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	_	1	_
貳	•	軟體資訊	_	2	_
參	•	結論	_ !	28	_

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」∘

MALLET module learning for language taulik	for LanguagE Toolkit
Home Tutorial slides / <u>video</u> Download API Quick Start Sponsors Mailing-List About Importing Data Classification Sequence Tagging Topic Modeling Qutimization Graphical Models MALLET is open source software [License]. For research use, please remember to <u>ate MALLET</u>	Current release: The following packaged release of MALLET 2.0 is available: mailet 2.0.8.tar.g: mailet 2.0.8.zig: Mindows installation: After unzipping MALLET, set the environment variable %MALLET_HOME% to point to the MALLET directory. In all command line examples, substitute balantiat for bala/mallet. Development release: To download the most current variation of MALLET 2.0, use our public GitHub repository: gt clone https://gtHub.com/mimo/Mallet.gtt from the command prompt to get the Mallet package. To build a Mallet 2.0 development release, you must have the <u>Apache ant</u> build tool installed. From the command prompt, first change to the mallet directory, and then type ant If ant finishes with "BUILD SUCCESSFUL", Mallet is now ready to use. If you would like to deploy Mallet as part of a larger application, it is helpful to create a single ".jar" file that contains all of the compiled code. Once you have compiled the individual Mallet class files, use the command: art jar The process will create a file "mallet.jar" in the "dist" directory within Mallet. Older release: MALLET version 0.4 is available for <u>download</u> , but is not being actively maintained. This release includes classes in the package "edu.umass.cs.mallet.base", while MALLET 2.0 contains classes in the package "edu.umass.cs.mallet.base", while MALLET 2.0 contains classes in the mailet-2.0-FCG.tar.og (notes) mailet-2.0-FCG.tar.og (notes) mailet-2.0-FCG.t
	回1 MALLE1 E 力 網 站 卜 載 貝 囬

(二)下載完成後,解壓縮「mallet-2.0.8.zip」。



mallet-2.0.8.zip 圖2 解壓縮 mallet-2.0.8.zip

(三) 解壓縮後畫面。

- □ ►							
組合管理 ▼ 加入至媒體櫃	▼ 共用對象 ▼	燒錄	新増資料夾	• *			
🗙 我的最愛							
🍃 煤體櫃	bin	class	dist	lib	sample-da		
🖳 電腦					ta		
👊 網路							
🗑 資源回收筒	src Makefile	stoplists	test README. md	build.xml	LICENSE		
13 個項目							

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。

					- • •
😋 🗢 🗣 🕌 « 10704 🖡	mallet-2.0.8 🕨 b	pin	▼ 4 / 授尋	tbin	Q
組合管理 ▼ 加入至媒體	櫃 ▼ 共用對	象▼ 燒錄	新増資料夾		- 🗌 🔞
🗙 我的最愛		6			
🧊 煤體櫃	classifier2i nfo	csv2classif y	csv2vector s	mallet	mallet.bat
[■] 電腦 ■] 本機磁碟(C:)					
抽取式磁碟 (J:)	mallethon	prepend-li cense.sh	svmlight2v ectors	text2classif y	text2vecto rs
🧣 網路					
🗑 資源回收筒	vectors2cl assify	vectors2inf o	vectors2to pics	vectors2ve ctors	
14 個項目					

