# 資訊檢索與文字探勘導論

# hw3 Report

B03705012 張晉華

## 1. 程式語言

Python 3.5.2

## 2. 執行環境

Linux OS (Ubuntu 16.04 LTS)

Python3 Packages Requirements:

- Python Natural Language Toolkit(nltk 3.2.2)
- Python Numpy(1.13.3)

## 3. 執行方式

- **Package Installation**
  - Python Natural Language Toolkit(NLTK)

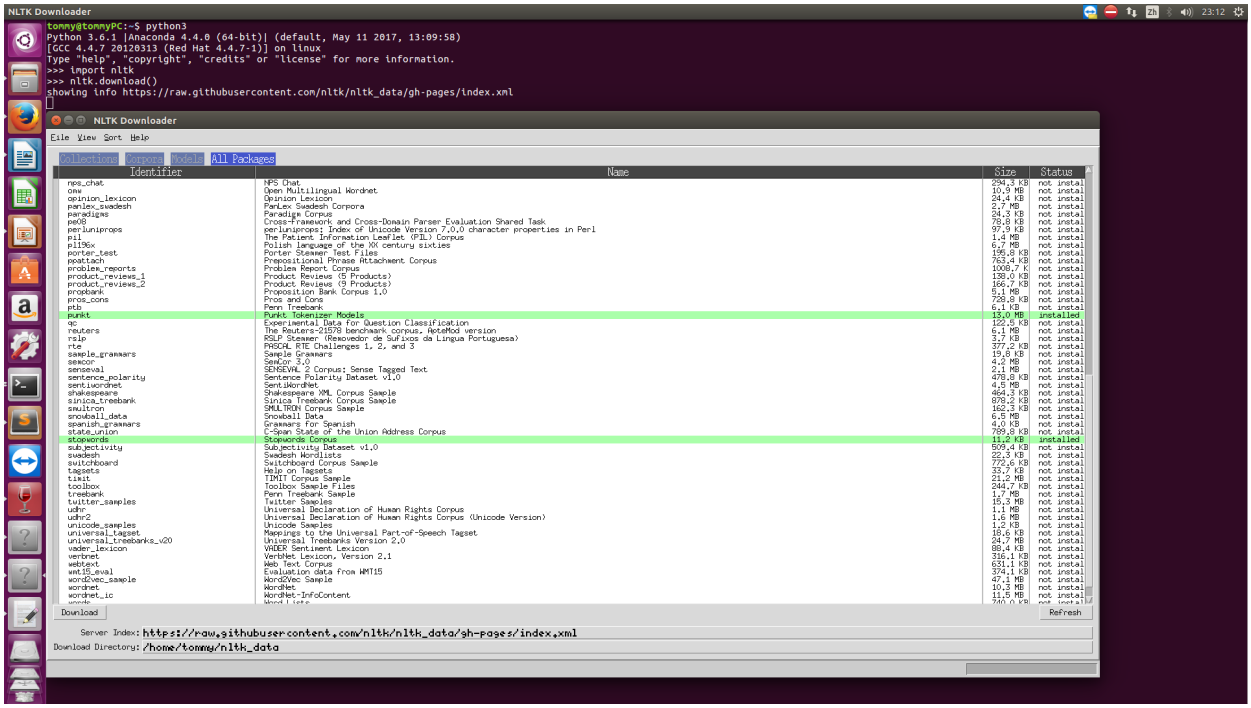
    Step1：Install Packages

    (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



- Python Numpy：可透過 pip3 install numpy 安裝

- **執行指令**
  - cd 進程式所在路徑，在 termial 執行 python3 classifier.py，檔案結果產出在同一路徑的 B03705012.txt，內容為所有 testing 文章的 class_id prediction.

## 4. 作業處理邏輯說明

- **classifier.py**

  分為 split_set,build_model,predict3 個大步驟

  - split_set

    將文章切成 train,test 和 validation(內部做 Evaluation 時使用）3 個 dataset，每個 dataset 為一個 dict{doc_id：class_id}

    （由於切割 dataset 時有加入隨機的變數，所以每次執行結果可能有所不同）

  - build_model

    先將每個 term 在每篇文章出現的次數先算好儲存成 dict{term：{doc_id:freq}}，然後計算每個 term 對每個 class 的 likelihood，選加總最高的 500 個為 feature，並計算 P(t|c)建立 NB classifier

  - predict

    利用建立好的 NB classifier 進行預測，並且對 validation 的結果執行 evaluation 計算 accuracy，並對 testing dataset 的結果執行 output 輸出成檔案。

  執行截圖如下頁圖：

```
tommy@tommyPC: ~/NTU/IR2017/hw3
tommy@tommyPC:~/NTU/IR2017/hw3$ python3 classifier.py
Number of Documents :: Training Dataset - 156 | Validation - 39 | Testing Dataset - 900
Number of Terms in Training Dataset - 4956
Terms after Feature Selection : ['state', 'us', 'thi', 'presid', 'say', 'two', 'peopl', 'hi', 'one', 'time', 'y
ear', 'last', 'offici', 'tri', 'said', 'new', 'also', 'first', 'week', 'american', 'kill', 'would', 'nation', '
sinc', 'news', 'report', 'call', 'mani', 'befor', 'power', 'hous', 'offic', 'open', 'ha', 'elect', 'clinton', '
countri', 'go', 'citi', 'like', 'help', 'onli', 'back', 'could', 'govern', 'becaus', 'take', 'monday', 'come',
'month', 'use', 'remain', 'leader', 'secur', 'dure', 'day', 'russian', 'presidenti', 'forc', 'unit', 'still', '
today', 'three', 'four', 'believ', 'author', 'polit', 'south', 'former', 'sunday', 'includ', 'work', 'sever', '
around', 'return', 'small', 'know', 'support', 'told', 'think', 'plan', 'death', 'least', 'polic', 'home', 'lea
v', 'end', 'next', 'way', 'close', 'get', 'opposit', 'gener', 'took', 'major', 'right', 'may', 'dont', 'veri',
'possibl', 'continu', 'democrat', 'investig', 'miss', 'white', 'night', 'begin', 'intern', 'came', 'saturday',
'friday', 'want', 'wa', 'mile', 'rule', 'men', 'found', 'public', 'station', 'ani', 'spokesman', 'anoth', 'expl
os', 'depart', 'capit', 'thursday', 'servic', 'arriv', 'court', 'die', 'tuesday', 'west', 'ago', 'made', 'find'
, 'prison', 'candid', 'minist', 'hope', 'good', 'run', 'hour', 'near', 'much', 'look', 'part', 'even', 'injur',
 'chief', 'bodi', 'appear', 'militari', 'left', 'visit', 'dead', 'convict', 'area', 'second', 'coast', 'campaig
n', 'step', 'meet', 'washington', 'hundr', 'away', 'see', 'famili', 'top', 'outsid', 'russia', 'earthquak', 'fo
reign', 'serv', 'ralli', 'street', 'emerg', 'appar', 'world', 'announc', 'man', 'head', 'question', 'put', 'set
', 'sign', 'thing', 'aid', 'law', 'constitut', 'build', 'son', 'worker', 'center', 'million', 'mr', 'im', 'acco
rd', 'follow', 'later', 'long', 'held', 'well', 'navi', 'war', 'vote', 'sea', 'figur', 'reach', 'case', 'point'
, 'talk', 'give', 'wife', 'five', 'effort', 'immedi', 'name', 'start', 'late', 'receiv', 'make', 'earlier', 'qu
