圖書資訊學研究19：1 (December 2024)：1-36
DOI:10.30177/JLISR.202412_19(1).0001


# 以集成學習方法改善LDA主題建模穩定性之研究

# Improving the Stability of LDA Topic Modeling With Ensemble Learning

**林頌堅***

**Sung-Chien Lin**

世新大學資訊傳播學系副教授
Associate Professor
Department of Information and Communications
Shih Hsin University

## 【摘要 Abstract】

主題建模可以應用在自動分析大量文本資料的主題結構。雖然已經發展出許多新的技術，但是主題建模的穩定性仍然是值得關注的問題之一。本研究針對Latent Dirichlet Allocation (LDA)主題建模演算法的穩定性，提出一個集成學習式的改善方法。這個方法的第一階段將從多個基本模型中選取穩定出現的有效主題，第二階段則根據這些有效主題的詞語出現資訊進行引導式主題建模產生穩定性較高的改良模型。本研究並以兩個不同的語料庫進行測試，結果證明所建議之改善方法能夠提高穩定性。並且為了節省整體的建模時間，本研究建議在建立每一個基本模型時只選用語料庫中的部分文本資料，實驗結果表示少量比例的文本資料即可獲得不錯的穩定性分數。

Topic modeling enables the rapid discovery of latent thematic structures within large amounts of unstructured textual data. Although many new techniques have been developed, the stability of topic modeling remains one of the noteworthy concerns. This study focuses

*通訊作者：林頌堅 scl@mail.shu.edu.tw
投稿日期：2023年9月13日；接受日期：2023年11月29日

on Latent Dirichlet Allocation (LDA) topic modeling and proposes a two-stage ensemble learning approach to improve the stability of topic modeling. The first stage of this method involves selecting stable and meaningful topics from multiple basic models, while the second stage utilizes the word occurrence information of these selected topics to guide the creation of improved models with higher stability. This research was tested on two different corpora, and the results demonstrate that the proposed improvement method can consistently improve the stability of models. Furthermore, in order to save overall modeling time, this study suggests sampling only a subset of the textual data from the corpora when building each basic model, as experimental results indicate that a small proportion of textual data can also yield satisfactory stability scores.

**【關鍵詞 Keywords】**

集成學習；潛在狄利克里分配（LDA）；穩定性；主題建模

Ensemble Learning; Latent Dirichlet Allocation (LDA); Stability; Topic Modeling

# 壹、前言

主題建模（topic modeling）是一種非監督式（unsupervised）的文本探勘方法（Chauhan & Shah, 2022）。這項方法假定語料庫（corpus）內的文本以及可觀察的詞語之間有一組隱藏的變數，稱為主題（topics），每個主題由一組彼此語意相關的詞語所組成，並且文本中大多只包含少數主題。主題建模的程序便是在輸入語料庫後，根據文本上出現的詞語次數，利用各種演算法推導出描述主題上的詞語組成以及各文本上的主題成分比例等模型參數。歷年來發展的主要主題建模演算法可參考表1。

由於可以迅速從大量非結構性的文本資料發現當中蘊含的主題結構，國內外已經有相當多主題建模應用的研究，包括提供研究人員分析語料庫內所有文本的主題，或是作為資訊檢索、情感分析等後續自然語言處理任務的文本特徵。國外的研究諸如應用在傳播領域之探討各種媒體（Elgesem, Steskal, & Diakopoulos, 2015; Jacobi, van Atteveldt, & Welbers, 2016）上討論的公共議題，或是追蹤新聞事件的發展（Kim & Oh, 2011），商業管理領域則有分析網路評論上對於產品設計與功能的評價與口碑（Tirunillai & Tellis, 2014），以及發現電影等娛樂產品的心

