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/1PTq1G45IaSRGoIdFQla4izk6KOImvwmM?usp=sharing

---

Student ID: B0928022

Name: 杜云驊

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

```
1 !wget -O Dcard.db https://github.com/cjwu/cjwu.github.io/raw/master
```

```
--2023-04-24 07:36:37--  https://github.com/cjwu/cjwu.github.io/raw/master/
Resolving github.com (github.com)... 20.27.177.113
Connecting to github.com (github.com)|20.27.177.113|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/cjwu/cjwu.github.io/master/cour
--2023-04-24 07:36:37--  https://raw.githubusercontent.com/cjwu/cjwu.github
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199
HTTP request sent, awaiting response... 200 OK
Length: 151552 (148K) [application/octet-stream]
Saving to: ‘Dcard.db’

Dcard.db            100%[===================>] 148.00K  --.-KB/s    in 0.03

2023-04-24 07:36:38 (5.20 MB/s) - ‘Dcard.db’ saved [151552/151552]
```

```
import sqlite3
import pandas as pd

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

	createdAt	title	excerpt	categories	topics	forum_en	fo
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			dressup	
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，（圖是草屯囝仔的校園演唱會截圖）	詢問	衣服	鞋子	衣物	男生穿搭	尋找	dressup	
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	退貨	售後服務	dressup	
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		衣服	尋找	日常穿搭	男生穿搭	dressup	

```
!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
```

```
docid = 355
texts = "[" + df['title'] + '] [' + df['topics'] + '] ' + df['excer
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，新的頻道改成有線條的，感覺大家好像比較喜歡這種風格，試試看新的風格，影
片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
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

```

```
32 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，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

```
1 new_df = df.drop(columns=['createdAt', 'categories', 'topics', 'for
2 new_df
```

	title	excerpt	forum_zh
0	專題需要數據🥺🥺幫填～	希望各位能花個20秒幫我填一下	穿搭
1	#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，（圖是草屯囝仔的校園演唱會截圖）	穿搭
2	#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...	穿搭
3	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	穿搭
4	#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...	穿搭
...	...	...	...
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，新的頻道改成有線條的，感覺大家好像比較喜歡...	YouTuber
356	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……	YouTuber

```
pip install jieba
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/cola
Requirement already satisfied: jieba in /usr/local/lib/python3.9/dist-packa
```

Implemement 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]

```
import jieba
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsClassifier

precision = 0
topk = 10

# YOUR CODE HERE!
# IMPLEMENTIG TRIE IN PYTHON


def chinese_tokenizer(text):
    words = jieba.cut(text)
    return " ".join(words)

df["text"] = df["title"].astype(str) + " " + df["excerpt"].astype(s
df["text_tokenized"] = df["text"].apply(chinese_tokenizer)


vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(df["text_tokenized"])
y = df["forum_zh"]


clf = KNeighborsClassifier(n_neighbors=5)
scores = cross_val_score(clf, X, y, cv=5)
precision = scores.mean()


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

```
precision =  0.04166666666666667
```

Colab 付費產品 - 按這裡取消合約

✓ 0 秒 完成時間：下午3:38