資訊檢索與文字探勘導論

hw1 Report

B03705012 張晉華

**1. 程式語言**

Python 3.6.1

**2. 執行環境**

Linux OS (Ubuntu 16.04 LTS)

Python3 Packages Requirements:

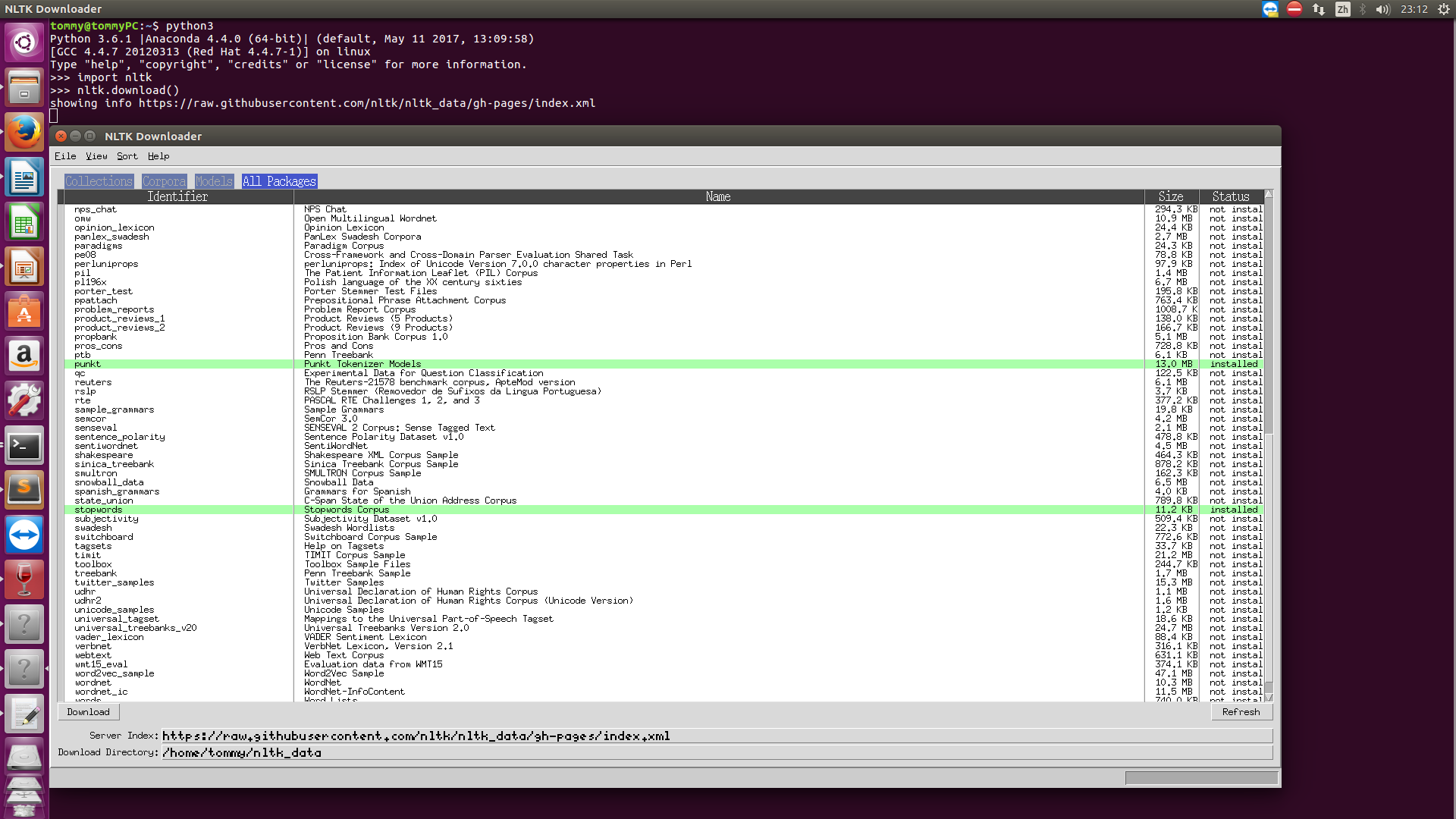
* Python Natural Language Toolkit(nltk 3.2.2)

**3. 執行方式**

Step1：Install Packages ： NLTK  
(可以透過pip3 install nltk安裝)

Step2：Corpus Download  
有兩個Corpus需要下載以供程式使用：punkt(tokenize需要)和stopwords

可以在python3環境下執行nltk.download()下載



Step3：cd進程式所在路徑，在termial執行python3 hw1.py，檔案結果產出在同一路徑的result.txt

**4. 作業處理邏輯說明**

Step1：將文本中的標點符號去除（ex:',.）

Step2：使用nltk.tokenize.word\_tokenize()，利用空白和punkt文本將出入的文章段分詞

Step3：使用nltk.PorterStemmer()實作stemming

Step4：利用stopwords文本將輸入文章段裡有在stopwords的set裡面的字去除。

各步驟結果如下圖：