**表1**

**歷年發展的各種主題建模演算法**

| 模型參數推導策略 | 演算法名稱與簡稱 | 參考文獻 |
|---|---|---|
| 矩陣分解 | 潛在語意索引（latent semantic index, LSI） | Deerwester, Dumais, Furnas, Landauer, & Harshman（1990） |
| | 非負矩陣分解（non-negative matrix factorization, NMF） | Lee & Seung（1999） |
| 統計推論 | 機率式潛在語意索引（probabilistic latent semantic index, PLSI） | Hofmann （1999） |
| | 潛在狄利克里分配（latent Dirichlet allocation, LDA） | Blei, Ng, & Jordan （2003） |
| 神經主題建模（neural topic modeling, NTM） | ProdLDA<br>Top2Vec<br>BERTopic | Srivastava & Sutton （2017）<br>Angelov（2020）<br>Grootendorst（2022） |

理主題特徵（psychological thematic features）並探討其與消費之間的關係（Toubia, Iyengar, Bunnell, & Lemaire, 2019）。以國內的研究而言：傅文成、黃瑎戩與顏瑞宏（2021）使用LDA分析新聞平臺上有關美軍在臺海周圍軍事巡弋活動新聞的民眾回文，瞭解回文中所承載的主題，並進而分析主題上的字詞使用頻率與情感偏向，剖析網路民意在「對美國戰略感知」、「兩岸軍事衝突風險」及「情感」等三項維度中的構成態樣。陳怡璇與劉桂君（2023）則蒐集臺灣疫情初期電子布告欄系統（bulletin board system, BBS）上有關COVID-19的發文與回文，利用LDA與網絡分析方法探勘網友的論述框架，以瞭解民眾的風險感知以及對風險的解讀。郝沛毅、歐仁彬、黃天受、林振穎與吳建生（2018）以LDA從新聞文章中萃取出隱含的主題模型，結合文章中的情緒資訊，輸入模糊支持向量機（fuzzy support vector machine）預測股價的漲跌趨勢。邱志洲、吳忠敏、簡德年、高凌菁與邱德生（2023）使用LDA分析病患的臨床紀錄、診斷資料及檢驗報告，結合電子健康紀錄中結構化資料，運用機器學習模型進行臨床結果預測。

然而在實際的應用上，除了LSI之外，LDA與其他主題建模演算法都有穩定性（stability）的問題（Abdelrazek, Eid, Gawish, Medhat, & Hassan, 2023），也就是在使用相同的參數設定，並且對相同語料庫重複進行多次建模的情況下，每次獲得的結果可能不一致。造成這些主題建模演算法的結果不穩定是因為它們使用了隨機性策略（stochastic strategies）從語料庫推導模型參數（Maier et al., 2018）。以LDA主題建模為例，LDA主題模型是一種描述語料庫內文件生成的機率模型，模型上的參數包含每一個文件上的主題組成比例以及構成主題的詞語機率分布。由於從語料庫推導的真實LDA主題模型參數是一個難解的問題（an intractable problem），實作上LDA主題建模技術通常假設主題數目為一個固定整數值，並利用變分推論（variational inference）（Blei et al., 2003）和吉布斯抽樣（Gibbs sampling）（Griffiths & Steyvers, 2004）等統計推論演算法找出模型參數的近似解（Yang, Pan, Lu, Topkara, & Song, 2016）。但因為在進行統計推論的過程中，需要使用隨機種子（random seed）產生每次建模初始的機率分布（Lancichinetti et al., 2015）以及重新推估較佳的模型參數（Koltcov, Nikolenko, Koltsova, Filippov, & Bodrunova, 2016），即便是相同的主題數目並且針對相同的語料庫，不同的隨機種子仍會讓每次建模的主題模型收斂到不同的局部最佳解（local optima）上，造成不穩定的情形（Yang et al., 2016）。

