

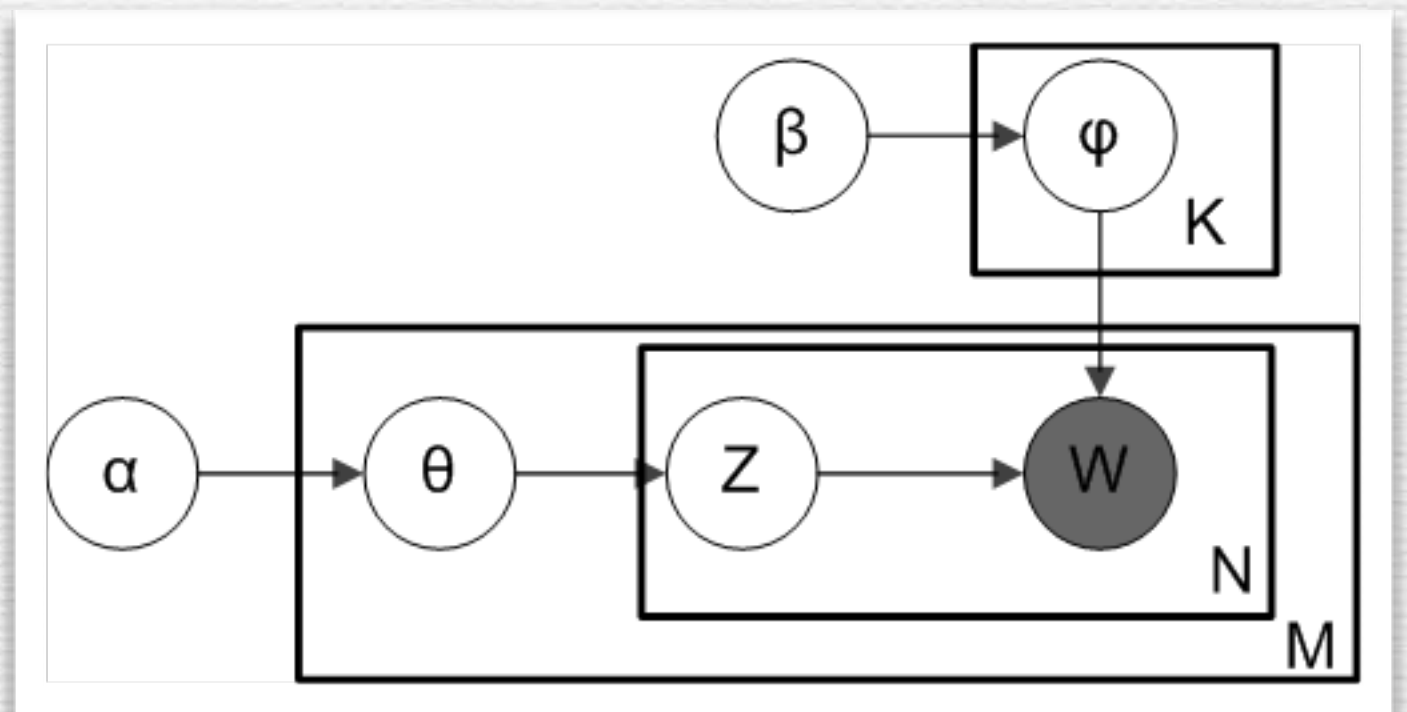


20180611課程內容

製作人：嘉真&炳宏

Agenda

- Latent Dirichlet allocation
- Evaluation of LDA
 - Perplexity
- Visualization
 - LDAvis



Latent Dirichlet allocation

Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) 是一種主題模型，它可以將文檔集中每篇文檔的**主題**按照**機率分布的形式**給出。同時它是一種**無監督學習算法**，在訓練時不需要手工標註的訓練集，需要的僅僅是文檔集以及指定**主題的數量 K** 即可。此外LDA的另一個優點則是，對於**每一個主題均可找出一些詞語來描述它**。

