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: ***https://colab.research.google.com/drive/11Mab5XNXeHcD42XbJPUxX5wAHvyvor8t?usp=sharing***

Student ID:B0928007

Name:余明昌

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/nlp2
```

```
--2023-04-24 08:52:20--  https://github.com/cjwu/cjwu.github.io/raw/master/
Resolving github.com (github.com)... 192.30.255.113
Connecting to github.com (github.com)|192.30.255.113|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/cjwu/cjwu.github.io/master/cour
--2023-04-24 08:52:20--  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 0.01

2023-04-24 08:52:20 (10.2 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	
			來源：覺得呱吉這襯衫好好		衣服 | 尋找 |		

```
!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-mult


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 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

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]

```
from nltk.corpus import stopwords
import string
import jieba

def tokenize_sentence(sentence):
    stop_words = stopwords.words("chinese")
    stop_words2 = ("\n", ", ", " ", "\r\n", "，", "。", "…", "★", "、", "《", "》"

    tokens = jieba.cut(sentence, cut_all = False, HMM = True)
    tokens = [i for i in tokens if i not in string.punctuation]
    tokens = [i for i in tokens if i not in stop_words]
    tokens = [i for i in tokens if i not in stop_words2]

    return tokens
```

```
import nltk

nltk.download("stopwords")
```

```
    [nltk_data] Downloading package stopwords to /root/nltk_data...
    [nltk_data]   Package stopwords is already up-to-date!
    True
```

```
import collections
import pandas as pd
import math

record = set()
idf_count = collections.defaultdict(int)
lengthOfArcticles = 0

for _, d in df.iterrows():
      tokens = tokenize_sentence(d["title"] + d["excerpt"])
      for token in set(tokens):
          record.add(token)
          idf_count[token] += 1
      lengthOfArcticles += 1

x = pd.DataFrame()
y = pd.DataFrame(columns = ["forum_zh"])

x_data, y_data = [], []
for _, d in df.iterrows():
    chart = collections.Counter(tokenize_sentence(d["title"] + d["excerpt"]))
    temp = {}
    w_length = sum(chart.values())
    for w, n in chart.items():
        tf = round(n / w_length, 4)
        idf = round(lengthOfArcticles / idf_count[w], 4)
        temp[w] = round(tf * math.log(idf, 10), 4)

    info_x = pd.DataFrame(temp, index = [len(x_data)])
    x_data.append(info_x)

    info_y = pd.DataFrame({"forum_zh": d["forum_zh"]}, index = [len(y_data)])
    y_data.append(info_y)

print("Concating x...")
x = pd.concat([x] + x_data, axis = 0)
x = x.fillna(0)
print("Concating y...")
y = pd.concat([y] + y_data, axis = 0)
```

```
Building prefix dict from the default dictionary ...
DEBUG:jieba:Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
DEBUG:jieba:Loading model from cache /tmp/jieba.cache
Loading model cost 1.263 seconds.
DEBUG:jieba:Loading model cost 1.263 seconds.
Prefix dict has been built successfully.
DEBUG:jieba:Prefix dict has been built successfully.
Concating x...
Concating y...
```

x

	專題	數據	🥺	幫填	～	希望	能花個	20	秒	幫	...
0	0.1879	0.2129	0.2756	0.2129	0.0795	0.123	0.2129	0.1377	0.1628	0.115	...
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
2	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
...	...	...	...	...	...	...	...	...	...	...	...
355	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
356	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
357	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...
358	0.0000	0.0000	0.0000	0.0000	0.0244	0.000	0.0000	0.0000	0.0000	0.000	...
359	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.000	...

360 rows × 5341 columns

y

	forum_zh
0	穿搭
1	穿搭
2	穿搭
3	穿搭
4	穿搭
...	...
355	YouTuber
356	YouTuber
357	YouTuber
358	YouTuber
359	YouTuber

360 rows × 1 columns

```
from sklearn.neighbors import KNeighborsClassifier

KNN = KNeighborsClassifier()
print("Training KNN...")
KNN.fit(x, y)
print("Predicting KNN...")
prediction = KNN.predict(x)
```

```
Training KNN...
Predicting KNN...
/usr/local/lib/python3.9/dist-packages/sklearn/neighbors/_classification.py
  return self._fit(X, y)
```

```
result = pd.DataFrame(columns = ["title", "excerpt", "forum_zh", "prediction"])
temp = []
for i, d in df.iterrows():
    info = pd.DataFrame({"title": d["title"], "excerpt": d["excerpt"], "forum_zh
    temp.append(info)

result = pd.concat([result] + temp, axis = 0)
```

```
result
```

	title	excerpt	forum_zh	prediction
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	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	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	感情
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	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	女孩
...	...	...	...	...
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	感情

```
precision = 0
topk = 10

# YOUR CODE HERE!
# IMPLEMENTIG TRIE IN PYTHON

from IPython.display import display

def find(n):
    target = [t for t in x.iloc[n]]
    _, index = KNN.kneighbors([target], 10)

    top10 = pd.DataFrame()

    temp = [result.iloc[i] for i in index]
    top10 = pd.concat([top10] + temp, axis = 1)

    display(top10)

    n = 0
    for _, r in top10.iterrows():
        if (r["forum_zh"] == r["prediction"]):
            n += 1

    return n

precision = find(355)

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

```
/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X
  warnings.warn(
```

	title	excerpt	forum_zh	prediction
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	感情
307	#詢問 求推薦父母喜歡的連續劇	如題，我媽媽最近身體不適，常常臥床開始看Netflix （他之前都沒看），她特別喜歡黑道律師...	Netflix	Netflix
59	喜歡上一個人	好奇問一下 想看大家有沒有相似的經驗，大學三年了 身邊的女生該熟悉的都熟了 不認識的也還是不...	感情	感情
140	水瓶男在想什麼⋯	我是雙魚 平常很獨立的 但喜歡一個人就會直接黏上去，跟一個水瓶男認識快一個月，一開始對他沒什...	星座	感情
49	完全搞不懂她在想什麼?有人能幫我解答看看嗎?(文長)	大家好 認識一個女生差不多1年了 互動感覺一直都不錯 對方據我所知是沒交過男友，雖然說認識一...	感情	感情
343	求問呱張新聞去哪了	之前呱張新聞真的是我日常調劑的好東西，但已經九個月沒更新了，有人知道怎	YouTuber	Netflix