國家發展委員會檔案管理局

107 年度電子檔案保存管理機制

委託服務案

電子文書檔案新知

電子檔案管理軟體簡介

以 MALLET 為例

(V 1.0)

中華民國 107 年 12 月

版本紀錄

版序	實施日期	修改內容
v1.0	1071231	初版

目錄

壹、MALLET 自然語言處理工具

MALLET（MAchine Learning for LanguaE Toolkit）是一個以機器學習語言的工具，以 Java 開發為基礎，透過 MALLET 工具，可以進行自然語言處理，包括文件分類、分群、建立主題模型、內文資訊擷取，以及其它與機器學習相關的應用。

MALLET 計算文件單詞出現機率形式，從而更有效的對文件進行主題分類。MALLET 包含了幾種文件分類的演算法，還有特徵提取的演算法等。文件分類的演算法如包括 NaïveBayes、Maximum Entropy 和 Decision Trees 等。

在建立主題模型上，MALLET 提供了一種分析大量未標記原文的簡單方法，而「主題」是由經常一起出現的一組詞所組成，分析上下文線索，主題模型可以連接具有相似含義的單詞，並區分具有多種含義的單詞使用。。

貳、軟體資訊

（一）軟體名稱：MALLET

（二）軟體版本：2.0.8

（三）支援系統：Windows OS、MAC OS

（四）軟體性質：自然語言學習

（五）支援語系：英語

（六）官方網站：http://mallet.cs.umass.edu/

一、安裝步驟

（一）http://mallet.cs.umass.edu/download.php，下載「mallet-2.0.8.zip」。

圖1　MALLET 官方網站下載頁面

（二）下載完成後，解壓縮「mallet-2.0.8.zip」。

圖2　解壓縮 mallet-2.0.8.zip

（三）解壓縮後畫面。

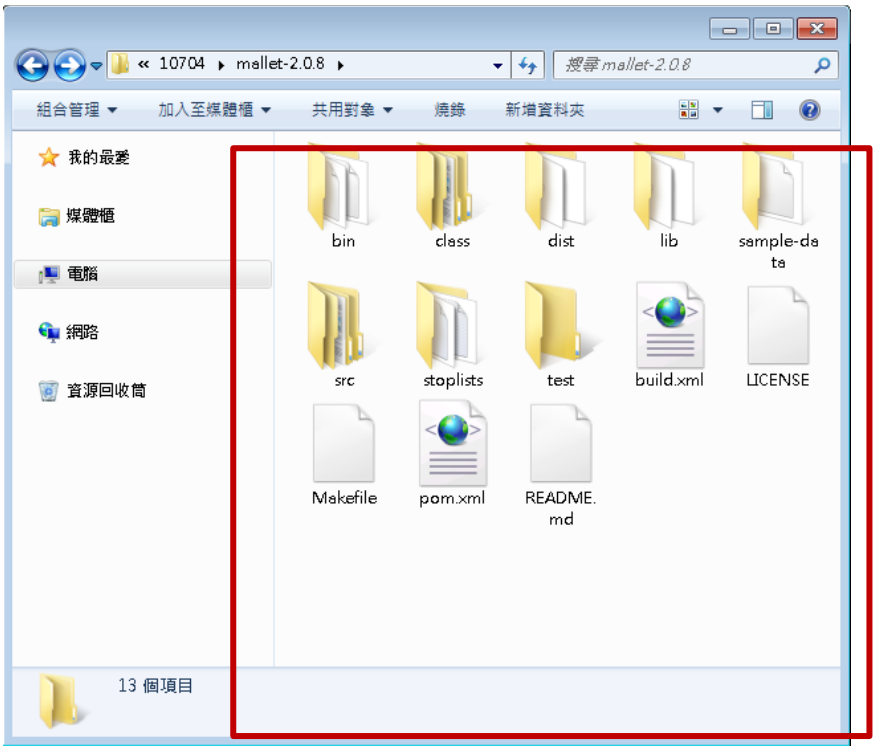


圖3　解壓縮後畫面

（四）進入「bin」資料夾，修改「mallet.bat」批次檔，可使用筆記本開啟。

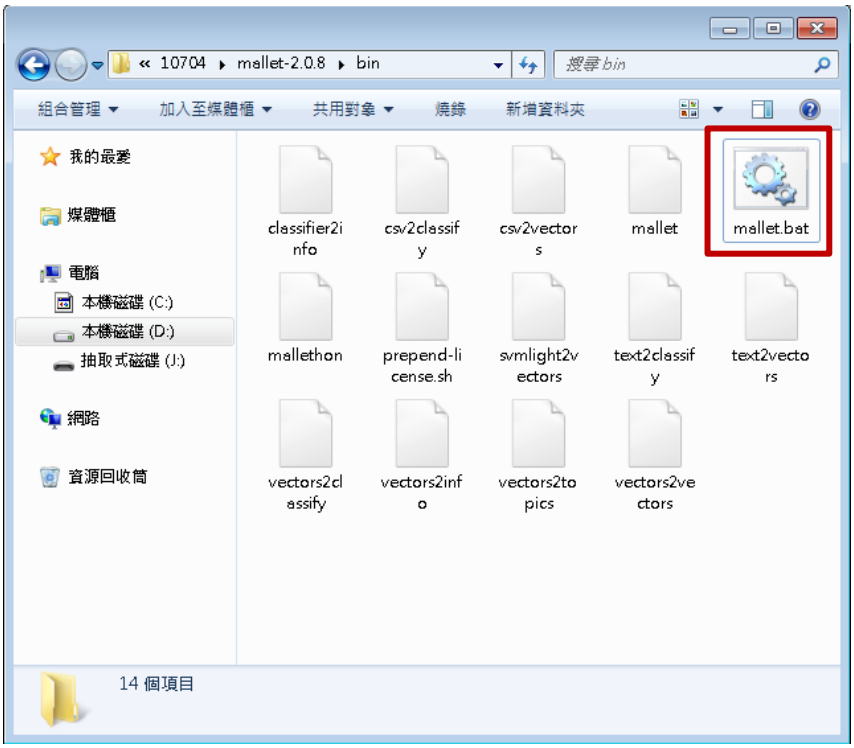


圖4　設定修改畫面

（五）將「%MALLET_HOME%」，修改為 MALLET 資料夾存放路徑，如圖 6。

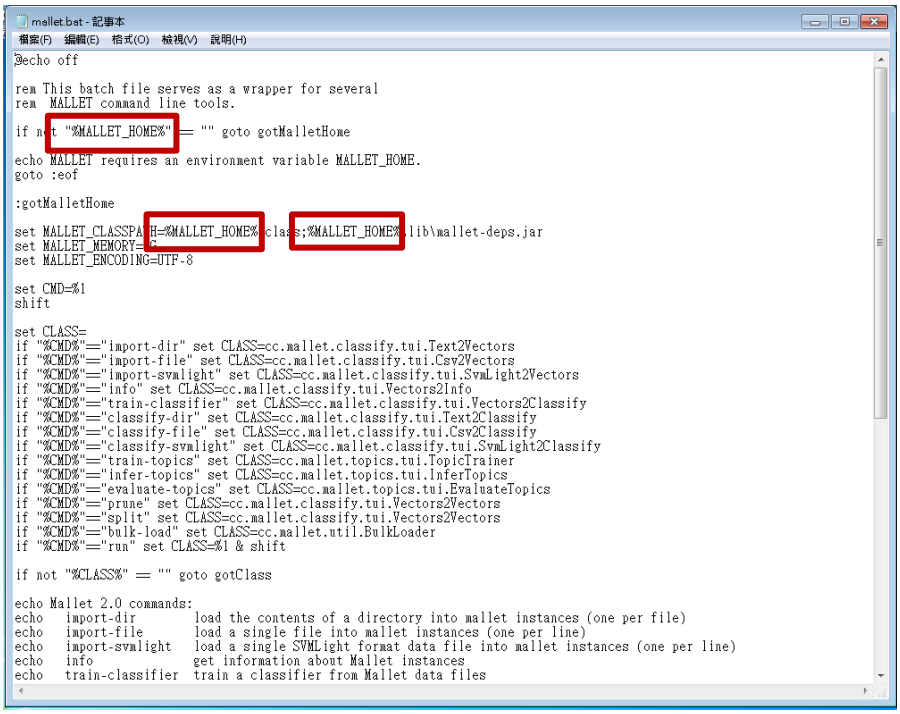


圖5　修改批次檔位置畫面

（六）修改為 MALLET 資料夾存放路徑。

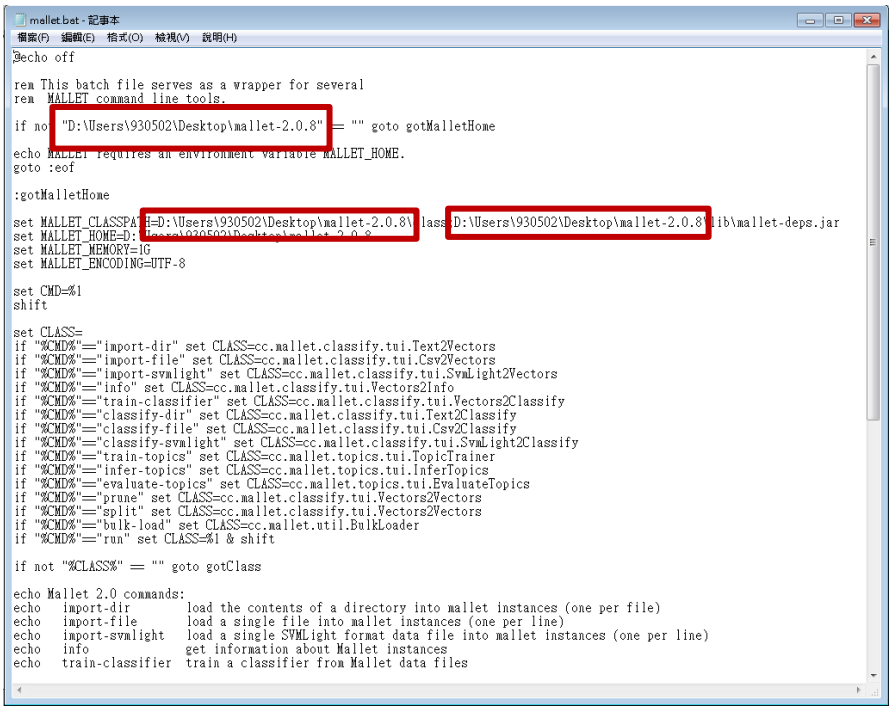


圖6　設定路徑位置畫面

（七）調整原始參數。

```
set CLASS=
if "%CMD%"=="import-dir" set CLASS=cc.mallet.classify.tui.Text2Vectors
if "%CMD%"=="import-file" set CLASS=cc.mallet.classify.tui.Csv2Vectors
if "%CMD%"=="import-svmlight" set CLASS=cc.mallet.classify.tui.SvmLight2Vectors
if "%CMD%"=="info" set CLASS=cc.mallet.classify.tui.Vectors2Info
if "%CMD%"=="train-classifier" set CLASS=cc.mallet.classify.tui.Vectors2Classify
if "%CMD%"=="classify-dir" set CLASS=cc.mallet.classify.tui.Text2Classify
if "%CMD%"=="classify-file" set CLASS=cc.mallet.classify.tui.Csv2Classify
if "%CMD%"=="classify-svmlight" set CLASS=cc.mallet.classify.tui.SvmLight2Classify
if "%CMD%"=="train-topics" set CLASS=cc.mallet.topics.tui.TopicTrainer
if "%CMD%"=="infer-topics" set CLASS=cc.mallet.topics.tui.InferTopics
if "%CMD%"=="evaluate-topics" set CLASS=cc.mallet.topics.tui.EvaluateTopics
if "%CMD%"=="prune" set CLASS=cc.mallet.classify.tui.Vectors2Vectors
if "%CMD%"=="split" set CLASS=cc.mallet.classify.tui.Vectors2Vectors
if "%CMD%"=="bulk-load" set CLASS=cc.mallet.util.BulkLoader
if "%CMD%"=="run" set CLASS=%1 & shift

if not "%CLASS%" == "" goto gotClass

echo Mallet 2.0 commands:
echo   import-dir         load the contents of a directory into mallet instances (one per file)
echo   import-file        load a single file into mallet instances (one per line)
echo   import-svmlight    load a single SVMLight format data file into mallet instances (one per line)
echo   info               get information about Mallet instances
echo   train-classifier   train a classifier from Mallet data files
echo   classify-dir       classify data from a single file with a saved classifier
echo   classify-file      classify the contents of a directory with a saved classifier
echo   classify-svmlight  classify data from a single file in SVMLight format
echo   train-topics       train a topic model from Mallet data files
echo   infer-topics       use a trained topic model to infer topics for new documents
echo   evaluate-topics    estimate the probability of new documents given a trained model
echo   prune              remove features based on frequency or information gain
echo   split              divide data into testing, training, and validation portions
echo   bulk-load          for big input files, efficiently prune vocabulary and import docs
echo Include --help with any option for more information
```

