

Word2Vec-以 gensim 訓練中文詞向量

參考及引用資料來源

- [1] [zake7749-使用 gensim 訓練中文詞向量](#)
- [2] [gensim/corpora/wikicorpus](#)
- [Word2Vec的簡易教學與參數調整指南](#)
- [zhconv](#)
- [jieba](#)

```
In [1]: %load_ext memory_profiler
!pip install -q zhconv
```

確認相關 Packages

```
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
```

準備中文訓練文本

訓練文本來源: [維基百科資料庫](#)

要訓練詞向量，第一步當然是取得資料集。由於 word2vec 是基於非監督式學習，訓練集一定一定要越大越好，語料涵蓋的越全面，訓練出來的結果也會越漂亮。[1]

- [zhwiki-20210101-pages-articles.xml.bz2](#) (1.9 GB)

```
wget "https://dumps.wikimedia.org/zhwiki/20210101/zhwiki-20210101-pages-articles.xml.bz2"
```

目前已經使用另一份 Notebook ([維基百科中文語料庫 zhWiki_20210101](#)) 下載好中文維基百科語料，並可以直接引用

```
In [3]: ZhWiki = "/Users/hungshihching/Desktop/uni/NLP/hw4/zhwiki-20230501-pages-art
!du -sh $ZhWiki
!md5sum $ZhWiki
!file $ZhWiki
```

```
2.6G      /Users/hungshihching/Desktop/uni/NLP/hw4/zhwiki-20230501-pages-artic
les-multistream.xml.bz2
zsh:1: command not found: md5sum
/Users/hungshihching/Desktop/uni/NLP/hw4/zhwiki-20230501-pages-articles-mult
istream.xml.bz2: bzip2 compressed data, block size = 900k
```

中文文本前處理

在正式訓練 `Word2Vec` 之前，其實涉及了文本的前處理，本篇的處理包括如下三點 (而實務上對應的不同使用情境，可能會有不同的前處理流程):

- 簡轉繁: `zhconv`
- 中文斷詞: `jieba`
- 停用詞

簡繁轉換

wiki 文本其實摻雜了簡體與繁體中文，比如「数学」與「數學」，這會被 `word2vec` 當成兩個不同的詞。[\[1\]](#)

所以我們在斷詞前，需要加上簡繁轉換的手續

以下範例使用了較輕量的 Package `zhconv`，
若需要更高的精準度，則可以參考 [OpenCC](#)

```
In [4]: zhconv.convert("这原本是一段简体中文", "zh-tw")
```

```
Out[4]: '這原本是一段簡體中文'
```

中文斷詞

使用 `jieba` `jieba.cut` 來進行中文斷詞，
並簡單介紹 `jieba` 的兩種分詞模式:

- `cut_all=False` **精確模式**，試圖將句子最精確地切開，適合文本分析；
- `cut_all=True` **全模式**，把句子中所有的可以成詞的詞語都掃描出來，速度非常快，但是不能解決歧義；

而本篇文本訓練採用**精確模式** `cut_all=False`

```
In [5]: 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 /Users/hungshihching/Desktop/uni/NLP/hw4/dict.txt.
big ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.u146ce28257808dd681bfe691c030addc.cache
Loading model cost 1.022 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 [6]: 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', 'interest
ing', '吧', '?']
```

引入停用詞表

停用詞就是像英文中的 **the, a, this**，中文的**你我他**，與其他詞相比顯得不怎麼重要，對文章主題也無關緊要的，

是否要使用停用詞表，其實還是要看你的應用，也有可能保留這些停用詞更能達到你的目標。[\[1\]](#)

- [Is it compulsory to remove stop words with word2vec?](#)
- [The Effect of Stopword Filtering prior to Word Embedding Training](#)

以下範例還是示範引入停用詞表，而停用詞表網路上有各種各樣的資源

剛好 `kaggle`，環境預設有裝 `spacy`，

就順道引用 `spacy` 提供的停用詞表吧 (實務上stopwords 應為另外準備好且檢視過的靜態文檔)

```
In [9]: 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)}")
```

