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")
import spacy

nlp_zh = spacy.load("zh_core_web_sm")
nlp_en = spacy.load("en_core_web_sm")
STOPWORDS =  nlp_zh.Defaults.stop_words | nlp_en.Defaults.stop_words | set(["\n", "\

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

In [2]:

```
from f import preprocess_and_tokenize
```

In [3]:

```
import fasttext
import fasttext.util
data = []
n = 0
with open("wiki_seg.txt") as f:
    for row in f.readlines():
        data.append(preprocess_and_tokenize(row))
```

```
Building prefix dict from /Users/kaichengliu/Desktop/University/大三下/
NLP2023/hw4/dict.txt.big ...
Loading model from cache /var/folders/2m/2yp8xh891bd67xxdsy56_6gc0000g
n/T/jieba.ud6625625ebbc2d2b90ddca0ff615d2a1.cache
Loading model cost 1.014 seconds.
Prefix dict has been built successfully.
```

In [6]:

```
from gensim.models import word2vec, fasttext


# Train
model = fasttext.FastText()
model.build_vocab(data)
model.train(data, epochs=model.epochs,total_examples=model.corpus_count, total_words
model.save('fasttext.model')
```

In [7]:

```
word = "這肯定沒見過 "

try:
    vec = model.wv[word]
except KeyError as e:
    print(e)
```

In [8]:

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

Out[8]:

```
[('輝劍', 0.9855037331581116),
 ('名松', 0.9701558947563171),
 ('飲料類', 0.9429479837417603),
 ('軟飲料', 0.936282753944397),
 ('飲品', 0.9102264642715454),
 ('經米濱', 0.8833635449409485),
 ('院卿', 0.8511365056037903),
 ('集滿時', 0.8151489496231079),
 ('港舊', 0.8033074140548706),
 ('牛奶', 0.7929983139038086)]
```

In [9]:

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

Out[9]:

```
[('hcar', 0.8902831673622131),
 ('jetcar', 0.8727462291717529),
 ('ccar', 0.8657267689704895),
 ('tramcar', 0.8570384979248047),
 ('motorcar', 0.8569256067276001),
 ('carcar', 0.8499772548675537),
 ('indycar', 0.8463096022605896),
 ('cab', 0.8461726903915405),
 ('camry', 0.8395930528640747),
 ('boxcar', 0.8376190066337585)]
```

In [10]:

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

Out[10]:

```
[('youtubefacebook', 0.9516081809997559),
 ('thefacebook', 0.9173453450202942),
 ('facebookpage', 0.9078859090805054),
 ('facebox', 0.8973644971847534),
 ('instagram', 0.8597090244293213),
 ('googleyoutube', 0.8427925109863281),
 ('twitteryoutube', 0.8049390912055969),
 ('youtube', 0.802020251750946),
 ('lnstagram', 0.7937968969345093),
 ('whatsapp', 0.7891119122505188)]
```

In [11]:

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

Out[11]:

```
[('happy8', 0.9779205918312073),
 ('happyend', 0.9232112765312195),
 ('unhappy', 0.9156875014305115),
 ('happytuk', 0.9103104472160339),
 ('happylive', 0.9087854027748108),
 ('happytime', 0.9048050045967102),
 ('happygo', 0.9044574499130249),
 ('happyface', 0.9038417339324951),
 ('happ3', 0.8902745842933655),
 ('chappy', 0.8894372582435608)]
```

In [12]:

```
model.wv.most_similar("合約")
```

Out[12]:

```
[('德康', 0.9534314870834351),
 ('倫戈縣', 0.8758705854415894),
 ('合同', 0.8432769775390625),
 ('續約', 0.8057641983032227),
 ('簽約', 0.8030670285224915),
 ('到期', 0.7902146577835083),
 ('隊辛', 0.7892029285430908),
 ('合同商', 0.7884400486946106),
 ('綠蠅', 0.7832315564155579),
 ('合同期', 0.7779042720794678)]
```

In [13]:

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

Out[13]:

```
0.9166099
```

In [14]:

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

Out[14]:

```
-0.035726584
```

In [ ]:

```

```