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。

福家(F) 編輯(E) 格式(O) 被視(V) 說明(H)	
jæcho off	~
rem This batch file serves as a wrapper for several rem MALLET command line tools.	
if no "D:\Users\930502\Desktop\mallet-2.0.8" = "" goto gotMalletHome	
echo MALLEI requires an environment variable MALLEI_HOME. goto :eof	
:gotMalletHome	
set MALLET_CLASSPA ^T =D:\Users\930502\Desktop\mallet-2.0.8\ set MALLET_HOWE-D: Tears\920502\Desktop\mallet-2.0.8 set MALLET_HOWE-D: Tears\920502\Desktop\mallet-2.0.8 set MALLET_ENCODING=UTF-8	s.jar ≡
set CMD≠%1 shift	
set CLASS= if "%UDW"="import-dir" set CLASS=cc.mallet.classify.tui.Text2Vectors if "%UDW"="import-swnlight" set CLASS=cc.mallet.classify.tui.Sv2Vectors if "%UDW"="info" set LLASS=cc.mallet.classify.tui.Sv2Vectors if "%UDW"="info" set LLASS=cc.mallet.classify.tui.Vectors2Classify if "%UDW"="classify-file" set CLASS=cc.mallet.classify.tui.Vectors2Classify if "%UDW"="classify-file" set LLASS=cc.mallet.classify.tui.Sv2Tassify if "%UDW"="classify-file" set CLASS=cc.mallet.classify.tui.Sv2Tassify if "%UDW"="classify-file" set CLASS=cc.mallet.classify.tui.Sv2Tassify if "%UDW"="classify-swlight" set CLASS=cc.mallet.classify.tui.Sv2Tassify if "%UDW"="rain-topics" set CLASS=cc.mallet.classify.tui.Sv2Tassify if "%UDW"="infer-topics" set CLASS=cc.mallet.topics.tui.TopicTrainer if "%UDW"="infer-topics" set CLASS=cc.mallet.topics.tui.FeTDpics if "%UDW"="infer-topics" set CLASS=cc.mallet.topics.tui.FeTDpics if "%UDW"="infer-topics" set CLASS=cc.mallet.topics.tui.FeTDpics if "%UDW"="infer-topics" set CLASS=cc.mallet.topics.tui.Sv2Text2Vectors if "%UDW"="puble.load" set CLASS=cc.mallet.topics.tui.Sv2Text2Vectors if "%UDW"="puble.classify.tui.Vectors2Vectors if "%UDW"="buble.classify.tui.Vectors2Vectors if "%UDW"="buble.classify.tui.Vectors2Vectors if "%UDW"="buble.classify.tui.Vectors2Vectors if "%UDW"="buble.classify.tui.Vectors2Vectors if "%UDW"="buble.classify.tui.Vectors2Vectors if "%UDW"="buble.classify.tui.Vectors2Vectors if %UDW"="buble.classify.tui.Vectors2Vectors if %UDW"="buble.classify.tui.Vector	
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	
	-

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							
"%CMD%"=="import-symlight" set CLASS=cc.mallet.classify.tui.SymLight2Vectors							
if "%CMD%"=="info" set CLASS=cc.mallet.classify.tui.Vectors2Info	f "%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-symlight" set CLASS=cc.mallet.classify.tui.SymLight2Classify							
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 po	er 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							
pho classify-file classify the contents of a directory with a saved classifier							
classify-file classify the contents of a difectory with a saved classifier							
echo classify-svmlight classify data from a single file in SVMLight format							
echo classify-symlight classify data from a single file in SVMLight format echo train-topics train a topic model from Mallet data files							
echo classify-svmlight classify dia 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							
<pre>cond classify-sile classify the concents of a diffectory with a saved classifier echo classify-symlight 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 mode </pre>	lel						
<pre>cond classify-sile classify the contents of a diffectory with a saved classifier echo classify-symlight 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 mode echo prune remove features based on frequency or information gain</pre>	lel						
ccho classify-sile classify the contents of a diffectory with a saved classifier echo classify-symlight 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 mode echo prune remove features based on frequency or information gain echo split divide data into testing, training, and validation portions	lel						
construction classify-symlight classify che concents of a diffective with a saved classifier echo classify-symlight 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	lel docs						

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-symlight" set class="cc".mallet.classify.tui.SymLight2Vectors if "%CHD%"=="train-classifier" set class="cc".mallet.classify.tui.Vectors2Classify if "<mark>«CMD»</mark>"=="classify-file" set class="cc".mallet.classify.tui.Csv2Classify if "ODD" == "classify-dir" set class="cc".mallet.classify.tui.Text2Classify if "%CMD%"=="classify-sym" set class="cc".mallet.classify.tui.SymLight2Classify 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.HierarchicalLD&TUI 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-symlight 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-symlight" set class="cc".mallet.classify.tui.SymLight2Vectors
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-symlight 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 Includehelp 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。

