In [1]:

```
import os
import gensim
import jieba
import zhconv

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 [2]:

```
import spacy

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

In [3]:

```
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 [4]:

```
def preprocess_and_tokenize(text, token_min_len = 1, token_max_len = 15, lower = 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 [7]:

```
import fasttext
import fasttext.util

tokenized_data = []
n = 0
with open("wiki_seg.txt", encoding="utf-8") as f:
    for row in f.readlines():
        tokenized_data.append(preprocess_and_tokenize(row))
```

```
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 1.383 seconds.
Prefix dict has been built successfully.
```

In [10]:

```
from gensim.models import FastText

model = FastText()

model.build_vocab(tokenized_data)
model.train(tokenized_data, total_examples = len(tokenized_data), epochs = 100)

model.save("fasttext.mdl")
```

In [12]:

```
model.wv.most_similar("飲料")
```

```
[('飲品', 0.9097651),
 ('果汁', 0.8754865),
 ('酒類', 0.7625548),
 ('啤酒', 0.7542788),
 ('可口可樂', 0.7510687),
 ('冰淇淋', 0.7476611),
 ('牛奶', 0.7422941),
 ('軟飲料', 0.7361464),
 ('罐裝', 0.7289342),
 ('口香糖', 0.7188053)]
```

In [13]:

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

```
[('motorcoach', 0.8330657),
 ('truck', 0.8212014),
 ('cab', 0.8202789),
 ('motorcar', 0.8142311),
 ('seat', 0.8125936),
 ('motor', 0.8124716),
 ('motorcycle', 0.8039779),
 ('roadster', 0.7971475),
 ('motorist', 0.7958347),
 ('automobile', 0.7922669)]
```

In [14]:

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

```
[('youtubefacebook', 0.8432591),
 ('instagram', 0.83393224),
 ('thefacebook', 0.7911547),
 ('專頁', 0.7892121),
 ('youtube', 0.7576334),
 ('lnstagram', 0.7363649),
 ('myspace', 0.7351787),
 ('linkedin', 0.7348244),
 ('telegram', 0.7286041),
 ('whatsapp', 0.7261969)]
```

In [15]:

```
model.wv.most_similar("詐欺")
```

```
[('欺詐', 0.7847247),
 ('詐騙', 0.6378833),
 ('竊盜', 0.5811350),
 ('殺人', 0.5783799),
 ('被害者', 0.5751656),
 ('詐欺罪', 0.5725432),
 ('誘拐', 0.5717222),
 ('委託人', 0.5662581),
 ('詐騙者', 0.5610893),
 ('信用調查', 0.5583631)]
```

In [16]:

```
model.wv.most_similar("合約")
```

```
[('合同', 0.7991111),
 ('簽約', 0.7823647),
 ('續約', 0.7787457),
 ('到期', 0.7484624),
 ('簽下', 0.7096284),
 ('租約', 0.7023362),
 ('買斷', 0.6730079),
 ('選擇權', 0.6726822),
 ('新東家', 0.6714864),
 ('解約', 0.6664419)]
```

In [17]:

```
model.wv.most_similar("飲料")
```

```
[('飲品', 0.9090546),
 ('果汁', 0.8486888),
 ('酒類', 0.7625674),
 ('啤酒', 0.7542456),
 ('可口可樂', 0.7510691),
 ('冰淇淋', 0.7476711),
 ('牛奶', 0.7422449),
 ('軟飲料', 0.7361465),
 ('罐裝', 0.7284132),
 ('口香糖', 0.7188053)]
```

In [19]:

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

```
0.91527834
```

In [20]:

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

```
0.05147627
```

In [ ]:

```

```