圖7　原始參數設定畫面

（八）調整為下圖參數。

```
set CLASS=
if "%CMD%"=="import-dir" set class="cc".mallet.classify.tui.Text2Vectors
if "%CMD%"=="import-file" set class="cc".mallet.classify.tui.Csv2Vectors
if "%CMD%"=="import-svmlight" set class="cc".mallet.classify.tui.SvmLight2Vectors
if "%CMD%"=="train-classifier" set class="cc".mallet.classify.tui.Vectors2Classify
if "%CMD%"=="classify-file" set class="cc".mallet.classify.tui.Csv2Classify
if "%CMD%"=="classify-dir" set class="cc".mallet.classify.tui.Text2Classify
if "%CMD%"=="classify-svm" set class="cc".mallet.classify.tui.SvmLight2Classify
if "%CMD%"=="train-topics" set class="cc".mallet.topics.tui.Vectors2Topics
if "%CMD%"=="infer-topics" set class="cc".mallet.topics.tui.InferTopics
if "%CMD%"=="estimate-topics" set class="cc".mallet.topics.tui.EstimateTopics
if "%CMD%"=="hlda" set class="cc".mallet.topics.tui.HierarchicalLDATUI
if "%CMD%"=="prune" set class="cc".mallet.classify.tui.Vectors2Vectors
if "%CMD%"=="split" set class="cc".mallet.classify.tui.Vectors2Vectors
if "%CMD%"=="bulk-load" set class="cc".mallet.util.BulkLoader
if "%CMD%"=="run" set CLASS=%1 & shift

if not "%CLASS%" == "" goto gotClass

echo Mallet 2.0 commands:
echo import-dir load the contents of a directory into mallet instances (one per file)
echo import-file load a single file into mallet instances (one per line)
echo import-svmlight load a single SVMLight format data file into mallet instances (one per line)
echo train-classifier train a classifier from Mallet data files
echo classify-file To apply a saved classifier to new unlabeled data (for one-instance-per-line data)
echo classify-dir To apply a saved classifier to new unlabeled data (for one-instance-per-file data)
echo classify-svm To apply a saved classifier to new svm data (for one-instance-per-line data)
echo train-topics train a topic model from Mallet data files
echo infer-topics use a trained topic model to infer topics for new documents
echo estimate-topics estimate the probability of new documents given a trained model
echo hlda train a topic model using Hierarchical LDA
echo prune remove features based on frequency or information gain
echo split divide data into testing, training, and validation portions
echo Include --help with any option for more information
```

圖8　修改後參數畫面

調整之參數

```
set CLASS=
if "%CMD%"=="import-dir" set class="cc".mallet.classify.tui.Text2Vectors
if "%CMD%"=="import-file" set class="cc".mallet.classify.tui.Csv2Vectors
if "%CMD%"=="import-svmlight" set class="cc".mallet.classify.tui.SvmLight2Vectors
if "%CMD%"=="train-classifier" set class="cc".mallet.classify.tui.Vectors2Classify
if "%CMD%"=="classify-file" set class="cc".mallet.classify.tui.Csv2Classify
if "%CMD%"=="classify-dir" set class="cc".mallet.classify.tui.Text2Classify
if "%CMD%"=="classify-svm" set class="cc".mallet.classify.tui.SvmLight2Classify
if "%CMD%"=="train-topics" set class="cc".mallet.topics.tui.Vectors2Topics
if "%CMD%"=="infer-topics" set class="cc".mallet.topics.tui.InferTopics
if "%CMD%"=="estimate-topics" set class="cc".mallet.topics.tui.EstimateTopics
if "%CMD%"=="hlda" set class="cc".mallet.topics.tui.HierarchicalLDATUI
if "%CMD%"=="prune" set class="cc".mallet.classify.tui.Vectors2Vectors
if "%CMD%"=="split" set class="cc".mallet.classify.tui.Vectors2Vectors
if "%CMD%"=="bulk-load" set class="cc".mallet.util.BulkLoader
if "%CMD%"=="run" set CLASS=%1 & shift

if not "%CLASS%" == "" goto gotClass

echo Mallet 2.0 commands:
echo import-dir load the contents of a directory into mallet instances (one per file)
echo import-file load a single file into mallet instances (one per line)
echo import-svmlight load a single SVMLight format data file into mallet instances
(one per line)
echo train-classifier train a classifier from Mallet data files
echo classify-file To apply a saved classifier to new unlabeled data (for
one-instance-per-line data)
echo classify-dir To apply a saved classifier to new unlabeled data (for
one-instance-per-file data)
echo classify-svm To apply a saved classifier to new svm data (for
one-instance-per-line data)
echo train-topics train a topic model from Mallet data files
echo infer-topics use a trained topic model to infer topics for new documents
echo estimate-topics estimate the probability of new documents given a trained model
echo hlda train a topic model using Hierarchical LDA
echo prune remove features based on frequency or information gain
echo split divide data into testing, training, and validation portions
echo Include --help with any option for more information
```

二、操作說明

（一）開啟命令提示字元，並進入\mallet-2.0.8\bin 目錄下。

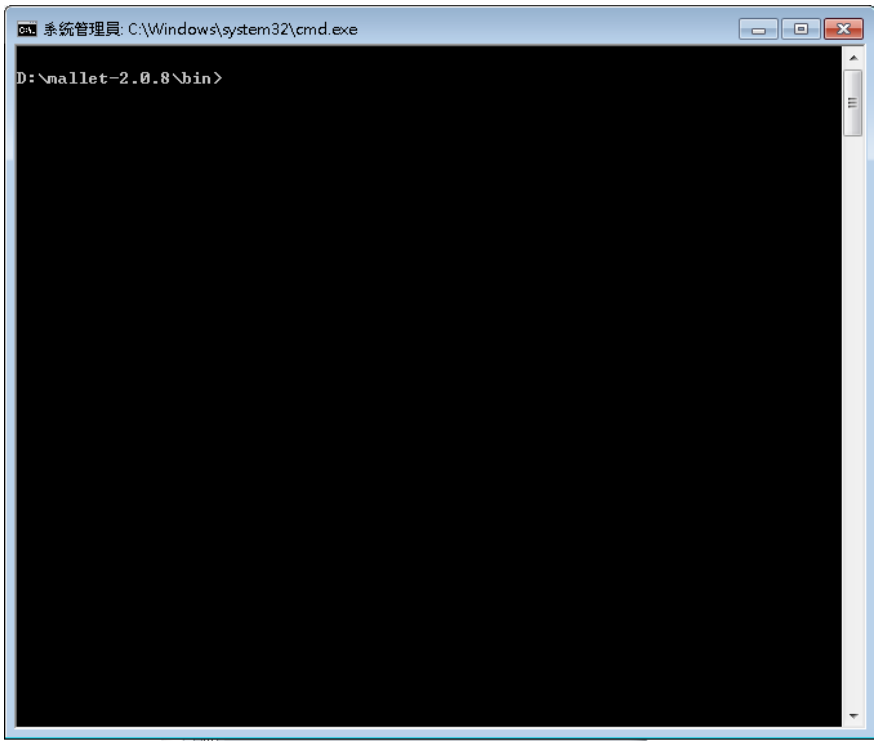


圖9　使用工具畫面

（二）可輸入「mallet」查詢指令參數。

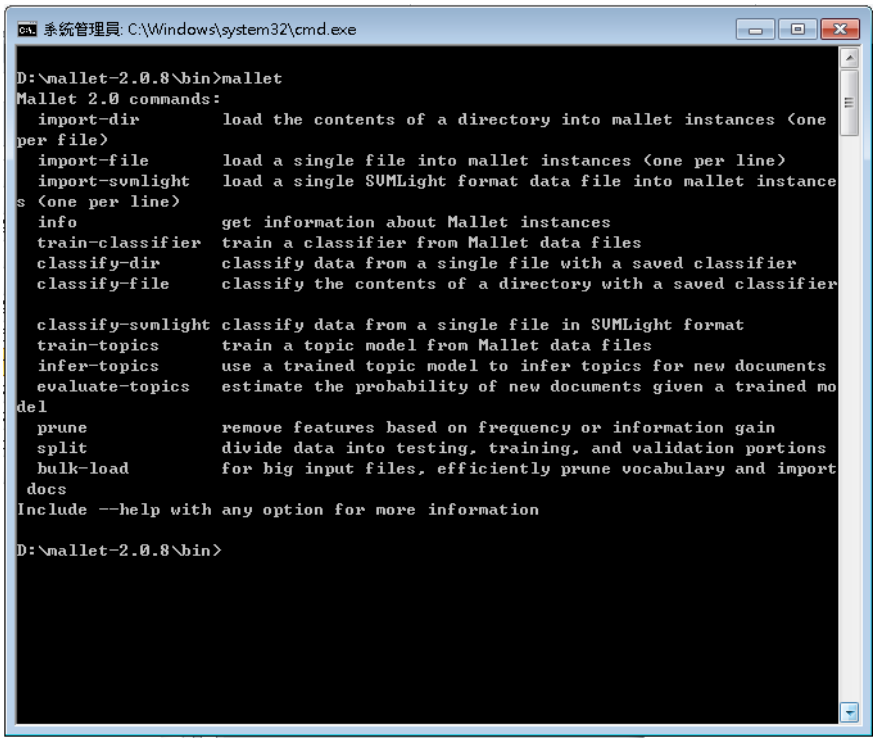


圖10　查詢參數畫面

三、操作示範

（一）導入數據操作

將資料導入 MALLET 格式有兩種主要方法，首先是匯入單一文件或匯入指定來源資料夾文件。

1、 匯入單一檔案，使用「import-file」指令。

(1)指令：D:\mallet-2.0.8\bin>Mallet import-file --input D:\mallet-2.0.8\sample-data\web\tw\tset.txt --output test.mallet。