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 (48.5 MB)

48.5/48.5 MB 4.7 MB/s eta 0:0

0:0000:0100:01

Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from zh-core-web-sm==3.5.0) (3.5.2)

Requirement already satisfied: spacy-pkuseg<0.1.0,>=0.0.27 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from zh-core-web-sm==3.5.0) (0.0.32)

Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (1.1.1)

Requirement already satisfied: pydantic!=1.8,!1.8.1,<1.11.0,>=1.7.4 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (1.10.7)

Requirement already satisfied: requests<3.0.0,>=2.13.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.30.0)

Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (3.0.12)

Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (1.0.4)

Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (3.3.0)

Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (1.0.9)

Requirement already satisfied: packaging>=20.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (23.0)

Requirement already satisfied: setuptools in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (66.0.0)

Requirement already satisfied: pathy>=0.10.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (0.10.1)

Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.4.6)

Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.0.7)

Requirement already satisfied: numpy>=1.15.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (1.24.3)

Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (3.0.8)

Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (8.1.10)

Requirement already satisfied: Jinja2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (3.1.2)

Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.0.8)

Requirement already satisfied: typer<0.8.0,>=0.3.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (0.4.1)

```

>zh-core-web-sm==3.5.0) (0.7.0)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (4.65.0)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (6.3.0)
Requirement already satisfied: typing-extensions>=4.2.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from pydantic!=1.8,!
=1.8.1,<1.11.0,>=1.7.4->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (4.5.0)
Requirement already satisfied: idna<4,>=2.5 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2023.5.7)
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (3.1.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.0.2)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (0.0.4)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (0.7.9)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (8.1.3)
Requirement already satisfied: MarkupSafe>=2.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from jinja2->spacy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (2.1.1)
✓ Download and installation successful
You can now load the package via spacy.load('zh_core_web_sm')
Collecting en-core-web-sm==3.5.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.5.0/en_core_web_sm-3.5.0-py3-none-any.whl (12.8 MB)
    12.8/12.8 MB 1.1 MB/s eta 0:00
0:0000:0100:01m
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from en-core-web-sm==3.5.0) (3.5.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.30.0)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (8.1.10)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.0.4)
Requirement already satisfied: setuptools in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (66.0.0)
Requirement already satisfied: packaging>=20.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (23.0)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.7.0)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->

```

```

>en-core-web-sm==3.5.0) (2.4.6)
Requirement already satisfied: Jinja2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.1.2)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.8)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (4.65.0)
Requirement already satisfied: pathy>=0.10.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.10.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.3.0)
Requirement already satisfied: pydantic!=1.8,!1.8.1,<1.11.0,>=1.7.4 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.10.7)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.1.1)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (6.3.0)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.12)
Requirement already satisfied: numpy>=1.15.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.24.3)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.0.9)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.7)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.0.8)
Requirement already satisfied: typing-extensions>=4.2.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from pydantic!=1.8,!1.8.1,<1.11.0,>=1.7.4->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (4.5.0)
Requirement already satisfied: idna<4,>=2.5 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.4)
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.1.0)
Requirement already satisfied: certifi>=2017.4.17 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2023.5.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.2)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from thinc<8.2.0,>=8.1.8->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (0.0.4)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from typer<0.8.0,>=0.3.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (8.1.3)

```


Requirement already satisfied: MarkupSafe>=2.0 in /Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from jinja2->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.1.1)

✓ Download and installation successful

You can now load the package via `spacy.load('en_core_web_sm')`

--

中文停用詞 Total=1891: ['古來', '不久', '(', '不會', '何必', '不力', '開外', '我是', '莫', '不止', '背地里', '尽然', '騰', '今后', ' [④d] ', '八成', '居然', ' [③F] ', '零', '即是说'] ...

