import jieba  
import pandas as pd  
from sklearn import tree  
from collections import Counter  
from pandas.core.frame import DataFrame  
import math  
  
#jieba.set\_dictionary('/dict.txt') #引用預設詞庫  
#stop = open('stop\_PTT.txt') #引用停用詞  
  
df = pd.read\_csv("ptt\_tag\_V1\_utf8.txt")  
df\_SI = df.loc[:103, ['titl', 'push', 'tag', 'stage', 'item', 'text']]  
# print(df\_SI)  
  
# 將ppt內文全部斷詞，將原始資料計算項目個數，再做成字串  
  
word\_list1 =[]  
for item1 in df\_SI['text']:  
 seglist = jieba.cut(item1, cut\_all=False)  
  
 word\_list2 = []  
 for item2 in seglist:  
 if len(item2)>1:  
 word\_list2.append(item2)  
 word\_list3 = Counter(word\_list2)  
 word\_list4 = pd.DataFrame.from\_dict(word\_list3,orient='index')  
 word\_list5 = pd.DataFrame.transpose(word\_list4)  
 #print(word\_list3)  
 #print(word\_list4)  
 #print(word\_list5)  
 word\_list1.append(word\_list5)  
df\_new = pd.concat(word\_list1,ignore\_index=True)  
header = list(df\_new)  
  
# TF-IDF  
  
df\_Bank=pd.DataFrame()  
list\_bank = ['花旗','中信','台新','國泰','玉山']  
for item\_bank in list\_bank:  
  
 arti\_bank = df\_new[df\_new[item\_bank]>0]  
 TF = arti\_bank.sum(axis=0) / sum(arti\_bank.sum(axis=0))  
 #print(TF)  
  
 shp = df\_new.shape[0]  
 cnt = df\_new.count()  
 sc = shp/cnt  
 #print(cnt)  
 #print(shp)  
 #print(sc)  
  
 IDF = []  
 for item\_sc in sc:  
 item\_idf = math.log(item\_sc)  
 IDF.append(item\_idf)  
 #print(IDF)  
  
 TFIDF = [i \* j for i, j in zip(TF, IDF)]  
 #print(TFIDF)  
 #print(header)  
  
 Dict\_TFIDF={item\_bank+'\_hd':header, item\_bank+'\_ti':TFIDF}  
 df\_TFIDF = DataFrame(Dict\_TFIDF)  
 df\_TFIDF1 = df\_TFIDF.sort\_values(item\_bank+'\_ti',ascending=False).reset\_index(drop=True)  
 df\_Bank = pd.concat([df\_Bank,df\_TFIDF1], axis=1)  
  
print(df\_Bank)