系統管理員: C:\Windows\system32\cmd.exe	- • ×
	~
D:\mallet-2.0.8\bin>mallet	
Mallet 2.0 commands:	=
import-dir load the contents of a directory into mallet ins	tances (one
per file)	
import-file load a single file into mallet instances (one pe	r line)
import-symlight load a single SVMLight format data file into mal	let instance
s (one per line)	
info get information about mallet instances	
alaosifu-din a classifier from hallet used files	popifion
classify-uir classify uata from a single file with a saveu of a dimectomy with a save	dssirier d classifiem
	a c1a3311161
classifu-sumlight classifu data from a single file in SUMLight for	mat
train-topics train a topic model from Mallet data files	
infer-topics use a trained topic model to infer topics for ne	w documents
evaluate-topics estimate the probability of new documents given	a trained mo
del	
prune remove features based on frequency or informatio	n gain
split divide data into testing, training, and validati	on portions
bulk-load for big input files, efficiently prune vocabular	y and import
docs	
Include —help with any option for more information	
D: Mallet-2.0.8 \bin>	
	-

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

😋 🕞 🗢 📙 « 本機磁碟 (D:) 🕨 mallet-2.0.8 🕨 bin ▼ 🍫 搜尋bin Q 組合管理 🔻 加入至媒體櫃 ▼ 共用對象 ▼ 燒錄 新增資料夾 ■ N ▼ ? 숨 我的最愛 ј 煤體櫃 test.mallett mallet.bat classifier2inf csv2classify csv2vectors 📑 電腦 0 嗿 網路 mallethon prepend-lice svmlight2vec text2classify text2vectors nse.sh tors vectors2info vectors2class vectors2topi vectors2vect ify cs ors 14 個項目

(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。。

 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
myl.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。

🔾 🔍 🗢 📕 🕨 🛪	\$構磁碟(D:) ▶ m	allet-2.0.8 🕨 bir	1	- 4 ∳ ∄	要尋 bin	Q
榴案(F) 編輯(E) 檢視(∀)	工具(T) 說明(H	-1)				
組合管理 ▼ 加入至煤業	目櫃▼ 共用對	象▼ 燒錄	新増資料夾			• 🔟 🔞
▶ ★ 我的最愛 ▶ (□) 煤酸						
4 👰 電腦	my1.classifie	my.classifier	en.mallet	Mallet	mallet.bat	mallet3.bat
▶ 団 本機磁碟 (C:) ▶ 📻 本機磁碟 (D:)						
> 🗣 網路	classifier2inf o	csv2classify	csv2vectors	mallethon	prepend-lice nse.sh	svmlight2vec tors
👿 資源回收商						
	text2classify	text2vectors	vectors2class ify	vectors2info	vectors2topi cs	vectors2vect ors
18 個項目						

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。

					0			
榴窯(F) 編輯(E) 檢視(∨)	欄窯(F) 編輯(E) 檢視(∨) 工具(T) 說明(H)							
組合管理 ▼ 加入至媒體	豊櫃 ▼ 共用對	象▼ 燒錄	新増資料夾			- 🔟 🔞		
☆ 我的最愛								
🍃 煤壁櫃	my2.classifie	my2.classifie	my2.classifie	my2.classifie	my2.classifie	my1.classifie		
📮 電腦	r.trial0	r.trial1	r.trial2	r.trial3	r.trial4	r		
□ 本機磁碟 (C:) □ 本機磁碟 (D:)								
🙀 網路	my.classifier	en.mallet	Mallet	mallet.bat	mallet3.bat	classifier2inf o		
🗑 資源回收筒								
	csv2classify	csv2vectors	mallethon	prepend-lice nse.sh	symlight2vec tors	text2classify		
	text2vectors	vectors2class ify	vectors2info	vectors2topi cs	vectors2vect ors			
23 個項目								

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 myl.classifier_o.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 myl.classifier_o.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。