--

英文停用詞 Total=326: ['twelve', 'n't', 'still', 'himself', 'with', 'my', 'in deed', 'make', 'sometimes', "'s", 'well', 'below', 'otherwise', 'ever', 'the', 'yet', 'beforehand', 'however', 'noone', 'nevertheless'] ...

```
In [10]: 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"))

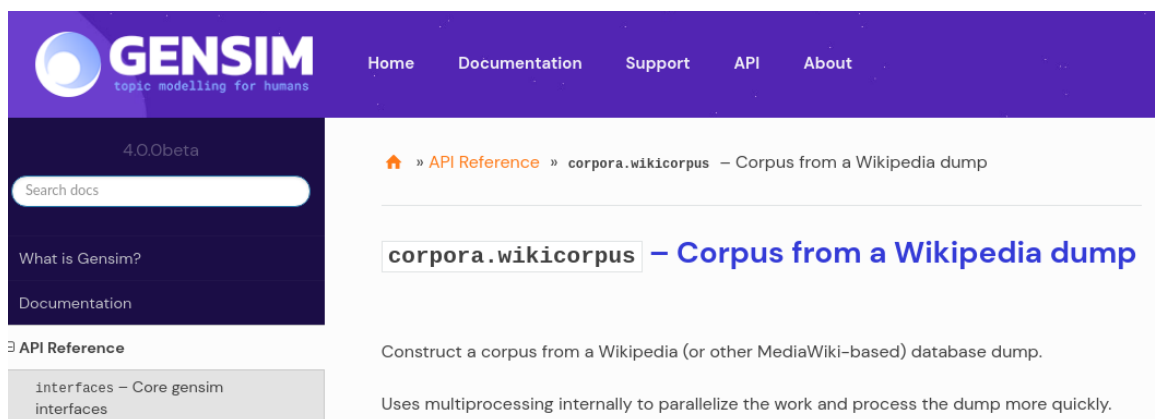
print(len(STOPWORDS))
```

2222

3005

讀取 wiki 語料庫，並且進行前處理和斷詞

維基百科 (wiki.xml.bz2) 下載好後，先別急著解壓縮，因為這是一份 xml 文件，裏頭佈滿了各式各樣的標籤，我們得先想辦法送走這群不速之客，不過也別太擔心，gensim 早已看穿了一切，藉由調用 `wikiCorpus`，我們能很輕鬆的只取出文章的標題和內容。[1]



The screenshot shows the Gensim website with a purple header. The main content area is white and displays the API Reference for `corpora.wikicorpus`. The title is `corpora.wikicorpus` – Corpus from a Wikipedia dump. The description states: "Construct a corpus from a Wikipedia (or other MediaWiki-based) database dump. Uses multiprocessing internally to parallelize the work and process the dump more quickly." The left sidebar contains links for Home, Documentation, Support, API, and About, as well as a search bar and a section for API Reference.

[2]

Supported dump formats:

- `<LANG>wiki-<YYYYMMDD>-pages-articles.xml.bz2`
- `<LANG>wiki-latest-pages-articles.xml.bz2`

The documents are extracted on-the-fly, so that the whole (massive) dump can stay compressed on disk.

```
In [11]: def preprocess_and_tokenize(
    text: str, token_min_len: int=1, token_max_len: int=15, lower: bool=True)
    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 [12]: 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
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 [16]: %%time
%%memit
from utils import preprocess_and_tokenize
from typing import List

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

Parsing /Users/hungshihching/Desktop/uni/NLP/hw4/zhwiki-20230501-pages-articles-multistream.xml.bz2...
```



```

/Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages/gensim/
utils.py:1333: UserWarning: detected OSX with python3.8+; aliasing chunki
ze to chunkize_serial
  warnings.warn("detected %s; aliasing chunkize to chunkize_serial" % entit
y)
Building prefix dict from the default dictionary ...
Building prefix dict from the default dictionary ...
Building prefix dict from the default dictionary ...
Building prefix dict from the default dictionary ...
Building prefix dict from the default dictionary ...
Building prefix dict from the default dictionary ...
Building prefix dict from the default dictionary ...
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Loading model cost 1.244 seconds.
Prefix dict has been built successfully.
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Loading model cost 1.228 seconds.
Prefix dict has been built successfully.
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Dumping model to file cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/
T/jieba.cache
Loading model cost 1.238 seconds.
Prefix dict has been built successfully.
Loading model cost 1.253 seconds.
Prefix dict has been built successfully.
Loading model cost 1.244 seconds.
Prefix dict has been built successfully.
Loading model cost 1.265 seconds.
Prefix dict has been built successfully.
Loading model cost 1.219 seconds.
Prefix dict has been built successfully.
peak memory: 2132.29 MiB, increment: 1419.94 MiB
CPU times: user 33min 1s, sys: 3min 43s, total: 36min 44s
Wall time: 1h 17min 35s