如果使用的主題建模演算法穩定性不高，文本內容分析結果的信度將會受到質疑（Ballester & Penner, 2022; Maier et al., 2018）。利用主題建模的結果於資訊檢索時，也會因為每次產生的主題標示可能不一致，影響使用者對資訊檢索系統的信任（Yang et al., 2016）。近年一些研究者開始著手解決主題建模的穩定性問題，根據主題建模過程上可能發生不穩定的各種原因提出改善方法。這些方法主要以LDA主題建模為對象，但也有研究對NMF的穩定性進行改善（Belford, Mac Namee, & Greene, 2018）。有些方法針對特定主題建模演算法，有些則在調整後可以適用於不同種類的演算法。目前常見的主題建模穩定性改善方法包括改變模型參數的推斷方法和選取較穩定的主題兩種主要策略。前者從語料庫中的詞語集群（term clustering）資訊，作為主題的初始資訊或後續的參數調整，引導主題模型參數的推斷（Koltcov et al., 2016; Lancichinetti et al., 2015; Maier et al., 2018）。但因為對詞語進行集群，每個詞語的分布將集中在某一個主題上，此一策略違反了主題建模方法中詞語可以出現在不同主題上的假設。後者以相同的參數設定，對同一語料庫重複建立多個主題模型，利用集群演算法（clustering algorithm）選取較穩定主題（Chuang et al., 2015; Mantyla, Claes, & Farooq, 2018）。然而這種策略沒有直接推導文本的主題成分比例，造成後續文本分析應用的不便與困難。為了解決前述兩種策略的問題，本研究嘗試整合它們的優點，提出針對LDA主題建模的兩階段集成式穩定性改善方法，本論文將實際驗證此一方法的可行性，並且也將討論此一方法對於其他主題建模方法的適用性。此外，由於本研究提出的方法運用集成學習的概念，需要比傳統LDA主題建模花費更多的計算時間。為了克服此一問題，本研究嘗試在第一階段只隨機取出語料庫中的部分文件進行建模。因此，也將針對這個技巧進行實驗與討論。

以下各節的內容包括：「文獻探討」先探討過去對於主題建模穩定性的研究以及各種改善方法。「兩階段主題建模的整合式穩定性改善技術」說明本研究所提出集成學習式主題建模改善方法與其特點。為了驗證本研究提出的方法可以改善LDA主題建模的穩定性，本研究進行了一序列的實驗，「研究過程與結果」即是說明這些實驗所使用的語料庫、實驗過程與結果。最後的「結論」總結本研究的發現並提出未來可以繼續研究的方向。

# 貳、文獻探討

本節首先簡要說明本論文所使用的術語與符號。為了比較主題建模穩定性的改善成效，需要先能夠測量主題建模的穩定性，因此接下來將先說明主題建模穩定性測量方法的相關研究，最後討論過去研究曾提及的主題建模穩定性改善方法。

## 一、本論文使用的術語與符號

假定語料庫C內共包含$D$個文本資料，以$w_{d1}, w_{d2}, ..., w_{dn}, ..., w_{dN_d}$)，表示第$d$個文本資料，其中$w_{dn}$是指這個文本的第$n$個詞語，而$N_d$則是第$d$個文本的長度，也就是這個文本上的詞語數目。此外，語料庫的詞彙量（vocabulary size）為$V$，也就是語料庫中包含$V$種詞語。

進行LDA主題建模時，研究者可以假定語料庫中具有$K$個主題，並且文本資料都是根據模型上的兩組參數產生。一組是由每個文本的主題組成比例$\theta_d$所構成的矩陣 $\theta = [\theta_1 \ \theta_2 \ ... \ \theta_d \ ... \ \theta_D]^T$，$\theta_d$是一個$K$維的向量，$\theta_{dk}$代表第$k$個主題在這個文本上的組成比例。另一組參數是代表語料庫中$K$個不同的主題，$\phi = [\phi_1 \ \phi_2 \ ... \ \phi_k \ ... \ \phi_K]^T$，每一個$\phi_k$的維度是$V$，表示語料庫中的詞語在第$k$個主題上出現機率。在每一個主題上的前面幾個出現機率較大的詞語稱之為這個主題的主題描述語（topic descriptor）。在LDA主題建模方法假設模型中的$\theta_d$與$\phi_k$參數分別是由超參數（hyperparameters）為α和β的狄利克里分布（Dirichlet distribution）抽出的多項分布。產生語料庫上的第$d$個文本時，以第$n$個詞語$w_{dn}$為例，首先自多項分布$\theta_d$抽取出這詞語的主題$z_{dn}$，再由主題對應的詞語分布$\phi_{z_{dn}}$選出最有可能的詞語$w_{dn}$。