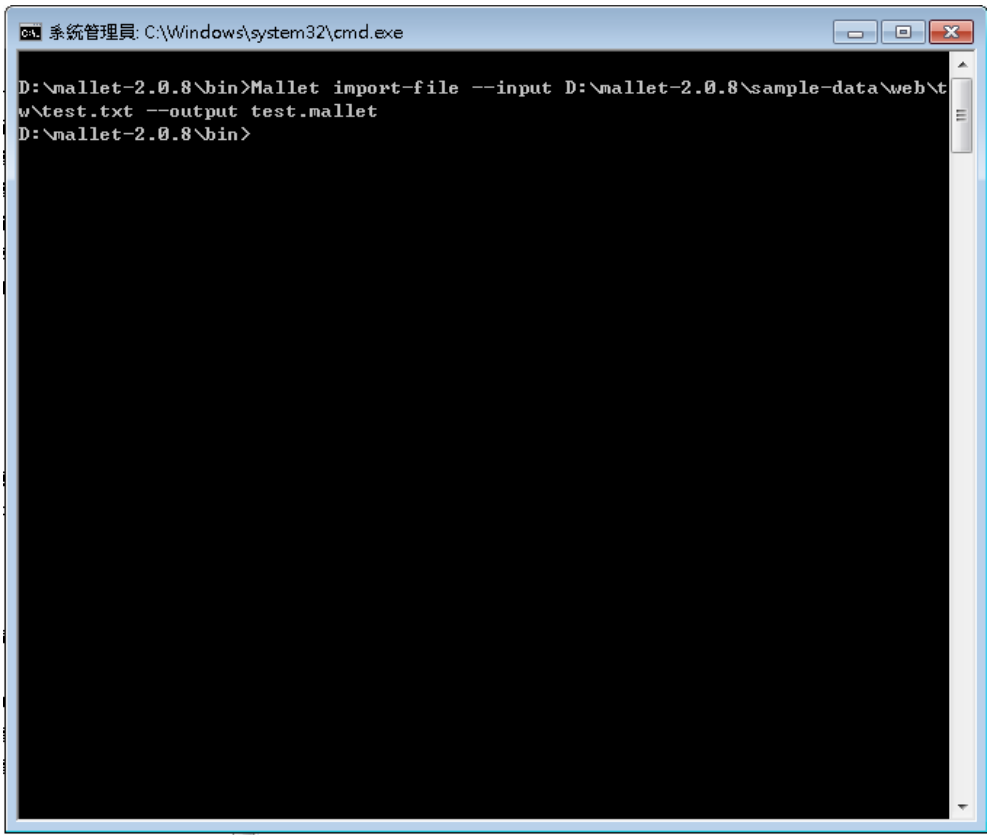


圖11　檔案匯入 Mallet 格式

(2)查看 Mallet 格式檔案，預設輸出資料夾為 bin 目錄。

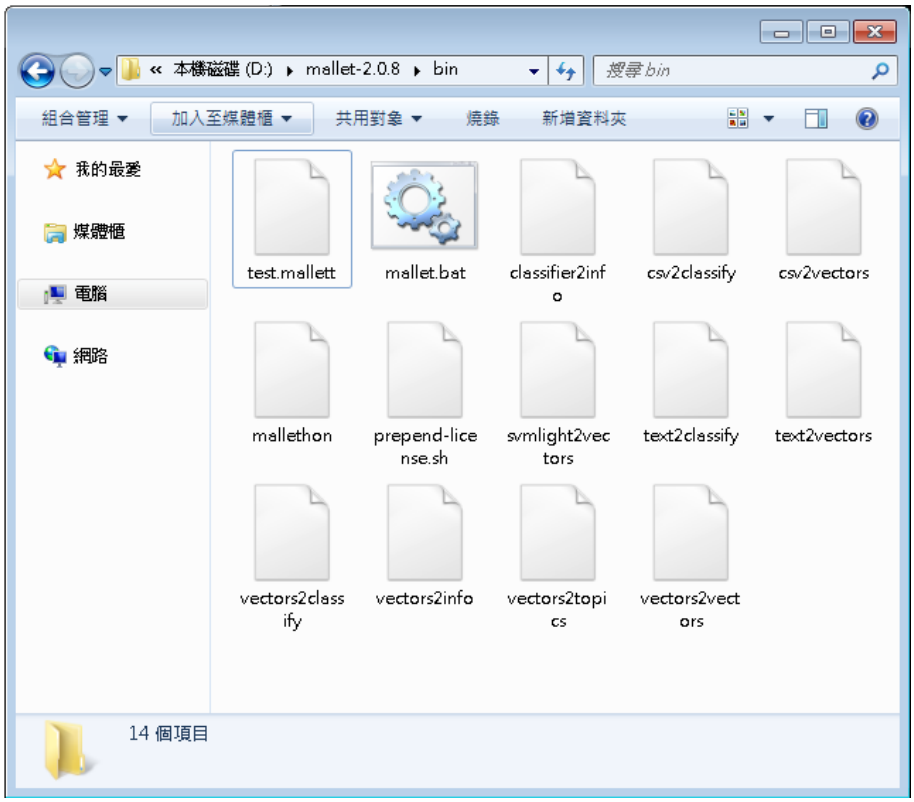


圖12　查看 Mallet 格式輸出位置

2、 匯入指定來源資料夾文件，使用「import-dir」指令。

(1)指令：D:\mallet-2.0.8\bin>路徑下輸入 Mallet import-dir --input D:\mallet-2.0.8\sample-data\web\en --output en.mallet 。

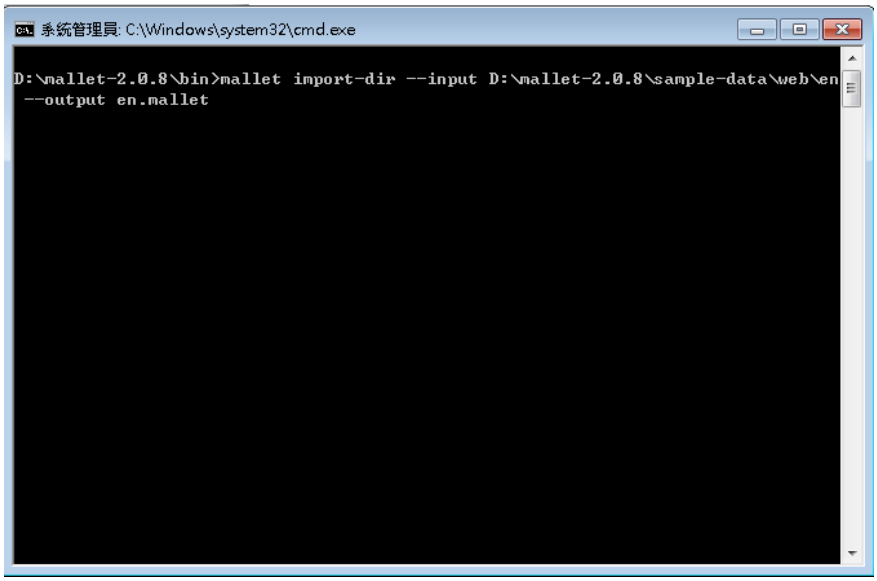


圖13　資料夾匯入 Mallet 格式

(1) 查看 Mallet 格式檔案，預設輸出資料夾為 bin 目錄。

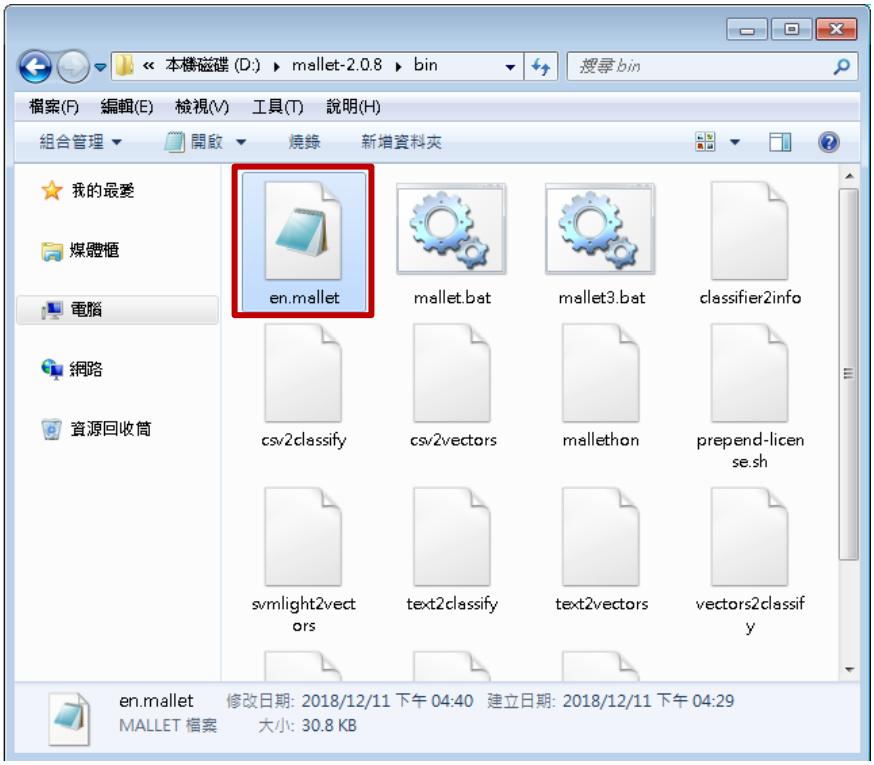


圖14　查看 Mallet 格式輸出位置

（二）文件分類操作

利用大量的訓練樣本訓練分類器，並利用測試樣本驗證分類器之性能，然後保存訓練好的分類器模型。當將新進文件輸入已訓練好的分類模型時，可輸出此文件所屬各個類別的概率。。

1、在 MALLET 數據文件上訓練分類器，使用「train-classifier」指令，演算法預設為 NaïveBayes，可使用其他分類演算法，加入參數「--trainer 驗算法名稱（如 MaxEnt、NaiveBayes、C4、DecisionTree 和其他)」。

(1)指令：「D:\mallet-2.0.8\bin>mallet train-classifier --input en.mallet --output-classifier my.classifier」。

