1. 執行環境 & 作業系統

   Mac Command Shell

2. 程式語言, 版本

   Python(2.7.10)

3. 執行方式

   (1) 安裝 pip 套件

   (2)安裝 nltk 與 tqdm(跑進度條)模組

```
~/Desktop/IRTM/hw/hw2 sudo pip install nltk tqdm
```

   (3)安裝 nltk 相關檔案

   下 python 指令進入 python shell 後再下以下指令

   (1. Import nltk 2.nltk.download() 3.d 4.stopwords)

```
~/Desktop/IRTM/hw/hw2 python
Python 2.7.10 (default, Jul 14 2015, 19:46:27)
[GCC 4.2.1 Compatible Apple LLVM 6.0 (clang-600.0.39)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download()
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
Downloader> d

Download which package (l=list; x=cancel)?
  Identifier> stopwords
    Downloading package stopwords to /Users/ken/nltk_data...
      Package stopwords is already up-to-date!
```

   (4)執行 hw2.py

   (先於某目錄下解壓縮, 並於該目錄下輸入以下指令)

```
~/Desktop/IRTM/hw/hw2 python hw2.py
```

4. 作業處理邏輯說明

   (1) 利用將作業一的程式(**切詞處理：token, 轉為小寫, stemming, 移除標點符號…**)移植至作業二，再從資料夾下的 **IRTM/**路徑下讀取所有的原始 txt 檔案，並利用 **TFIDFSave** 函式將其每一個出現的 term 做 frequency 加總

   (2) 利用 **TermToIdxMake** 函式建立 term 與 idx 的映射 dictionary—**term_idx**

   (3) 利用 **TFIDFGet** 函式與 **TransferTermToIdx** 函式建立 tf-idf 的 dictionary—**tf_idf_dicts**

   (4) 根據此 **term_idx** 內容，將其每個 term 的 frequency 內容寫入 dictionary.txt

   (5) 根據 **tf_idf_dicts** 的內容，將第一個位置(doc1)的裡所包含的 term 與 tf_idf 內容寫入 1.txt

(6) 建立 **cosine** 函式，先利用 **vectorize** 函式將兩個輸入的 doc 裡所包含的 term 變為同樣長度經過正規劃的 vector，再將其做相乘回傳 cosine 值

5. 使用 nltk 中的套件做 stopword，移除標點符號後得到共 14984 個 term，而最後對 document1 及 document2 算出的 cosine 值為 0.174425354169