LDA主題建模便是經由語料庫文本上出現的詞語次數以及給定的主題數目與α和β等超參數推斷主題模型的參數θ和ϕ以及每個詞語被指定的主題$z$。由於這是相當難解的問題，所以目前大都是利用吉布斯抽樣（Griffiths & Steyvers, 2004）或是變分推斷（Blei et al., 2003）求得近似解。以吉布斯抽樣來說，首先以隨機的方式指定語料庫中每一個文本上各個詞語的主題，並且給定的超參數α和β，估計θ和ϕ，然後基於θ和ϕ的條件分布（conditional distributions）重新指定每一個詞語的主題，根據詞語所指定的主題，再次估計θ和ϕ，一直重覆到收斂或指定次數為止。由於每次進行LDA主題建模，程式一般預設使用不同的隨機種子，所以每次每個詞語一開始隨機指定的主題便不相同。另外重新將文本上的詞語基於條件分布指定主題，估計較精確的模型參數值，也是一個隨機過程（Maier et al., 2018）。因此，每一次的建模推斷出的結果將會有某種

程度的差異。根據以上討論，可以瞭解穩定性是應用LDA主題建模於文本分析時需要考慮的重要問題。

## 二、主題建模穩定性測量的相關研究

在測量主題建模的穩定性時，通常是對相同的語料庫並且以相同超參數進行多次主題建模，產生多個主題模型，然後比較每一對模型，計算它們之間的一致性分數（agreement score）。

由於每次建模所得到的主題次序與內容不大可能完全相同，在模型A上編號為$k$的主題可能與模型B上同樣編號$k$的主題相差很大，但與編號$k'$的另一個主題較相似。因此在估計模型之間的一致性，de Waal與Barnard（2008）、Greene、O'Callaghan與Cunningham（2014）、Yang等人（2016）及Belford等人（2018）建議先利用匈牙利演算法（Hungarian algorithm）（Kuhn, 1955）進行主題對齊（topic alignment），找出兩個模型上相似主題的最佳配對組合。在主題對齊之後，de Waal與Barnard（2008）利用文件上的主題組成比例（也就是$\theta_d$），比較兩個模型之間在所有文件上的相關性，以相關性的高低表示一致性的大小。Yang等人（2016）計算一致性分數的方式則是在主題建模後，先定義文件與詞語所指定的主題標示為該文件或詞語最大組成比例的主題；然後在主題對齊後，比較兩個模型上文件或詞語標示相同的比例。Greene等人（2014）及Belford等人（2018）將主題間的相似性則定義為主題描述語（在主題上出現機率最大的前幾個詞語）之間的Jaccard分數。進行主題對齊之後，將兩個模型之間的一致性最佳配對主題描述詞之Jaccard分數取平均值。

也有研究提出不需進行主題對齊的一致性估計方法。Ballester與Penner（2022）利用文件上的主題組成比例測量語料庫中任何一對文件之間的相似性，然後計算所有模型所測得相似性的標準差（standard deviation），標準差的值如果很大表示該對文件在不同模型的相似性計算結果相差很大。最後以所有文件對的相似性標準差之平均值作為主題建模的穩定性估計值。Maier等人（2018）測量任何一對主題之間詞語出現機率的餘弦值超過0.7的比例作為一致性的估計分數。Belford等人（2018）除了前述以主題對齊最佳配對主題描述語之Jaccard分數平均值之外，還提出其他兩種一致性估計方法：一種是以所有主題的主題描述語集合作為模型的代表特徵，利用兩個模型的集合差（set difference）作為它們之間的差異。如果兩個詞語集合相差愈大，這對模型的差異愈大，也就是一致性愈小。另一種估計方法則是將主題模型上每個文件