```

初始化 WikiCorpus 後，能藉由 `get_texts()` 可迭代每一篇文章，它所回傳的是一個 `tokens list`，我以空白符將這些 `tokens` 串接起來，統一輸出到同一份文字檔裡。這邊要注意一件事，`get_texts()` 受 `article_min_tokens` 參數的限制，只會回傳內容長度大於 50 (default) 的文章。

- **article_min_tokens** (*int, optional*) – Minimum tokens in article. Article will be ignored if number of tokens is less.

秀出前 3 篇文章的前10 個 token

```

In [17]: 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/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/ry/phxc250s3lx_m2h646xkjb1h0000gn/T/ji
eba.cache
Loading model cost 1.146 seconds.
Prefix dict has been built successfully.
Loading model cost 1.090 seconds.
Prefix dict has been built successfully.
Loading model cost 1.072 seconds.
Prefix dict has been built successfully.
Loading model cost 1.099 seconds.
Prefix dict has been built successfully.
Loading model cost 1.091 seconds.
Prefix dict has been built successfully.
Loading model cost 1.094 seconds.
Prefix dict has been built successfully.
Loading model cost 1.110 seconds.
Prefix dict has been built successfully.
[2023-05-12 01:00:35] 已寫入 99999 篇斷詞文章
[2023-05-12 01:08:13] 已寫入 199999 篇斷詞文章
[2023-05-12 01:15:11] 已寫入 299999 篇斷詞文章
[2023-05-12 01:22:08] 已寫入 399999 篇斷詞文章
[2023-05-12 01:29:20] 已寫入 499999 篇斷詞文章
[2023-05-12 01:35:26] 已寫入 599999 篇斷詞文章
[2023-05-12 01:42:28] 已寫入 699999 篇斷詞文章
[2023-05-12 01:49:28] 已寫入 799999 篇斷詞文章

```

訓練 Word2Vec

```

In [23]: %%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})

# 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_

```

```
# 儲存模型
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 58min 1s, sys: 2min 4s, total: 3h 6s
 Wall time: 35min 22s

儲存的模型總共會產生三份檔案

```
In [24]: ! ls word2vec.zh*
```

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

```
In [25]: !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
```

查看模型以及詞向量實驗

模型其實就是巨大的 Embedding Matrix

```
In [26]: print(model.wv.vectors.shape)
model.wv.vectors
```

```
Out[26]: (1801313, 300)
array([[ 1.7665801, -1.1271235, -0.6275652, ..., -1.9071926,
         1.9823675, -0.59206486],
       [ 0.54720587, -0.77660155, -2.002865, ..., -0.8983154,
         1.8231131, -0.75861055],
       [ 1.9373907, -2.5447996, -0.2674481, ..., -0.4668254,
         1.485772, -0.46241406],
       ...,
       [ 0.00992353, 0.06258205, -0.00734942, ..., 0.01078969,
        -0.08407021, 0.01095441],
       [ 0.04854058, 0.01850454, -0.00345393, ..., 0.03257393,
        -0.03420385, 0.02309467],
       [ 0.03171879, 0.03774663, 0.04955105, ..., 0.05260075,
        -0.04511588, -0.01988711]], dtype=float32)
```

收錄的詞彙