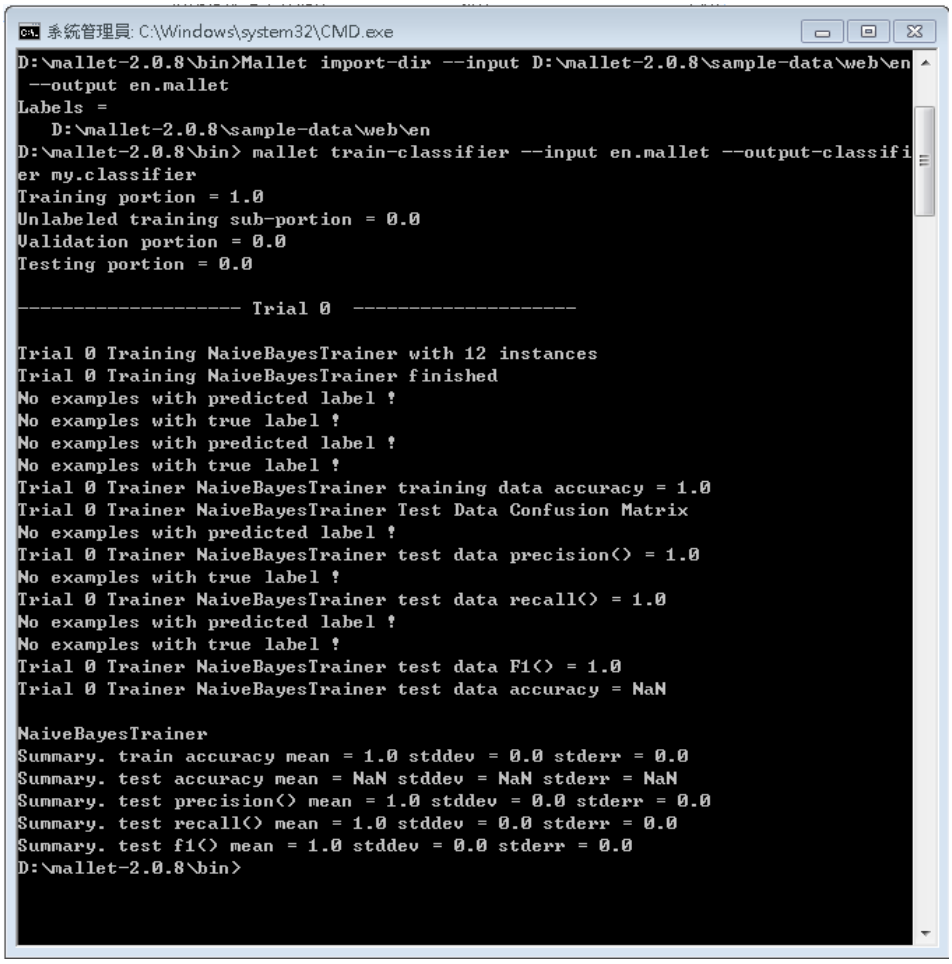


圖15　訓練分類器命令畫面

(2)查看產生之分類器格式檔案，預設輸出資料夾為 bin 目錄。

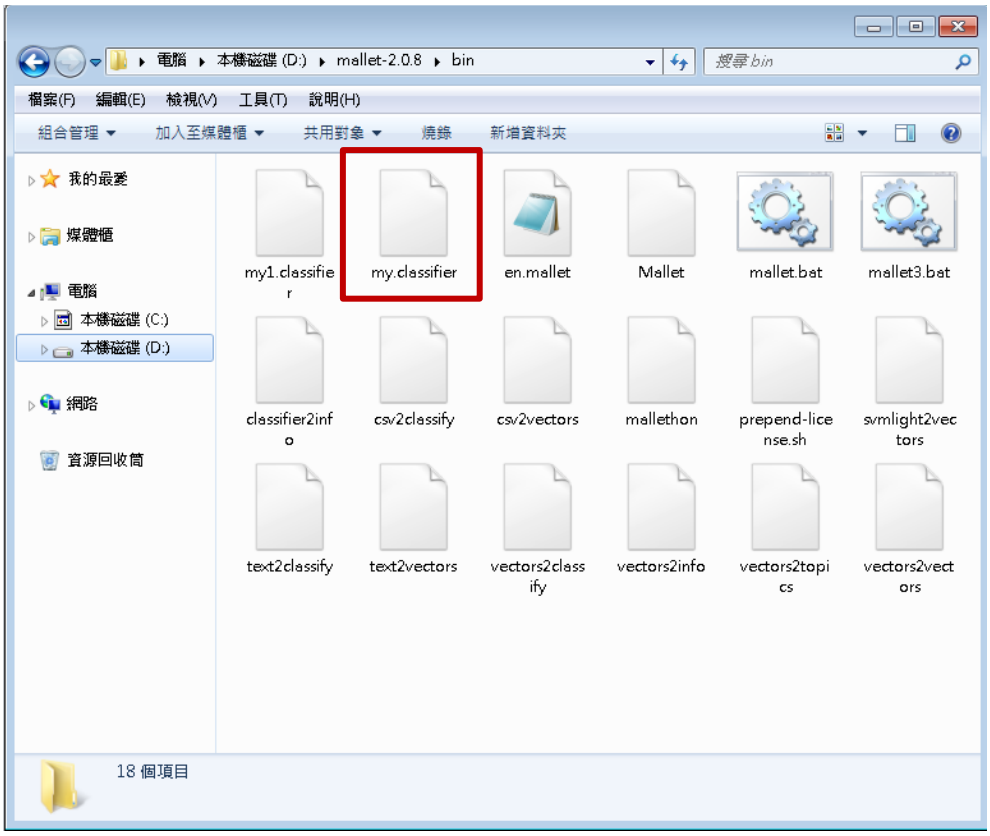


圖16　分類器格式產出畫面

2、訓練分類器，加入參數「--trainer（演算法）--training-portion（參數值）」。

(1)指令：「D:\mallet-2.0.8\bin> mallet train-classifier --input en.mallet --trainer NaiveBayes --training-portion 0.8 --output-classifier my1.classifier」。此命令隨機抽取 80%的訓練實例和剩下的 20%為測試實例，用於測試已訓練好的分類器的準確性。

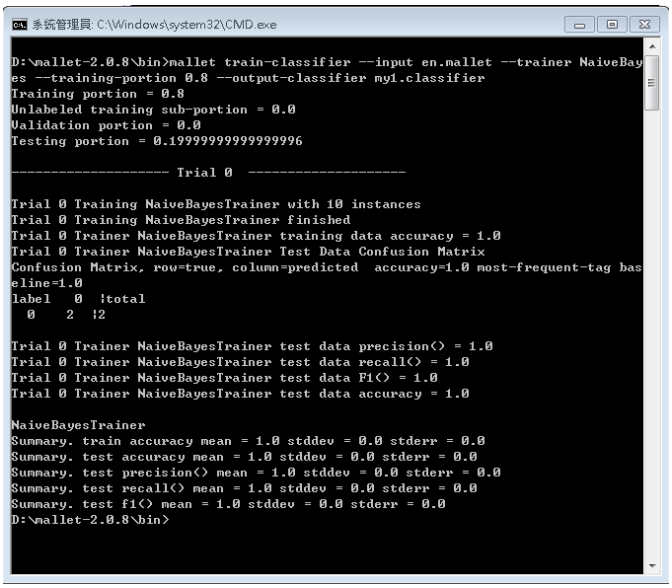


圖17　隨機訓練命令畫面

(2)查看產生之分類器格式檔案，預設輸出資料夾為 bin 目錄。

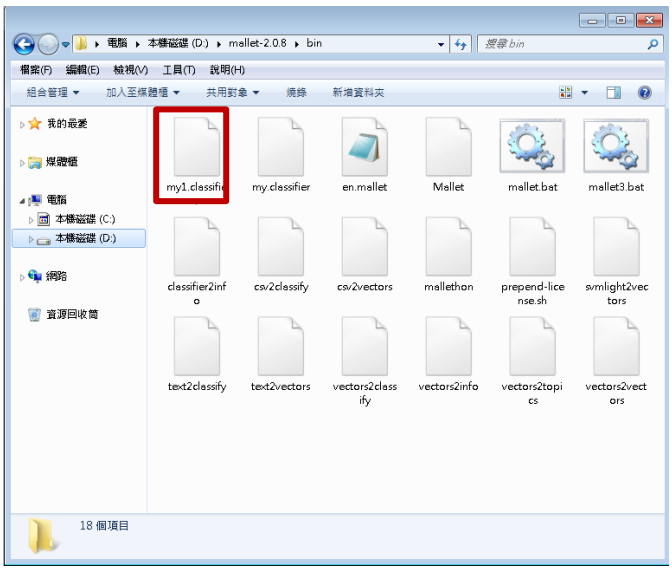


圖18　分類器格式產出畫面

3、 分割分類器，加入參數「--training-portion（參數值）--num-trials（分割數量）」。

(1)指令：「D:\mallet-2.0.8\bin> mallet train-classifier --input en.mallet --trainer NaiveBayes --training-portion 0.8 --num-trials 5 --output-classifier my2.classifier」。此命令拆分 5 個隨機抽取 80%的訓練實例和剩下的 20%為測試實例。

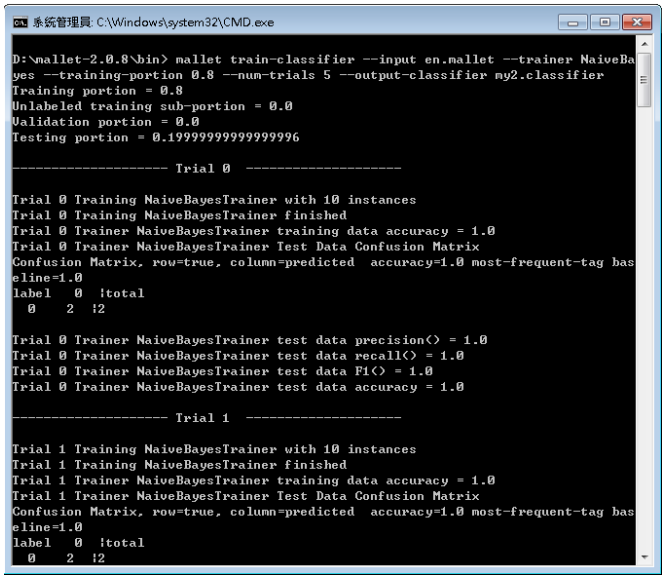


圖19　隨機拆分命令畫面

(2)查看分割產生之分類器格式檔案，預設輸出資料夾為 bin 目錄。

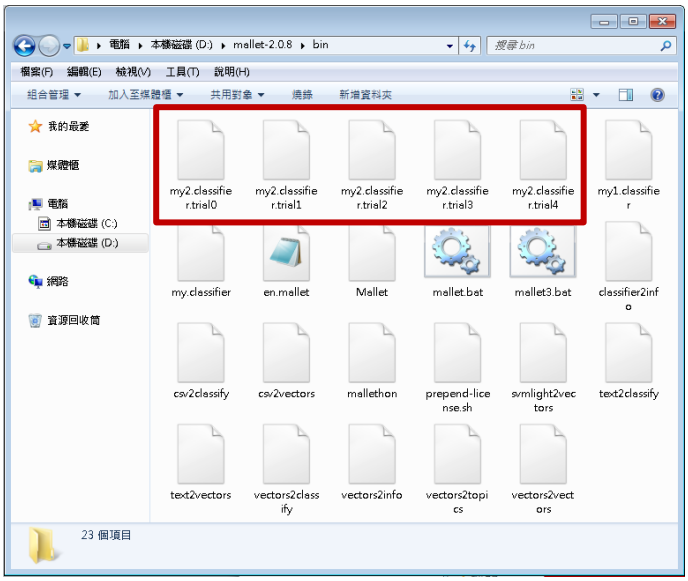


圖20　分類器格式產出畫面

4、利用分類器對未分類的資料進行分類。

(1)指令：「D:\mallet-2.0.8\bin>mallet classify-file --input D:\mallet-2.0.8\sample-data\web\en\hill.txt --output D:\test\hill2.txt --classifier my1.classifier」。此命令是對一未知類別文件進行分類。