<pre>4 集(C) 編輯(E) 格式(O) 稅稅(v) 說明(H) frile:/D:/mallet-2.0.8/sample-data/web/en/elizabeth_needham.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/mild.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/hawes.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/hill.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/shiloh.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/tata/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/tata/web/en/thylacine.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/tata/web</pre>	📄 hill3.txt - 記事本	
<pre>file:/D:/mallet-2.0.8/sample-data/web/en/equipartition_theorem.txt file:/D:/mallet-2.0.8/sample-data/web/en/equipartition_theorem.txt file:/D:/mallet-2.0.8/sample-data/web/en/maves.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/haves.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/shiloh.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/shiloh.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/tylacine.txt 1.0</pre>	檔案(F) 編輯(E) 榕式(O) 檢視(V) 說明(H)	
	<pre>matc() matc() matc() matc() matc() file:/):/mallet-2.0.8/sample-data/web/en/elizabeth_needham.txt file:/):/mallet-2.0.8/sample-data/web/en/gunnhild.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/hill.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/hill.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/shiloh.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/thespis.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/thespis.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/thespis.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/thylacine.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/tylacine.txt 1.0 file:/):/mallet-2.0.8/sample-data/web/en/yard.txt 1.0 file:/D:/mallet-2.0.8/sample-data/web/en/yard.txt 1.0</pre>	1.0
		- • · ·

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。

- 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。

	欉磁碟 (D:) ▶ msllet-2.0.8 ▶	bin	•	▶ 搜尋 bin		٩
檔案(F) 編輯(E) 檢視(V)	工具(T) 說明(H)					
組合管理 ▼	烧錄 新增資料夾					
🗙 我的最愛		m 2 doceif	nu 2 eleccif	nu 2 slossif	m 2 closeif	nu 2 slossif
🍃 煤燈櫃	et	ier.trial0	ier.trial1	ier.trial2	ier.trial3	ier.trial4
[點] 電腦 □ 本機磁碟 (C:) 本機磁碟 (D:)	my1.classif my.classifi ier er	en.mallet	inallet.bat	mallet3.ba t	classifier2i nfo	csv2classif y
🙀 網路	csv2vector mallethon s	prepend-li cense.sh	svmlight2v ectors	text2classif v	text2vecto rs	vectors2cl assify
資源回收筒	vectors2inf vectors2to o pics	vectors2ve ctors		,		
Topic.mallet 修 MALLET 檔案	改日期: 2018/12/13 上午 10:27 大小: 22.9 KB	建立日期: 20)18/12/13 上午	10:27		

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。

						-	
	機磁碟(D:) ▶ m	allet-2.0.8 🕨	bin	•	ℓ → 授 尋 bin		٩
檔案(F) 編輯(E) 檢視(V)	工具(T) 說明(⊢	Ð					
組合管理 ▼ 🦳 開啟 ▼	焼錄 新	f 増資料夾				***	
🗙 我的最愛							
演 煤體櫃	Topic.mall et	Mallet	my2.classif ier.trial0	my2.classif ier.trial1	my2.classif ier.trial2	my2.classif ier.trial3	my2.classif ier.trial4
▶■ 電腦 ■ 本機磁碟(C:)	my1.classif	my.classifi	en.mallet	inallet.bat	inallet3.ba	classifier2i	csv2classif
本機磁碟 (D:)	ier	er			t	nfo	У
📬 網路	csv2vector s	mallethon	prepend-li cense.sh	svmlight2v ectors	text2classif y	text2vecto rs	vectors2cl assify
🗑 資源回收简	vectors2inf o	vectors2to pics	vectors2ve ctors				
Topic.mallet 修 MALLET 檔案	改日期: 2018/12/ 大小: 22.9 KB	13 上午 10:27	'建立日期: 20)18/12/13 上午	= 10:27		

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)。

🔟 🛃 🤊 • 🔍 - 🚨 🗋 🎲 I=	Topic.txt -	Microsof	t E×ce	ł						_ 0 ;	23
留窯 常用 插入 版画配置 公式 資料	枝開 械視	開發人員	Ę						~	() – Ø	23
ボタン 新細明酸 × 12 × A* A* K上 ジョン ・	■ ■ ■ ■ ■ ■ ■ ■ 事 ■ ■ ■ 事 ■ ■ 事 ■ 事 ■ ■ の *	通用格式 \$ ~ % ***********************************	, ,	 ・ ・	U化的 a表相 g式 - g	條件 ▼ 📑 * 📑 節 	插入 刪除 格式 諸存桔	×Σ×/ *⊌* * ②* 排月	。 梁 辞 编輯	武具 選取 ▼	
A1 - 🗲 #d	oc name topic pro	portion									~
AB		-	С	D	E	F	G	Н	Ι	J	E
1 #doc name topic proportion			-	_	1	_			-		T
2 0 file:/D:/mallet-2.0.8/sample-data/web/en/el	izabeth_needham.	txt	13	0.179612	11	0.121359	3	0.063107	6	0.053398	
3 1 file:/D:/mallet-2.0.8/sample-data/web/en/ed	uipartition_theore	m.t	17	0.375862	8	0.141379	16	0.058621	6	0.044828	
4 2 file:/D:/mallet-2.0.8/sample-data/web/en/gu	unnhild.txt		9	0.234615	18	0.073077	10	0.073077	8	0.057692	
5 3 file:/D:/mallet-2.0.8/sample-data/web/en/ha	wes.txt		4	0.248387	12	0.093548	3	0.080645	19	0.06129	
6 4 file:/D:/mallet-2.0.8/sample-data/web/en/hi	ll.txt		5	0.251678	3	0.07047	1	0.07047	13	0.063758	
7 5 file:/D:/mallet-2.0.8/sample-data/web/en/sh	uloh.txt		12	0.256345	4	0.109137	8	0.073604	2	0.068528	
8 6 file:/D:/mallet-2.0.8/sample-data/web/en/su	inderland_echo.tx	t i	18	0.188356	15	0.14726	1	0.071918	3	0.058219	
9 7 file:/D:/mallet-2.0.8/sample-data/web/en/th	espis.txt		13	0.106897	3	0.093103	19	0.086207	10	0.07931	
10 8 file:/D:/mallet-2.0.8/sample-data/web/en/th	ylacine.txt		16	0.298883	18	0.075419	0	0.069832	6	0.053073	
11 9 file:/D:/mallet-2.0.8/sample-data/web/en/ur	anus.txt		2	0.222581	6	0.125806	10	0.06129	1	0.06129	=
12 10 file:/D:/mallet-2.0.8/sample-data/web/en/ya	ard.txt		- 7	0.257764	11	0.07764	8	0.065217	5	0.052795	
13 11 file:/D:/mallet-2.0.8/sample-data/web/en/zi	nta.txt		14	0.252778	19	0.136111	13	0.080556	5	0.058333	
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											Ŧ
H 4 F H Topic 😒											1
「統治 🎦								100% -		· +) _{.:}

圖29 主題模型文字檔畫面

		- (a 🗅 🦚	↓						Торі	c-key.txt - M	icrosoft Exce	ł							= X3
楣	s:	常用	插入	版画證	公式	資料	枝開	械視	開發人員										ه 🕜 ه	- # 23
『 貼」		朝下 複載・ 【複載格式	新細明銀 B/I	1 U - I	- 1: - 👌 -	2 - v 4 <u>A</u> -	А` ∧` фž -	= <mark>=</mark> = 5 3 3	≫- 律律	■ 自動換列 ■ 防欄置中、	通用結式 \$ - %	, <u>*.0</u> .0	· 設定格式化	構成化為 儲存	格 插入	副除 格式	Σ 自動加 ■ 填漏・ ② 清除・	124 × 2000 , 排序9	「「「」」 (前温 尋找奥	
	剪貼	海 G		亨	P型		5		對齊方式		動	值	N SINGET -	横式		儲存格		编輯	325.40.	
		K23		- (0	ţ	*														~
	Α	В	С	D	E	_	F	G	Н	I	J	K	L	М	N	0	P	0	R	5
1	0	2.5 cre	ating end	ed year la	te marke	ts portr	ayal mil	single dil	non-gove	rnment educa	ational agen	cy collidin;	g million rad	liation-proces	sed ice tel	escope mink	running op	ossum		
2	1	2.5 sou	th female	thought :	sullivan s	cored 1	na wes a	ccomplish	ments ima	ge born asso	ciation grad	uated majo	- rity compos	ed albedo bor	nd extrem	ely outer exti	nction ?底a	防l?sa防n	works	
3	2	2.5 ring	gs ring du	ıst uraniar	n particles	s theatr	e april c	eremony h	neld premi	ier performer	columns st	ability dens	e fragmentat	tion understoo	od lack op	aque exceed	exist			
4	3	2.5 wai	gilbert r	elative ma	aj leading	world	runs cai	dos whig b	eague boy	rfriend veer-z	aara fiction	annual ide	als tension p	ublisher princ	eton plan	et confines				
5	4	2.5 uni	on confe	derate ker	ntucky ha	wes co	nfedera	tes general	states job	inston don fo	rces commo	onwealth ne	utrality civil	l commerciali	sm promo	te diameter o	rganics ape	x bounties	cynocephal	.us
6	5	2.5 test	cricket a	ustralian l	hill caree:	record	l death l	kehna worl	ked innin;	gs run played	husband p	unjab play [.]	variety telug	u committee :	support st	ephen				
7	6	2.5 inc	luding ed	ward mod	ons additi	on disc	overed	online npa	caused p	ublicize sterli	ing confine	ł strictly ex	isted collisic	onal young du	e bodies o	dark? he ind	complete sp	acecraft		
8	7	2.5 yar	d nationa	l years pa	rks park	standar	ds mov	ement rest	figure per	rformance fili	ms importar	it members	forest relatio	onship execut	ive initiati	ves led head	legislative			
9	8	2.5 wil	derness a	merican f	ighting la	unched	l spent o	classical th	ermal soc	iety noted ed	ucate aerod	ynamic mai	in reproducti	ive times rule	rs paper's	rules retreat a	assaulting w	allace's		
10	9	2.5 gur	nhild no	rway life l	king serv	ice jou	malist re	eturn erik e	stablishe	l character su	ibsequently	selection b	roke elliot se	ettlement foss	il views ty	neside artille	ry reported			
11	10	2.5 nur	nerous n	umber ura	nus relate	ed prote	ection li	ne orkney	news alvi	da star-crosse	ed consecut	ive foundin	g activist op	tically consis	ts eminen	t elderly trade	e headed su	llivan's		=
12	11	2.5 nee	dham ac	ted critical	l presider	nt day r	nother c	:o-owner le	ead movie	s helping pee	ers called in	dustrializati	on concentr	ate served frie	end twenty	/ exosphere?	orona ex	tended ban	ıds	
13	12	2.5 hat	tle time a	rmv gen t	ennessee	united	written	nosition m	cant'e gran	it heauregard	landing nit	ishura huel	shiloh gove	ernor success	soldier or	nosed organs				
1	13	2.5 lon	don year:	s debut na	mow mai	supial	actors th	nespis rece	ived serie	s died wome	n biggest in	dia heroine	earned hind	i zinta discor	d nps resu	lted				
15	14	2.5 101	actress i	nin mulai	i zima ac	ing iop	-grossi	ng miniare	innar me	tependent pro	oductions st	age regular	asia cinema	KaDIII IIdiiida	ie gaya ku	n science				
16	15	2.5 ech	o areas e	ast forced	evening	men er	ngland r	ichard cen	tury accla	im romance (credited pre	ity masterp	ieces univer:	sity tragedian	s pinafore	operatic und	amaged bu	ilding		
17	16	2.5 thy	lacine tas	manian b	ack tiger	law wa	ter pouc	ch devil sp	ecies exti	nct tasmania t	thylacinus e	arly austral	ia generally	america's aes	thetic zon	es sheath ma	rsupials			
18	17	2.5 ave	rage syst	em equipa	artition th	eorem	energy	kinetic vir	ginia cons	idered effects	s heat motic	n equilibriu	um original e	energies temp	erature an	nericans math	ner tazzy ma	arch ohio		
19	18	2.5 sun	derland s	survived g	reek deo	ember j	paper da	aily thomas	s storey ne	ewspaper bas	ed high cor	ception ori	ginated mec	hanism large	kilometre	s james conve	ergent mam	mals name	s	
20	19	2.5 naa	award n	nodern go	ds opera	kings v	vadia ne	ess bbc mo	vie overse	eas salaam co	mmercial c	hanging ky	a recognised	l made englisl	h-languag	e bollywood	appeared			
21	_																			
22	_																			
23	_																			-
14.4	► H	Topic-ke	y 😕	,		1			1											
就建	1	3																100% 😑		- ÷ "

圖30 主題組合畫面

l	#doc source pos typeindex type topic	
2	#alpha : 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5	.5 2.5 2.5 2.5 2.5 2.5 2.5
З	#beta : 0.01	
4	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 0 0	elizabeth 13
5	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 1 1	needham ll
6	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 2 2	died 13
7	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth_needham.txt 3 3	mother 11
8	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 4 1	needham ll
9	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 5 4	english ll
10	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 6 5	procuress ll
11	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 7 6	- brothel-keeper 13
12	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 8 7	th-century 13
13	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 9 8	london 13
14	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 10	dentified ll
15	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt ll	0 bawd 6
16	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 12	l greeting 3
17	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 13	2 moll 1
18	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.txt 14	3 hackabout 6
19	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needbam txt 15	4 plate 19
20	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needbam tvt 16	5 william 13
21	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needbam tyt 17	6 hogarth's 0
22	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 18	7 series 13
23	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 19	8 satirical 13
24	0 D:\mailet-2 0 8\cample_data\web\en\eliyabeth needham tyt 20	9 etchings 3
25	0 D:\mailet-2 0 8\cample_data\web\en\eliyabeth_needham_tyt_2]	0 harlot's 12
26	0 D:\mallet-2 0 8\sample-data\web\en\elizabeth needham tyt 22	1 progress 5
27	0 D:\mailet-2 0 8\sample-data\web\en\eliyabeth needham tyt 23	needham 11
28	0 D:\mailet -2 0 8\comple_data\web\en\elizabeth needham tvt 24	2 notorious 15
29	0 D:\mailet-2 0 8\cample_data\web\en\eliyabeth needham tyt 25	london 13
30	0 D:\mailet 2.0.0(sample data\web\en\elizabeth needham.txt 26	$3 \pm i m = 12$
31	0 D:\mailet-2 0 8)sample data\web\en\elizabeth needham.txt 27	4 recorded 15
32	0 D:\mallet-2 0 8\sample-data\web\en\elizabeth needham tyt 28	5 life 9
33	0 D:\mailet-2 0 8\cample_data\web\en\eliyabeth needham tyt 29	6 genuine 1
34	0 D:)mallet-2 0 8)sample-data)web)en)elizabeth needham tyt 30	7 nortraits 5
35	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 3]	8 survive 10
36	0 D:\mallet-2 0 8\sample-data\web\en\elizabeth needham tyt 32	9 house 10
37	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 33	0 exclusive 13
38	0 D:\mailet-2 0 8\cample_data\web\en\elizabeth needham tyt 34	london 13
39	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 35	l customers 3
40	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 36	2 highest 4
41	0 D:\mallet-2 0 8\sample-data\web\en\elizabeth needham tyt 37	3 strata 5
42	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needbam tvt 38	4 fashionable 11
43	0 D:\mailet-2 0 8\cample_data\web\en\elizabeth needham tyt 39	5 society 8
44	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 40	6 eventually 7
45	0 D:\mailet-2 0 8\sample-data\web\en\elizabeth needham tyt 4]	7 crossed 6
46	0 D:\mailet-2 0 8\cample_data\web\en\elizabeth_needham.tvt 42	8 morel 13
47	0 D:\mailet 2.0.0(sample data\web\en\elizabeth_needham.txt 42	9 reformers 2
4.9	0 D:\mailet 2.0.0(sample data\web\en\elizabeth needham.txt 40	0 dev 11
49	0 D:\mailet 2.0.0(sample data\web\en\elizabeth needham.txt 45	diad 13
50	0 D:)mallet-2 0 8)sample-data)web(en(elizabeth_neednam.txt 43	l result 13
51	0 D.)mallet - 2 0 8)cample data/web/en/elizabeth needham.txt 40	2 comoro 3
52	0 D:)mallet-2 0 8)sample-data)web/en/elizabeth needham.twt 49	3 treatment 11
53	0 D.)mallet - 2 0 8)cample data;web(en(elizabeth_needham.txt 40	4 received 13
54	0 D:)mallet-2 0 8)sample-data)web(en(elizabeth_needham.txt 4)	5 sentenced 13
55	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham.tvt 51	6 stand 2
56	0 D:\mallet-2.0.8\sample-data\web\en\elizabeth needham tvt 52	7 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。

			- • •
	ample-data 🕨 web 🕨 pe	earl ▼ *y 搜尋/	pearl 🔎
檔案(F) 編輯(E) 檢視(∀) 工具(T) 說明(H)		
組合管理 ▼ 加入至媒體櫃 ▼	共用對象 ▼ 焼錄	新増資料夾	• • • •
★ 我的最愛 ▶ 下戦 ■ 桌面 3 最近的位置	E.	E.	
🈭 煤體櫃	MediaDestruction Area.txt	SystemPersistence ExerciseArea.txt	TechnicalServicesA rea.txt
🖳 電腦			
🖬 本欉磁碟(C:)			
👝 本機磁碟 (D:)			
4 網路			
3 個項目			

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「software」、文 件2機率最高為主題2之組合,出現最多之單詞為「 inigration」,以上數值統計是參照 Topic-state 檔案內之單 詞數量(圖39)。

B	C	D	E	F	G	H	Ι	J	K	L	M	N	0	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.t	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	Ø	0.192488	8	0.131455	1	0.084507	4	0.070423	3	0.070423	б	0.042254	9	0.037559	5	0.032864	7	0.028169

圖37 主題模型文字檔畫面

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 - MALLET

A	B C	D	E	F	G	H	Ι	J	K	L	М	N	0	P	Q	R	S	T	U	V
0	5 records	outputs	include	mpeg	developed	doc	record	technology	certificates	refers	dvd	includes	mechanism	n vinyl/tape/l	jpge	encapsulat	replicated	integrity	institutions	compliant
1	5 format	hard	usability	software	tapes	place	openoffice	ppt	docx	proprietary	png	multimedia	suite	repair	break	re-formatti	r completely	actual	destroying	cassette
2	5 migration	pdf/a	formats	tiff	data	jpeg	issued	component	shared	migrated	quality	verification	risk	jpeg/tiff/pd	ipdfcreator	WaV	gif	university	taipei	national
3	5 optical	disks	disk	dvd	windows	tools	includes	scope	ffmpeg	XIS	postscript	incorporates	library	jVC	kiosk	solaris	shredder	destroy	provided	recovery
4	5 storage	operating	methods	preserving	files	pdf	steps	destroyed	ensure	validation	forms	odt	WINY	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	hard ware	inch	servers	earthquake	produced	odt	signify	encapsulate	knowledge	email	online	creation	testing	workstation	nhouses	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	tremoving	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			1
C	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
C	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	4	4	years	7
C	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	5	5	destroyed	4
C	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	б	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
C	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	9	9	legal	1
C	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
C	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	13	5	destroyed	4
C	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	б
0	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	16	14	scope	3
C	D:\mallet-2.0.8\sample-data\web\pearl\MediaDestructionArea.txt	17	0	destruction	8
C	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
-		1			· · · ·

圖39 單詞統計畫面

参、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。。