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/dic

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

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

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

In [4]:
```
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 [5]:
```
import fasttext
import fasttext.util

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

```
Building prefix dict from /Users/hsiu/Desktop/cgu/nlp/HW/hw4/dict.
txt.big ...
Dumping model to file cache /var/folders/8q/s5rn_lld0qj8zg63cy8t3x
tw0000gn/T/jieba.u362f17834e03b0390f888de340f1b114.cache
Loading model cost 1.067 seconds.
Prefix dict has been built succesfully.
```

In [6]:
```
from gensim.models import FastText

model = FastText()

model.build_vocab(tokenized_data)
model.train(tokenized_data, total_examples = len(tokenized_data), e

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

In [7]:
```
model.wv.most_similar("飲料")
```

Out[7]:
```
[('飲品', 0.9117832183837891),
 ('果汁', 0.8205680847167969),
 ('酒類', 0.7712692618370056),
 ('可口可樂', 0.7413931488990784),
 ('冰淇淋', 0.7404088973999023),
 ('啤酒', 0.7330322265625),
 ('牛奶', 0.7243913412094116),
 ('口香糖', 0.7166147828102112),
 ('食品', 0.7139566540718079),
 ('軟飲料', 0.712602972984314)]
```

In [8]:
```
model.wv.most_similar("car")
```

Out[8]:
```
[('truck', 0.8288451433181763),
 ('motor', 0.8246199488639832),
 ('seat', 0.8153491020202637),
 ('motorcoach', 0.8148860931396484),
 ('motorist', 0.8130364418029785),
 ('motorcar', 0.8084441423416138),
 ('cab', 0.806026816368103),
 ('motorcycle', 0.803566038608551),
 ('automobile', 0.8030974864959717),
 ('tractor', 0.8025482892990112)]
```

In [9]:
```
model.wv.most_similar("facebook")
```

Out[9]:
```
[('youtubefacebook', 0.8336827158927917),
 ('instagram', 0.8332372307777405),
 ('專頁', 0.7994852066040039),
 ('thefacebook', 0.7915207743644714),
 ('youtube', 0.7496374845504761),
 ('linkedin', 0.7484966516494751),
 ('lnstagram', 0.7386245131492615),
 ('blogger', 0.724703311920166),
 ('whatsapp', 0.7205793857574463),
 ('telegram', 0.716116189956665)]
```

In [10]:
```
model.wv.most_similar("詐欺")
```

Out[10]:
```
[('欺詐', 0.7753905057907104),
 ('詐騙', 0.6466060876846313),
 ('竊盜', 0.615557849407196),
 ('被害者', 0.5958162546157837),
 ('盜領', 0.590490996837616),
 ('詐欺罪', 0.5892660617828369),
 ('殺人', 0.5864850282669067),
 ('信用調查', 0.5851697325706482),
 ('誘拐', 0.5669840574264526),
 ('犯罪', 0.5627837181091309)]
```

In [11]:
```
model.wv.most_similar("合約")
```

Out[11]:
```
[('合同', 0.7965822815895081),
 ('簽約', 0.7835227847099304),
 ('續約', 0.7830365896224976),
 ('到期', 0.7624660134315491),
 ('簽下', 0.7064452767372131),
 ('租約', 0.7030776143074036),
 ('解約', 0.7006136775016785),
 ('買斷', 0.6862468719482422),
 ('新東家', 0.671778678894043),
 ('籤下', 0.6657853722572327)]
```

In [12]:
```
model.wv.most_similar("飲料")
```

Out[12]:
```
[('飲品', 0.9117832183837891),
 ('果汁', 0.8205680847167969),
 ('酒類', 0.7712692618370056),
 ('可口可樂', 0.7413931488990784),
 ('冰淇淋', 0.7404088973999023),
 ('啤酒', 0.7330322265625),
 ('牛奶', 0.7243913412094116),
 ('口香糖', 0.7166147828102112),
 ('食品', 0.7139566540718079),
 ('軟飲料', 0.712602972984314)]
```

In [13]:
```
model.wv.similarity("連結", "鏈結")
```

Out[13]:
```
0.9339111
```

In [14]:
```
model.wv.similarity("連結", "陰天")
```

Out[14]:
```
0.041523255
```