標示的主題（也就是該文件的最大成分主題）視為文件集群（document clustering）的結果，利用比較集群結果相似性常用的正規化交互資訊（normalized mutual information）（Strehl & Ghosh, 2002）作為主題模型之間的一致性分數。

根據以上的討論，本研究依據「是否需要先進行主題對齊」與「使用的主題模型參數」等兩種條件，將上述研究提出的各種穩定性測量方法整理為表2。但事實上在前人的研究當中，不論是上述的哪種方法都可以發現主題建模的穩定性是相當嚴重的問題，需要提出改善的技術或方法。

在各種測量模型之間的一致性方法當中，Greene等人（2014）及Belford等人（2018）計算主題對齊後主題之間的相似性平均值是較直觀且容易實作的方法，因此本研究選擇此一方法。但由於利用Jaccard分數對主題上的詞語出現機率計算時需要先行決定代表各個主題的關鍵詞語數目，然而在不同語料、甚至不同主題當中，關鍵詞語的數目並不能夠事先確定。因此本研究改採用餘弦測量（cosine measure）計算主題相似性。

## 三、主題建模穩定性改善的相關研究

近年來，研究人員嘗試運用多種不同的經驗法則（heuristics）作為策略來改善主題建模的穩定性。由於各種建模方法大多使用隨機性策略來推斷模型的參數，有些研究認為這是造成主題建模不穩定的主要原因，因此使用較穩定的模型參數推斷策略來改善；另一些主題建模穩定性的改善方法則是緣由於許多研究人員觀察到在多次建模的結果當中，有些主題在各個模型之間只有略微不同，但有些主題則只出現在相當少數的模型上，因此嘗試透過選取在多次建模結果中較穩定的主題或模型來達成改善主題建模穩定性的目的。

在改變模型參數推斷方法的穩定性改善策略當中，Agrawal、Fu

**表2**

**主題建模穩定性測量方法之分析**

| 一致性估計方法分類 | 使用文件的主題組成比例（$\theta_d$） | 使用主題上的詞語出現機率（$\phi_k$） |
|---|---|---|
| 先行進行主題對齊 | de Waal & Barnard（2008）；Yang et al.（2016） | Greene et al.（2014）；Belford et al.（2018） |
| 不需進行主題對齊 | Belford et al.（2018）；Ballester & Penner（2022） | Belford et al.（2018）；Maier et al.（2018） |

與Agrawal（2018）將LDA主題建模的統計推斷方法改以利用基因演算法（genetic algorithm）中的差分演化演算法（differential evolution algorithm），尋找較佳模型的超參數，包括$K$、θ和ϕ等，增加不同次產生的LDA主題模型上的主題相似性。Lancichinetti等人（2015）及Maier等人（2018）的構想是依據主題模型的概念：主題是由一組語意相關的詞語所構成，這些詞語經常在文本中共同出現。因此他們建議可以預先利用集群演算法聚集彼此經常共同出現的詞語形成群組，然後在進行主題建模預設詞語在主題上的出現機率分布（也就是ϕ）時，利用共同出現的詞語群組，以較佳的起始值來避免建模開始時的不確定性。同樣是利用詞語共同出現的資訊來改善主題建模穩定性，在Koltcov等人（2016）的方法中則是將共同出現的詞語群組運用在指派主題的過程。當重新將文本上的詞語基於條件分布指定主題，估計較精確的模型參數值時，不僅只針對當前文本上的詞語指定主題，也同時調整這個詞語所屬群組的其他相關詞語，藉以避免在進行模型推斷時落入局部極大值（local maximum）。Yang等人（2016）認為在進行多次主題建模的過程當中，每次文件的主題標示（文件最大組成比例的主題）應該都能夠產生相同的結果。也就是標示為同一主題的一組文件，在重新建模之後，仍然可以標示為相同的主題，並且任何一對標示為不同主題的文件，也能夠維持不同的主題標示。因此他們建議利用成對文件之間的關係作為穩定性限制加入主題建模的過程中，藉以減少模型重建前後同一文件標示為不同主題的不穩定情形。在重新估計模型的文件的主題組成比例（θ）時，將標示為要保留主題的文件的主題組成比例作為增強的部分，並且以標示為其他主題且與這些文件差異最大的文件之主題組成比例作為減弱的部分。在實際運作上，Lancichinetti等人（2015）、Maier等人（2018）、Koltcov等人（2016）及Yang等人（2016）的主題建模穩定性改善方法可以藉由提供詞語或文件的資訊進行引導式主題建模（guided topic modeling）（Jagarlamudi, Daumé, & Udupa, 2012）而達成。

