.997 seconds.
Prefix dict has been built successfully.
Loading model cost 1.011 seconds.
Prefix dict has been built successfully.
Loading model cost 1.027 seconds.
Prefix dict has been built successfully.
Loading model cost 1.019 seconds.
Prefix dict has been built successfully.
Loading model cost 0.993 seconds.
Prefix dict has been built successfully.
Loading model cost 0.999 seconds.
Prefix dict has been built successfully.
Loading model cost 1.076 seconds.
Prefix dict has been built successfully.


[2023-05-12 00:49:33] 已寫入 99999 篇斷詞文章
[2023-05-12 00:59:42] 已寫入 199999 篇斷詞文章
[2023-05-12 01:08:39] 已寫入 299999 篇斷詞文章
[2023-05-12 01:16:44] 已寫入 399999 篇斷詞文章
[2023-05-12 01:24:49] 已寫入 499999 篇斷詞文章
[2023-05-12 01:31:20] 已寫入 599999 篇斷詞文章
[2023-05-12 01:39:23] 已寫入 699999 篇斷詞文章
[2023-05-12 01:47:13] 已寫入 799999 篇斷詞文章
```

In [16]:

```
%%time

from gensim.models import word2vec
import multiprocessing

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


sentences = word2vec.LineSentence(WIKI_SEG_TXT)

model = word2vec.Word2Vec(sentences, vector_size=word_dim_size, workers=max_cpu_cour

output_model = f"word2vec.zh.{word_dim_size}.model"
model.save(output_model)
```

```
Use 8 workers to train Word2Vec (dim=300)
CPU times: user 2h 43min 18s, sys: 2min 5s, total: 2h 45min 23s
Wall time: 33min 41s
```

In [17]:

```
! ls word2vec.zh*
```

```
word2vec.zh.300.model                word2vec.zh.300.model.wv.vectors.
npy
word2vec.zh.300.model.syn1neg.npy
```

In [18]:

```
! du -sh word2vec.zh*
```

```
 57M    word2vec.zh.300.model
2.0G    word2vec.zh.300.model.syn1neg.npy
2.0G    word2vec.zh.300.model.wv.vectors.npy
```

In [19]:

```
print(model.wv.vectors.shape)
model.wv.vectors
```

```
(1795845, 300)
```

Out[19]:

```
array([[-8.9804405e-01, -6.1582553e-01, -2.0657480e+00, ...,
         3.0532640e-01, -2.0539918e+00,  2.4737072e-01],
       [-6.8318689e-01, -2.7498755e-01, -2.3455980e+00, ...,
         5.1383954e-01, -4.1464081e+00,  7.0632869e-01],
       [ 5.9505785e-01, -3.8984559e+00, -3.6878426e+00, ...,
        -5.4230469e-01, -8.8972461e-01, -1.7220721e+00],
       ...,
       [-3.9683364e-02,  5.5359375e-02, -1.9255705e-02, ...,
        -4.8501860e-02, -1.8312057e-02, -4.9458582e-02],
       [-3.6725014e-02,  2.8391223e-04,  4.6228167e-02, ...,
         1.3471413e-03,  2.9617954e-02,  8.6252647e-04],
       [-7.0929840e-02,  5.5286088e-03, -2.8707324e-02, ...,
        -1.4371433e-02, -8.0098964e-02,  3.3185694e-02]], dtype=float3
2)
```

In [21]:

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

```
(300,)
```

Out[21]:

```
array([ 0.52072793, -1.6518036 ,  0.36288863,  1.135915  , -0.7030483
5,
        1.8427767 ,  2.8063443 ,  0.3151409 , -0.24242602,  1.7130088
,
       -2.3480964 , -0.33735946,  2.3137958 ,  2.378177  ,  1.5551311
,
        1.6362667 , -0.363144  ,  0.9505343 , -3.3144307 , -0.5378922
,
       -0.9810759 , -3.0546303 , -0.5947663 ,  0.49536592, -1.0434953
,
       -1.6851808 ,  0.05314718, -2.4895465 , -1.3067619 ,  0.3629197
5,
       -0.29339752, -1.2593311 ,  2.8497407 ,  1.3100216 ,  0.3798682
,
       -0.01096008,  0.41875076,  1.2550446 , -0.4058333 ,  0.5590972
,
       -1.0038521 , -2.8408015 , -2.3901646 , -1.503638  ,  0.67122
,
        0.44948995, -1.7289228 ,  3.6393576 , -1.0319693 , -0.9767573
5,
       -0.20535083,  1.4012731 , -1.7907532 , -0.22726612, -1.309933
,
       -0.594253  ,  1.0532155 ,  0.2436298 ,  3.8445463 , -1.1118541
,
        0.58783054, -0.18030244,  2.5362673 , -0.16973338, -0.833415
,
        0.15504768, -0.30341446, -0.31407633,  2.1811786 , -1.0813227
,
        1.1266313 , -0.09310023,  0.5643399 ,  0.04777153,  2.2637548
,
        3.0563805 ,  0.9590055 ,  0.9550096 ,  1.5268732 , -0.1665690
2,
        0.37066597,  1.8510469 ,  0.23027307, -1.3373727 ,  1.1370852
,
        0.08197025, -2.4749882 , -0.84463763,  0.6633244 ,  1.5305537
,
        0.47227532, -2.1427178 , -1.9209265 ,  0.21254745,  0.8459894
,
       -0.6183559 , -1.3738403 ,  1.7637836 ,  1.4814557 ,  1.2859105
,
       -0.13078265, -2.544083  , -1.8408682 ,  1.3910108 , -0.3713553
8,
       -1.5476817 ,  0.5845316 , -1.9944983 , -0.8271785 ,  1.0413675
,
        3.478696  , -0.03621124, -0.00833122, -2.2498865 ,  0.4214513
6,
        0.87444913,  1.7574667 , -2.261779  , -1.9068464 ,  0.5304237
,
       -1.3953812 ,  1.0734229 , -2.247176  ,  0.67458254, -1.2645268
,
       -0.6227138 , -0.88734484,  1.5191737 , -0.01728725,  0.1839151
,
```

```
       -1.0788031 , -0.36952817,  0.7667618 , -0.2826106 , -2.45538  
,
        1.2123202 , -0.80780125,  0.9454721 ,  0.69413257,  2.3255794 
,
       -2.2571428 ,  0.13410896,  0.62153715,  3.108441  ,  0.8032581 
,
        0.06835987, -0.15267569, -0.01495016,  0.06116261, -1.2217599 
,
        0.10699552, -0.83401656, -2.2781394 , -1.2594914 ,  0.7629080
4,
        3.789427  , -1.2343781 , -0.36470515,  0.6075259 , -1.1035403 
,
        0.5488379 , -3.0323946 ,  0.35713112, -1.6817954 , -1.8494017 
,
       -1.8565239 ,  0.646627  ,  1.8296556 , -1.7754846 , -0.0996441
5,
       -2.2628238 ,  0.72121954,  0.76173884, -0.17306203,  0.3310746
8,
       -2.9879072 ,  1.4218646 , -1.3931118 , -1.0621144 , -0.546853  
,
       -1.9520303 ,  0.47924942,  1.0256658 ,  1.2909127 ,  1.8193763 
,
       -0.8595991 , -1.3042834 ,  0.40604672, -0.17011577, -1.3653984 
,
        1.7019204 , -4.461056  ,  0.5317203 ,  0.40313295,  4.3814926 
,
        1.7012216 , -0.15868495,  0.768121  , -0.40178722, -0.2687057 
,
        3.5909367 , -2.1233828 ,  1.4476473 , -1.1029419 ,  0.4228113
6,
        2.0539784 , -2.700973  , -1.3358669 , -0.7420144 ,  1.8476152 
,
       -0.00638305,  0.00814767, -0.13205753,  1.4886452 ,  0.4700526
3,
        1.1000513 , -1.4177285 , -0.50822717, -0.9965469 ,  1.8610064 
,
       -1.1591445 ,  0.17966165,  0.1506077 , -0.07027501,  0.1284336
6,
       -0.65926605, -0.28592187, -1.5827903 ,  1.6404732 , -1.2585398 
,
       -1.2994282 , -0.28035772, -2.1745868 ,  3.900518  , -0.9090284 
,
        0.44346768,  1.8960009 , -0.8785855 ,  0.8574798 , -2.5014231 
,
        1.7498785 ,  0.36479485, -1.464113  , -0.15043321, -1.5082169 
,
        1.2897736 , -2.9128196 , -0.9943601 ,  1.7838794 ,  0.6131538 
,
        0.41155478, -0.7988246 ,  0.61847734,  0.42134854,  0.2911720
3,
        1.8067286 , -3.093319  , -1.0415454 ,  0.04076328, -0.7647599 
,
        0.58683854,  1.062825  , -0.5477041 , -1.2421118 , -1.6041924 
,
       -0.5225819 ,  1.4514538 ,  0.72582275, -0.5801124 ,  0.3154741 
,
        0.10474767, -1.6237147 , -1.056725  , -0.17724873, -2.2321765 
,
        0.12747808,  1.7704551 , -1.666156  ,  1.9570284 , -3.0715609 
,
       -0.248687  ,  2.6662476 ,  0.59851825, -1.1653717 , -1.5669887 
```

```
,
        -2.4165962 ,  3.4639125 ,  1.0410556 ,  2.0990322 , -2.1556349
,
        -1.3394848 , -3.17473   ,  0.80243057, -1.5371408 , -1.80441
,
        -1.1453549 , -0.53952533,  0.70419157, -0.02630014, -3.437272
],
      dtype=float32)
```

In [23]:

```python
word = "這肯定沒見過 "

try:
    vec = model.wv[word]
except KeyError as e:
    print(e)
```

```
"Key '這肯定沒見過 ' not present"
```

In [24]:

```python
model.wv.most_similar("飲料", topn=10)
```

Out[24]:

```
[('飲品', 0.8073071837425232),
 ('酒類', 0.6659426689147949),
 ('含酒精', 0.6561747193336487),
 ('瓶裝', 0.6519318222999573),
 ('果汁', 0.648198127746582),
 ('類飲料', 0.6387940645217896),
 ('罐裝', 0.6385535001754761),
 ('無糖', 0.6312093734741211),
 ('中下藥', 0.6241925358772278),
 ('熱飲', 0.6208222508430481)]
```

In [25]:

```python
model.wv.most_similar("car")
```

Out[25]:

```
[('truck', 0.6935850977897644),
 ('seat', 0.687899112701416),
 ('saloon', 0.6666427254676819),
 ('tikita', 0.6554520130157471),
 ('wagon', 0.6397509574890137),
 ('videna', 0.6372187733650208),
 ('cars', 0.6315094232559204),
 ('driving', 0.6306036710739136),
 ('chevrolet', 0.627369225025177),
 ('pickup', 0.627359926700592)]
```

In [26]:

```
model.wv.most_similar("facebook")
```

Out[26]:

```
[('臉書', 0.7975596785545349),
 ('專頁', 0.7508485317230225),
 ('instagram', 0.733160138130188),
 ('面書', 0.7203158140182495),
 ('貼文', 0.7138100266456604),
 ('twitter', 0.6872799396514893),
 ('推特', 0.6858128309249878),
 ('粉絲團', 0.6795867085456848),
 ('粉絲頁', 0.677693784236908),
 ('臉書粉', 0.6592646837234497)]
```

In [34]:

```
model.wv.most_similar("happy")
```

Out[34]:

```
[('birthday', 0.6514894366264343),
 ('joyful', 0.6371191740036011),
 ('xmas', 0.6153290867805481),
 ('lucky', 0.6140486598014832),
 ('wish', 0.6123074889183044),
 ('unhappy', 0.6120163798332214),
 ('lovely', 0.6077404618263245),
 ('merry', 0.6035582423210144),
 ('ending', 0.6033361554145813),
 ('kappy', 0.6021612882614136)]
```

In [29]:

```
model.wv.most_similar("合約")
```

Out[29]:

```
[('合約將', 0.7606512308120728),
 ('合同', 0.7543858289718628),
 ('新合約', 0.7510531544685364),
 ('年合約', 0.7280263900756836),
 ('簽約', 0.7179040908813477),
 ('合約並', 0.7054702043533325),
 ('其合約', 0.7034028172492981),
 ('合約期', 0.6976761221885681),
 ('續約', 0.6844313740730286),
 ('因合約', 0.6735947728157043)]
```

In [30]:

```
model.wv.similarity("連結", "鏈結")
```

Out[30]:

```
0.79744154
```

In [31]:

```
model.wv.similarity("連結", "陰天")
```

Out[31]:

```
-0.0012724159
```

In [32]:

```
print(f"Loading {output_model}...")
new_model = word2vec.Word2Vec.load(output_model)
```

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

In [33]:

```
model.wv.similarity("連結", "陰天") == new_model.wv.similarity("連結", "陰天")
```

Out[33]:

```
True
```

In [ ]:

```
```