Latent Dirichlet allocation

- 要學會 LDA，你可能需要知道：
 - 一個函數：gamma函數
 - 四個分佈：二項分佈、多項分佈、beta分佈、Dirichlet分佈
 - 一個概念和一個理念：共軛先驗和貝葉斯框架
 - 兩個模型：pLSA，LDA
 - 一個採樣：Gibbs Sampling

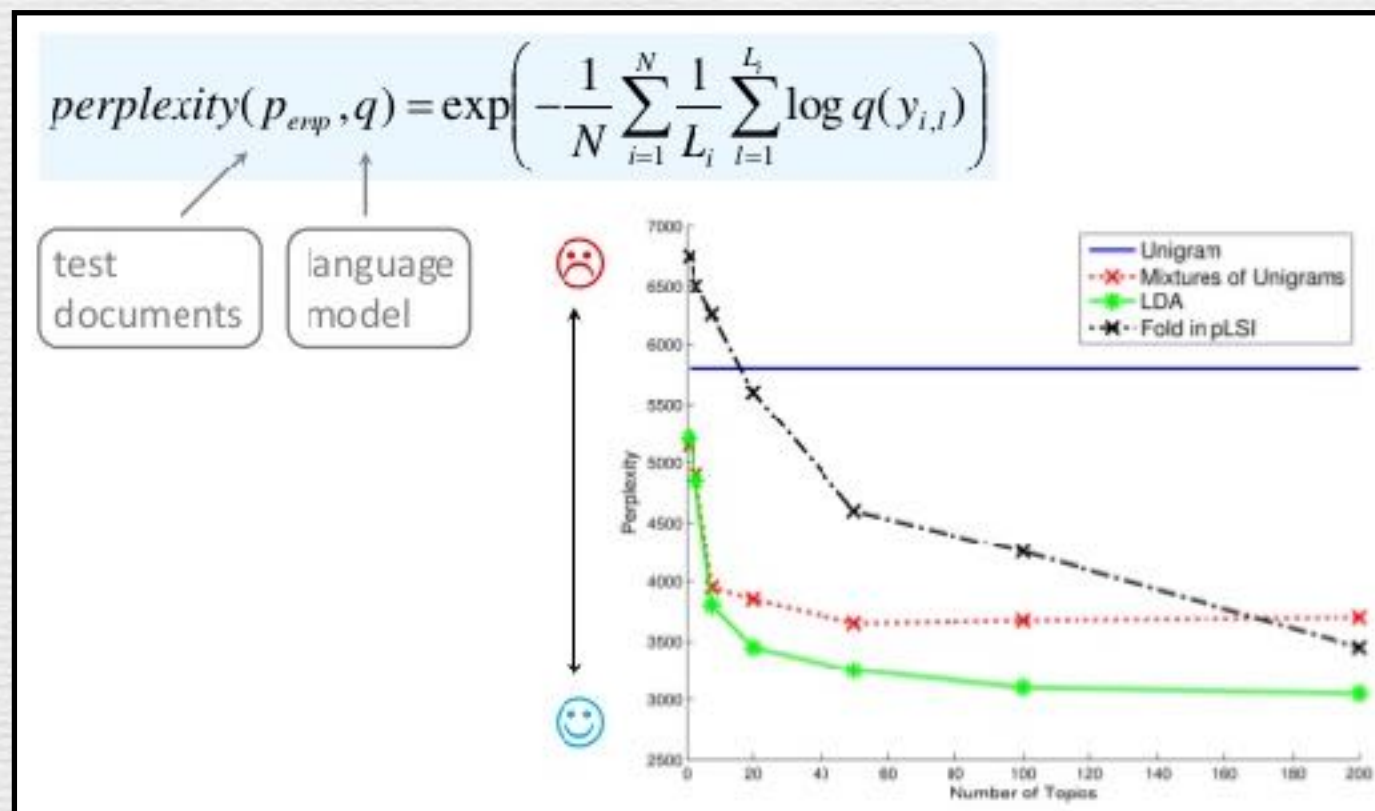
太過複雜

Latent Dirichlet allocation

- 使用現成的 LDA 套件，最重要在於：
 - Topic 的數量 $K \Rightarrow$ Perplexity
 - 每個 Topic 均可找出詞語來描述 \Rightarrow LDAvis
- LDA 常用套件：(說明)
 - topicmodels + **lda** + **LDAvis**