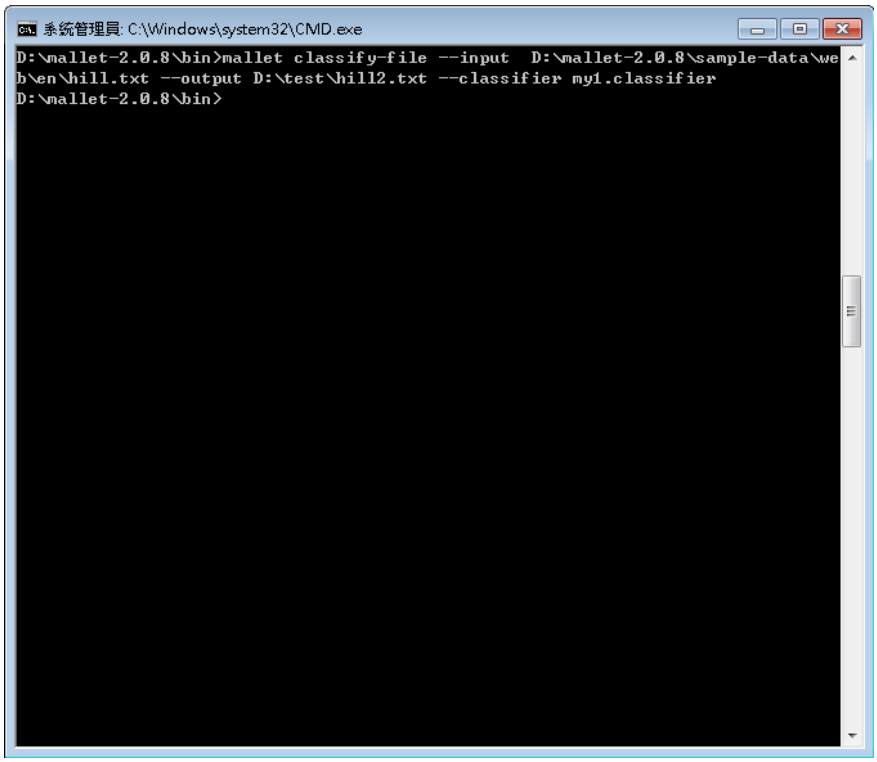


圖21　對一未知類別文件進行分類命令畫面

(2)查看分類結果。

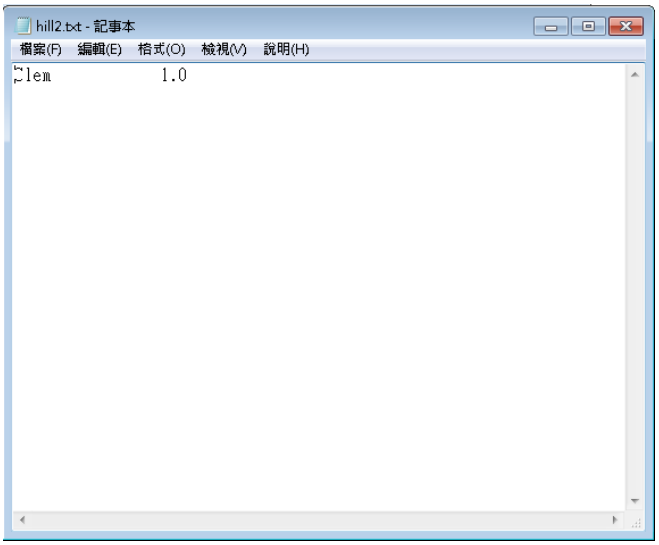


圖22　分類結果畫面

(3)指令：「D:\mallet-2.0.8\bin>mallet classify-dir --input D:\mallet-2.0.8\sample-data\web\en --output D:\test\hill2.txt --classifier my1.classifier」。此命令是對資料夾所有未知類別文件進行分類。

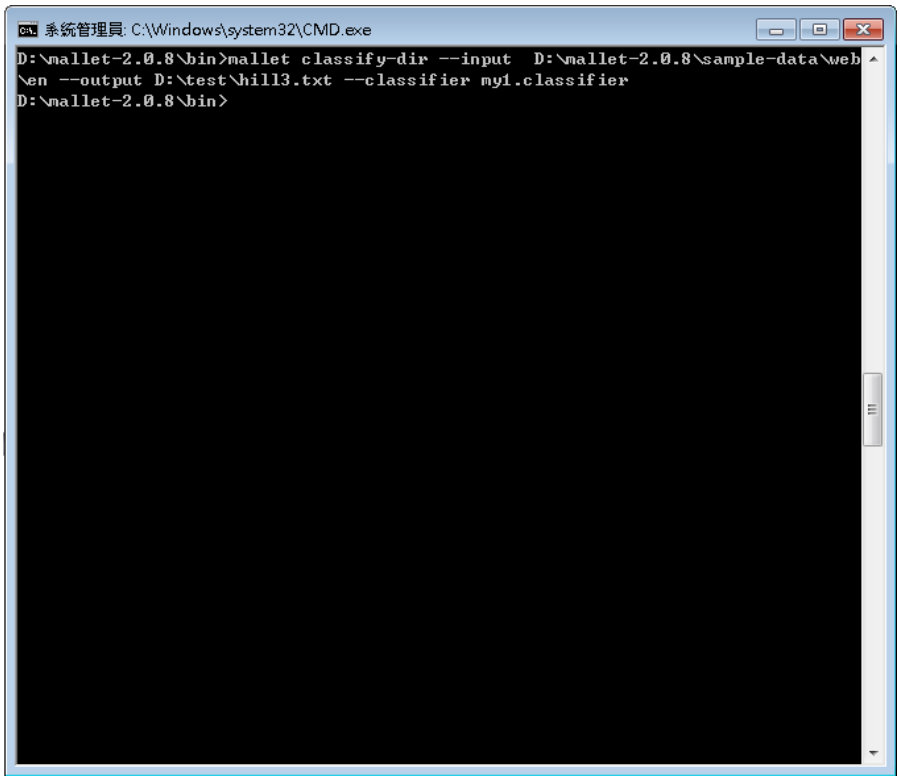


圖23　對資料夾所有未知類別文件進行分類命令畫面

(4)查看分類結果。

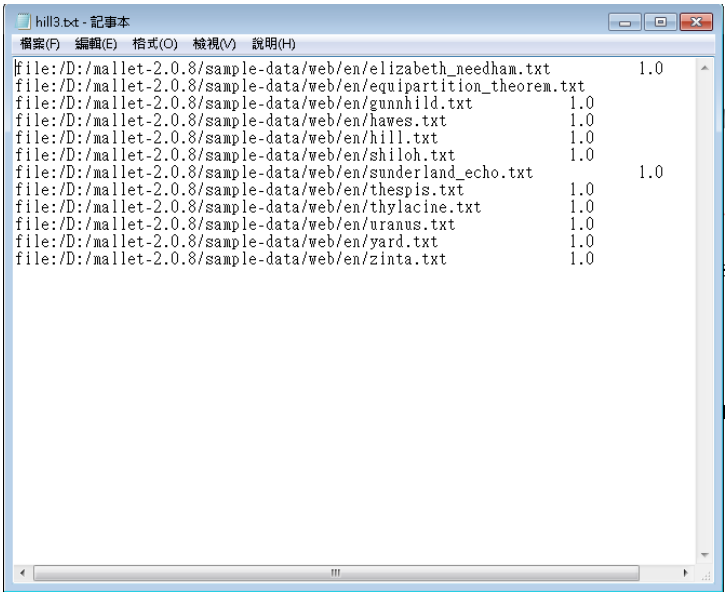


圖24　分類結果畫面

（三）建立主題模型 1

指的是一種從文件中抽取隱藏「主題」結構的技術方法，此方法用於分析大量的未標示或未知類別的文件，透過分析這些文件，可以得出一些「主題」。而每個「主題」會由一些經常出現在一起的詞所組成。

1、將文件轉置為 MALLET 格式，使用 「train-topics」命令並加入「--keep-sequence --remove-stopwords」參數。

(1)指令：「D:\mallet-2.0.8\bin>mallet import-dir --input D:\mallet-2.0.8\sample-data\web\en --output Topic.mallet --keep-sequence --remove-stopwords」，此命令是將來源資料「en 」目錄下的所有文件轉換為特徵序列，因建立主題模型的資料格式為特徵序列，非特徵向量，所以必須使用--keep-sequence 參數來限制轉換資料的格式，而--remove-stopwords 參數為移除停用詞。

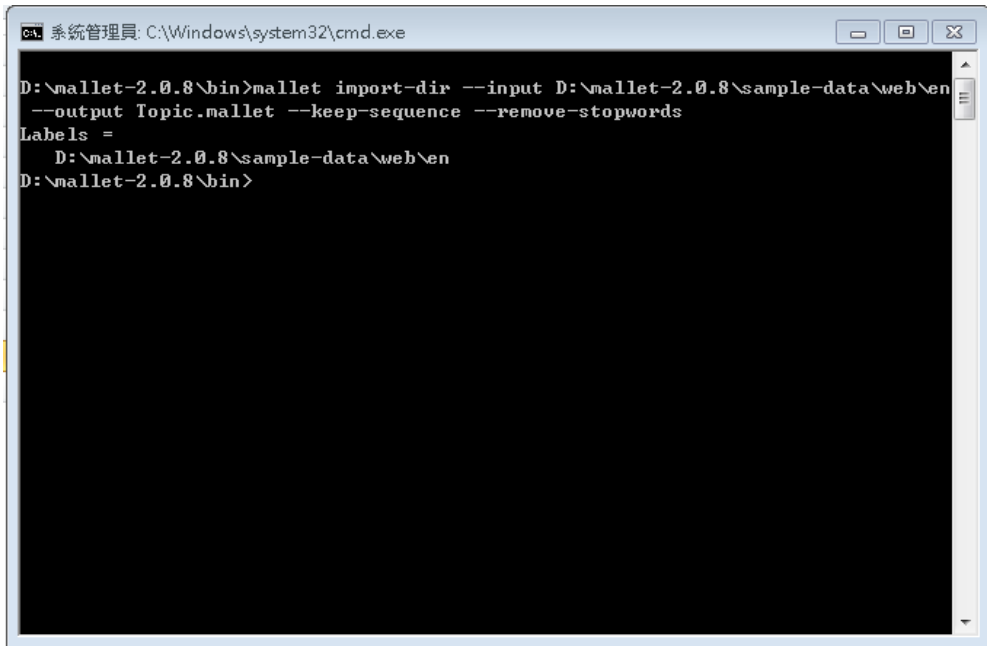


圖25　將資料轉置為 MALLET 格式命令畫面

(2)查看轉置 Mallet 格式檔案，預設輸出資料夾為 bin 目錄。

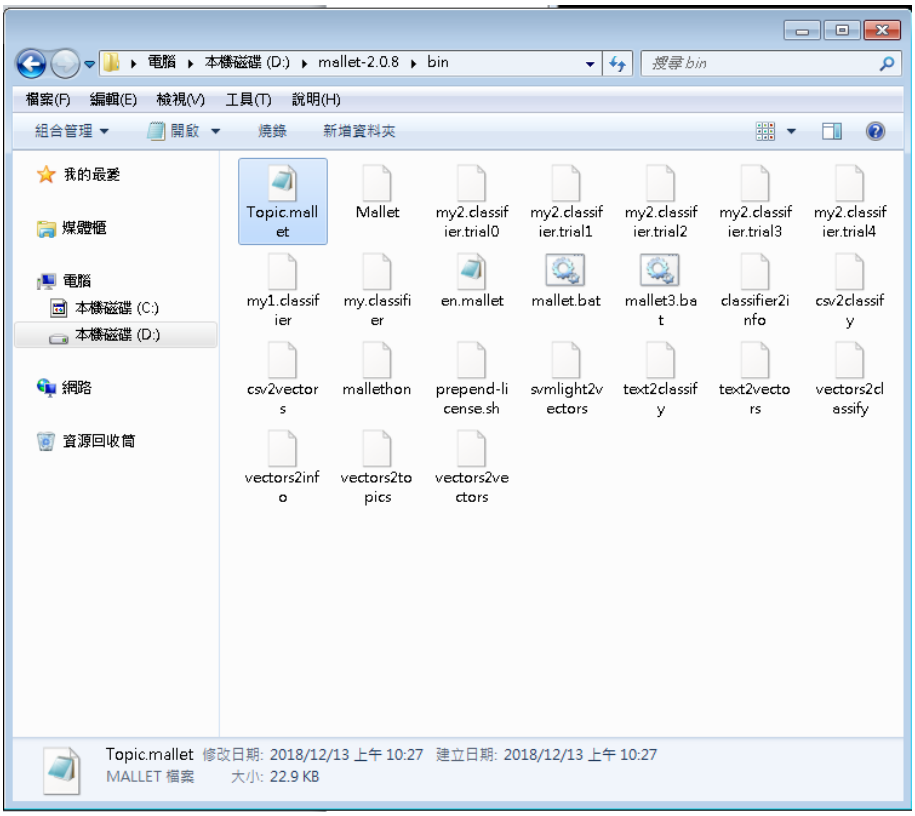


圖26　查看轉置 Mallet 格式輸出位置

2、建立主題模型

(1)指令：「D:\mallet-2.0.8\bin>mallet train-topics --input D:\mallet-2.0.8\bin\Topic.mallet --num-topics 20 --output-doc-topics Topic.txt --output-topic-keys Topic-key.txt --output-state Topic-state.gz --inferencer-filename Topic.inferencer」。此命令是將原先 MALLET 數據來建立主題模型。

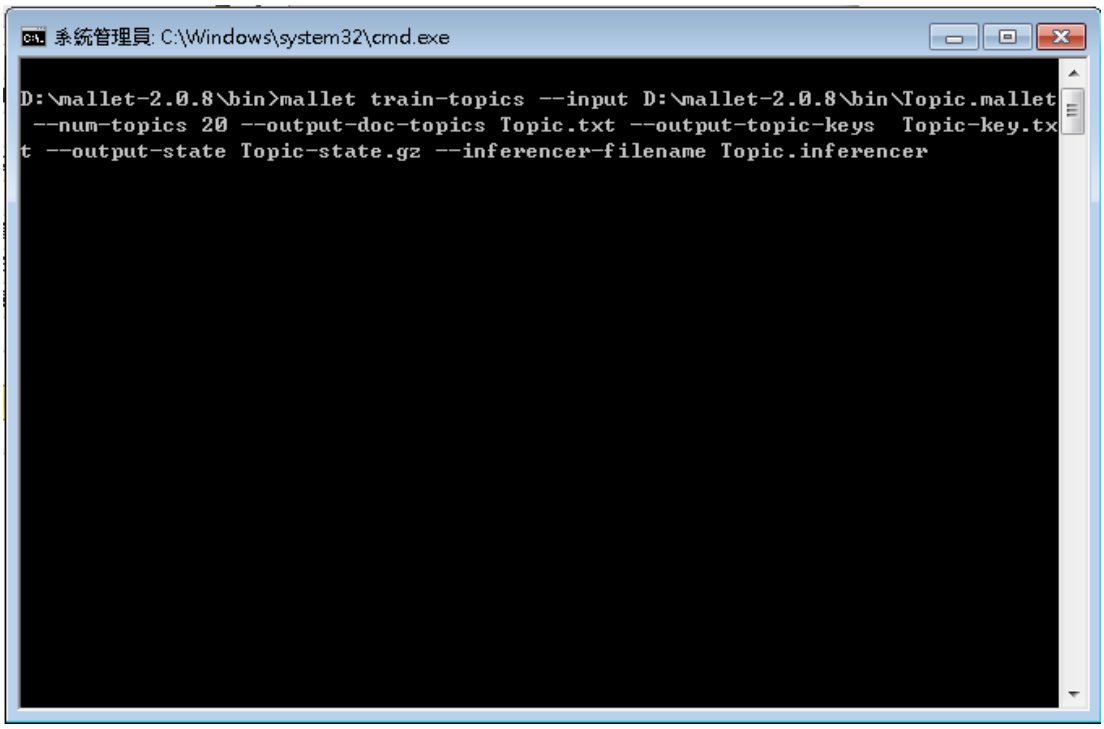


圖27　建立主題模型指令畫面

(2)參數說明：

A. 參數「--num-topics 20」

意思是限定主題個數為 20(預設為 10)，提供語言資料庫內容的大致概述。

B. 參數「--output-doc-topics」

此參數輸出主題矩陣儲存至文字檔。

C. 參數「--output-topic-keys 」

此參數輸出可用於檢查模型是否正常工作以及顯示模型的結果。

D. 參數「--output-state」

此參數可輸出一個壓縮文字檔，與輸出模型類似，包含語言資料庫中的單詞及其所分配主題組合，可使非 Java 架構的軟體可以輕鬆地解析和使用此文件格式。

E. 參數「--inferencer-filename」

將用已訓練好的模型創建一個主題推理工具。

(3)查看轉出檔案，預設輸出資料夾為 bin 目錄。

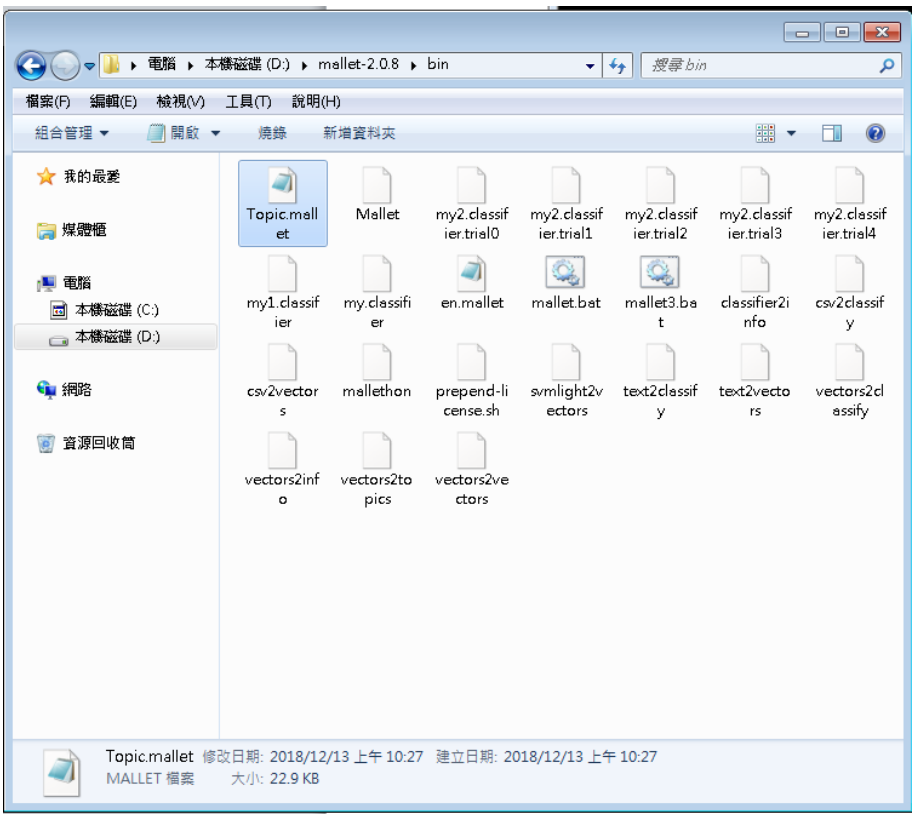
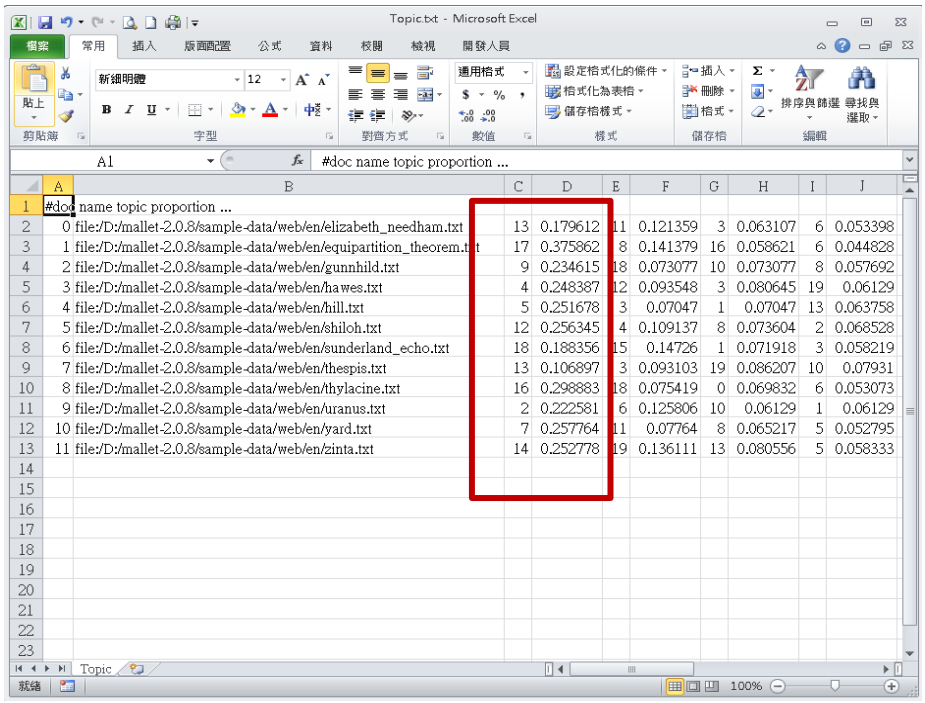


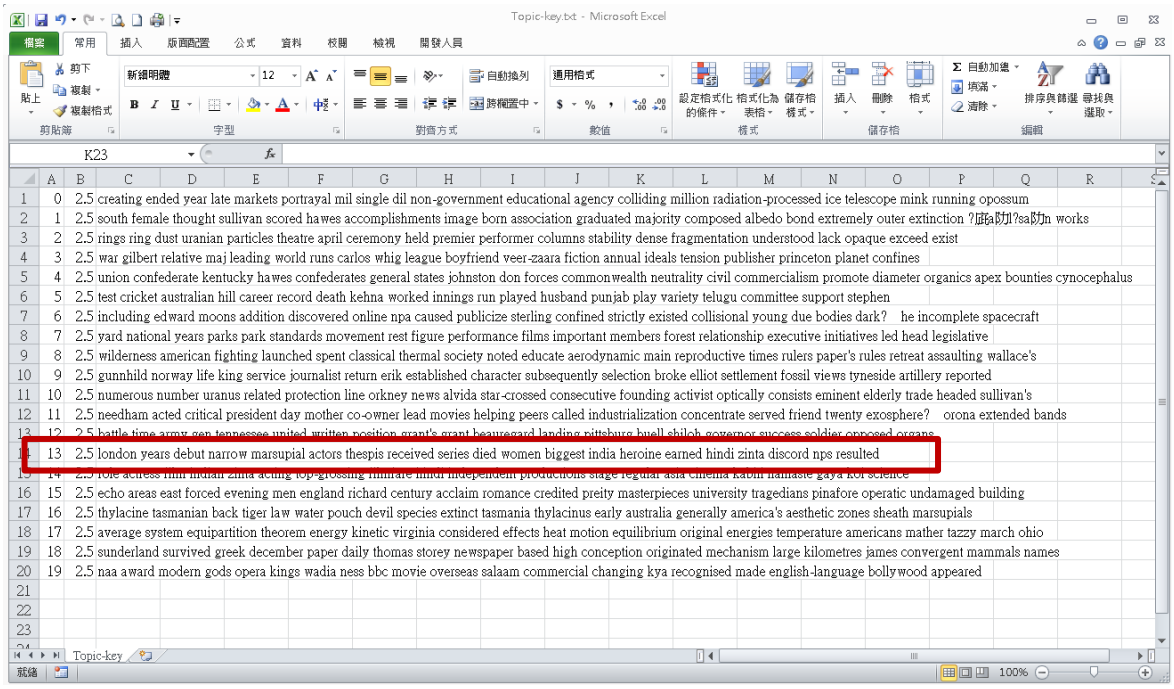
圖28　查看轉出檔案畫面

(4)可使用 Excel 開啟舊檔方式，開啟主題模型文字檔，來源資料共有 12 個文件，而在下圖 Excel 表格內欄位(紅框處)表示所創建的第 13 個主題組合(可參考 Topic-state 檔，圖 30)在該文件出現最多之單詞機率(可參考 Topic-key.txt，圖 31)。



