資訊檢索與文字探勘導論_HW1_Report

B06705030 資管三 蕭昀豪

1. 執行環境

Jupyter Notebook

2. 程式語言 (請標明版本)

Python3.7.3

3. 執行方式

。使用的非原生套件:

- punkt (原生套件 nltk 的衍生套件)
- stopwords (原生套件 nltk 的衍生套件)

在 Source Code 中，我已附上安裝指令，如下圖

```
In [ ]:
#download the necessary packages
import nltk
nltk.download('punkt')
nltk.download('stopwords')
```

如果您們沒有 nltk，可依以下指令安裝

pip install --user -U nltk

4. 作業處理邏輯說明

我採用 OOP 編程方式，創建一進行文字處理的類別 textMachine。使用者首先需要在宣告一 textMachine 時，將欲處理的 document 檔名當作參數傳入 textMachine 初始化。textMachine 會將該檔案中的文字取

出，存入變數成員 self.__document 中，作為保留原始資料之用。此次作業我預設檔案名稱為"document.txt"，在此前提執行。

接著初始化一新變數成員 self.__result，使其值等同 self.__document，作為文字處理對象。然後利用 self.processing(self)涵式對 self.__result 進行處理。處理流程為

1. lowercasing everything
2. tokenization
3. stemming using Porter's Algorithm
4. remove stop words

   其中在步驟 2~4，我借助 nltk 涵式庫中的涵式進行處理。

最終利用 textMachine 的成員涵式 saveResult 將擷取下的 term 以"\n"為間隔存入 result.txt。