```
In [29]: print(f"總共收錄了 {len(model.wv.key_to_index)} 個詞彙")

print("印出 20 個收錄詞彙:")
print(list(model.wv.key_to_index.keys())[:20])
'''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])'''
```

總共收錄了 1801313 個詞彙

印出 20 個收錄詞彙:

```
['年', '月', '日', '於', '為', '「', '與', '後', '臺', '中', '10', '對', '12',
 '11', '軍', '參考', '來', '時', '小行星', '馬']
```

詞彙的向量

```
In [32]: vec = model.wv['數學家']  
         print(vec.shape)  
         vec  
  
(300,)
```

```
Out[32]: array([ 1.74011815e+00,  1.95135880e+00,  1.76903471e-01,  1.34683633e+00,
 -7.78381288e-01,  1.07351613e+00, -2.26528168e+00, -4.39695454e+00,
  1.60824370e+00,  6.74168885e-01, -2.01055431e+00, -2.95352602e+00,
 -8.38819623e-01, -3.47166671e-03, -1.45592725e+00,  1.05842268e+00,
 -4.00823742e-01, -1.79298186e+00,  1.10139084e+00, -1.31101739e+00,
  9.98532593e-01, -7.65580595e-01, -1.52170980e+00, -1.07541192e+00,
  5.05351782e-01,  6.10343277e-01,  1.01701379e-01,  1.28162038e+00,
  6.18267477e-01,  2.23264647e+00, -1.12130570e+00, -8.34661648e-02,
  7.94168174e-01,  1.98880661e+00, -8.38796020e-01, -1.05934846e+00,
  3.47467804e+00,  1.67059243e+00, -7.47771740e-01,  3.98761630e-02,
  3.01819730e+00, -1.66275930e+00,  6.04396045e-01,  4.48689371e-01,
  6.60387725e-02, -3.01543951e-01,  7.61732697e-01,  1.09199500e+00,
 -2.13081741e+00, -6.27408743e-01,  6.92147017e-01,  1.69998899e-01,
 -3.98554385e-01,  7.33857036e-01,  2.33825660e+00, -1.19615030e+00,
 -2.03371406e+00, -6.08632565e-01, -1.57995969e-01,  3.07275444e-01,
 -4.85512316e-01,  8.29940677e-01,  3.26558375e+00,  1.14788294e+00,
  2.57875361e-02,  5.74688971e-01, -3.29742312e-01, -2.50302958e+00,
 -7.66694844e-01,  1.31902292e-01, -5.29394436e+00, -1.17092896e+00,
  1.01205087e+00,  2.59167600e+00, -1.71093142e+00, -1.47353439e-03,
 -6.30553905e-03,  1.90181756e+00,  2.46364966e-01, -1.40823638e+00,
  5.01711130e-01, -5.86952567e-01,  3.93168256e-02,  1.15085673e+00,
 -4.19566870e+00,  1.94269884e+00, -3.54155362e-01, -3.19538474e+00,
 -1.05039704e+00,  2.68587518e+00,  1.01687586e+00,  5.89128971e-01,
  7.79885948e-01,  1.18194306e+00, -7.34340191e-01, -3.16617787e-02,
 -4.20457602e-01,  1.14089525e+00, -2.33355626e-01,  2.41921559e-01,
  2.91108394e+00,  1.71129167e+00,  1.23318231e+00, -1.11128199e+00,
 -8.56924355e-02,  6.10284805e-01, -1.36568451e+00,  4.20828247e+00,
  2.24304691e-01,  2.87969375e+00, -1.86118329e+00,  9.24999714e-01,
 -6.43376186e-02, -1.32477522e+00,  2.87615478e-01,  1.06900573e+00,
 -2.30301067e-01,  1.81920052e+00, -9.53970030e-02,  1.09469128e+00,
 -1.54376340e+00, -1.16368973e+00, -1.98198950e+00, -1.50783825e+00,
 -5.73561311e-01,  3.03979897e+00,  1.24124348e+00,  1.31517875e+00,
  4.67630893e-01, -2.68420005e+00,  1.37751293e+00, -2.22385359e+00,
 -1.08808935e+00, -8.84614110e-01, -1.57952249e-01, -9.60997224e-01,
  1.98923934e+00,  4.27921712e-01,  3.17066121e+00,  5.77173591e-01,
  2.37883832e-02, -1.28705013e+00, -3.57446879e-01,  9.46666360e-01,
 -1.63065088e+00, -2.81343031e+00,  1.34118044e+00,  2.81249762e+00,
 -2.98990190e-01,  1.45472479e+00, -1.72422922e+00,  1.92167664e+00,
  5.59086204e-01,  9.96241331e-01, -1.91745687e+00, -1.13945782e+00,
  2.00155878e+00,  3.02035417e-02,  2.31300369e-01,  2.43352675e+00,
 -7.04378262e-02,  2.36642909e+00,  1.25676356e-02, -8.33705604e-01,
  1.85923982e+00, -8.01342010e-01, -9.57680196e-02, -1.13436770e+00,
  2.42100373e-01, -4.59679455e-01, -7.11464584e-01, -4.23632205e-01,
 -5.92963994e-01, -1.59384704e+00,  4.54612195e-01, -1.34560573e+00,
 -4.34021759e+00,  6.50437832e-01,  2.03798819e+00, -6.24823689e-01,
 -1.64691651e+00,  8.67111802e-01, -6.84349775e-01, -6.94151342e-01,
  2.68483758e-01,  3.59034598e-01, -2.41840816e+00, -3.39427686e+00,
 -1.27024853e+00,  1.58476830e+00,  3.24839801e-01, -1.45843554e+00,
  1.86541474e+00,  2.10164738e+00, -1.48921281e-01,  2.25778270e+00,
 -1.26199830e+00,  1.20313659e-01, -8.85447025e-01,  4.46925431e-01,
  1.76088738e+00, -1.98065913e+00,  1.67166710e+00, -1.16331327e+00,
 -1.12037575e+00,  7.55044758e-01, -1.35273576e-01,  1.28910780e+00,
  2.11596417e+00,  3.80887575e-02, -3.76932144e-01, -5.79229176e-01,
  3.75835323e+00,  2.78065205e-01,  2.17380166e+00,  1.57285690e+00,
 -1.33583844e+00,  1.76706111e+00,  8.60062987e-03,  1.29166603e+00,
 -2.16684639e-01,  4.16935310e-02, -1.06292140e+00,  8.72782350e-01,
  2.45772386e+00, -1.14450395e+00, -5.71814716e-01,  4.97798324e-01,
  1.38657641e+00,  3.26260424e+00,  1.75127149e-01,  7.46975183e-01,
  2.23785067e+00, -1.71099758e+00, -8.47493649e-01, -1.05481640e-01,
 -3.57688576e-01,  1.81262434e+00, -2.26629451e-01,  9.73981261e-01,
 -2.93491453e-01,  2.00788021e+00,  1.78873396e+00, -3.36846799e-01,
  6.34928703e-01,  1.76742232e+00, -1.96257517e-01,  8.09297711e-02,
 -1.53312340e-01,  2.88582993e+00, -3.31068397e-01,  5.93693733e-01,
  7.73647785e-01,  2.05837440e+00,  2.96037793e+00,  4.99703318e-01,
```