#doc	name	topic	proportion ...								
0	file:/D:/mallet-2.0.8/sample-data/web/en/elizabeth_needham.txt	13	0.179612	11	0.121359	3	0.063107	6	0.053398
1	file:/D:/mallet-2.0.8/sample-data/web/en/equipartition_theorem.txt	17	0.375862	8	0.141379	16	0.058621	6	0.044828
2	file:/D:/mallet-2.0.8/sample-data/web/en/gunnhild.txt	9	0.234615	18	0.073077	10	0.073077	8	0.057692
3	file:/D:/mallet-2.0.8/sample-data/web/en/hawes.txt	4	0.248387	12	0.093548	3	0.080645	19	0.06129
4	file:/D:/mallet-2.0.8/sample-data/web/en/hill.txt	5	0.251678	3	0.07047	1	0.07047	13	0.063758
5	file:/D:/mallet-2.0.8/sample-data/web/en/shiloh.txt	12	0.256345	4	0.109137	8	0.073604	2	0.068528
6	file:/D:/mallet-2.0.8/sample-data/web/en/sunderland_echo.txt	18	0.188356	15	0.14726	1	0.071918	3	0.058219
7	file:/D:/mallet-2.0.8/sample-data/web/en/thespis.txt	13	0.106897	3	0.093103	19	0.086207	10	0.07931
8	file:/D:/mallet-2.0.8/sample-data/web/en/thylacine.txt	16	0.298883	18	0.075419	0	0.069832	6	0.053073
9	file:/D:/mallet-2.0.8/sample-data/web/en/uranus.txt	2	0.222581	6	0.125806	10	0.06129	1	0.06129
10	file:/D:/mallet-2.0.8/sample-data/web/en/yard.txt	7	0.257764	11	0.07764	8	0.065217	5	0.052795
11	file:/D:/mallet-2.0.8/sample-data/web/en/zinta.txt	14	0.252778	19	0.136111	13	0.080556	5	0.058333

圖29　主題模型文字檔畫面



0	2.5	creating ended year late markets portrayal mil single dil non-government educational agency colliding million radiation-processed ice telescope mink running opossum
1	2.5	south female thought sullivan scored hawes accomplishments image born association graduated majority composed albedo bond extremely outer extinction ?庇a阝加?sa阝加n works
2	2.5	rings ring dust uranian particles theatre april ceremony held premier performer columns stability dense fragmentation understood lack opaque exceed exist
3	2.5	war gilbert relative maj leading world runs carlos whig league boyfriend veer-zaara fiction annual ideals tension publisher princeton planet confines
4	2.5	union confederate kentucky hawes confederates general states johnston don forces commonwealth neutrality civil commercialism promote diameter organics apex bounties cynocephalus
5	2.5	test cricket australian hill career record death kehna worked innings run played husband punjab play variety telugu committee support stephen
6	2.5	including edward moons addition discovered online npa caused publicize sterling confined strictly existed collisional young due bodies dark?　he incomplete spacecraft
7	2.5	yard national years parks park standards movement rest figure performance films important members forest relationship executive initiatives led head legislative
8	2.5	wilderness american fighting launched spent classical thermal society noted educate aerodynamic main reproductive times rulers paper's rules retreat assaulting wallace's
9	2.5	gunnhild norway life king service journalist return erik established character subsequently selection broke elliot settlement fossil views tyneside artillery reported
10	2.5	numerous number uranus related protection line orkney news alvida star-crossed consecutive founding activist optically consists eminent elderly trade headed sullivan's
11	2.5	needham acted critical president day mother co-owner lead movies helping peers called industrialization concentrate served friend twenty exosphere?　orona extended bands
12	2.5	battle time army gen tennessee united written position grant's grant beauregard landing pittsburg buell shiloh governor success soldier opposed organs
13	2.5	london years debut narrow marsupial actors thespis received series died women biggest india heroine earned hindi zinta discord nps resulted
14	2.5	role actress film indian zinta acting top-grossing filmfare hindi independent productions stage regular asia cinema kabhi hamaste gaya kol science
15	2.5	echo areas east forced evening men england richard century acclaim romance credited preity masterpieces university tragedians pinafore operatic undamaged building
16	2.5	thylacine tasmanian back tiger law water pouch devil species extinct tasmania thylacinus early australia generally america's aesthetic zones sheath marsupials
17	2.5	average system equipartition theorem energy kinetic virginia considered effects heat motion equilibrium original energies temperature americans mather tazzy march ohio
18	2.5	sunderland survived greek december paper daily thomas storey newspaper based high conception originated mechanism large kilometres james convergent mammals names
19	2.5	naa award modern gods opera kings wadia ness bbc movie overseas salaam commercial changing kya recognised made english-language bollywood appeared

圖30　主題組合畫面

```
#doc source pos typeindex type topic
#alpha : 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5
#beta : 0.01
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 0 0 elizabeth 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 1 1 needham 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 2 2 died 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 3 3 mother 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 4 1 needham 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 5 4 english 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 6 5 procuress 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 7 6 brothel-keeper 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 8 7 th-century 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 9 8 london 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 10 9 identified 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 11 10 bawd 6
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 12 11 greeting 3
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 13 12 moll 1
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 14 13 hackabout 6
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 15 14 plate 19
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 16 15 william 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 17 16 hogarth's 0
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 18 17 series 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 19 18 satirical 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 20 19 etchings 3
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 21 20 harlot's 12
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 22 21 progress 5
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 23 1 needham 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 24 22 notorious 15
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 25 3 london 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 26 23 time 12
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 27 24 recorded 15
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 28 25 life 9
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 29 26 genuine 1
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 30 27 portraits 5
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 31 28 survive 10
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 32 29 house 10
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 33 30 exclusive 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 34 3 london 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 35 31 customers 3
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 36 32 highest 4
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 37 33 strata 5
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 38 34 fashionable 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 39 35 society 8
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 40 36 eventually 7
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 41 37 crossed 6
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 42 38 moral 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 43 39 reformers 2
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 44 40 day 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 45 2 died 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 46 41 result 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 47 42 severe 3
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 48 43 treatment 11
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 49 44 received 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 50 45 sentenced 13
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 51 46 stand 2
0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 52 47 pillory 4
```

圖31　單詞統計畫面

（四）建立主題模型 2

使用本局電子檔案保存實驗室英文版網站之環境介紹來建立主題模型 2。

1、將原始文件轉置為 MALLET 格式。

(1)將原始文件存放於 PEARL 資料夾內。

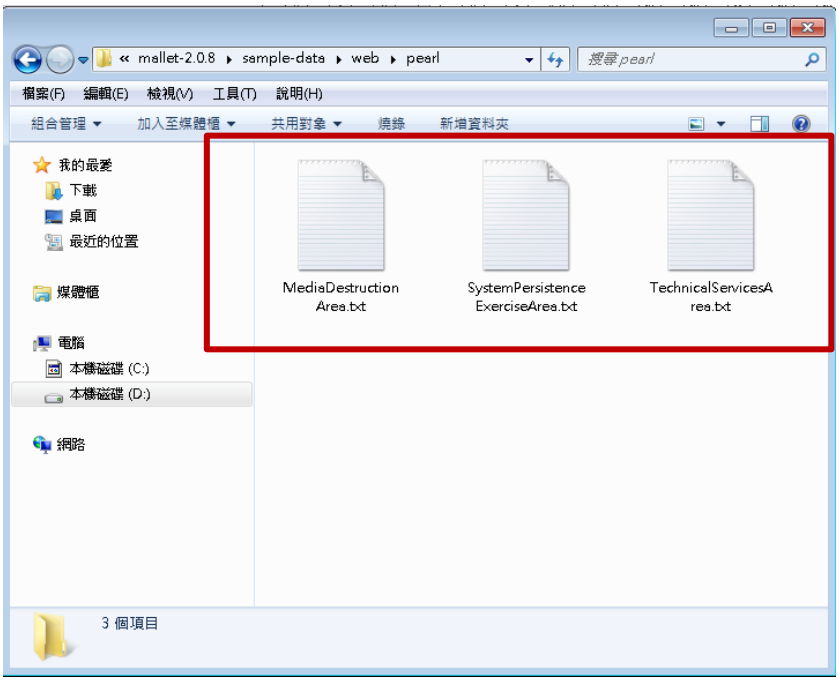


圖32　準備原始文件畫面

(2)將原始文件導入 MALLET 格式。

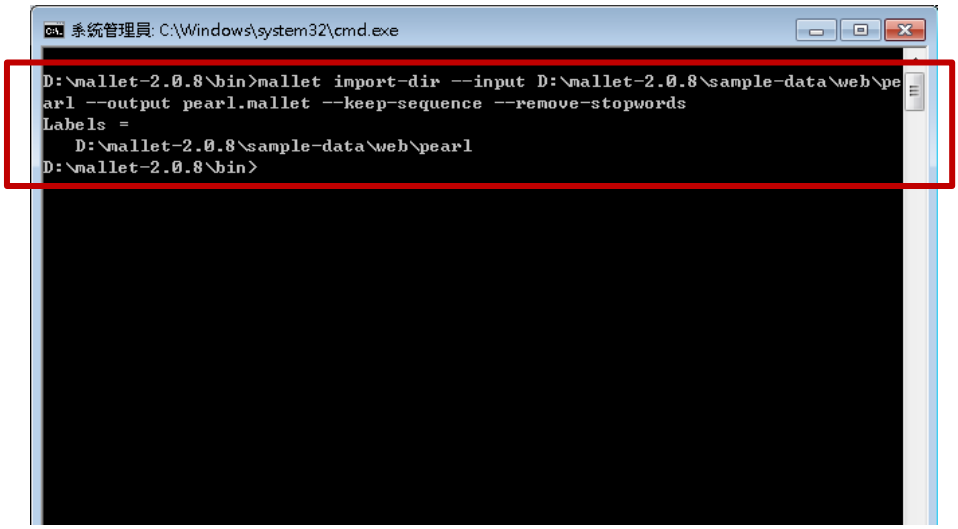


圖33　導入 MALLET 格式畫面

(3)建立主題模型。

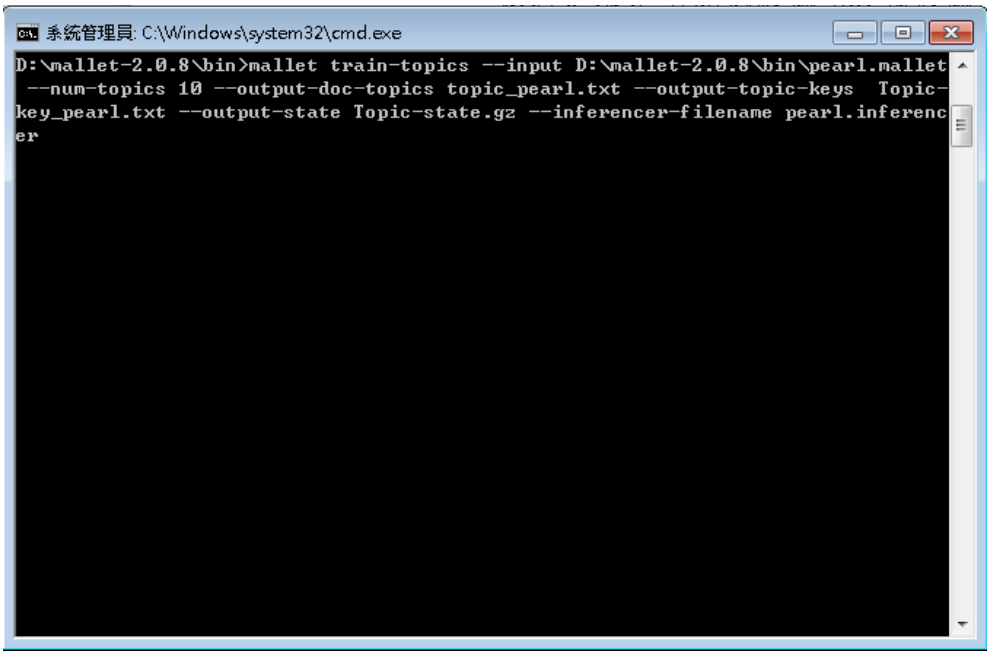


圖34　建立主題模型指令畫面

(4)查看轉出檔案，預設輸出資料夾為 bin 目錄，成功產生相關參數所產出報告。

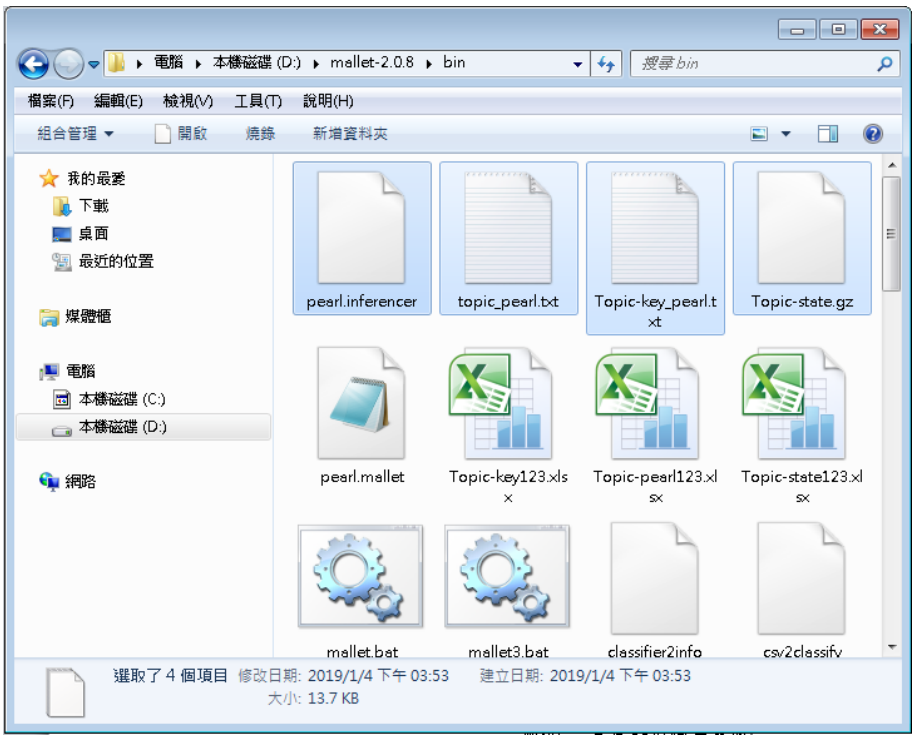


圖35　查看轉出檔案畫面

(5)解壓縮「Topic-state.gz」檔案。

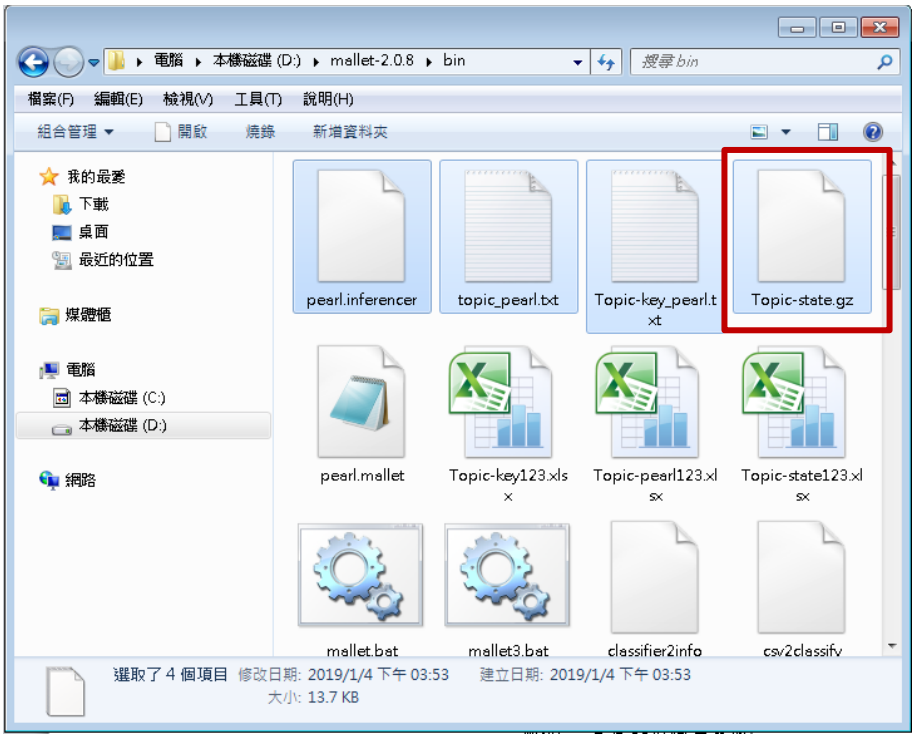


圖36　解壓縮畫面

(6)分別查看文件內容，使用 Excel 開啟舊檔方式開啟，來源資料共有 3 個文件，共創建 10 個主題組合(參考 Topic-key_pearl 檔，圖 38)及各文件單詞機率分析(圖 37)，其中文件 0 機率最高為主題 4 之組合，而出現最多之單詞為「media」、文件 1 機率最高為主題 6 之組合，出現最多之單詞為「software」、文件 2 機率最高為主題 2 之組合，出現最多之單詞為「migration」，以上數值統計是參照 Topic-state 檔案內之單詞數量(圖 39)。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
#doc name topic proportion ...																					
0	file:/D:/mallet-2.0.8/sample-data/web/pearl/MediaDestructionArea.txt	4	0.153333	8	0.146667	3	0.14	1	0.14	7	0.113333	9	0.08	0	0.073333	6	0.053333	5	0.053333	2	0.046667
1	file:/D:/mallet-2.0.8/sample-data/web/pearl/SystemPersistenceExerciseArea.txt	6	0.222656	5	0.195313	9	0.136719	7	0.101563	4	0.097656	3	0.078125	8	0.0625	0	0.050781	2	0.027344	1	0.027344
2	file:/D:/mallet-2.0.8/sample-data/web/pearl/TechnicalServicesArea.txt	2	0.309859	0	0.192488	8	0.131455	1	0.084507	4	0.070423	3	0.070423	6	0.042254	9	0.037559	5	0.032864	7	0.028169

圖37　主題模型文字檔畫面

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
0	5	records	outputs	include	mpeg	developed	doc	record	technology	certificates	refers	dvd	includes	mechanism	vinyl/tape/f	jpge	encapsulati	replicated	integrity	institutions	compliant
1	5	format	hard	usability	software	tapes	place	openoffice	ppt	docx	proprietary	png	multimedia	suite	repair	break	re-formattir	completely	actual	destroying	cassette
2	5	migration	pdf/a	formats	tiff	data	jpeg	issued	component	shared	migrated	quality	verification	risk	jpeg/tiff/pdf	pdfcreator	wav	gif	university	taipei	national
3	5	optical	disks	disk	dvd	windows	tools	includes	scope	ffmpeg	xls	postscript	incorporates	library	jvc	kiosk	solaris	shredder	destroy	provided	recovery
4	5	storage	operating	methods	preserving	files	pdf	steps	destroyed	ensure	validation	forms	odt	wmv	stored	cabinet	terms	centos	ibm	divided	ertsc's
5	5	software	preserved	server	preserve	hardware	archive	correspondence	directory	management	nec	types	preserves	network	box	reader	cassettes	house	viewing	equipment	related
6	5	preservation	floppy	hardware	inch	servers	earthquake	produced	odt	signify	encapsulate	knowledge	email	online	creation	testing	workstation	houses	rolls	museum	effectively
7	5	disc	system	ertsc	applications	physical	discs	removed	copying	bank	e-correspondence	red	unix	total	consists	including	costs	ensuring	fully	pristine	high-frequency
8	5	media	electronic	file	destruction	pears	result	losses	microfilm	vhs/beta/betacam	ghostscript	eml	features	png	migrating	overland	lto	sun	characterist	removing	professional
9	5	area	tape	system	drive	introduced	systems	magnetic	prone	wdl	archives	mini-size	range	diskette	virtual	aix	hat	needed	build	large	high

圖38　主題組合畫面

#doc	source	pos	typeindex	type	topic
#alpha	:	5	5	5	5
#beta	:	0.01			
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	0	0	destruction	8
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	1	1	files	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	2	2	required	9
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	3	3	number	5
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	4	4	years	7
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	5	5	destroyed	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	6	6	added	7
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	7	7	preserving	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	8	8	approved	7
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	9	9	legal	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	10	10	procedures	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	11	1	files	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	12	11	removed	7
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	13	5	destroyed	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	14	12	methods	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	15	13	deemed	6
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	16	14	scope	3
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	17	0	destruction	8
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	18	15	includes	3
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	19	16	cassette	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	20	17	tapes	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	21	18	floppy	6
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	22	19	disks	3
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	23	20	optical	3
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	24	21	discs	7
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	25	22	hard	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	26	19	disks	3
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	27	23	magnetic	9
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	28	17	tapes	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	29	24	steps	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	30	25	destroying	1
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	31	26	storage	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	32	27	media	8
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	33	12	methods	4
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	34	28	include	0
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	35	29	physical	7
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	36	0	destruction	8
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	37	22	hard	1

圖39　單詞統計畫面

參、結論

為了讓機器能判斷二個句子是否具關聯性，只能利用文章內的重複詞語來做判斷，與主題關係越密切的詞語，它的出現機率越大，反之則越小。

自然語言學習工具種類繁多，MALLET 只是其中的一種，透過 MALLET 工具可衡量文章之間的語義相似性，任何語言只要能夠對它進行分詞，就可以進行訓練，並得到它的主題模型。。