有些研究人員觀察到在同一語料庫與相同參數下進行主題建模時，多數產生的模型上會有許多相當相似甚至相同的主題，僅有少數主題可能會分裂成兩個主題、因出現次數過少而消失，或與其他主題合併（Chen et al., 2020; Chuang et al., 2015; Mantyla et al., 2018）。Chen等人（2020）建議可以檢視與比較多次主題建模的結果，也就是「主題模型的集成」（an ensemble of topic models），藉由多個模型的主題所形成的空間中觀察主題之間的異同。具體而言，不同模型上相似的主題在空間中彼此靠近，不相似的主題則會彼此遠離。他們認為一群多個彼此

靠近的主題雖然在不同的模型上略有不同，但可能是來自同一個典型主題（archetype topic）。其他在空間上散落於廣闊區域的少數主題可能是源於演算法不完美而產生於少數模型上的計算產物（computational artifacts），也就是主題建模過程中偶然產生的雜訊，不是真正存在的主題，Chen等人（2020）認為這些與其他主題相去甚遠的主題可以忽略。因此，可以利用此一現象，作為主題建模穩定性的改善策略，例如Rieger、Rahnenführer與Jentsch（2020）從多次建模結果中選取一個和其他模型最一致的中心LDA模型。他們首先建立多個LDA主題模型，然後根據兩個模型之間彼此主題相似的程度，測量每一對模型之間的一致性，根據一致性的測量結果，選取中心LDA模型，將這個中心模型視為是最穩定的主題模型輸出。Chuang等人（2015）及Mantyla等人（2018）則嘗試找出穩定出現在多數模型上相似或相同的主題。他們分別利用凝聚階層集群（agglomerative hierarchical clustering, AHC）和*K*-中心點（*K*-medoids）演算法將不同模型內相似的主題聚集起來，將集群中聚集較多彼此相似的主題視為較穩定的主題集群，捨棄無法聚集成群的雜訊。與前述之各個針對LDA主題建模的研究不同，Belford等人（2018）的研究則是以NMF主題建模為穩定性改善對象，提出集成式NMF方法。Belford等人（2018）先對語料庫進行多次NMF主題建模，產生多對詞語主題矩陣和主題詞語矩陣，接著取出各次的主題詞語矩陣，集成為一個巨大的主題詞語矩陣，最後對此一巨大主題詞語矩陣進行NMF，最後所得的主題詞語矩陣，便是較穩定的主題建模結果。

圖1整理上述各種主題建模的穩定性改善方法。首先依照這種方法屬於改變模型參數的推斷方法或是選取較穩定的主題或模型進行區分。改變模型參數的推斷方法又再區分為利用基因演算法尋找最佳超參數組合，或是利用詞語、文件的相似資訊進行引導式主題建模，引導式主題建模又可分為根據詞語相似資訊預設或調整詞語在主題上的出現機率分布，或是根據文件相似性資訊調整文件的主題組成比例。在選取較穩定的主題或模型這類的方法，則分為選取中心模型或是以集成式方法聚集並發現較穩定的主題等兩種次分類。

## 參、兩階段主題建模的整合式穩定性改善技術

### 一、研究問題

從前一節對於前人提出之各種主題建模穩定性改善方法的討論中，本研究發現這些方法存在若干問題：

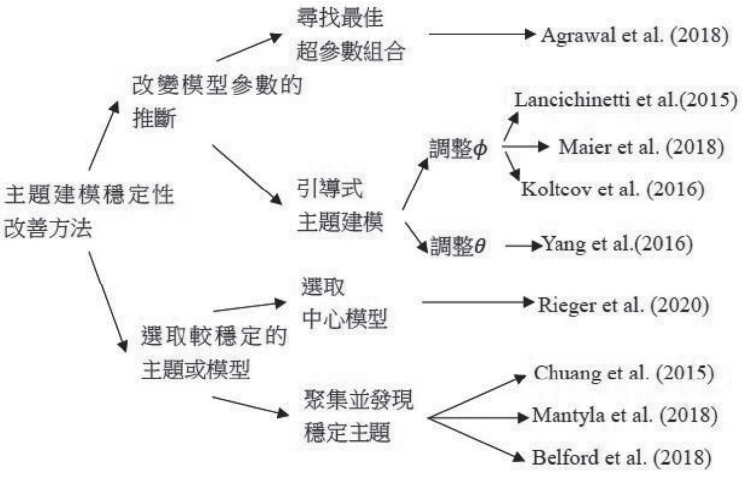


**圖1　主題建模的穩定性改善方法分析**

1. 以改變模型參數推斷的策略而言，目前Lancichinetti等人（2015）、Maier等人（2018）及Koltcov等人（2016）對語料庫中的詞語進行集群，根據集群結果，作為詞語在主題上出現機率分布的預設值或調整的參考資訊，進行引導式主題建模。但是在詞語的集群結果，每個詞語都只能指定在唯一的一個群組內。若是利用詞語群組引導主題建模的結果，此時詞語的分布將集中出現在某一個主題上。但主題建模認為某些詞語在不同主題上具有不同的意義，這些詞語可以會出現在多個不同主題當中。因此這種策略違反了原先主題建模的假設。
2. 主題建模在文本內容分析的應用，除了根據語料庫內包含之每個主題，瞭解整個語料庫的內涵之外，還包括透過比較文件上主題組成比例以及計算其在時間上的累積，藉以進行後續的應用，例如檢索相似的文件、文件群組，以及統計分析每個主題的趨勢、模式以及特定事件（Alexander & Gleicher, 2016）。因此完整的主題建模結果不僅包括詞語在各主題上的出現機率分布ϕ，以每一個主題上最關鍵的詞語，來分析語料庫包含的主題以及其意涵，也需要利用文件上的主題組成比例θ，作為文件的主題摘要，計算文件相似性與主題在時間上的分布。然而和Belford等人（2018）、Chuang等人（2015）及Mantyla等人（2018）只選取模型中較穩定主題，缺乏文件上各主題的組成比例，在後續應用上會有困難。

為此，本論文嘗試運用兩種策略的優點作為互補，提出一個兩階段

的整合式技術。第一階段稱為「有效主題選取」：首先建立多個主題模型，比對這些主題模型上所有主題的相似性，然後運用集群演算法，將彼此相似的主題聚集形成群組，區別穩定出現在多個模型上的有效主題和僅出現在極少數模型上的雜訊。第二階段稱為「引導式主題建模」：針對第一階段挑選出有效主題的群組，以群組成員組成代表各群組的主題特徵，取代先前研究常用的詞語集群結果，進行引導式主題建模，產生完整而較穩定的主題模型。簡而言之，這個方法的第一階段嘗試從多次主題建模結果中選取穩定出現的有效主題，在此一階段產生的結果，語料庫中的詞語依據它們在不同主題上的重要性，可以出現在多個有效主題當中，符合主題建模的假設，作為詞語在主題上出現機率分布預設值的參考資訊，在第二階段進行引導式主題建模，建立完整而較穩定的主題模型。圖2是這個方法的示意圖。雖然本研究以LDA主題建模作為第一階段的基本模型建模以及第二階段引導式主題建模的演算法，但是事實上這些部分可以運用其他的主題建模演算法，本節將說明如何在本研究所提出的主題建模改善方法中應用其他演算法。此外，由於在第一階段需要建立多個基本模型，比傳統LDA主題建模需要花費更多時間，因

