大數據分析之應用：自動化電子公文分類系統與專利文件推薦系統

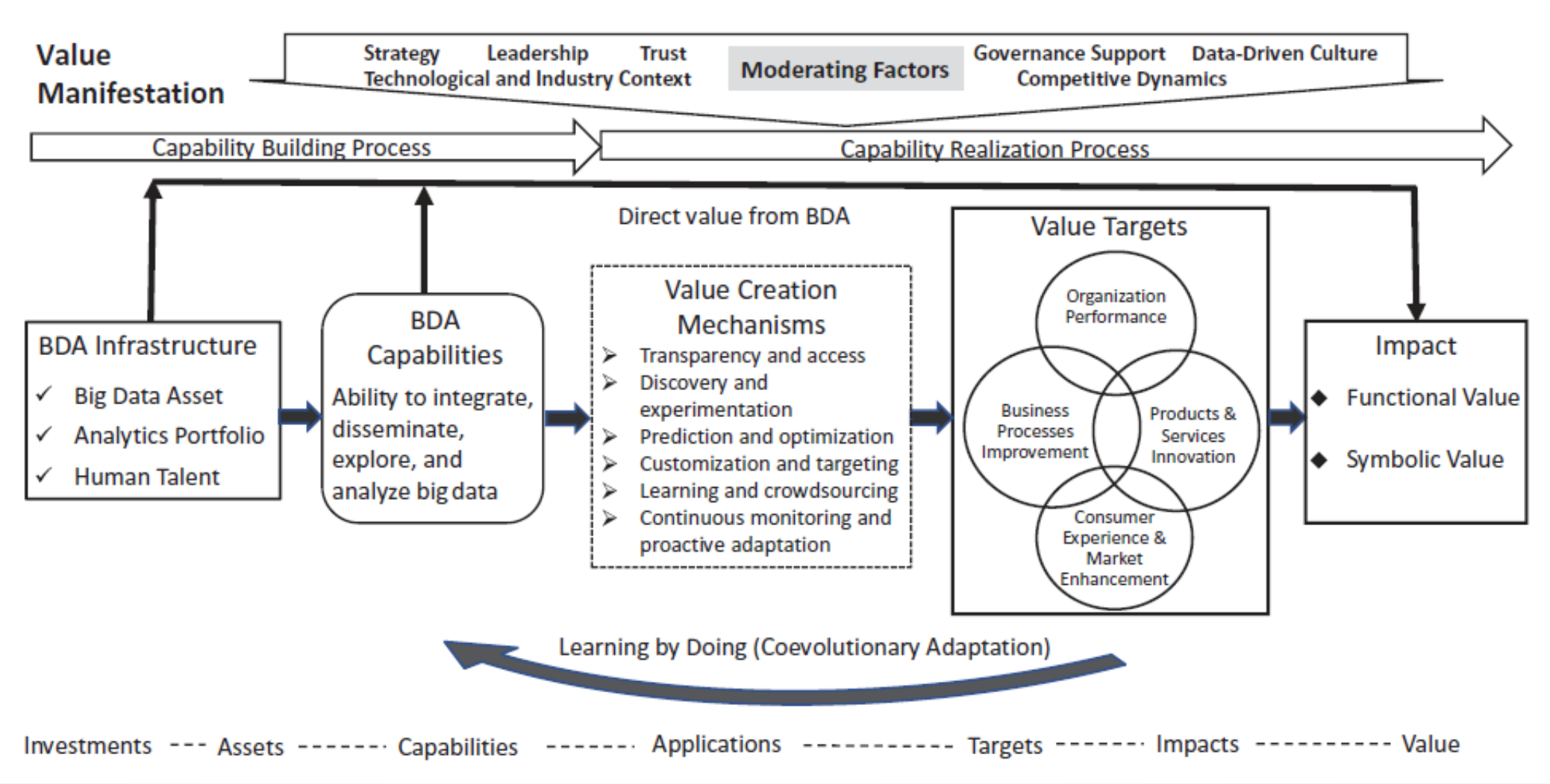
國立屏東科技大學
資訊管理系
陳灯能

大數據分析(Big Data Analysis, BDA)

- 何謂大數據(Big Data)
  - 大數據意指資料量的規模巨大到無法人工處理，甚至一般電腦無法在合理的時間內進行分析處理，以轉換成人類所能解讀形式的資訊。
- 大數據的5V特色
  - Volume：巨量性
  - Velocity：速度性、時效性
  - Variety：多樣性
  - Veracity：真實性
  - Value：價值性

→大數據是數量極大、變化速度快、資料來源/型態多樣，能夠反正真正事實，並且能夠產生價值的資料資產。

Value Creation by BDA

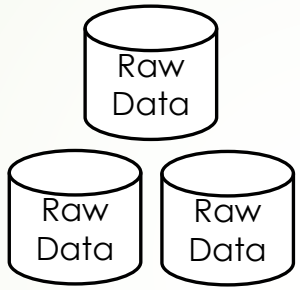


Varun Grover, Roger H.L. Chiang, Ting-Peng Liang & Dongsong Zhang (2018)

資料分析程序

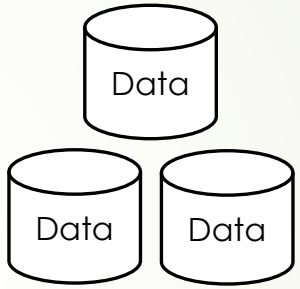
資料來源

資料選擇



目標資料

資料預處理



預處理後資料

資料轉換

轉換後資料

資料分析

資料視覺化

結果判讀

知識

資料前處理
Data Preprocessing
Data Cleaning
Data Munging
Data Wrangling

80%時間

資料分析
統計分析
機器學習

20%時間

文字探勘(Text Mining)