```
-1.26696539e+00, -1.00073421e+00, -3.78241420e-01, 3.03533959e+00,
-8.97487760e-01, 4.59813654e-01, -1.08569950e-01, -1.32857192e+00,
1.46145737e+00, -3.13317490e+00, 5.84039330e-01, -5.17652631e-01,
-2.82364416e+00, -7.90281773e-01, -4.04611863e-02, 3.08757752e-01,
9.66587424e-01, 7.34959096e-02, 2.19430280e+00, -1.23225641e+00,
8.60014975e-01, 3.44372302e-01, 6.61219954e-01, -8.44181716e-01,
2.37105799e+00, -1.08591938e+00, -4.03119564e-01, 3.47950011e-01,
2.92703599e-01, 1.67907536e+00, -1.33794749e+00, 2.28555813e-01,
-4.16531324e-01, 1.08214569e+00, 1.53097737e+00, -3.28087425e+00,
-1.89150643e+00, -5.89538991e-01, 1.05205834e-01, -2.05275640e-01,
-3.89074981e-01, 3.94866228e-01, 2.08917117e+00, -1.26067090e+00],
dtype=float32)
```

沒見過的詞彙

```
In [33]: 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"

查看前 10 名相似詞

`model.wv.most_similar` 的 `topn` 預設為 10

```
In [34]: model.wv.most_similar("飲料", topn=10)
```

```
Out[34]: [('飲品', 0.824095606803894),
('軟飲料', 0.6708950996398926),
('罐裝', 0.6678032875061035),
('果汁', 0.6509614586830139),
('酒類', 0.6470938920974731),
('含酒精', 0.6469045281410217),
('瓶裝', 0.6460524797439575),
('熱飲', 0.6239790916442871),
('無糖', 0.6226282715797424),
('利口酒', 0.6170544624328613)]
```

```
In [35]: model.wv.most_similar("car")
```

```
Out[35]: [('seat', 0.6847618818283081),
('truck', 0.678520143032074),
('tikita', 0.6600988507270813),
('chevrolet', 0.6424673199653625),
('saloon', 0.6400571465492249),
('automobile', 0.6392562389373779),
('cab', 0.6391147375106812),
('cars', 0.6383315324783325),
('wagon', 0.6346812844276428),
('luxury', 0.6337199211120605)]
```

```
In [36]: model.wv.most_similar("facebook")
```



```
Out[36]: [('臉書', 0.8104088306427002),
          ('專頁', 0.7592917680740356),
          ('instagram', 0.7500594854354858),
          ('面書', 0.717880368232727),
          ('貼文', 0.708476722240448),
          ('twitter', 0.6982824802398682),
          ('推特', 0.6892275214195251),
          ('臉書粉', 0.6869004964828491),
          ('臉書上', 0.6860032081604004),
          ('粉絲團', 0.6766610741615295)]
```

```
In [46]: model.wv.most_similar("欺詐")
```

```
Out[46]: [('逃稅', 0.6541138291358948),
          ('詐騙', 0.6440050601959229),
          ('欺詐性', 0.6029208302497864),
          ('敲詐', 0.6014794111251831),
          ('不道德', 0.5992634892463684),
          ('漏稅', 0.598671019077301),
          ('詐', 0.597423791885376),
          ('舞弊', 0.5942248702049255),
          ('偷稅', 0.5866773128509521),
          ('不誠實', 0.5776230096817017)]
```

```
In [38]: model.wv.most_similar("合約")
```

```
Out[38]: [('合同', 0.7676588296890259),
          ('新合約', 0.7592244744300842),
          ('年合約', 0.7211328148841858),
          ('簽約', 0.718211829662323),
          ('合約將', 0.6907707452774048),
          ('續約', 0.6809772253036499),
          ('之合約', 0.6638931632041931),
          ('合約並', 0.662800669670105),
          ('租約', 0.6360508799552917),
          ('其合約', 0.6279857158660889)]
```

計算 Cosine 相似度

```
In [39]: model.wv.similarity("連結", "鏈接")
```

```
Out[39]: 0.74120027
```

```
In [40]: model.wv.similarity("連結", "陰天")
```

```
Out[40]: 0.020136362
```

讀取模型

```
In [41]: print(f"Loading {output_model}...")
          new_model = word2vec.Word2Vec.load(output_model)

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

```
In [42]: model.wv.similarity("連結", "陰天") == new_model.wv.similarity("連結", "陰天")
```

```
Out[42]: True
```

In []: