fastText-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 0.990 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]:
```python
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 [7]:
```python
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.Defau
print("--")
print(f"英文停用詞 Total={len(nlp_en.Defaults.stop_words)}: {list(nlp_en.Defau
```

```
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 3.3 MB/s eta 0:0
0:0000:0100:01m
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /Users/hungshihching/o
pt/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/hungshi
hching/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: smart-open<7.0.0,>=5.2.1 in /Users/hungshihch
ing/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: tqdm<5.0.0,>=4.38.0 in /Users/hungshihching/o
pt/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: srsly<3.0.0,>=2.4.3 in /Users/hungshihching/o
pt/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: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in /Use
rs/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from sp
acy<3.6.0,>=3.5.0->zh-core-web-sm==3.5.0) (1.10.7)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /Users/hungshihching/o
pt/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: typer<0.8.0,>=0.3.0 in /Users/hungshihching/o
pt/anaconda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0-
>zh-core-web-sm==3.5.0) (0.7.0)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /Users/hungshihchi
ng/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: setuptools in /Users/hungshihching/opt/anacon
da3/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: spacy-loggers<2.0.0,>=1.0.0 in /Users/hungshi
hching/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: spacy-legacy<3.1.0,>=3.0.11 in /Users/hungshi
hching/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: numpy>=1.15.0 in /Users/hungshihching/opt/ana
conda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-co
re-web-sm==3.5.0) (1.24.3)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /Users/hungshihc
hing/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: requests<3.0.0,>=2.13.0 in /Users/hungshihchi
ng/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: cymem<2.1.0,>=2.0.2 in /Users/hungshihching/o
pt/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: preshed<3.1.0,>=3.0.2 in /Users/hungshihchin
g/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: pathy>=0.10.0 in /Users/hungshihching/opt/ana
conda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->zh-co
re-web-sm==3.5.0) (0.10.1)
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: packaging>=20.0 in /Users/hungshihching/opt/a
naconda3/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: langcodes<4.0.0,>=3.2.0 in /Users/hungshihchi
ng/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: typing-extensions>=4.2.0 in /Users/hungshihch
ing/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/anac
onda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->sp
acy<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/op
t/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.1
3.0->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/hungshihch
ing/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/op
t/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.1
3.0->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/hungshihch
ing/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/op
t/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/o
pt/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/a
naconda3/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 22.1 MB/s eta 0:
00:0000:010:01
Requirement already satisfied: spacy<3.6.0,>=3.5.0 in /Users/hungshihching/o
pt/anaconda3/envs/nlp/lib/python3.9/site-packages (from en-core-web-sm==3.5.
0) (3.5.2)
Requirement already satisfied: pathy>=0.10.0 in /Users/hungshihching/opt/ana
conda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-co
re-web-sm==3.5.0) (0.10.1)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /Users/hungshihchi
ng/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: cymem<2.1.0,>=2.0.2 in /Users/hungshihching/o
pt/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: srsly<3.0.0,>=2.4.3 in /Users/hungshihching/o
pt/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: packaging>=20.0 in /Users/hungshihching/opt/a
naconda3/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: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in /Use
rs/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages (from sp
acy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (1.10.7)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /Users/hungshihching/o
pt/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: typer<0.8.0,>=0.3.0 in /Users/hungshihching/o
pt/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: setuptools in /Users/hungshihching/opt/anacon
da3/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: catalogue<2.1.0,>=2.0.6 in /Users/hungshihchi
ng/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: requests<3.0.0,>=2.13.0 in /Users/hungshihchi
ng/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: 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: smart-open<7.0.0,>=5.2.1 in /Users/hungshihch
ing/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: numpy>=1.15.0 in /Users/hungshihching/opt/ana
conda3/envs/nlp/lib/python3.9/site-packages (from spacy<3.6.0,>=3.5.0->en-co
re-web-sm==3.5.0) (1.24.3)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /Users/hungshihchin
g/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: thinc<8.2.0,>=8.1.8 in /Users/hungshihching/o
pt/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/hungshi
hching/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: murmurhash<1.1.0,>=0.28.0 in /Users/hungshihc
hing/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: spacy-legacy<3.1.0,>=3.0.11 in /Users/hungshi
hching/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: 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: typing-extensions>=4.2.0 in /Users/hungshihch
ing/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/anac
onda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.13.0->sp
acy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (3.4)
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/hungshihch
ing/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: urllib3<3,>=1.21.1 in /Users/hungshihching/op
t/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.1
3.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2.0.2)
Requirement already satisfied: certifi>=2017.4.17 in /Users/hungshihching/op
t/anaconda3/envs/nlp/lib/python3.9/site-packages (from requests<3.0.0,>=2.1
3.0->spacy<3.6.0,>=3.5.0->en-core-web-sm==3.5.0) (2023.5.7)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /Users/hungshihching/op
t/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/hungshihch
ing/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/o
pt/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/a
naconda3/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：['长话短说', '哎呀', '任', ' ［④d］', '一般', '不仅仅', '其
余', '尤其', '即如', '看来', '通过', '届时', '反过来', ' ｛－', '连声', '宁肯', '这
么样', '零', '从此以后', '比方'] ...
--
英文停用詞 Total=326：['six', 'herein', 'these', 'call', 'can', 'you', 'min
e', 'where', 'elsewhere', 'somewhere', 'often', 'throughout', 'never', ''s',
'another', 'anywhere', 'put', 'seems', 'into', 'it'] ...
```

In [9]:
```python
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))
```
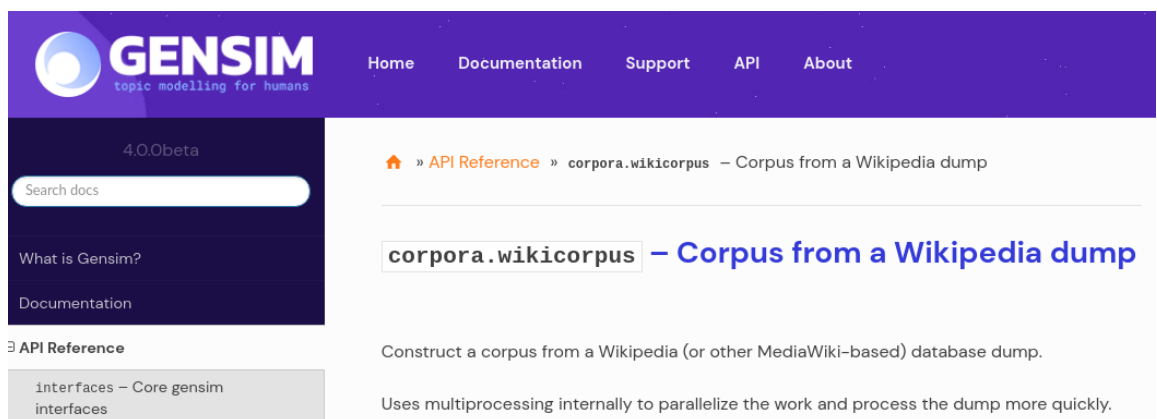
```
2222
3005
```

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]

[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.