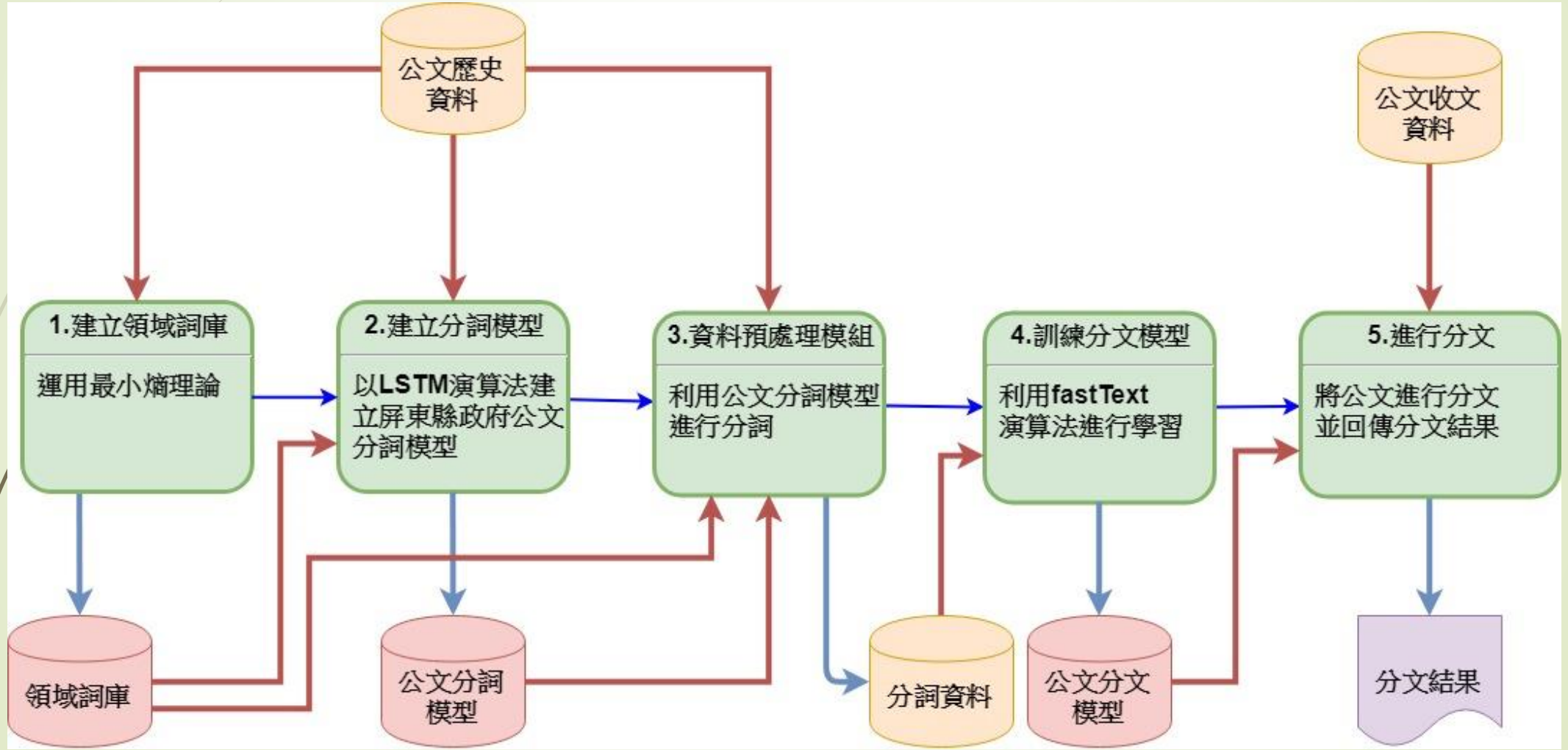
- 定義
  - 從非結構化的文件資料中，尋找模式並將訊息與雜訊分離，而自動發現以前未知的新知識 (M. Hearst , 2003)
  - 文字探勘中最重要的因素在於文件中被提取的訊息，而形成新的事實或假設，其目的是發現迄今未知的知識領域。
- 相關技術
  - 中文斷詞、詞頻計算、統計分析、機器學習.....
- 應用領域
  - 網路口碑、社群網站、情緒分析、專利文件分析.....

自動化電子公文分類系統

系統概述

- 實際應用在屏東縣政府電子公文系統，針對每天約五六百件的電子公文自動分派到各局處，有效的提升公文分派效率，準確度約90%。
- 每年處理約20萬件公文。
- 核心技術
  - 領域詞庫
    - 利用最小熵原理建立領域詞庫(NLP zero)。
  - 分詞模型
    - 利用LSTM建立分詞模型(FoolNLTK)。
  - 利用FastText進行公文分類。

系統架構


公文歷史資料
公文收文資料
1.建立領域詞庫
運用最小熵理論
2.建立分詞模型
以LSTM演算法建立屏東縣政府公文分詞模型
3.資料預處理模組
利用公文分詞模型進行分詞
4.訓練分文模型
利用fastText演算法進行學習
5.進行分文
將公文進行分文並回傳分文結果
領域詞庫
公文分詞模型
分詞資料
公文分文模型
分文結果

中文分詞問題

- 中文分詞技術
  - 基於詞典
  - 基於統計
    - 監督式
      - 優點：解決大量的未登錄詞問題(Out of Vocabulary, OOV)
      - 缺點：需要大量訓練資料
    - 非監督式
      - 優點：需要較少的資料量
      - 缺點：準確率與召回率較低
- 專業領域分詞
  - 缺乏大量已標註資料進行模型訓練➔非監督式學習

NLP zero

- 利用文字信息熵最小化處理無監督的NLP的方法

$H_W/\bar{s}$ is the word entropy per letter

$$H_W = -\sum_{w \in W} P_w \log(P_w)$$

$$\bar{s} = \sum P_w s_w$$

FoolNLTK

- 特點
  - 目前準確度最高的中文分詞框架
  - 基於BiLSTM模型訓練而成
  - 包含分詞、詞性標註、實體識別
  - 允許自訂詞庫

FastText

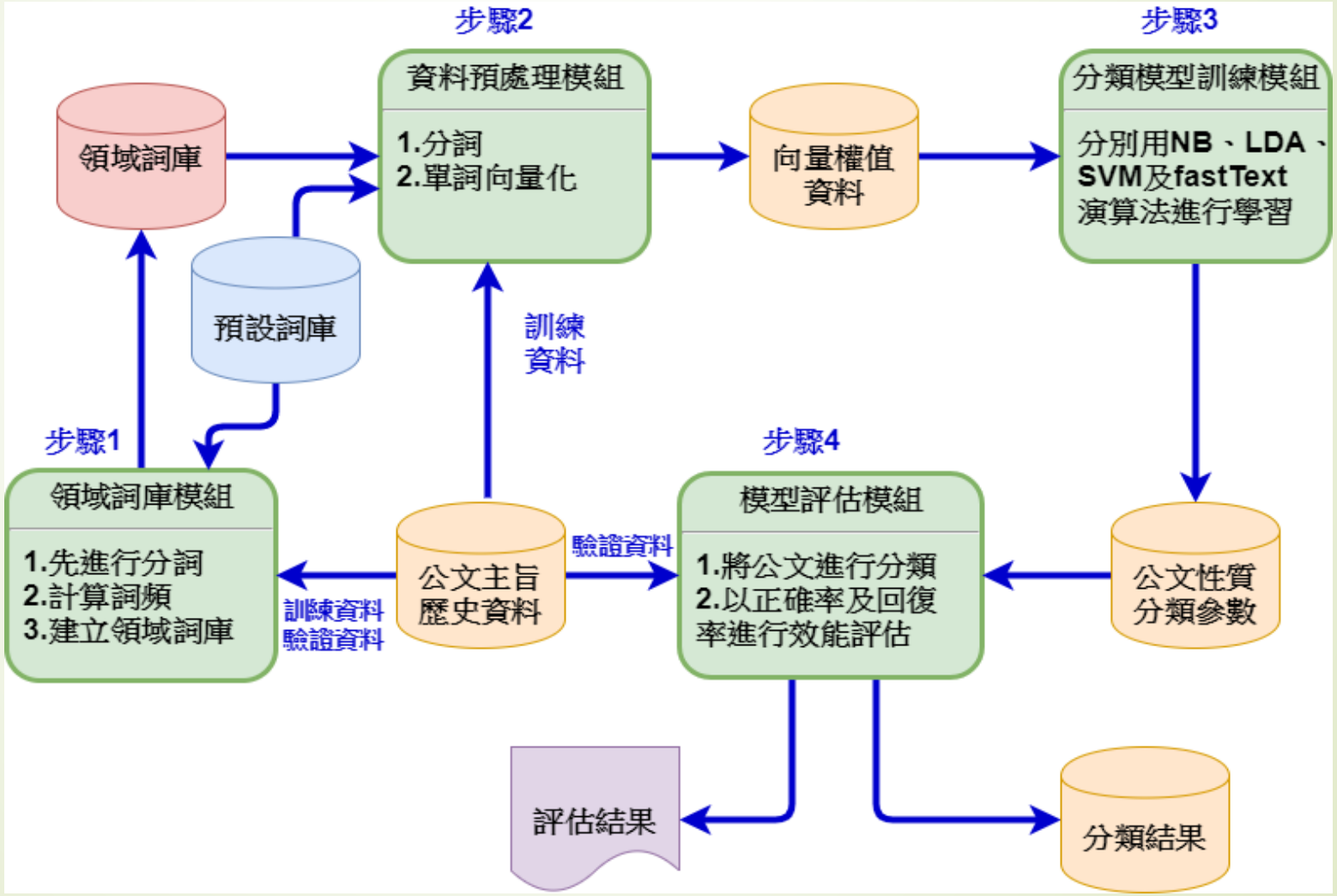
- facebook於2016年開放源始碼的文件分類工具，結合自然語言處理和機器學習技術，用來處理文字分類和學習單詞向量表示（Word Vector Representation）
- FastText在標準多核CPU的環境中，可以在10分鐘內訓練機器學習模型超過10億個單詞，訓練機器學習模型的時間也從幾天大幅降低為幾秒鐘。

模型驗證

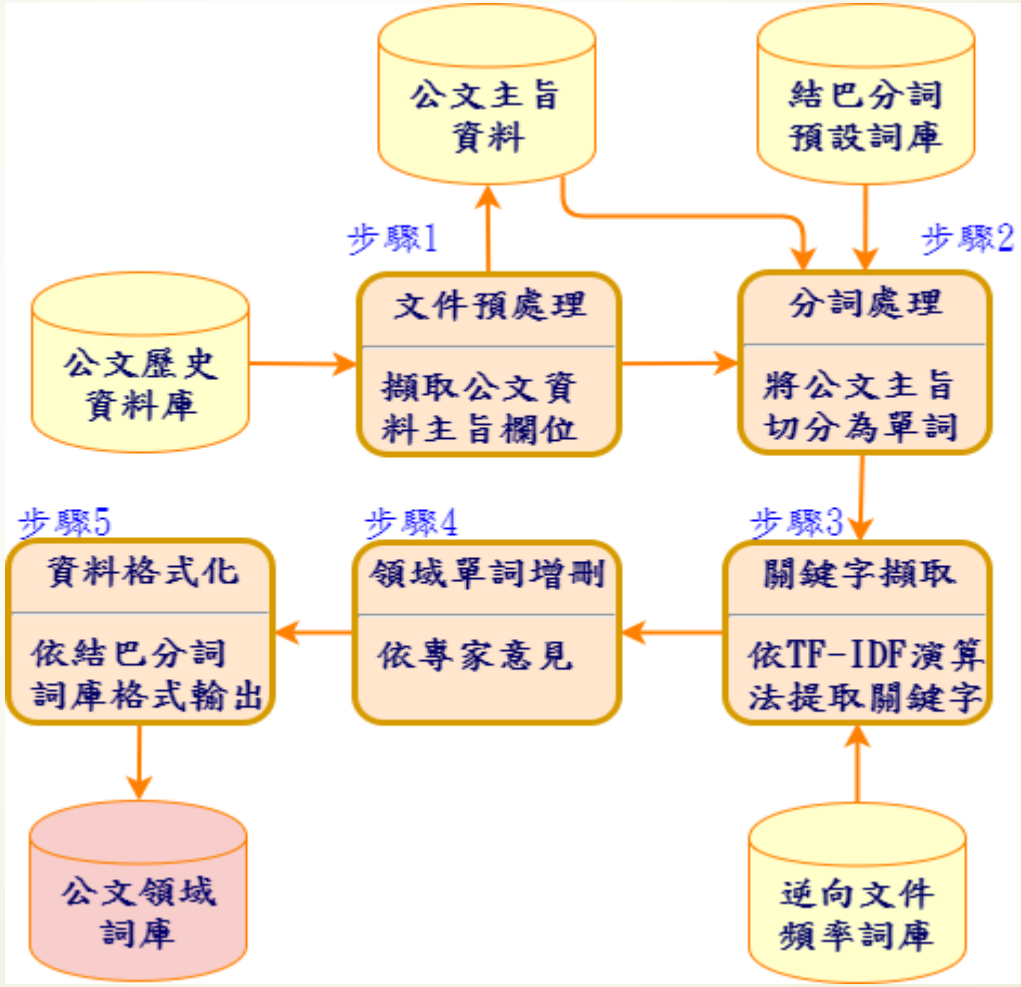
- 屏東縣府公文系統每年收文文量大約20萬筆，本系統以103年至106年的公文資料建構公文分文模型，訓練資料約80萬筆。
- 測試資為107年1-5月份資料約8萬筆。

	FastText	KNN	SVM	NB
訓練資料	0.896	0.897	0.907	0.810
107年1-5月資料	0.889	0.758	0.846	0.813

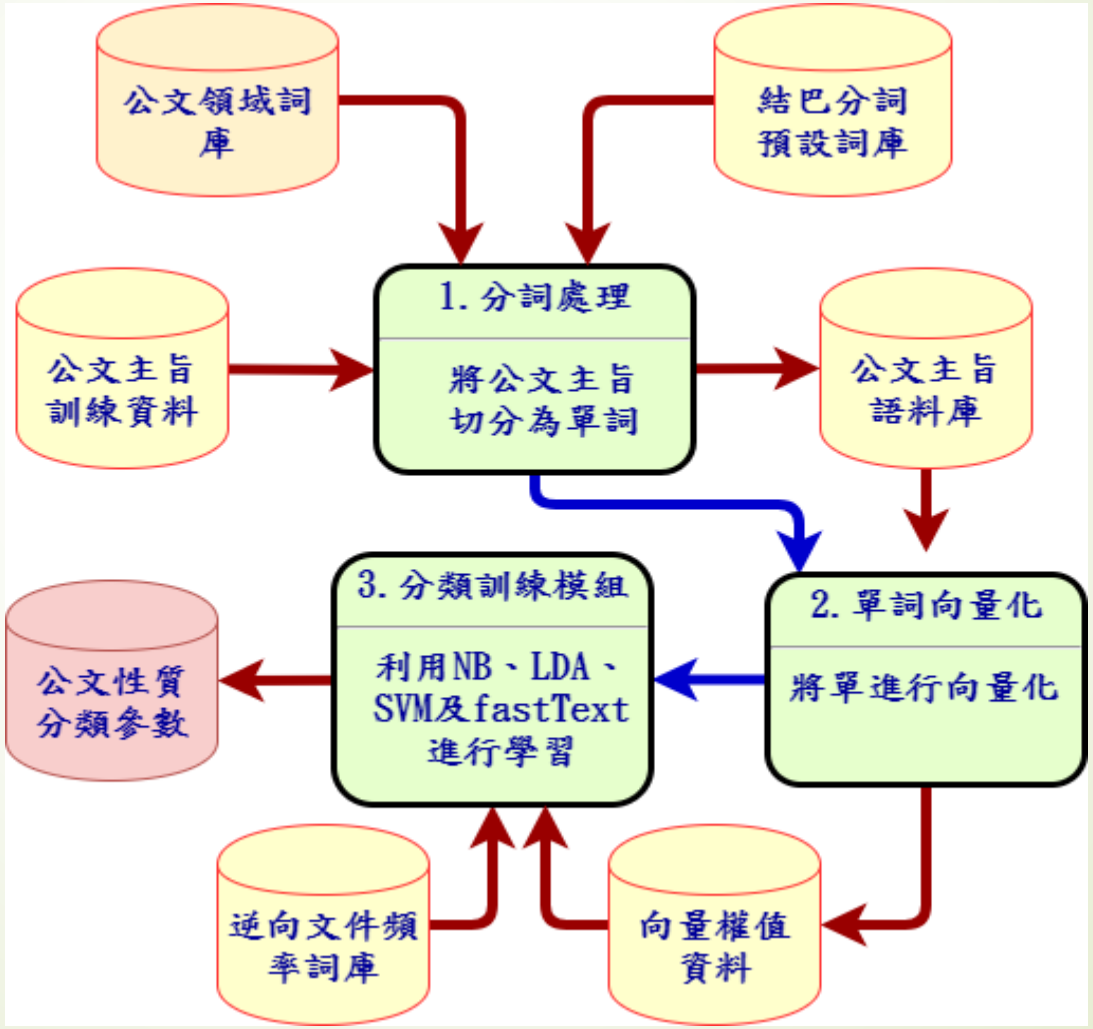
前期實驗—系統架構


步驟2
步驟3
資料預處理模組
1.分詞
2.單詞向量化
領域詞庫
預設詞庫
向量權值
資料
分類模型訓練模組
分別用NB、LDA、
SVM及fastText
演算法進行學習
訓練
資料
步驟1
步驟4
領域詞庫模組
1.先進行分詞
2.計算詞頻
3.建立領域詞庫
公文主旨
歷史資料
驗證資料
模型評估模組
1.將公文進行分類
2.以正確率及回復
率進行效能評估
公文性質
分類參數
訓練資料
驗證資料
評估結果
分類結果

前期實驗—領域詞庫建立方法



前期實驗—分類模組


公文領域詞庫
結巴分詞預設詞庫
公文主旨訓練資料
1. 分詞處理
將公文主旨切分為單詞
公文主旨語料庫
2. 單詞向量化
將單進行向量化
3. 分類訓練模組
利用NB、LDA、SVM及fastText進行學習
公文性質分類參數
逆向文件頻率詞庫
向量權值資料

前期實驗—各方法分類效能 (利用2013-2015年公文資料進行分類)

分類模型	資料切割	訓練量	測試量	正確率	精確率	回復率
LDA	十折法	637,535	71,285	0.946	0.893	0.893
SVM		637,535	71,285	0.998	0.997	0.997
NB		637,535	71,285	0.920	0.841	0.841
FastText		637,535	71,285	0.996	0.992	0.992
LDA	隨機抽樣	47,058	708,820	0.924	0.849	0.849
SVM		47,058	708,820	0.892	0.785	0.785
NB		47,058	708,820	0.936	0.872	0.872
FastText		47,058	708,820	0.992	0.992	0.992

前期實驗—各方法預測效能 (利用建立模型預測2016年公文資料)

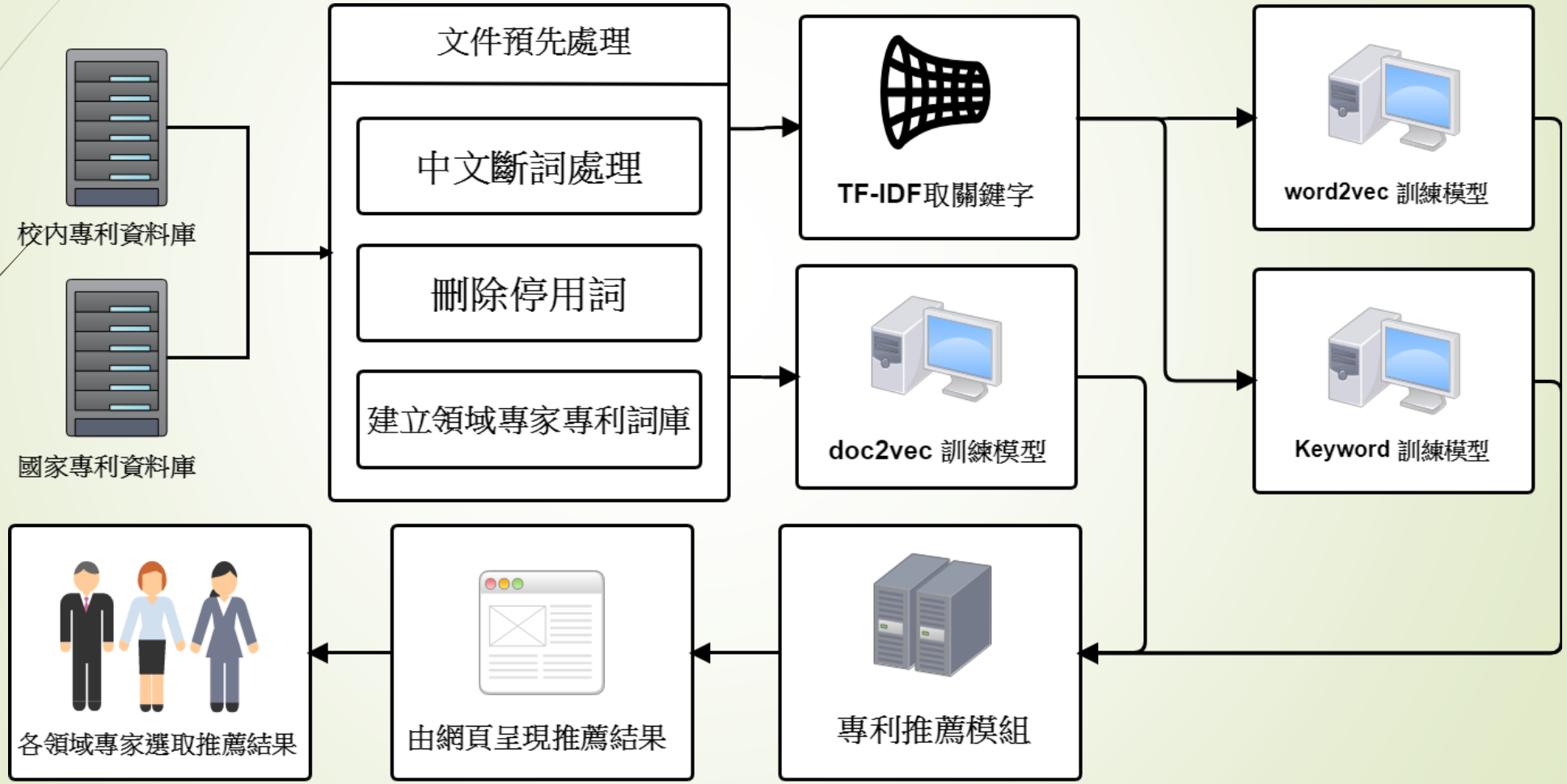
分類模型	測試資料量	正確率	精確率	回復率
LDA	272,941	0.908	0.817	0.817
SVM	272,941	0.909	0.819	0.819
NB	272,941	0.945	0.889	0.889
FastText	272,941	0.912	0.823	0.823

專利文件推薦系統

系統概述

- 應用於屏科大專利技術媒合平台，提供查詢、管理專利文件時之輔助。
  http://patech.npust.edu.tw/
- 將校內專利(863筆)與國家專利文件(178萬筆)進行相似度比對並推薦最相關專利給查詢者。
- 核心技術
  - 文件前處理
    - 詞向量技術(word2vec)、文本向量技術(doc2vec)
  - 相似度計算
    - 產生專利文件的向量模型，利用cosine相似度進一步運算出專利相似度高的專利。

系統架構


校內專利資料庫
國家專利資料庫
文件預先處理
中文斷詞處理
刪除停用詞
建立領域專家專利詞庫
TF-IDF取關鍵字
doc2vec 訓練模型
word2vec 訓練模型
Keyword 訓練模型
專利推薦模組
由網頁呈現推薦結果
各領域專家選取推薦結果

文件預先處理

- 專利文件本身是沒有經過特殊處理的文件檔，本研究分析的標的包含專利標題及專利摘要。

```
<twpat>
    <專利編號>201644285</專利編號>
    <專利名稱>揚聲器;SPEAKER</專利名稱>
    <公告>20161216</公告>
    <申請日>20150604</申請日>
    <申請號>104118129</申請號>
    <國際分類>H04R-009/04(2006.01);H04R-009/06(2006.01)</國際分類>
    <公報卷期>1424</公報卷期>
    <發明人名>楊宗隆 YANG, BILL</發明人名>
    <專利權人名>
        <name>固昌通訊股份有限公司 COTRON CORPORATION</name>
        <addr>臺北市士林區承德路4段150號12樓 TW</addr>
    </專利權人名>
    <代理人>
        <name>葉璟宗</name>
        <name>詹東穎</name><name>劉亞君</name>
    </代理人>
    <摘要>
        一種揚聲器，包括一殼體、一初級線圈、一振膜以及一導電環。殼體具有一腔室。初級線圈固定於殼體，用以接收外部音源訊號。振膜固定於殼體且
        位於腔室。導電環可以是次級線圈，其固定於振膜且僅接觸振膜，用以感應初級線圈的磁場變化而帶動振膜振動。
    </摘要>
</twpat>
</patent>
```

文件預先處理(cont.)