Evaluation of LDA

- Q: 我的 LDA 的主題 (Topic) 要分成幾群？
- A: (其中一種方法) 使用 Perplexity 來衡量。



Perplexity

期望越低越好

Perplexity

第二步:選擇輸出欄位

輸出欄位

- 文章
- ✓ 文章+詞彙
- 文章+類別
- 文章+情緒
- 文章+詞彙+詞性
- 文章+詞彙+詞頻
- 文章+詞彙+類別
- 文章+類別+情緒
- 文章+詞彙+詞性+詞頻
- 詞彙+詞頻
- 詞彙+類別
- 詞彙+詞性+詞頻

word

artDate	artTime	word
2018/03/06	08:45:00	示範
2018/03/07	08:45:31	詞彙

1. 選擇文章+詞彙

3. 獲得_artWord.csv

2. 下載檔案

Ma_artWord.csv

資料預覽 圖形試做 檔案下載

Show 10 entries

artTitle	artDate	artTime
00001	18/05/20	02:05:10
00001	18/05/20	02:05:10
00001	18/05/20	02:05:10
00001	18/05/20	02:05:10

Perplexity

	A	B	C	D	
1	artTitle	artDate	artTime	artUrl	word
2	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	啊啊啊
3	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	TVBS
4	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	生平
5	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	第一次
6	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	腳底
7	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	按摩
8	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	獻給
9	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	東吳
10	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	神父
11	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	完全
12	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	印證

Viewer	
Rename	More ▾
Desktop > Coursera-SwiftKey > final > en_US	

- ☐ en_US.blogs.txt
- ☐ en_US.news.txt
- ☐ en_US.twitter.txt
- ☒ Ma_artWord.csv
- ☐ test.segment.2018-06-09_08_55_06.txt
- ☐ test.segment.2018-06-09_08_55_37.txt
- ☐ test.segment.2018-06-09_09_14_23.txt
- ☐ test.segment.2018-06-09_09_15_02.txt
- ☐ test.segment.2018-06-09_09_17_38.txt
- ☐ test.txt
- ☐ test.segment.2018-06-11_00_07_28.txt
- ☐ Perplexity.R
- ☐ LDavis.R

確認路徑

Perplexity

```
Perplexity.R x LDAvis.R x
Source on Save
1 install.packages('rjson')
2 install.packages('data.table')
3 install.packages('text2vec')
4 install.packages('lda')
5 library('data.table')
6 library('text2vec')
7 library('lda')
8
9 data <- fread("Ma_artWord.csv", encoding = "UTF-8")
10 column <- names(data)[1:ncol(data)]
11
12 temp <- data[1]$artTitle
13 mystr = "" # Set default string
14 text = c()
15 id = c()
16 id <- append(id, temp) # Set start title_id
17
52
53 # Parameters for LDA
54 # topics = c(10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
55 topic <- seq(from = 5, to = 100, by = 5)
56 perplexity = c()
57 for(n_topic in topic){
58   n_iter = 10
59   model = LDA$new(n_topic, doc_topic_prior = 0.1, topic_word_prior = 0.01)
60   doc_topic_distr =
61     model$fit_transform(dtm, n_iter = n_iter, n_check_convergence = 1,
62                         convergence_tol = -1, progressbar = FALSE)
63   topic_word_distr_10 = model$topic_word_distribution
64   perplexity <- c(perplexity, perplexity(dtm, topic_word_distr_10, doc_topic_distr))
65 }
66 plot(topic, perplexity, type = 'b', col = "blue", pch = 19)
67
```

