```
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"))


def preprocess_and_tokenize(text, token_min_len = 1, token_max_len = 15, lower =
    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 STOPWOR
    ]


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/ming/Desktop/大三/自然語言處理/hw4/dict.txt.b
Loading model from cache /var/folders/ly/4jgkxntx463ghnpv_mjkrb600000gn/T/j
Loading model cost 0.440 seconds.
Prefix dict has been built successfully.
```

```
from gensim.models import FastText

model = FastText()

model.build_vocab(tokenized_data)
model.train(tokenized_data, total_examples = len(tokenized_data), epochs = 300)

model.save("fasttext.mdl")


model.wv.most_similar("飲料")

     [('飲品', 0.9090189933776855),
      ('果汁', 0.8486908674240112),
      ('酒類', 0.7625470161437988),
      ('啤酒', 0.7542792558670044),
      ('可口可樂', 0.7510690689086914),
      ('冰淇淋', 0.747668981552124),
      ('牛奶', 0.7422364950180054),
      ('軟飲料', 0.7361465096473694),
      ('罐裝', 0.7284132838249207),
      ('口香糖', 0.7188053131103516)]


model.wv.most_similar("car")

     [('motorcoach', 0.8330457806587219),
      ('truck', 0.8212042450904846),
      ('cab', 0.8201040029525757),
      ('motorcar', 0.8142408728599548),
      ('seat', 0.8125919699668884),
      ('motor', 0.8123607039451599),
      ('motorcycle', 0.8039779663085938),
      ('roadster', 0.7971475720405579),
      ('motorist', 0.7958347797393799),
      ('automobile', 0.7922669053077698)]


model.wv.most_similar("facebook")

     [('youtubefacebook', 0.8432331681251526),
      ('instagram', 0.8339320421218872),
      ('thefacebook', 0.7910415530204773),
      ('專頁', 0.7892942428588867),
      ('youtube', 0.7576664090156555),
      ('lnstagram', 0.7364949584007263),
      ('myspace', 0.7351047396659851),
      ('linkedin', 0.7348244786262512),
      ('telegram', 0.7286040782928467),
      ('whatsapp', 0.728196918964386)]
```

```
model.wv.most_similar("詐欺")

    [('欺詐', 0.7847247123718262),
     ('詐騙', 0.6378833055496216),
     ('竊盜', 0.581135094165802),
     ('殺人', 0.5783799886703491),
     ('被害者', 0.5751656889915466),
     ('詐欺罪', 0.5725432634353638),
     ('誘拐', 0.5717222094535828),
     ('委託人', 0.5662581920623779),
     ('詐騙者', 0.5610893368721008),
     ('信用調查', 0.5583631992340088)]


model.wv.most_similar("合約")

    [('合同', 0.7991208434104919),
     ('簽約', 0.7823731303215027),
     ('續約', 0.7787519097328186),
     ('到期', 0.7484908103942871),
     ('簽下', 0.7096284031867981),
     ('租約', 0.7023362517356873),
     ('買斷', 0.6730079054832458),
     ('選擇權', 0.6726821064949036),
     ('新東家', 0.6714859008789062),
     ('解約', 0.6664419770240784)]


model.wv.most_similar("飲料")

    [('飲品', 0.9090189933776855),
     ('果汁', 0.8486908674240112),
     ('酒類', 0.7625470161437988),
     ('啤酒', 0.7542792558670044),
     ('可口可樂', 0.7510690689086914),
     ('冰淇淋', 0.747668981552124),
     ('牛奶', 0.7422364950180054),
     ('軟飲料', 0.7361465096473694),
     ('罐裝', 0.7284132838249207),
     ('口香糖', 0.7188053131103516)]


model.wv.similarity("連結", "鏈結")

    0.93718535


model.wv.similarity("連結", "陰天")

    0.12764695
```

Colab 付費產品 - 按這裡取消合約

● ×