

▼ Lab#4, NLP@CGU Spring 2023

This is due on 2023/04/20 16:00, commit to your github as a PDF (lab4.pdf) (File>Print>Save as PDF).

IMPORTANT: After copying this notebook to your Google Drive, please paste a link to it below. To get a publicly-accessible link, hit the *Share* button at the top right, then click "Get shareable link" and copy over the result. If you fail to do this, you will receive no credit for this lab!

**LINK: paste your link here** <https://colab.research.google.com/drive/1UfXlvOSY8atIGPc8D11HKFvX2eixFHu4#scrollTo=pKEvBE5INLVI-->

**Student ID:**B0928029

**Name:**毛謙鎧

▼ Word Embeddings for text classification

請訓練一個 kNN或是SVM 分類器來和 Google's Universal Sentence Encoder (a fixed-length 512-dimension embedding) 的分類結果比較

```
!wget -O Dcard.db https://github.com/cjwu/cjwu.github.io/raw/master/courses/nlp2023/lab4-Dcard-Dataset.db

import sqlite3
import pandas as pd

conn = sqlite3.connect("Dcard.db")
df = pd.read_sql("SELECT * FROM Posts;", conn)
df
```

	createdAt	title	excerpt	categories
0	2022-03-04T07:54:19.886Z	專題需要數據🥹🥹 幫填～	希望各位能花個20秒幫我填一下	
1	2022-03-04T07:42:59.512Z	#詢問 找衣服🥹	想找這套衣服🥹，但發現不知道該用什麼關鍵字找，（圖是草屯囡仔的校園演唱會截圖）	詢問 衣服   鞋子   衣物
2	2022-03-04T07:24:25.147Z	#黑特 網購50% FIFTY PERCENT請三思	因為文會有點長，先說結論是，50%是目前網購過的平台退貨最麻煩的一家，甚至我認為根本是刻意刁...	黑特   網購   三思
3	2022-03-04T06:39:13.017Z	尋衣服	來源：覺得呱吉這襯衫好好看～，或有人知道有類似的嗎	衣服   尋找   日常穿
4	2022-03-04T06:28:06.137Z	#詢問 想問	各位，因為這個證件夾臺灣買不到，是美國outlet 的限量版貨，所以在以下的這間蝦皮上買，但...	詢問 穿搭   閒聊版   閒
...	...	...	...	...
355	2022-03-03T03:41:10.972Z	開了新頻道	昨天上了第一支影片，之前有發過沒有線條的動畫影片，新的頻道改成了有線條的，感覺大家好像比較喜歡...	Youtuber   頻道   有
356	2022-03-03T02:26:58.821Z	估計某個YTUBER又有陰謀論可以寫了	今天全台灣大停電，應該過幾天就會有個戴面具的出來說，一定是中共.....，我從上個影片就預測了.....	陰謀
357	2022-03-02T21:25:51.080Z	#問 阿神和放火發生過什麼嗎？	想問有沒有人知道阿神和放火是認識還是有結過什麼仇之類的嗎？首先我個人基本沒關注過放火，但是最...	Youtuber
358	2022-03-02T20:33:47.713Z	#文長 我眼中的Rice&Shine	無意引戰，單純分享我的觀察與個人想法～這幾天看了Dcard幾篇關於Rice& Shine的貼...	Riceandshine   You Vlog
359	2022-03-02T17:52:26.967Z	一個隨性系YouTube頻道	哈哈哈哈哈，沒錯我就是親友團來介紹一個我覺得很北七的頻道，現在觀看真的低的可憐，也沒事啦，就多...	Youtuber

360 rows x 7 columns

```
!pip3 install -q tensorflow_text
!pip3 install -q faiss-cpu

import tensorflow_hub as hub
import numpy as np
import tensorflow_text
import faiss

embed_model = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual/3")
```

```
docid = 355
texts = "[" + df['title'] + ' ] [' + df['topics'] + ' ] ' + df['excerpt']
texts[docid]

'[開了新頻道] [Youtuber | 頻道 | 有趣 | 日常 | 搞笑] 昨天上了第一支影片，之前有發過沒有線條的動畫影片，新的頻道改成了有線條的，感覺大家好像比較喜歡...
內容主要是分享自己遇到的小故事，不知道這樣的頻道大家是否會想要看呢？喜歡的話也'
```

```
embeddings = embed_model(texts)
embed_arrays = np.array(embeddings)
index_arrays = df.index.values
topk = 10
# Step 1: Change data type
embeddings = embed_arrays.astype("float32")

# Step 2: Instantiate the index using a type of distance, which is L2 here
index = faiss.IndexFlatL2(embeddings.shape[1])

# Step 3: Pass the index to IndexIDMap
index = faiss.IndexIDMap(index)

# Step 4: Add vectors and their IDs
index.add_with_ids(embeddings, index_arrays)

D, I = index.search(np.array([embeddings[docid]]), topk)

plabel = df.iloc[docid]['forum_zh']

cols_to_show = ['title', 'excerpt', 'forum_zh']
plist = df.loc[I.flatten(), cols_to_show]

precision = 0
for index, row in plist.iterrows():
    if plabel == row["forum_zh"]:
        precision += 1

print("precision = ", precision/topk)
precision = 0

df.loc[I.flatten(), cols_to_show]
```

precision =	0.8			
		title		excerpt forum_zh
355	開了新頻道	昨天上了第一支影片，之前有發過沒有線條的動畫影片，新的頻道改成了有線條的，感覺大家好像比較喜歡...		Youtuber
359	一個隨性系YouTube頻道	哈哈哈哈哈，沒錯我就是親友團來介紹一個我覺得很北七的頻道，現在觀看真的低的可憐，也沒事啦，就多...		Youtuber
330	《庫洛魔法使》（迷你）服裝製作	又來跟大家分享新的作品了~，頻道常常分享 {縫紉} {服裝製作} 等相關教學，大家對服裝製...		Youtuber
342	自己沒搞清楚狀況就不要亂黑勾惡	勾惡幫主在自己頻道簡介跟每部影片的下方都已經說明了，要分會會長以上才能看全部影片，這個說明已...		Youtuber
338	廚師系Youtuber	友人傳了這篇文給我，我一看，十大廚師系Youtuber，就猜一定有MASA，果不其然，榜上有...		Youtuber
243	毀我童年的家人	小時候都很喜歡看真珠美人魚和守護甜心，但是！！，每次晚餐看電視的時候，只要有播映到這種場景...		有趣
349	喜歡看寵物頻道的有嗎？🐶			Youtuber
332	#安利 翎遇嗎 采翎	如題啦！ 最近突然超愛采翎，以前就很喜歡了，最近越來越愛~~，從之前的呱張新聞到新資料夾到翎...		Youtuber

▼ Implement Your kNN or SVM classifier Here!

請比較分類結果中選出 topk 相近的筆數，並計算 forum\_zh 是否都有在 query text 的 forum\_zh 中

```
[開了新頻道] [Youtuber | 頻道 | 有趣 | 日常 | 搞笑]
```

```
precision = 0
topk = 10

from sklearn.neighbors import KNeighborsClassifier

# Step 1: Train the kNN classifier
classifier = KNeighborsClassifier(n_neighbors=5, metric='euclidean')
classifier.fit(embeddings, df['forum_zh'])

# Step 2: Make predictions on a test set
test_embeddings = embed_model(test_texts)
test_embeddings = np.array(test_embeddings).astype("float32")
predicted_labels = classifier.predict(test_embeddings)
```

```
# # DO NOT MODIFY THE BELOW LINE!  
print("precision = ", precision/topk)
```

