1. 執行環境 & 作業系統

**Mac Command Shell**

1. 程式語言, 版本

**Python(2.7.10)**

1. 執行方式

(1) [安裝pip套件](https://packaging.python.org/tutorials/installing-packages/)

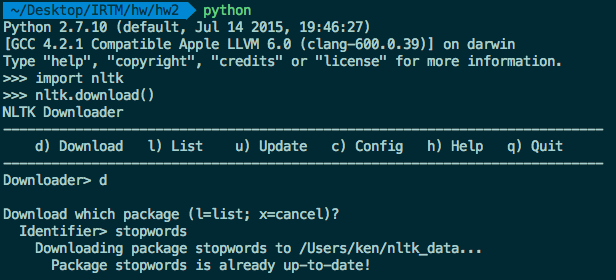
(2)安裝nltk與tqdm(跑進度條)模組



(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)



(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)



1. 作業處理邏輯說明
2. 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
3. 利用**TermToIdxMake**函式建立term與idx的映射dictionary—**term\_idx**
4. 利用**TFIDFGet**函式與**TransferTermToIdx**函式建立tf-idf的dictionary—**tf\_idf\_dicts**
5. 根據此**term\_idx**內容，將其每個term的frequency內容寫入dictionary.txt
6. 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
7. 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
8. 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