

In [1]:

```
!pip install memory_profiler
```

Requirement already satisfied: memory\_profiler in c:\users\user\anaconda3\lib\site-packages (0.61.0)  
Requirement already satisfied: psutil in c:\users\user\anaconda3\lib\site-packages (from memory\_profiler) (5.9.0)

In [2]:

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

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 4.3.0  
jieba 0.42.1

In [ ]:

```
import urllib.request  
url = "https://github.com/fxsjy/jieba/raw/master/extra_dict/dict.txt.big"  
filename = "dict.txt.big"  
urllib.request.urlretrieve(url, filename)
```

In [ ]:

```
import urllib.request  
url = "https://dumps.wikimedia.org/zhwiki/20230501/zhwiki-20230501-pages-articles.xml.bz2"  
  
filename = "zhwiki-20230501-pages-articles.xml.bz2"  
urllib.request.urlretrieve(url, filename)
```

In [12]:

```
import os

ZhWiki = r"C:\Users\user\Desktop\notebook\nlp\HW4\zhwiki-20230501-pages-articles.xml.bz2"

print(f"File size: {os.path.getsize(ZhWiki) / (1024*1024):.2f} MB")
```

File size: 2509.82 MB

In [3]:

```
zhconv.convert("这原本是一段简体中文", "zh-tw")
```

Out[3]:

'這原本是一段簡體中文'

In [4]:

```
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + " / ".join(seg_list)) # 全模式

seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + " / ".join(seg_list)) # 精確模式
```

Building prefix dict from C:\Users\user\Desktop\notebook\nlp\HW4\dict.txt.  
big ...  
Loading model from cache C:\Users\user\AppData\Local\Temp\jieba.u1860406d2  
d6aafb868e1ddf4bccba943.cache  
Loading model cost 0.891 seconds.  
Prefix dict has been built successfully.

Full Mode: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

Default Mode: 我/ 来到/ 北京/ 清华大学

In [5]:

```
print(list(jieba.cut("中英夾雜的example · Word2Vec應該很interesting吧?")))
```

['中', '英', '夾雜', '的', 'example', '·', 'Word2Vec', '應該', '很', 'interesting', '吧', '?']

In [ ]:

In [ ]:

In [ ]:

In [7]:

```
import spacy

# # 下載語言模組
# 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

nlp_zh = spacy.load("zh_core_web_sm") # 載入 spacy 中文模組
nlp_en = spacy.load("en_core_web_sm") # 載入 spacy 英文模組

# 印出前20個停用詞
print('--\n')
print(f"中文停用詞 Total={len(nlp_zh.Defaults.stop_words)}: {list(nlp_zh.Defaults.stop_words)}")
print("--")
print(f"英文停用詞 Total={len(nlp_en.Defaults.stop_words)}: {list(nlp_en.Defaults.stop_words)}")

--
```

中文停用詞 Total=1891: ['一个', '总是', '这种', '以致', '一时', '归根到底', '以后', '即若', '逐步', '反之', '唯有', '``', '仅', '清楚', '却不', '大概', '专门', '仅仅', '的话', '着呢'] ...

--  
英文停用詞 Total=326: ['thence', 'could', 'alone', 'just', 'regarding', 'wh ether', 'herself', 'meanwhile', 'noone', 'herein', 'something', 'forty', 'last', 'themselves', 'at', 'ever', 'few', 'amongst', 'on', 'is'] ...

In [8]:

```
STOPWORDS = nlp_zh.Defaults.stop_words | \
            nlp_en.Defaults.stop_words | \
            set(["\n", "\r\n", "\t", " ", ""])
print(len(STOPWORDS))

# 將簡體停用詞轉成繁體·擴充停用詞表
for word in STOPWORDS.copy():
    STOPWORDS.add(zhconv.convert(word, "zh-tw"))

print(len(STOPWORDS))
```

2222

3005

In [9]:

```
def preprocess_and_tokenize(
    text: str, token_min_len: int=1, token_max_len: int=15, lower: bool=True) -> List[str]:
    if lower:
        text = text.lower()
    text = zhconv.convert(text, "zh-tw")
    return [
        token for token in jieba.cut(text, cut_all=False)
        if token_min_len <= len(token) <= token_max_len and \
           token not in STOPWORDS
    ]
```

In [10]:

```
print(preprocess_and_tokenize("歐幾里得，西元前三世紀的古希臘數學家，現在被認為是幾何之父，此畫
print(preprocess_and_tokenize("我来到北京清华大学"))
print(preprocess_and_tokenize("中英夾雜的example，Word2Vec應該很interesting吧?"))
```

```
['歐幾', '裡得', '西元前', '世紀', '古希臘', '數學家', '幾何', '父', '此畫',
'拉斐爾']
['來到', '北京', '清華大學']
['中', '英', '夾雜', 'example', 'word2vec', 'interesting']
```

In [13]:

```
print(f"Parsing {ZhWiki}...")
wiki_corpus = WikiCorpus(ZhWiki, token_min_len=1)
```

```
Parsing C:\Users\user\Desktop\notebook\nlp\HW4\zhwiki-20230501-pages-artic
les.xml.bz2...
```

In [14]:

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

# print(jieba.lcut("".join(next(g))[:50]))
# print(jieba.lcut("".join(next(g))[:50]))
```

```
['歐幾里得', '西元前三世紀的古希臘數學家', '現在被認為是幾何之父', '此畫為拉斐爾
的作品', '雅典學院', '数学', '是研究數量', '屬於形式科學的一種', '數學利用抽象化
和邏輯推理', '從計數']
['蘇格拉底之死', '由雅克', '路易', '大卫所繪', '年', '哲學', '是研究普遍的',
'基本问题的学科', '包括存在', '知识']
['文學', '在狭义上', '是一种语言艺术', '亦即使用语言文字为手段', '形象化地反映客
观社会生活', '表达主观作者思想感情的一种艺术', '文学不仅强调传达思想观念', '更强
调传达方式的独特性', '且讲究辞章的美感', '文学']
```

In [15]:

```
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")
```

```
[2023-05-14 02:47:24] 已寫入 99999 篇斷詞文章
[2023-05-14 02:51:14] 已寫入 199999 篇斷詞文章
[2023-05-14 02:58:02] 已寫入 299999 篇斷詞文章
[2023-05-14 03:02:42] 已寫入 399999 篇斷詞文章
```

In [17]:

```
%%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_counts)

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

Use 12 workers to train Word2Vec (dim=300)  
 CPU times: total: 36min 29s  
 Wall time: 9min 28s

In [19]:

```
!dir word2vec.zh*
```

磁碟區 C 中的磁碟是 OS  
 磁碟區序號: 2E58-FFAB

C:\Users\user\Desktop\notebook\nlp\HW4 的目錄

```
2023/05/14 上午 03:41      58,888,453 word2vec.zh.300.model
2023/05/14 上午 03:41    1,894,270,928 word2vec.zh.300.model.syn1neg.npy
2023/05/14 上午 03:40    1,894,270,928 word2vec.zh.300.model.wv.vectors.n
py
          3 個檔案    3,847,430,309 位元組
          0 個目錄    233,171,443,712 位元組可用
```

In [21]:

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

(1578559, 300)

Out[21]:

```
array([[ -1.6653749e+00,  1.0125446e+00, -2.3497075e-01, ...,
         6.5279901e-01, -6.3151971e-02, -3.9201072e-01],
       [ -1.1377993e+00,  3.5012981e-01, -1.2351167e+00, ...,
         2.3135342e-01,  1.4836991e-01, -2.0512626e+00],
       [ -1.2004058e+00,  2.7550453e-01, -1.2185031e+00, ...,
        -7.0264214e-01,  2.3253256e-01, -1.2694845e+00],
       ...,
       [ -6.8653323e-02,  5.9258785e-02,  2.8251331e-02, ...,
        -3.3067100e-02,  1.9669712e-02,  7.4995020e-03],
       [ -2.5511291e-02,  3.4906086e-02,  3.3190895e-03, ...,
        -2.7539186e-02, -6.2455032e-03,  1.2487747e-03],
       [ -2.2402661e-02, -4.6018749e-02,  1.7832810e-02, ...,
         8.2145467e-02, -2.6185357e-03,  3.2317400e-02]], dtype=float32)
```

In [22]:

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

(300,)

Out[22]:

In [23]:

```
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]:

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

Out[24]:

```
[('飲品', 0.8991072773933411),
 ('服飾', 0.8648388981819153),
 ('化妝品', 0.8595173954963684),
 ('零食', 0.8388224840164185),
 ('冰淇淋', 0.8376069664955139),
 ('手錶', 0.8360370993614197),
 ('食品', 0.8342810869216919),
 ('咖啡', 0.8305132389068604),
 ('炸雞', 0.8282167911529541),
 ('家電', 0.8261498212814331)]
```

```

array([[1.73590153e-01, 1.05861150e-01, 3.86546999e-01, 1.94412321e-01,
7.13969707e-01, -1.70435771e-01, -7.49925196e-01, 6.25101864e-01,
model.wv.most_similar('car'), 6.235216e-01, -3.97005975e-01, 1.40192702e-01,
-3.94991428e-01, 3.90102625e-01, -9.47430074e-01, -1.05168450e+00,
Out[25]: 6.88838840e-01, 8.04549605e-02, -8.90006572e-02, -1.12779295e+00,
[('truck', 2.23094583e-01, 3.04401278e-01, -1.58859994e-02, 3.81029606e-01,
6.79669797e-01, 2.04536229e-01, 2.28729635e-01, -8.49799871e-01,
('motor', 0.72244606e-01, 4.11285411e-01, 6.18912280e-01, -8.74497294e-01, -2.23667756e-01,
('seat', 0.72219462e-01, 0.76080333e-01, 2.630227e+00, -2.07038015e-01, -4.31582600e-01,
('wagon', 0.71978807e-01, 1.12210088e-01, 5.08446206e-01, 4.24137592e-01, -7.33164847e-01,
('saloon', 0.51020658e-01, 5.05088043e-01, 2.30213711e-01, -7.49031484e-01, -4.94351327e-01,
('convertible', 0.62770885e-01, 0.78074788e-01, 0.60239692e-01, -1.03193270e-01, 1.33800149e-01, -3.36436629e-01,
('cadillac', 0.31932703e-01, 0.68251031e-01, 0.8117196e-02, 1.33800149e-01, -3.36436629e-01,
('cab', -1.22809696e+00, 0.41382124e-01, 2.93678939e-01, -5.99473953e-01,
('coupe', 0.63600788e-02, 0.62879312e-01, 3.61034691e-01, 4.62778509e-01,
('volkswagen', 0.32185233e-02, 0.53500372e-01, -1.22598958e+00, -3.50774705e-01,
3.17766547e-01, -8.32236469e-01, -1.27992025e-02, -7.29819611e-02,
-1.62444770e-01, -1.00119698e+00, -2.85759956e-01, -1.24853623e+00,
In [26]: 2.90425457e-02, 3.56662571e-01, 4.54109460e-02, -2.17015579e-01,
-9.15125012e-01, -6.47101223e-01, -5.15292466e-01, -3.99453014e-01,
model.wv.most_similar('facebook'), 1.51281304e+00, 7.53818372e-01, -4.24622357e-01, 1.17260683e+00,
Out[26]: 3.16786021e-02, 2.06181437e-01, 1.61036134e+00, -5.78405261e-01,
-8.51928413e-01, 8.43687952e-02, -4.49963510e-01, 1.23113357e-01,
[('instagram', 7.94604059e-01, 3.41637837e-01, -2.09587976e-01, -4.49379459e-02,
('臉書', 5.07824695e-01, 1.36519447e-01, -2.02043995e-01, -1.49788216e-01,
('專頁', 0.67447304e-01, 0.82745411e-01, -1.99506998e-01, -5.50231695e-01,
('twitter', 4.82006728e-01, 0.10340171e-01, -4.61400062e-01, -5.81640720e-01,
('myspace', 1.44017608e-01, 0.50872802e-01, -1.03228383e-01, 1.06252098e+00,
('facebook', 0.6950004e-02, 0.34270287e-01, -2.92005807e-01, 1.03278942e-01,
('新浪微博', 9.74491369e-01, 1.38034311e-01, -1.02526166e-01, -6.23101771e-01,
('微博', 2.04825340e-01, 0.57220531e-01, 2.64037788e-01, 1.61983166e-02,
('blog', 5.91807714e-01, 0.16642208e-01, 1.34670877e+00, 3.28007489e-01,
('推特', 5.60946849e-01, 0.53656750e-01, 1.27209783e+00, 1.63879126e-01,
-4.99410152e-01, -7.86466122e-01, -3.09436738e-01, -1.07434607e+00,
3.24648440e-01, 4.84539241e-01, 1.60598442e-01, 3.43937129e-01,
In [27]: 7.64592946e-01, 3.53211552e-01, 6.96989894e-01, -1.54561117e-01,
model.wv.most_similar('詐欺'), 6.05747402e-01, 0.55276591e-01, 1.76422620e+00, -7.13210106e-01,
1.13959730e-01, 5.69492638e-01, 5.33497274e-01, 8.00196409e-01,
Out[27]: 7.35539973e-01, -1.50320217e-01, 8.22590470e-01, 1.20513044e-01,
3.83800447e-01, 7.49204934e-01, 8.36285353e-01, -1.52555555e-01,
[('盜竊', 2.01328251e-01, 0.69223833e-01, 2.49911606e-01, 3.16327870e-01,
('賣淫', 0.86136309e-01, 0.32255744e-01, -2.73749411e-01, -7.68475235e-01,
('欺詐', 0.85443328e-01, 0.30639617e-01, 4.92964953e-01, -1.20616567e+00,
('洗錢', 0.85520381e-01, 0.52014168e-01, 6.21262938e-02, -2.04415902e-01,
('民事訴訟', 5.40074366e-01, 1.71929718e-01, -6.49453253e-02, -5.63927472e-01,
('性騷擾', 0.84501027e-01, 0.24945077e-01, -4.88689452e-01, 1.36920583e+00,
('解決問題', 1.03118789e+00, 1.96478754e-01, 1.72282353e-01, 1.24037065e-01,
('竊盜', 0.84155446e-01, 0.69698538e-01, -7.31955171e-01, 6.96158707e-01,
('和理非', 0.84014451e-01, 0.375366e-01, -6.69988871e-01, -9.62584987e-02,
('誇張', 0.83983457e-01, 0.470461e-01, 4.35587279e-02, 1.55492082e-01,
-8.23529959e-01, -1.94951549e-01, 6.33303523e-02, 2.46968761e-01,
-2.25162521e-01, -4.88768257e-02, 1.43074006e-01, 8.27811882e-02,
-1.29040927e-01, 3.74063522e-01, -2.93814331e-01, -2.68902749e-01,
-3.19289029e-01, 4.01403487e-01, 2.28707001e-01, -3.59716356e-01,
-4.98536706e-01, 4.63828802e-01, 7.16557920e-01, -4.69643474e-01,
-7.36510932e-01, 3.96820866e-02, 3.90475124e-01, 7.03137159e-01,
2.07976341e-01, -5.70604265e-01, -1.39502332e-01, -1.55387014e-01,
-5.03716350e-01, -1.28435120e-01, 6.68973029e-02, -5.92034400e-01,
3.32382619e-01, -1.05778563e+00, 5.46518207e-01, 1.06559169e+00,
-5.38995536e-03, -2.74318933e-01, 4.60235596e-01, 8.44479680e-01,
-5.21656930e-01, -2.53093421e-01, -1.29771680e-01, -2.67023832e-01,
-7.79551446e-01, -2.38149717e-01, -7.04342782e-01, 5.23818374e-01,

```



```

-5.01899838e-01, -4.01827455e-01, -1.32295892e-01, 1.27258420e+00,
In [28]: 4.06621426e-01, -6.86595798e-01, -6.35920823e-01, 1.81315448e-02,
model.wv.most_similar("合約")
3.04722726e-01, -5.97079575e-01, 1.26926899e-01, -5.36140323e-01,
-9.94297341e-02, 4.50762063e-01, 3.16474331e-03, 1.09551442e+00,
Out[28]: 2.55481273e-01, -6.15745068e-01, -4.74356450e-02, 4.97288316e-01,
3.33022565e-01, -4.93229240e-01, 3.45670253e-01, 1.51718944e-01,
[('總值', 2.3781860728898696), ('年內', 5.01899838e-01, -4.01827455e-01, -1.32295892e-01, 1.27258420e+00,
('年內', 5.01899838e-01, -4.01827455e-01, -1.32295892e-01, 1.27258420e+00,
('耗資超過', 5.13736574e-01, -7.41752545e-01, 4.57816347e-02, 1.82818204e-01,
('預算為', 6.2899435e-01, -3.07924746773e-01, -2.99062490e-01, -8.25757682e-01,
('並被罰款', 4.49030697e-01, -1.479493159e-01, -2.16443747e-01, -7.52396941e-01,
('億新台幣', 1.03600447e-01, 3.62300508e-01, -4.12097313e-02, 2.62476176e-01,
('據了解', 8.7097108e-01, -3.5542823e-01, -5.69173276e-01, -3.56761992e-01,
('花費', 3.62850126e-01, 3.4777815691275e-01, 3.02805975e-02, -5.12886643e-0
2), ('被罰款', 0.7666801810264587),
('萬美金', -4.1616029977798462)]
dtype: float32

```

In [29]:

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

Out[29]:

0.5345687

In [30]:

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

Out[30]:

0.3341626

In [31]:

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

Loading word2vec.zh.300.model...

In [32]:

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

Out[32]:

True

In [ ]: