In [1]:
```
%load_ext memory_profiler
```

In [2]:
```
import os
import gensim
import jieba
import zhconv
from gensim.corpora import WikiCorpus
from datetime import datetime
from typing import List

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

jieba.set_dictionary("dict.txt.big")
```

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) = 27e78ff
901bcd3803955d9373537dd3f
zhwiki-20230501-pages-articles-multistream.xml.bz2: bzip2 compress
ed data, block size = 900k
```

In [4]:
```
import spacy
spacy.cli.download("zh_core_web_sm")
spacy.cli.download("en_core_web_sm")

nlp_zh = spacy.load("zh_core_web_sm")
nlp_en = spacy.load("en_core_web_sm")
```

```
Collecting zh-core-web-sm==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/zh_core_web_sm-3.5.0/zh_core_web_sm-3.5.0-py3-none-any.whl (https://github.com/explosion/spacy-models/releases/download/zh_core_web_sm-3.5.0/zh_core_web_sm-3.5.0-py3-none-any.whl) (48.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 48.5/48.5 MB 52.1 kB/s eta 0:00:00
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /Users/hsiu/opt/anaconda3/lib/python3.9/site-packages (from zh-core-web-sm==3.5.0) (3.5.2)
Collecting spacy-pkuseg<0.1.0,>=0.0.27
  Downloading spacy_pkuseg-0.0.32-cp39-cp39-macosx_11_0_arm64.whl (2.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.3/2.3 MB 56.0 kB/s eta 0:00:00
Requirement already satisfied: jinja2 in /Users/hsiu/opt/anaconda3/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.11.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /Use
```

In [5]:
```
STOPWORDS =  nlp_zh.Defaults.stop_words | nlp_en.Defaults.stop_word

for word in STOPWORDS.copy():
    STOPWORDS.add(zhconv.convert(word, "zh-tw"))
```

In [6]:
```
def preprocess_and_tokenize(text, token_min_len = 1, token_max_len
    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 n
    ]
```

In [7]:
```
%%time
%%memit

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

```
Parsing zhwiki-20230501-pages-articles-multistream.xml.bz2...

/Users/hsiu/opt/anaconda3/lib/python3.9/site-packages/gensim/utils
.py:1332: UserWarning: detected OSX with python3.8+; aliasing chun
kize to chunkize_serial
  warnings.warn("detected %s; aliasing chunkize to chunkize_serial
" % entity)

peak memory: 2102.64 MiB, increment: 1444.09 MiB
CPU times: user 11min 15s, sys: 1min 52s, total: 13min 8s
Wall time: 15min 21s
```

In [8]:
```
g = wiki_corpus.get_texts()
print(next(g)[:10])
print(next(g)[:10])
print(next(g)[:10])
```

```
['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 [9]:
```
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(datetime.now()):.19}] 已寫入 {texts_num} 篇
```

```
[2023-05-11 10:24:06] 已寫入 99999 篇斷詞文章
[2023-05-11 10:26:05] 已寫入 199999 篇斷詞文章
[2023-05-11 10:29:21] 已寫入 299999 篇斷詞文章
[2023-05-11 10:31:46] 已寫入 399999 篇斷詞文章
```

In [10]:
```
%%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_d


sentences = word2vec.LineSentence(WIKI_SEG_TXT)

model = word2vec.Word2Vec(sentences, vector_size=word_dim_size, wor

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

```
Use 8 workers to train Word2Vec (dim=300)
CPU times: user 29min 2s, sys: 1min 59s, total: 31min 1s
Wall time: 7min 56s
```

In [11]:
```
! ls word2vec.zh*
```

```
word2vec.zh.300.model                word2vec.zh.300.model.wv.vect
ors.npy
word2vec.zh.300.model.syn1neg.npy
```

In [12]:
```
! du -sh word2vec.zh*
```

```
 57M    word2vec.zh.300.model
1.8G    word2vec.zh.300.model.syn1neg.npy
1.8G    word2vec.zh.300.model.wv.vectors.npy
```

In [13]:
```
print(model.wv.vectors.shape)
model.wv.vectors
```

```
(1578559, 300)
```

Out[13]:
```
array([[-1.7076457e+00,  1.7358593e+00, -3.4208825e-01, ...,
         6.5092337e-01,  6.4788365e-01,  2.2596502e-01],
       [-8.6641109e-01,  1.0497972e+00,  4.8340130e-01, ...,
         2.2776024e-01, -5.6819314e-01,  3.1535363e-01],
       [-1.3289380e+00,  1.2796842e+00,  3.3163098e-01, ...,
         6.8806994e-01, -4.8488963e-01,  5.2340209e-01],
       ...,
       [-3.0601058e-02,  5.1672857e-02,  1.3539110e-02, ...,
        -1.4815503e-02,  5.7537202e-02, -1.9980976e-02],
       [-3.5888821e-02,  4.1780226e-02,  6.7084683e-03, ...,
        -2.1981372e-02,  7.2206617e-03, -9.2937918e-03],
       [-4.5324679e-02, -1.6357239e-02, -9.1987170e-02, ...,
         4.6356138e-02, -4.8235804e-03,  1.1394625e-03]], dtype=fl
oat32)
```

In [14]:
```
vec = model.wv['數學家']
print(vec.shape)
vec
```

```
(300,)
```

Out[14]:
```
array([ 0.53049093,  0.01827557,  0.2647456 ,  0.30384144,  0.8471
7506,
       -0.3160529 , -0.84586823,  0.37918493, -0.4042048 ,  0.0628
6512,
       -0.43317407,  0.15615006,  0.1487516 ,  0.7061684 , -0.9255
093 ,
       -0.9253154 , -0.759502  ,  0.18732086, -0.22279754, -1.3600
307 ,
       -0.13007402,  0.33115828,  0.2514567 ,  0.3103663 ,  0.5041
195 ,
        0.5105871 ,  0.275075  , -0.99158734, -0.8061154 ,  0.6094
9665,
       -0.97135323, -0.36245635,  0.5228062 , -0.9929437 , -0.3768
951 ,
       -0.27796376,  0.32308862,  0.17225985, -0.15119103, -0.5225
684 ,
        0.6778689 ,  0.5649924 , -0.4203485 , -0.3324206 , -0.8110
069
```

In [15]:
```
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 [16]:
```
model.wv.most_similar("飲料", topn=10)
```

Out[16]:
```
[('飲品', 0.9133813977241516),
 ('炸雞', 0.8783719539642334),
 ('冰淇淋', 0.8746067881584167),
 ('服飾', 0.8678401708602905),
 ('化妝品', 0.8615073561668396),
 ('零食', 0.8536527156829834),
 ('啤酒', 0.8466242551803589),
 ('珠寶', 0.8449875116348267),
 ('電子產品', 0.8367608785629272),
 ('食品', 0.8348352313041687)]
```

In [17]:
```
model.wv.most_similar("car")
```

Out[17]:
```
[('truck', 0.7824944257736206),
 ('brake', 0.7262388467788696),
 ('貨卡車', 0.716031551361084),
 ('motorcycle', 0.7159311771392822),
 ('motor', 0.7153626680374146),
 ('volkswagen', 0.7126927971839905),
 ('hybrid', 0.7122430205345154),
 ('saloon', 0.7067892551422119),
 ('cadillac', 0.7067204713821411),
 ('convertible', 0.7063345313072205)]
```

In [18]:
```
model.wv.most_similar("facebook")
```

Out[18]:
```
[('instagram', 0.879693329334259),
 ('臉書', 0.8340373039245605),
 ('專頁', 0.7984201312065125),
 ('twitter', 0.779519259929657),
 ('xanga', 0.7731849551200867),
 ('facebook專頁', 0.7496058344841003),
 ('myspace', 0.7484684586524963),
 ('微博', 0.746212363243103),
 ('推特', 0.7453634142875671),
 ('新浪微博', 0.7418176531791687)]
```

In [19]:
```
model.wv.most_similar("詐欺")
```

Out[19]:
```
[('恐嚇', 0.8686190843582153),
 ('盜竊', 0.8680432438850403),
 ('販毒', 0.8651722073554993),
 ('性騷擾', 0.8630328178405762),
 ('洗錢', 0.856837272644043),
 ('脅迫', 0.851883590221405),
 ('毆打', 0.848781943321228),
 ('的質疑', 0.8473358154296875),
 ('欺詐', 0.846764326095581),
 ('搶劫', 0.8438825607299805)]
```

In [20]:
```
model.wv.most_similar("合約")
```

Out[20]:
```
[('年內', 0.8128535151481628),
 ('總值', 0.8031351566314697),
 ('據了解', 0.7691568732261658),
 ('卻在', 0.7609128355979919),
 ('耗資超過', 0.7589302659034729),
 ('鑑於', 0.7558416128158569),
 ('聯盟於', 0.751262366771698),
 ('億台幣', 0.7507293820381165),
 ('億美金', 0.74944007396698),
 ('往後', 0.7488248348236084)]
```

In [21]:
```
model.wv.similarity("連結", "鏈結")
```

Out[21]:
```
0.5377023
```

In [22]:
```
model.wv.similarity("連結", "陰天")
```

Out[22]:
```
0.30575657
```

In [23]:
```
print(f"Loading {output_model}...")
new_model = word2vec.Word2Vec.load(output_model)
```

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

In [24]:
```
model.wv.similarity("連結", "陰天") == new_model.wv.similarity("連結"
```

Out[24]:
```
True
```