ak', 'inform', 'secretari', 'caus', 'prime', 'charg', 'across', 'oper', 'went', 'earli', 'live', 'expect', 'mem
ber', 'ask', 'need', 'round', 'employe', 'damag', 'act', 'moscow', 'radio', 'process', 'red', 'felt', 'peac', '
hold', 'parti', 'chang', 'team', 'show', 'wednesday', 'turn', 'line', 'becom', 'yet', 'stay', 'urg', 'began', '
fear', 'collaps', 'fbi', 'accus', 'third', 'senat', 'injuri', 'search', 'play', 'hit', 'robert', 'john', 'issu'
, 'televis', 'respons', 'pull', 'stop', 'seen', 'high', 'recent', 'decid', 'far', 'suffer', 'place', 'san', 're
publican', 'william', 'protest', 'financi', 'victim', 'releas', 'destroy', 'race', 'express', 'reason', 'hand',
 'plane', 'import', 'side', 'town', 'along', 'strong', 'fail', 'insid', 'attack', 'six', 'prepar', 'morn', 'whe
ther', 'vice', 'defens', 'trade', 'suspect', 'east', 'north', 'corrupt', 'trial', 'term', 'vladimir', 'heard',
'decis', 'director', 'congress', 'car', 'neighbor', 'soldier', 'execut', 'thousand', 'consid', 'lot', 'chanc',
'known', 'store', 'lead', 'larg', 'whose', 'togeth', 'southern', 'crash', 'lost', 'number', 'develop', 'enough'
, 'group', 'past', 'robberi', 'summit', 'recov', 'cooper', 'agent', 'identifi', 'christma', 'strike', 'america'
, 'ground', 'crew', 'shook', 'account', 'buri', 'control', 'victori', 'torpedo', 'el', 'middl', 'retir', 'warn'
, 'local', 'statement', 'pope', 'taken', 'never', 'condit', 'sent', 'phone', 'seek', 'life', 'behind', 'tell',
'base', 'thought', 'alreadi', 'clear', 'confirm', 'agenc', 'ship', 'rescu', 'brunei', 'western', 'economi', 'so
und', 'georg', 'assist', 'slam', 'loui', 'associ', 'japan', 'st', 'texa', 'nearli', 'seven', 'insist', 'danger'
, 'milosev', 'slobodan', 'attend', 'appeal', 'region', 'normal', 'eight', 'net', 'got', 'organ', 'better', 'mov
e', 'send', 'stand', 'attempt', 'gather', 'within', 'kind', 'rel', 'sailor', 'toll', 'aboard', 'advis', 'weathe
r', 'parent', 'dig', 'putin', 'justic', 'consecut', 'rush', 'win', 'mexico', 'eve', 'arm', 'shot', 'deni', 'fli
', 'determin', 'doe', 'evid', 'tie', 'survivor', 'weapon', 'defend', 'challeng', 'whi', 'bush', 'feder', 'promi
s', 'pm', 'problem', 'air', 'rout', 'rest', 'face', 'wait', 'though', 'david', 'site', 'armi', 'tonight', 'subm
arin', 'mel', 'sport', 'vietnam', 'citizen', 'spread', 'join', 'didnt', 'realli', 'longer', 'given', 'among', '
despit', 'feel', 'result', 'carri', 'magnitud', 'landslid', 'asiapacif', 'trip', 'civilian', 'sentenc', 'crimin
', 'decemb', 'yugoslavia', 'econom', 'violenc', 'dozen', 'port', 'bid', 'juli', 'forward', 'murder', 'alli', 'c
ancel', 'medic', 'terror']
Accuracy : 100.000000
Prediction of Testing Dataset Finished.
tommy@tommyPC:~/NTU/IR2017/hw3$
```

## 5. 心得

這次的作業其實大致上就是跟著老師投影片上的算法實作一次，印象比較深刻的就是實作的過程中忘記老師曾經提醒過 P(t|c)連乘的結果可能會過小被當成 0 要改成 log 加總的技巧，導致一度讓很多篇字數較多的文章無法被成功的分類而因此苦惱，以後應該要更提醒自己多加注意這方面的細節。