Homework 1

Yu-Chieh Kuo B07611039

1 Overview

這份作業要求我們使用python package求出給定輸入資料的TF-IDF vectors，並計算前兩份TF-IDF vectors之間的cosine similarity。我針對給定的輸入文件進行一些前處理，包括轉成小寫、去除英文的stop words，並簡單去除掉標點符號後利用sklearn套件內的函數求出給定資料的TFIDF vector，並求出前兩個vector之間的cosine similarity約為0.200。

2 Code Explanation

以下是我的原始程式碼，此份程式碼也會包在壓縮檔內一同繳交。

```
from nltk.tokenize import word_tokenize
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np
import os
import nltk

all_file = list(range(1, 1096))

for i in range(1, 1096):
    ### read file
    doc_name = "~/NTUCourse/NTU-IM-ITM/HW-01/PA1-data/" + str(i) + ".txt"
    file = os.path.expanduser(doc_name)
    f = open(file)
    docs = f.read()
    #print(docs)

    ### lowerize
    token_lower = docs.lower()
    #print(token_lower)

    ### stop punctuation
    token_sequence = word_tokenize(token_lower)
    stop_punc = [',',';','.','\'','?']
    tokens_wo_stop_punc = [x for x in token_sequence if x not in stop_punc]
    #print(tokens_wo_stop_punc)

    ### stop words in english
```

```
    stop_words = nltk.corpus.stopwords.words('english')
    tokens_wo_stop_words = [x for x in tokens_wo_stop_punc if x not in
   stop_words]
    #print(tokens_wo_stop_words)

    lst = ' '.join([str(elem) for elem in tokens_wo_stop_words])
    final_docs = []
    final_docs.append(lst)
    #print(final_docs)

    all_file[i-1] = lst

### calculate TFIDF vectors
TFIDF_vectorizer = TfidfVectorizer()
TFIDF_vectors = TFIDF_vectorizer.fit_transform(all_file)

### outfile the 1.vec and 2.vec
for i in range(1,3):
    filename = "doc" + str(i) + ".txt"
    outF = open(filename, "w")
    for line in TFIDF_vectors.toarray()[i - 1]:
        outF.write(str(line))
        outF.write("\n")
    outF.close()

### print the cosine similarity
print("The cosine similarity of 1.vec and 2.vec is",cosine_similarity(
   TFIDF_vectors[0], TFIDF_vectors[1]).flatten()[0])
```

Listing 1: Python code

以下將講解我如何處理資料，以及如何求出TFIDF vectors與cosine similarity。

1. **Import packages:** 讀入本次作業需要的套件，如nltk、sklearn、numpy、os等，示於第一行至第八行，其中os用於讀檔時需指定檔案路徑，其餘套件則是如上課所教。

2. **Read files:** 第十四至第十七行是用來將檔案讀入python中，由於讀檔需要透過絕對路徑找尋檔案，因此需要用到os這個套件。我透過迴圈將檔案存進doc這個變數中。

3. **Text pre-processing:** 第二十一至第三十七行則是對文字進行前處理，我將文件中所有英文字轉成小寫，並去除標點符號與英文中的停止單字。值得一提的是透過套件內的方程式所產生出來的文字會是很多element的list，所以我於第三十五至三十七行將這些切割好的單字重新連成一個list，並將這些處理好的文字assign給預先定義好的list中，以利後續計算TFIDF vectors。

4. **TFIDF vectors:** 第四十三、四十四行就是透過上課所教的方法計算TFIDF vectors。

5. **Output files:** 第四十七行至第五十三行是將TFIDF vector輸出成兩份文字檔1.vec與2.vec，流程就只是讀$TFIDF_vectors$ 這個變數中的內容，並將他寫入欲輸出的檔案中。

6. **Cosine similarity:** 第五十六行就是印出前兩個vector之間的cosine similarity，也是透過上課教的方法，用套件內的函數直接計算兩份vector之間的cosine similarity。最後算出來的結果約為0.200。