```python
In [10]: 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
             ]
```

```python
In [11]: 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']
```

```python
In [12]: %%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-artic
les-multistream.xml.bz2...
```

```
/Users/hungshihching/opt/anaconda3/envs/nlp/lib/python3.9/site-packages/gens
im/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 ...
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 1.153 seconds.
Prefix dict has been built successfully.
Loading model cost 1.161 seconds.
Prefix dict has been built successfully.
Loading model cost 1.086 seconds.
Prefix dict has been built successfully.
Loading model cost 1.112 seconds.
Prefix dict has been built successfully.
Loading model cost 1.056 seconds.
Prefix dict has been built successfully.
Loading model cost 1.051 seconds.
Prefix dict has been built successfully.
Loading model cost 1.084 seconds.
Prefix dict has been built successfully.
peak memory: 1858.68 MiB, increment: 1202.60 MiB
CPU times: user 32min 39s, sys: 3min 50s, total: 36min 30s
Wall time: 1h 16min 50s
```

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 偏文章的前10 個 token

```
In [13]:  g = wiki_corpus.get_texts()
          print(next(g)[:10])
          print(next(g)[:10])
          print(next(g)[:10])
```

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

```
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 1.171 seconds.
Loading model cost 1.174 seconds.
Prefix dict has been built successfully.
Prefix dict has been built successfully.
Loading model cost 1.185 seconds.
Prefix dict has been built successfully.
Loading model cost 1.195 seconds.
Prefix dict has been built successfully.
Loading model cost 1.187 seconds.
Prefix dict has been built successfully.
Loading model cost 1.175 seconds.
Prefix dict has been built successfully.
Loading model cost 1.150 seconds.
Prefix dict has been built successfully.
```
['歐幾裡', '西元前', '三世', '紀的', '古希臘', '數學家', '現在', '認為', '幾何',
'之父']
['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', '為', '手段', '形象化', '客
觀']

將處理完的語料集存下來，供後續使用

```python
In [14]:  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/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 1.176 seconds.
Prefix dict has been built successfully.
Loading model cost 1.177 seconds.
Prefix dict has been built successfully.
Loading model cost 1.187 seconds.
Prefix dict has been built successfully.
Loading model cost 1.170 seconds.
Prefix dict has been built successfully.
Loading model cost 1.176 seconds.
Prefix dict has been built successfully.
Loading model cost 1.151 seconds.
Prefix dict has been built successfully.
Loading model cost 1.172 seconds.
Prefix dict has been built successfully.
```

```
[2023-05-12 12:42:42] 已寫入 99999 篇斷詞文章
[2023-05-12 12:50:17] 已寫入 199999 篇斷詞文章
[2023-05-12 12:57:15] 已寫入 299999 篇斷詞文章
[2023-05-12 13:04:22] 已寫入 399999 篇斷詞文章
[2023-05-12 13:11:57] 已寫入 499999 篇斷詞文章
[2023-05-12 13:18:14] 已寫入 599999 篇斷詞文章
[2023-05-12 13:26:09] 已寫入 699999 篇斷詞文章
[2023-05-12 13:34:22] 已寫入 799999 篇斷詞文章
```

訓練 fastText

In [18]:
```python
%%time

from gensim.models import FastText
from gensim.models.word2vec import LineSentence
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 FastText (dim={word_dim_size})

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

# 訓練 FastText 模型
model = FastText(sentences, vector_size=word_dim_size, workers=max_cpu_count
```

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

```
Use 8 workers to train FastText (dim=300)
CPU times: user 6h 16s, sys: 11min 52s, total: 6h 12min 9s
Wall time: 1h 9min 8s
```

儲存的模型總共會產生三份檔案

In [19]:
```
! ls fasttext.zh*
```

```
fasttext.zh.300.model
fasttext.zh.300.model.syn1neg.npy
fasttext.zh.300.model.wv.vectors_ngrams.npy
fasttext.zh.300.model.wv.vectors_vocab.npy
```

In [20]:
```
!du -sh fasttext.zh*
```

```
 57M    fasttext.zh.300.model
2.0G    fasttext.zh.300.model.syn1neg.npy
2.2G    fasttext.zh.300.model.wv.vectors_ngrams.npy
2.0G    fasttext.zh.300.model.wv.vectors_vocab.npy
```

查看模型以及詞向量實驗

模型其實就是巨大的 Embedding Matrix

In [21]:
```
print(model.wv.vectors.shape)
model.wv.vectors
```

```
(1801313, 300)
```
Out[21]:
```
array([[ 3.5294087e+00, -2.4728107e+00, -3.3658781e+00, ...,
         5.5365825e-01,  6.0241776e+00, -4.3863277e+00],
       [ 5.5733305e-01, -1.7858478e-01, -5.0626736e+00, ...,
        -2.0252099e+00,  4.3347752e-01, -7.0453277e+00],
       [ 4.8810065e-01,  3.2869332e+00, -9.2730266e-01, ...,
        -6.1990862e+00,  5.7194448e+00, -8.1582327e+00],
       ...,
       [-3.6278430e-02,  9.2022851e-02, -2.0581124e-02, ...,
        -2.0813271e-03,  3.1496827e-02,  1.5000439e-02],
       [ 5.6029115e-02,  2.4121341e-01, -1.5050855e-01, ...,
         1.8107755e-01, -1.9315310e-02, -2.8615113e-02],
       [-2.1139957e-01,  3.0734321e-01,  9.7958821e-01, ...,
        -1.0782071e-02, -7.7970199e-02, -1.5616444e-01]], dtype=float32)
```

收錄的詞彙

In [22]:
```
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']
```

Out[22]: `'print(f"總共收錄了 {len(model.wv.vocab)} 個詞彙")\n\nprint("印出 20 個收錄詞彙:")\nprint(list(model.wv.vocab.keys())[:10])'`

詞彙的向量

In [23]:
```python
vec = model.wv['數學家']
print(vec.shape)
vec
```

(300,)

```
Out[23]:  array([-1.59387505e+00,  2.50691628e+00,  3.75720054e-01,  1.66334236e+00,
                  3.78292277e-02,  9.63003576e-01, -2.19865227e+00, -2.73550653e+00,
                  3.14566463e-01, -5.24781644e-01, -9.82077479e-01,  1.40853776e-02,
                 -1.81269252e+00,  1.67503846e+00, -9.20567155e-01, -6.71969533e-01,
                 -2.48001263e-01, -2.76921630e+00, -1.15092449e-01, -2.82676309e-01,
                 -1.79702890e+00,  6.06646240e-01,  2.07849121e+00,  1.72411072e+00,
                 -9.98204112e-01, -4.29534972e-01,  8.45891893e-01,  5.95484078e-01,
                 -1.82911471e-01, -3.74818772e-01, -2.43063927e+00,  9.64546502e-01,
                 -1.49381176e-01, -1.32062685e+00,  9.50988710e-01, -2.22288084e+00,
                  1.17434299e+00, -4.69304979e-01, -2.93430352e+00,  1.36387572e-01,
                  3.21646482e-01,  1.49611264e-01, -6.07900918e-01, -1.34132192e-01,
                  1.08165944e+00, -4.11090183e+00,  2.85739017e+00,  6.17159545e-01,
                  3.97191569e-02,  6.50986493e-01, -1.89106330e-01, -1.00554109e+00,
                 -3.10830379e+00,  3.28530401e-01, -1.52698323e-01, -8.44725311e-01,
                  9.56938386e-01,  2.72862744e+00, -1.95729524e-01, -2.10608506e+00,
                  9.52691436e-01,  4.40533876e-01, -2.46478394e-01,  4.50502425e-01,
                 -1.29843616e+00,  8.22740614e-01, -4.75307778e-02, -1.33183861e+00,
                 -7.03986049e-01,  1.58372235e+00, -3.38559687e-01, -7.62892723e-01,
                  5.10872304e-01,  2.65607285e+00, -4.97602038e-02, -1.39200187e+00,
                 -1.66156483e+00,  6.11999154e-01, -6.87287897e-02,  1.67237854e+00,
                 -2.56772017e+00, -4.60918605e-01,  2.53625536e+00,  1.39643896e+00,
                 -1.39784110e+00, -1.46155328e-01, -1.57716453e+00, -3.47367907e+00,
                 -1.32986367e+00, -1.38915944e+00, -2.74625808e-01,  3.19288112e-02,
                  1.37352735e-01, -6.92346096e-02, -6.63332522e-01, -3.45693022e-01,
                 -1.27403900e-01,  7.70505369e-01,  1.34264314e+00, -1.88267171e+00,
                  1.18355346e+00,  2.04772639e+00, -3.73724522e-03,  1.08498871e+00,
                  1.86692512e+00, -2.62517381e+00, -1.47760832e+00,  4.47082472e+00,
                  1.08710992e+00,  1.04343033e+00, -2.43925735e-01,  2.45133543e+00,
                  6.88490450e-01, -1.42014337e+00, -6.17439486e-02,  7.62792230e-01,
                 -1.06854844e+00, -1.76054919e+00,  1.46899152e+00, -1.07032967e+00,
                  1.72993112e+00, -2.41574740e+00,  1.50274885e+00,  8.49633276e-01,
                 -3.17127973e-01,  2.38300347e+00,  6.30162060e-01,  1.96526337e+00,
                 -1.13477334e-01, -1.46750048e-01,  2.48623103e-01, -7.79943764e-01,
                  2.15136075e+00,  9.59935844e-01, -1.91290140e-01, -1.23067498e+00,
                  1.87161788e-01,  1.06132679e-01,  2.10292125e+00, -1.45100906e-01,
                 -7.62847781e-01,  9.67496872e-01,  1.45733619e+00,  2.18957469e-01,
                  2.90596080e+00, -2.09353185e+00,  4.08733398e-01,  1.61074996e+00,
                  6.43477976e-01,  2.97597945e-01,  2.27632612e-01,  2.58670092e-01,
                  1.18777585e+00,  1.71221924e+00,  1.43743217e+00,  9.67478096e-01,
                  7.44253248e-02, -3.08438271e-01,  9.24532592e-01, -3.04136965e-02,
                 -2.29417777e+00,  2.56111741e+00,  1.39130867e+00,  5.10467827e-01,
                 -3.69881898e-01, -6.19951785e-01,  1.34217334e+00, -1.72916806e+00,
                 -1.49066104e-02,  3.43515545e-01,  3.96899074e-01,  2.38401294e+00,
                  1.39472783e+00,  1.29092455e-01,  3.11960697e-01, -1.27326882e+00,
                 -1.00506699e+00,  2.56291389e+00, -1.42586720e+00,  5.03195338e-02,
                  1.51455963e+00,  9.05661583e-01,  5.39308965e-01, -4.20009047e-01,
                  2.26594496e+00,  1.25655353e+00, -1.35127008e+00,  9.02198732e-01,
                  1.31912804e+00,  3.57353234e+00,  6.35365188e-01,  3.47347236e+00,
                  1.94280064e+00,  2.55723906e+00,  7.31840968e-01,  2.62270164e+00,
                 -2.09922504e+00,  1.17788756e+00,  7.35983074e-01, -3.85721952e-01,
                  3.79986852e-01, -1.97034371e+00, -5.82545161e-01,  1.15764022e+00,
                  5.99241674e-01, -2.49447390e-01,  6.15953743e-01,  1.98849666e+00,
                  3.17865157e+00, -8.57991517e-01,  2.74693298e+00,  8.93612087e-01,
                  2.63120008e+00,  8.18389893e-01, -1.05498683e+00, -2.08067730e-01,
                 -1.22313154e+00,  1.19683838e+00,  1.44286764e+00, -1.56527030e+00,
                 -1.54964006e+00,  3.71310204e-01,  1.29893959e+00,  6.31124198e-01,
                  1.43484664e+00, -5.89890480e-02, -2.98322350e-01, -5.33336341e-01,
                  1.47559965e+00,  8.97132695e-01,  7.89537802e-02,  6.52392805e-01,
                 -2.43649229e-01, -1.62721324e+00, -1.74335802e+00, -5.43126225e-01,
                 -2.02359605e+00,  2.13031602e+00, -1.59767854e+00,  1.44884336e+00,
                  7.85594642e-01,  6.62480712e-01,  1.64054108e+00, -4.51099649e-02,
                 -1.75535071e+00,  1.12059247e+00,  1.94820952e+00,  1.03447127e+00,
                  3.86004150e-01, -1.02474296e+00,  3.09551746e-01,  1.85309005e+00,
                  2.24413133e+00,  5.58408558e-01,  1.54626215e+00,  8.42069030e-01,
```

```
       1.74492371e+00,  2.99126245e-02,  5.35367131e-01,  2.76648498e+00,
      -1.91627598e+00, -5.49638510e-01, -1.64982527e-01, -2.03246689e+00,
       3.37515497e+00, -6.27814651e-01,  1.21679437e+00, -4.88097399e-01,
      -6.75468862e-01, -2.31950313e-01, -3.72293091e+00, -1.10152209e+00,
      -8.88435006e-01,  8.51761639e-01,  2.04803157e+00,  1.05621934e+00,
       6.86107159e-01, -1.26635754e+00, -9.66005325e-02,  5.03416121e-01,
       4.50871557e-01, -9.18411434e-01,  7.48757839e-01, -8.08944851e-02,
      -6.71347082e-01, -2.22347662e-01,  4.16042864e-01,  8.35543573e-01,
      -1.30313134e+00, -1.00410378e+00,  4.89792041e-02, -1.21605980e+00,
      -1.51601946e+00,  1.99778175e+00,  2.55595267e-01,  9.77041841e-01,
       1.82202923e+00,  3.35812896e-01,  2.45464563e+00, -1.98096371e+00],
    dtype=float32)
```

沒見過的詞彙

In [25]:
```python
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)
```

查看前 10 名相似詞

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

In [26]:
```python
model.wv.most_similar("飲料", topn=10)
```

Out[26]:
```
[('精飲料', 0.9644166231155396),
 ('輝劍', 0.9633467197418213),
 ('名松', 0.961124837398529),
 ('種飲料', 0.9519960284233093),
 ('飲料業', 0.9483360052108765),
 ('飲料則', 0.948043704032898),
 ('自飲料', 0.9471467137336731),
 ('搖飲料', 0.9469248652458191),
 ('飲料類', 0.9441595077514648),
 ('軟飲料', 0.929158091545105)]
```

In [27]:
```python
model.wv.most_similar("car")
```

Out[27]:
```
[('hcar', 0.8683251738548279),
 ('carcar', 0.8598149418830872),
 ('ccar', 0.836090624332428),
 ('jetcar', 0.8307980298995972),
 ('tramcar', 0.8291366696357727),
 ('boxcar', 0.8262635469436646),
 ('cargru', 0.8123184442520142),
 ('zipcar', 0.8119108080863953),
 ('cars', 0.8107830882072449),
 ('motorcar', 0.8080064058303833)]
```

In [28]:
```python
model.wv.most_similar("facebook")
```

Out[28]:
```
[('youtubefacebook', 0.9296781420707703),
 ('thefacebook', 0.896470943502808),
 ('facebookpage', 0.8949395418167114),
 ('facebox', 0.8620952367782593),
 ('instagram', 0.8215540051460266),
 ('twitteryoutube', 0.775385856628418),
 ('twitter', 0.7578170299530029),
 ('googleyoutube', 0.7560174465179443),
 ('youtube', 0.7499861121177673),
 ('lnstagram', 0.7471879124641418)]
```

In [40]:
```python
model.wv.most_similar("欺詐")
```

Out[40]:
```
[('抱出', 0.94637531042099),
 ('杭郡', 0.9386997818946838),
 ('反欺詐', 0.9242189526557922),
 ('欺詐法', 0.9105420708656311),
 ('欺詐案', 0.908968448638916),
 ('性兒', 0.9085018038749695),
 ('欺詐性', 0.8954063057899475),
 ('欺詐者', 0.8413706421852112),
 ('證欺詐', 0.8390339612960815),
 ('床帳', 0.82300865650177)]
```

In [32]:
```python
model.wv.most_similar("合約")
```

Out[32]:
```
[('合約爭', 0.9561173319816589),
 ('合約機', 0.9552879929542542),
 ('合約員', 0.9533215761184692),
 ('員合約', 0.9518018960952759),
 ('止合約', 0.9515566229820251),
 ('價合約', 0.950880765914917),
 ('僱合約', 0.9487921595573425),
 ('購合約', 0.9484802484512329),
 ('商合約', 0.9482670426368713),
 ('號合約', 0.9477567076683044)]
```

計算 Cosine 相似度

In [33]:
```python
model.wv.similarity("連結", "鍵接")
```

Out[33]:
```
0.40969405
```

In [34]:
```python
model.wv.similarity("連結", "陰天")
```

Out[34]:
```
0.0027875535
```

讀取模型

In [37]:
```python
print(f"Loading {output_model}...")
new_model = FastText.load(output_model)
```

```
Loading fasttext.zh.300.model...
```

In [38]:
```python
model.wv.similarity("連結", "陰天") == new_model.wv.similarity("連結", "陰天")
```

Out[38]:
```
True
```

In [ ]: