學號:B0928031

姓名:鄭茹云

學系:人工智慧三甲

連結: https://colab.research.google.com/drive/1SOP44AhXZ0dNf_ZgzKyw1N9VO7Vc_Dcn?usp=share_link

```
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")
```

+ 程式碼 + 文字

```
STOPWORDS = nlp_zh.Defaults.stop_words | nlp_en.Defaults.stop_words | set(["\n", "\r\n", "\t", " ", ""])

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 = 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
    ]
```

儲存成功! ×

```
    # Remove any non-alphanumeric characters and convert to lowercase
    processed_row = re.sub(r"[^a-zA-Z0-9\s]", "", row.lower())
    # Tokenize the processed row by splitting on whitespace
    tokens = processed_row.split()
    return tokens

tokenized_data = []
n = 0

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

```
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")
```

```
[('blaxploitation', 0.4243933856487274),
 ('outcome', 0.407603919506073),
 ('propositions', 0.399801462888871765),
 ('symptoms', 0.39865773916244507),
 ('physalis', 0.394120454788208),
 ('outcomes', 0.3824048936367035),
 ('righteous', 0.38179323077201843),
 ('intruders', 0.377671480178833),
 ('emerged', 0.37718498706817627),
 ('protests', 0.3754940629005432)]
```

```
model.wv.most_similar("car")
      [('sportscar', 0.577189028263092),
       ('vicar', 0.5765175819396973),
       ('sport', 0.5706731677055359),
       ('carrier', 0.5491780638694763),
       ('carel', 0.5439125895500183),
       ('pant', 0.5433809161186218),
       ('hire', 0.532235324382782),
       ('racer', 0.5291739106178284),
       ('sprinter', 0.5249852538108826),
       ('saloon', 0.5234070420265198)]
```

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

      [('youtube', 0.7343934774398804),
       ('twitter', 0.6572933197021484),
       ('instagram', 0.640592098236084),
       ('2016', 0.6034048795700073),
       ('2015', 0.5937403440475464),
       ('2014', 0.5867704153060913),
       ('myradio', 0.5784428715705872),
       ('google', 0.5735743641853333),
       ('2017', 0.5628921985626221),
       ('2019', 0.5565149784088135)]
```

```
model.wv.most_similar("詐欺")

      [('enviro400', 0.6054567694664001),
       ('v107', 0.4900737404823303),
       ('20082005', 0.4782906770706177),
       ('v106', 0.42807525396347046),
       ('cs55', 0.4104488790035248),
       ('tamelerdeamani', 0.40449631214141846),
       ('cs5', 0.39534831047058105),
       ('kobushi', 0.3933582901954651),
       ('jwp', 0.3882831931114197),
       ('v108', 0.3882051706314087)]
```

```
model.wv.most_similar("合約")

      [('0593', 0.32641032338142395),
       ('89148', 0.31937190890312195),
       ('inbox', 0.31526505947113037),
```

儲存成功！ ×

```
       ('30k', 0.31312167644500073),
       ('5657', 0.3108726143836975),
       ('x55px', 0.3064502775669098),
       ('mp3264x24481', 0.3006058633327484),
       ('v104', 0.2967165410518646)]
```

```
model.wv.most_similar("飲料")

      [('blaxploitation', 0.4243933856487274),
       ('outcome', 0.407603919506073),
       ('propositions', 0.39980146288871765),
       ('symptoms', 0.39865773916244507),
       ('physalis', 0.394120454788208),
       ('outcomes', 0.3824048936367035),
       ('righteous', 0.38179323077201843),
       ('intruders', 0.377671480178833),
       ('emerged', 0.37718498706817627),
       ('protests', 0.3754940629005432)]
```

```
model.wv.similarity("連結", "鏈結")

      -0.041576713
```

```
model.wv.similarity("連結", "陰天")

      -0.07708004
```

儲存成功！