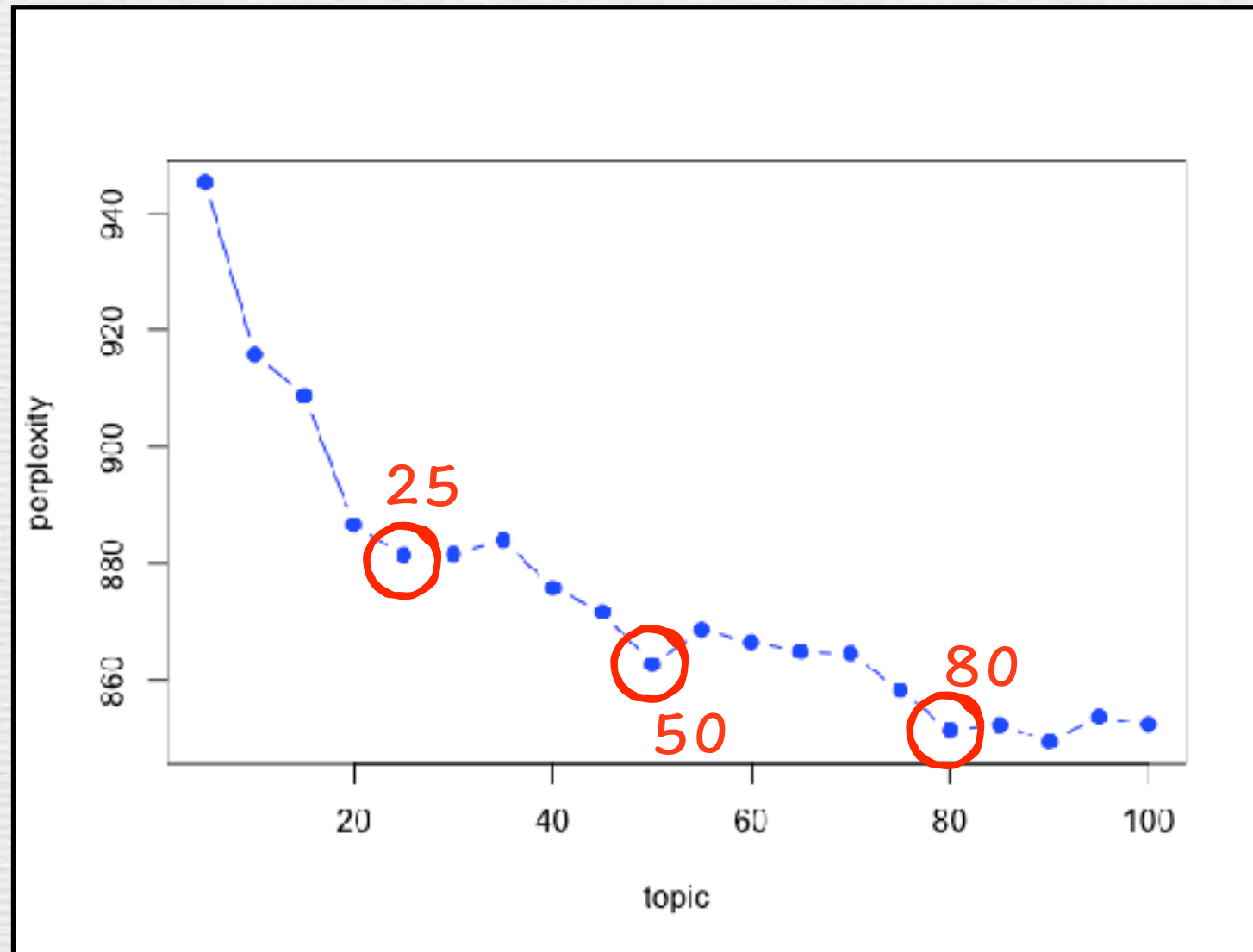
若text2vec套件出問題請點我

設定Topic數量測試

設定輸出樣式

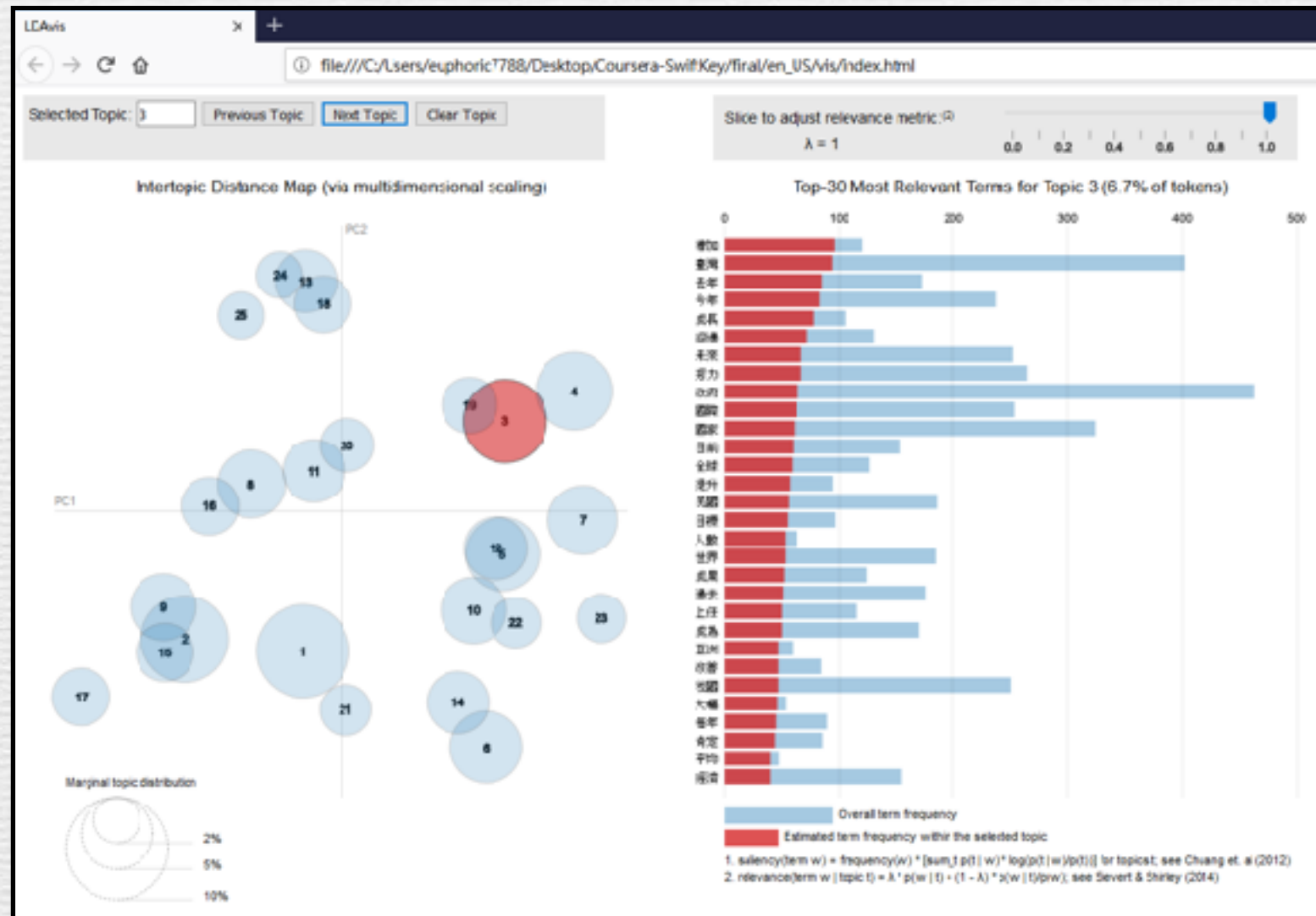
Perplexity

- 根據 Perplexity 衡量 LDAvis 的 Topic 數量



Visualization

- LDA分群 (以馬英九貼文為例) : 25 群
















LDAvis

- LDAvis 是一個交互式的主題模型可視化包，可以將主題模型建模後的結果，利用D3.js封裝好的一個可視化模板，製作成一個網頁交互版的結果分析工具。
- 目前僅支援：**Firefox**瀏覽器（請先安裝）

LDAvis

	A	B	C	D	
1	artTitle	artDate	artTime	artUrl	word
2	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	啊啊啊
3	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	TVBS
4	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	生平
5	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	第一次
6	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	腳底
7	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	按摩
8	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	獻給
9	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	東吳
10	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	神父
11	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	完全
12	1	2018/5/20	2:05:10	https://www.facebook.com/MaYingjeou/posts/1938159356246187	印證

Viewer
Rename  More ▾
Desktop > Coursera-SwiftKey > final > en_US

<input type="checkbox"/>		en_US.blogs.txt
<input type="checkbox"/>		en_US.news.txt
<input type="checkbox"/>		en_US.twitter.txt
<input type="checkbox"/>		Ma_artWord.csv
<input type="checkbox"/>		test.segment.2018-06-09_08_55_06.txt
<input type="checkbox"/>		test.segment.2018-06-09_08_55_37.txt
<input type="checkbox"/>		test.segment.2018-06-09_09_14_23.txt
<input type="checkbox"/>		test.segment.2018-06-09_09_15_02.txt
<input type="checkbox"/>		test.segment.2018-06-09_09_17_38.txt
<input type="checkbox"/>		test.txt
<input type="checkbox"/>		test.segment.2018-06-11_00_07_28.txt
<input type="checkbox"/>		Perplexity.R
<input type="checkbox"/>		LDAvis.R

確認路徑

LDavis

```
Perplexity.R x LDavis.R x
Source on Save
1 install.packages('later')
2 install.packages('LDavis')
3 library('data.table')
4 library('text2vec')
5 library('later')
6
7 data <- fread("Ma_artword.csv", encoding = "UTF-8")
8 column <- names(data)[1:ncol(data)]
9
10 temp <- data[1]$artTitle
11 mystr = "" # Set default string
12 text = c()
13
43 get.terms <- function(x) {
44   index <- match(x, vocab)
45   index <- index[!is.na(index)]
46   rbind(as.integer(index - 1), as.integer(rep(1, length(index))))
47 }
48 documents <- lapply(doc.list, get.terms)
49 |
50 K <- 25 # Topics
51 G <- 5000 # iteration times
52 alpha <- 0.10
53 eta <- 0.02
54
55 # LDA
56 library(lda)
57 set.seed(357)
58 fit <- lda.collapsed.gibbs.sampler(documents = documents, K = K, \
59
```

設定參數

LDavis

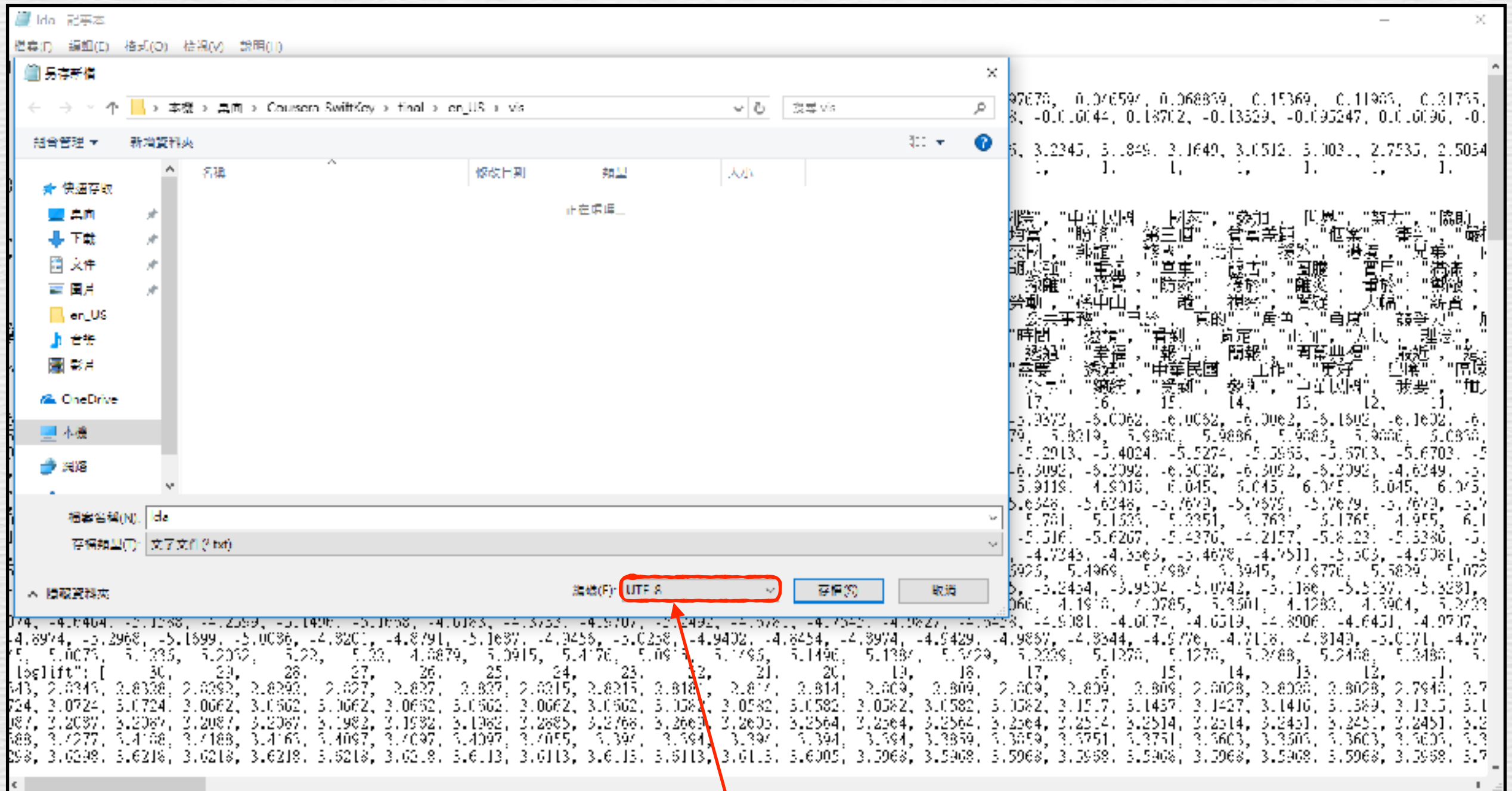
多出vis資料夾

名稱	修改日期	類型	大小
vis	2018/6/11 上午 1...	檔案資料夾	
.Rhistory	2018/6/11 上午 0...	RHISTORY 檔案	22 KB
en_US.blogs	2018/6/8 下午 06...	文字文件	205,235 KB
en_US.news	2018/6/8 下午 06...	文字文件	200,989 KB
en_US.twitter	2018/6/8 下午 06...	文字文件	163,189 KB
LDavis	2018/6/11 上午 1...	R 檔案	3 KB
Ma_artWord	2018/6/8 下午 05...	CSV 檔案	11,933 KB
Perplexity	2018/6/11 上午 0...	R 檔案	3 KB

名稱	修改日期	類型	大小
d3.v3	2018/5/11 上午 1...	JavaScript 指令檔	302 KB
index	2018/5/11 上午 1...	Chrome HTML D...	1 KB
ldc	2018/5/11 上午 1...	階層式樣式表文件	1 KB
lde	2018/5/11 上午 1...	JSON 檔案	173 KB
ldevis	2018/5/11 上午 1...	JavaScript 指令檔	52 KB

以「記事本」開啟

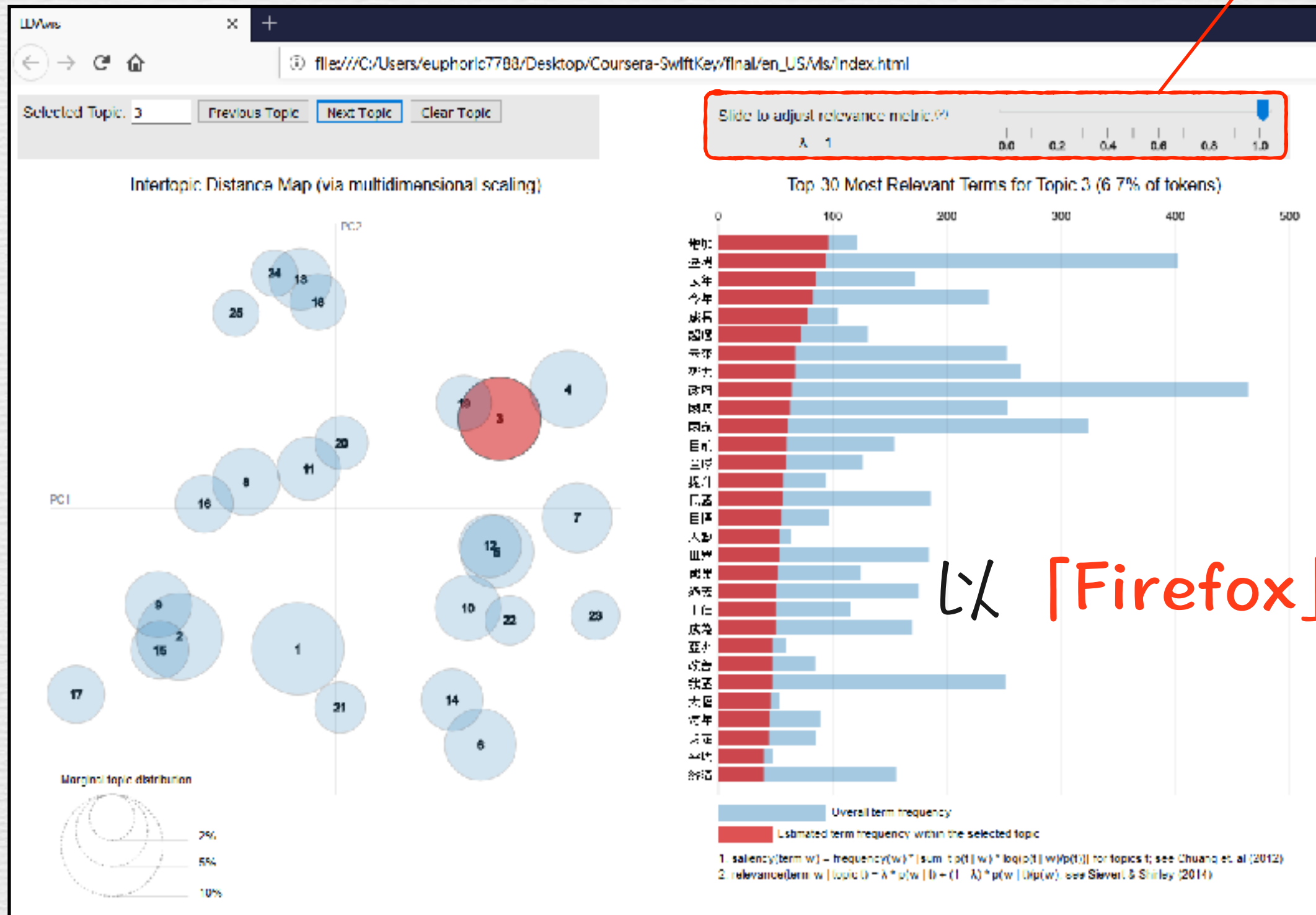
LDavis



選擇UTF-8，並存檔覆蓋原來檔案

LDavis

λ 參數



以「Firefox」開啟

LDavis

- 關於 λ 參數：
 - 相關性公式： $\text{relevance}(\text{term } w \mid \text{topic } t) = \lambda * p(w \mid t) + (1 - \lambda) * p(w \mid t)/p(w)$
 - 某個詞語主題的相關性，由 λ 參數來調節。如果 λ 接近1，那麼在該主題下更頻繁出現的詞，跟主題更相關；如果 λ 越接近0，那麼該主題下更特殊、更獨有的詞，跟主題更相關。

參考資料

- https://blog.csdn.net/sinat_26917383/article/details/51547298
- <https://rdrr.io/github/dselivanov/text2vec/man/perplexity.html>
- <https://computational-communication.com/%E5%8F%AF%E8%A7%86%E5%8C%96/ldavis-intro/>