學號：0653440

（一）步驟：

1.利用 jieba 套件將檔案的 postContent 欄位代表各文章之內容切割並各取出 200 個代表關鍵字。

```
data=pd.read_excel("FDATA.xlsx")
cutdata=[]
for i in range(data.shape[0]):
    x = jieba.analyse.extract_tags(data["postContent"][i], 200)
    cutdata.append(' '.join(x))
```

2.計算 tfidf

```
vectorizer = CountVectorizer()
transformer = TfidfTransformer()
tfidf = transformer.fit_transform(vectorizer.fit_transform(cutdata))
word = vectorizer.get_feature_names()
datatfidf = tfidf.toarray()
```

3. 分群、分類(k-means, -Decision Tree Classifier):

-K-means:(輸出各群判斷結果及正確率，最後輸出總正確率。)

```
kmeans = KMeans(n_clusters=5, random_state=0).fit(datatfidf)
kmeanslabels = kmeans.labels_
labelsslabel=[[],[],[],[],[]]

for i in range(len(kmeanslabels)):
    labelsslabel[kmeanslabels[i]].append(data['mainTag'][i])

allcorrect=0
for i in range(5):
    print('第',i+1,'群有',len(labelsslabel[i]),'篇文章。')
    setlabel=list(set(labelsslabel[i]))
    y=np.zeros(len(setlabel))
    for x in labelsslabel[i]:
        y[setlabel.index(x)]+=1
    print('包含了：')
    for z in range(len(setlabel)):
        print(int(y[z]),'篇為',setlabel[z],'分類。', end='')
    print()
    print('判斷此群為',setlabel[np.argmax(y)])
    print('正確率為：',max(y)/sum(y))
    allcorrect=allcorrect+max(y)
    print()
print('xxxxxxxxxxxxx')
print('總正確率:',allcorrect/len(data))
print('xxxxxxxxxxxxx')
```

-Decision Tree Classifier

先將 mainTag 轉為數字 1~5

```
data['mainTag'] = pd.Categorical.from_array(data.mainTag).labels
```

帶入 Decition Tree 套件，切割 75%訓練集、25%測試集，並計算測試的 Accuracy。

```
X = datatfidf
Y = data['mainTag']
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.25)
clf=DecisionTreeClassifier(max_depth=10)
clf=clf.fit(X_train,Y_train)

from sklearn.metrics import accuracy_score
Y_predict=clf.predict(X_test)
print('xxxxxxxxxxxxxxxxxxxxxxxxx')
print("Accuracy=",accuracy_score(Y_test,Y_predict))
print('xxxxxxxxxxxxxxxxxxxxxxxxx')
```

（三）分析：

1.輸出結果：

-k-means:

```
第 1 群有 13 篇文章。
包含了：
1 篇為 美食 分類。1 篇為 財經 分類。11 篇為 運動 分類。
判斷此群為 運動
正確率為： 0.846153846154

第 2 群有 39 篇文章。
包含了：
3 篇為 美食 分類。10 篇為 科技 分類。16 篇為 財經 分類。3 篇為 運動 分類。7 篇為 天氣 分類。
判斷此群為 財經
正確率為： 0.410256410256

第 3 群有 43 篇文章。
包含了：
13 篇為 美食 分類。8 篇為 科技 分類。12 篇為 運動 分類。10 篇為 財經 分類。
判斷此群為 美食
正確率為： 0.302325581395

第 4 群有 27 篇文章。
包含了：
3 篇為 美食 分類。1 篇為 科技 分類。23 篇為 天氣 分類。
判斷此群為 天氣
正確率為： 0.851851851852

第 5 群有 18 篇文章。
包含了：
1 篇為 科技 分類。3 篇為 財經 分類。14 篇為 運動 分類。
判斷此群為 運動
正確率為： 0.777777777778
```

```
xxxxxxxxxxxxxxxxxxxxxxx
k-means總正確率: 0.55
xxxxxxxxxxxxxxxxxxxxxxx
```

-Decision Tree Classifier:

```
xxxxxxxxxxxxxxxxxxxxxxxxxxx
Accuracy= 0.828571428571
xxxxxxxxxxxxxxxxxxxxxxxxxxx
```

2.討論：

-k-means:

觀察結果發現，科技被分在各群中，而產生了兩群為運動分類的結果，應該是因為運動的篇數較多的原因。

總正確率為 0.55，已比 1/5 亂猜高出許多，故此分類還是有用的。

-Decision Tree Classifier:

相較 kmeans，利用訓練後的 decision tree 來預測，可達到 7、8 成的正確率。