- 斷詞
  - 使用Python為基礎的「結巴斷詞」系統
- 刪除停用詞
  - Stop Words(停用詞) 例如：的、了、之。
  - 透過濾除無異議之字詞，訓練後的效果也更加準確
- 建立領域專家專利詞庫
  - 因專利經常包含專業術語與特殊名詞，因此為提升斷詞的準確度，透過尋找領域專家協助建置該專業領域之詞庫，本研究再加以整合成專利領域詞庫。

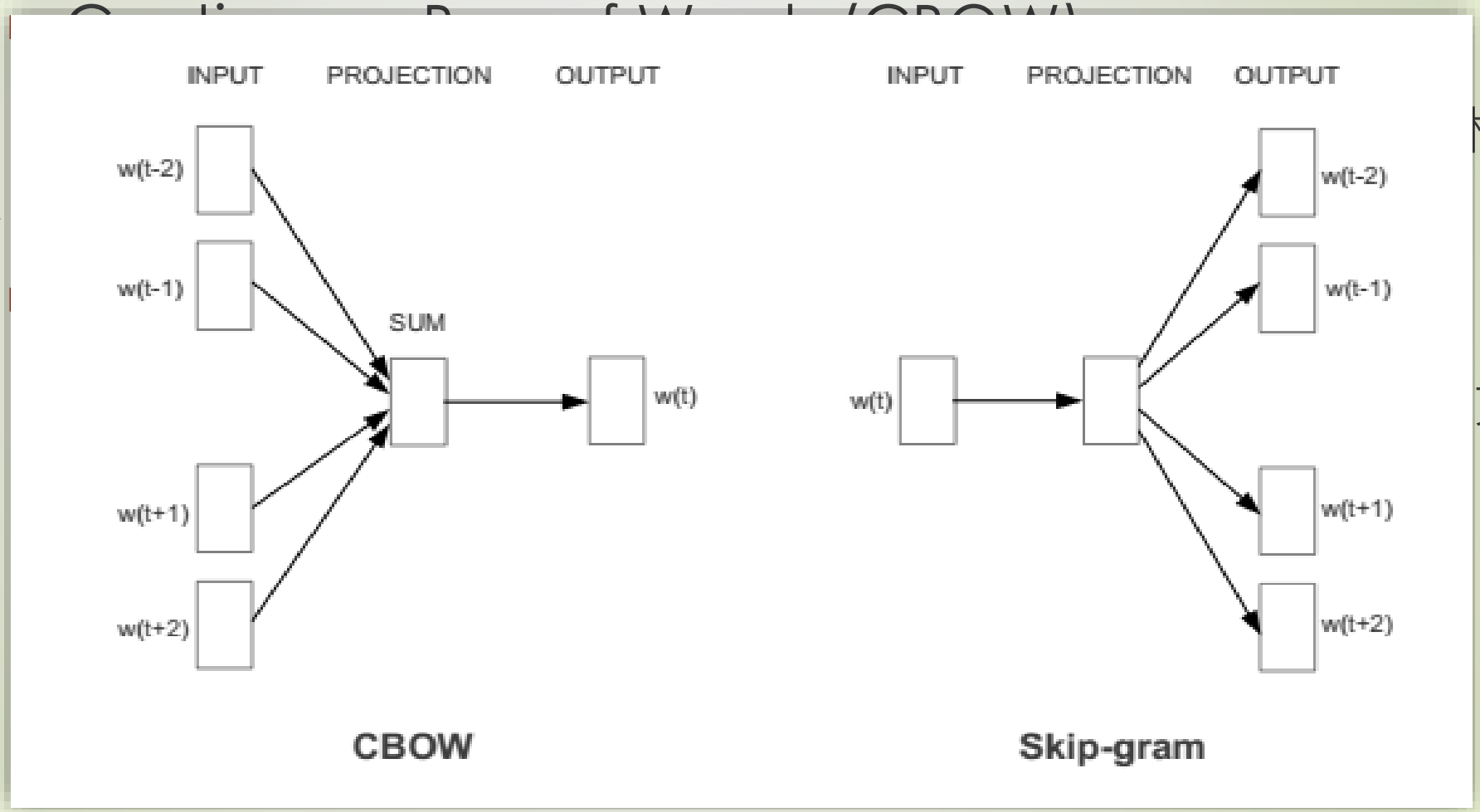
詞頻分析(TF-IDF)

- TF-IDF可用以計算文件中的關鍵詞，已被證明在訊息檢索領域中具有非常高的可信度，廣泛運用在大數據與文件分析。
- TF (Term Frequency)
  TF = 該詞出現次數 / 該專利文件所有單詞
- IDF (Inverse Document Frequency)
  IDF = 該詞在所有專利文件中出現的頻率
- TF-IDF = TF * IDF =可以產生出高權重的TF-IDF

Word2vec(詞向量)

- 詞向量廣泛應用於自然語言處理(NLP)領域。
- 基本概念
  - 利用一個向量來表示每一個詞，如此就能把一段由許多詞組成的文句，轉換成一個個詞向量來表示，並把這樣數值化的資料，送到模型裡做後續的應用。
  - 後續常利用Cosine Similarity計算文本相似度。

Word2vec(詞向量) (cont.)



Word2vec(詞向量) (cont.)

- 訓練資料
  - 我們 去 中正 讀 碩士班 好不好？
  - 我們 去 中山 讀 碩士班 好不好？
  - 我們 去 台大 讀 碩士班 好不好？
  - 我們 去 元智 讀 碩士班 好不好？
- 預測資料
  - 我們 去 屏科 讀 碩士班 好不好？

Doc2Vec(文本向量)

- 基於Word2Vec，Doc2vec給文本配置了向量，並在訓練過程中更新，將連續的句子或者是整個文本計算成一定的向量值進而比較出文章與文章間的相關程度。
- 可以獲得sentences/paragraphs/documents的向量。
- 影響模型準確率的因素
  - 語料庫大小、文檔數量

Doc2Vec(文本向量)(cont.)

- dbow (distributed bag of words)

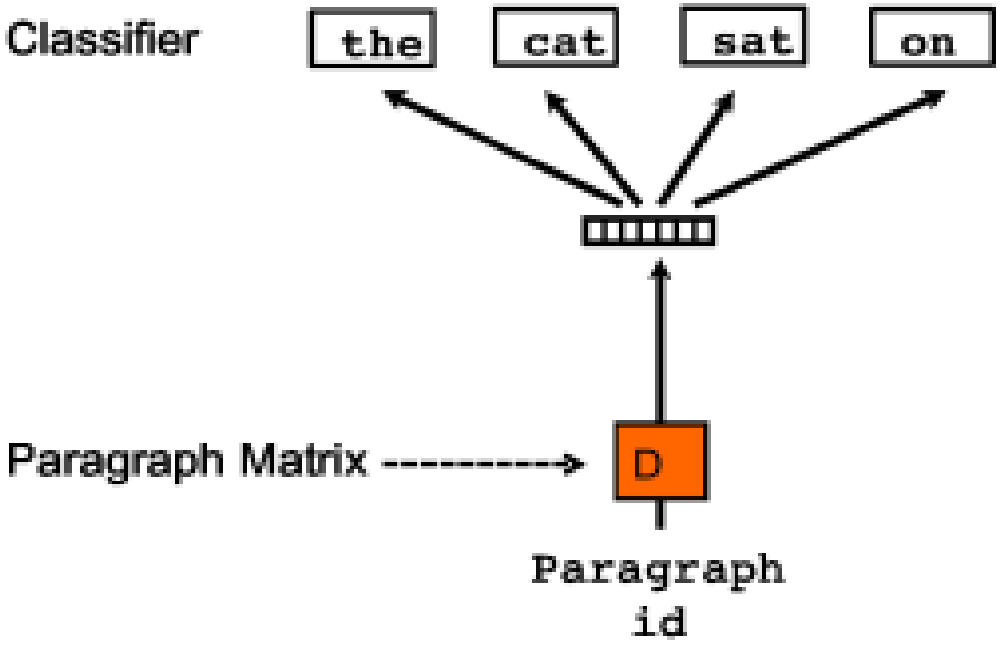


Figure 3. Distributed Bag of Words version of paragraph vectors. In this version, the paragraph vector is trained to predict the words in a small window.

Doc2Vec(文本向量)(cont.)

- dm (distributed memory)

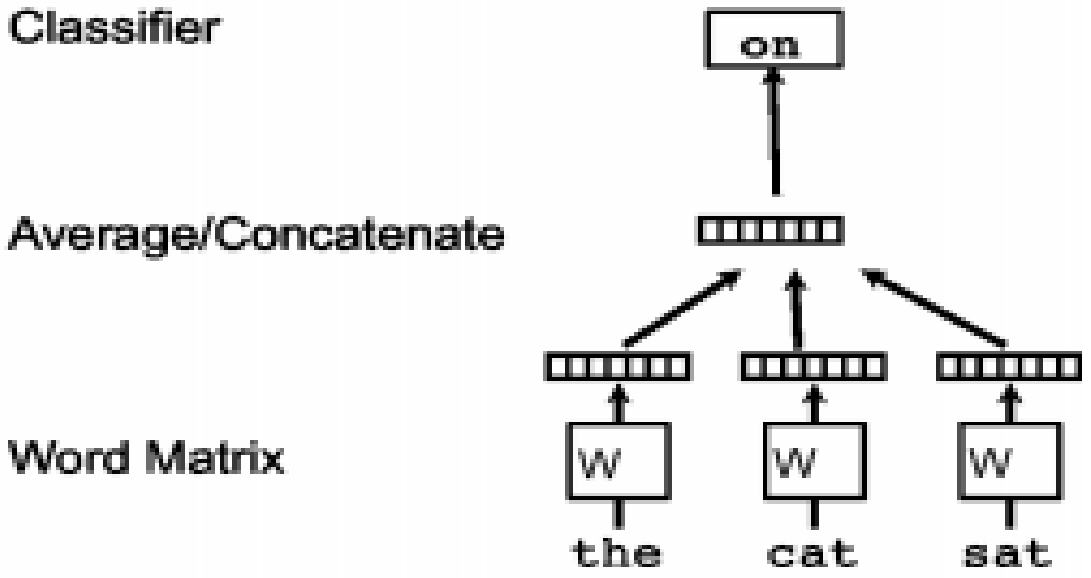
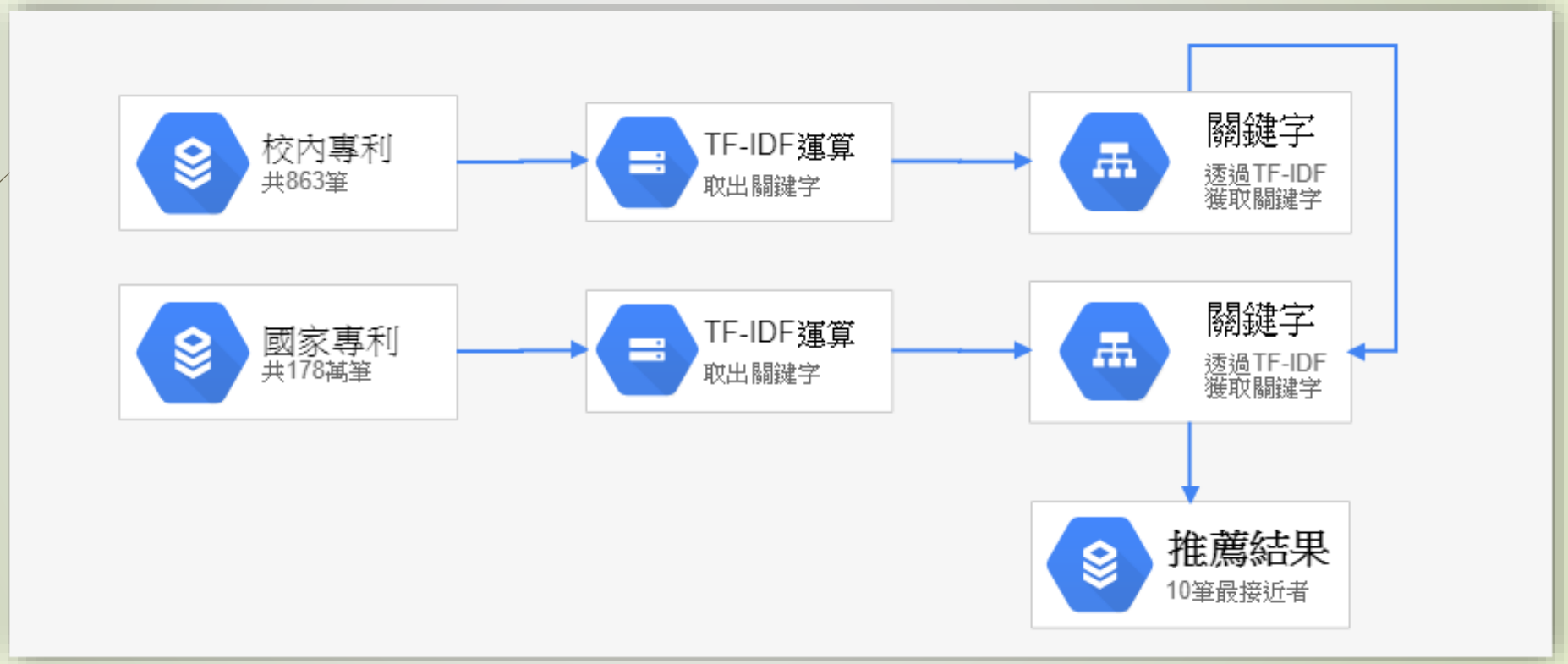


Figure 1. A framework for learning word vectors. Context of three words ("the," "cat," and "sat") is used to predict the fourth word ("on"). The input words are mapped to columns of the matrix W to predict the output word.

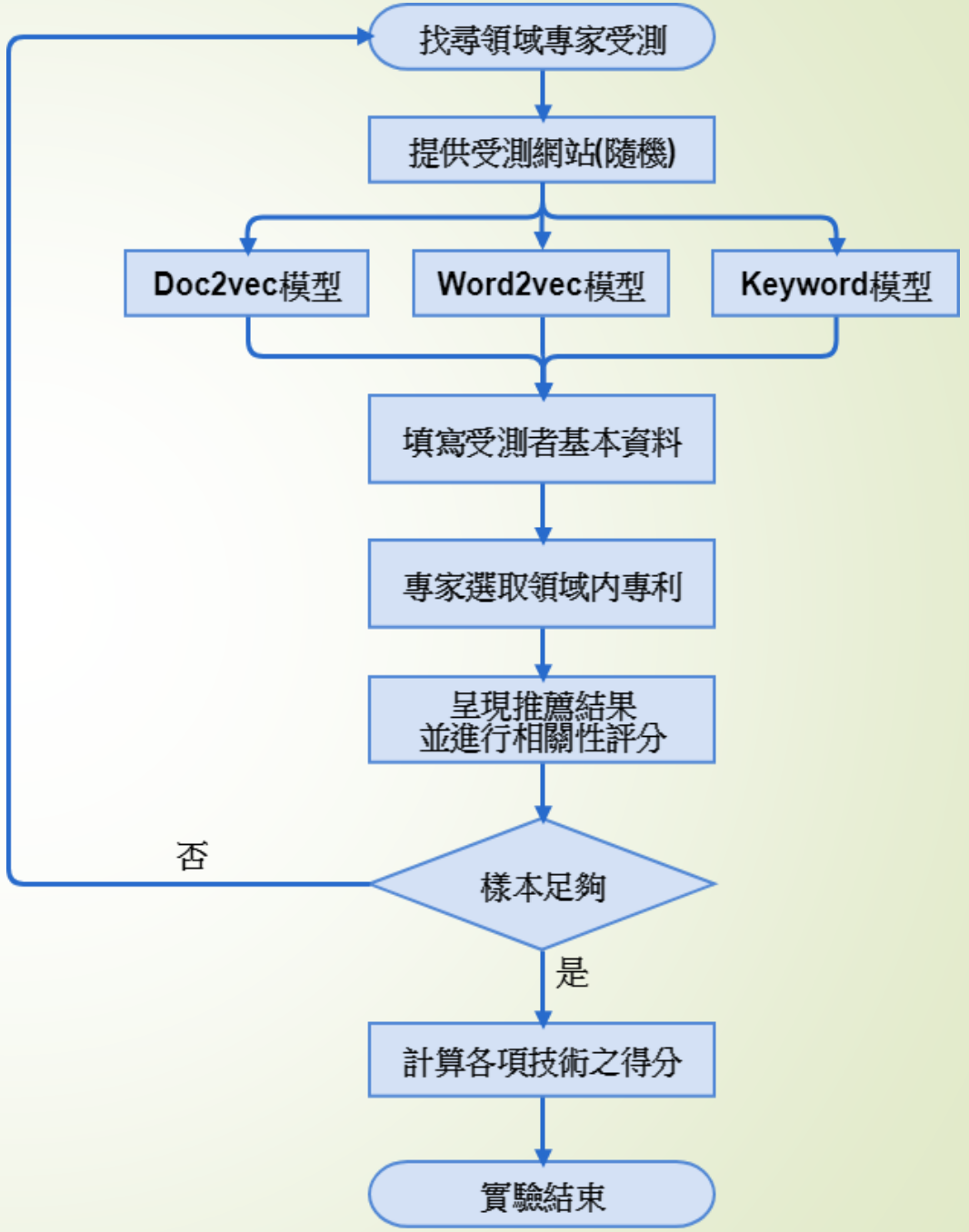
推薦模型1：Keyword模型



推薦模型2：Word2Vec模型

推薦模型3：Doc2Vec模型

實驗流程


找尋領域專家受測
提供受測網站(隨機)
Doc2vec模型
Word2vec模型
Keyword模型
填寫受測者基本資料
專家選取領域內專利
呈現推薦結果
並進行相關性評分
樣本足夠
否
是
計算各項技術之得分
實驗結束

實驗結果(1/3)

技術類型	總平均	受測者數量
Doc2Vec	3.396	36
Word2Vec	2.702	32
Keyword	2.076	34

實驗結果(2/3)

技術類型	總平均	受測者數量
未導入詞庫之Doc2Vec	2.937	36
導入詞庫之Doc2Vec	3.396	36

實驗結果(3/3)

- 透過相關分析得知Doc2Vec的權重與使用者評分具有相關性。

	Score	ModelWeight
1	5.0	.9096143
2	5.0	.9285521
3	5.0	.9309729
4	4.0	.9113946
5	5.0	.9415802
6	5.0	.9452369
7	4.0	.9046406
8	2.0	.9075798
9	4.0	.8912207
10	4.0	.8953187

相關

		相關性得分	相關權重
相關性得分	Pearson 相關	1	.321**
	顯著性 (雙尾)		.000
	個數	358	358
相關權重	Pearson 相關	.321**	1
	顯著性 (雙尾)	.000	
	個數	358	358

**. 在顯著水準為0.01時 (雙尾)，相關顯著。

Q & A