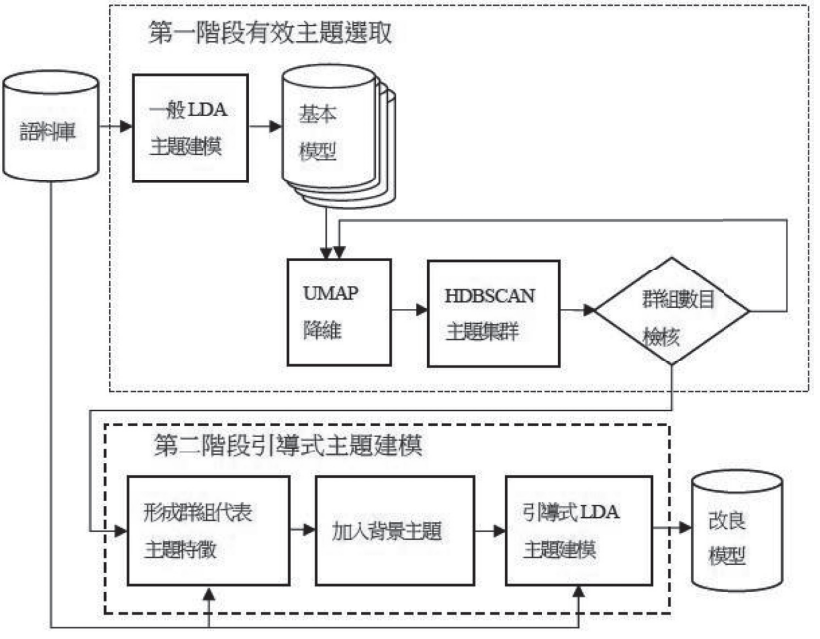


**圖2　本研究所提出的兩階段主題建模整合式穩定性改善技術示意圖**

此本節也將提出此主題建模穩定性改善方法的減少時間技巧。以下詳細說明每一階段中應用的概念與本研究使用的技術與技巧。

## 二、有效主題選取

第一階段的主題集群，其目的是從多次建模的結果模型裡選取經常出現的「有效主題」。Chuang等人（2015）及Mantyla等人（2018）針對主題建模演算法的不確定性（non-deterministic），分別使用AHC和*K*-medoids集群演算法進行主題集群，區分穩定出現的有效主題或偶然出現的雜訊，本研究也將採用集群演算法自動進行主題集群選取有效主題。Chuang等人（2015）及Mantyla等人（2018）的研究在應用於主題集群時，有以下的問題：

1. 主題集群能夠區別有效主題與雜訊的主要因素是緣由於有效主題之間有較近的距離，在雜訊的附近則少有相似的主題。所有模型上的主題在空間上的分布並不均勻，形成多個在空間上密度較高的主題群組以及不在群組中的雜訊，群組在空間上具有任意形狀。然而*K*-medoids則是依據反覆計算群組的範圍和中心，使中心和群組內的其他成員之間有最小的距離，因為這樣的特性，*K*-medoids最適合應用在不同方向上具有相同直徑的群組，並不適合於主題在空間上分布不均勻的主題集群應用。
2. 如同上述的說明，主題在空間上分布不均勻，這種情形將造成要產生的主題群組數目或條件經常無法事先預知。然而AHC和*K*-medoids都需要先設定集群的數目或條件。
3. AHC和K-medoids都會強迫將所有的集群對象歸入某一個群組當中，但是將雜訊強行歸入任何一個群組勢必影響後續進一步的處理與結果的解讀。

基於以上三個問題，本研究嘗試利用McInnes與Healy（2017）提出的Hierarchical Density Based Spatial Clustering of Applications with Noise（以下簡稱HDBSCAN）演算法進行主題集群。HDBSCAN是一種依據空間上的資料分布密度為基礎的集群演算法（density-based clustering algorithm），這個演算法適用於任意形狀的群組，並且只需指定最少成群的主題數目，不需要事先決定要產生的群組數目或條件，而且不會強迫將其他相差較多的集群對象歸入任一個群組內。此外，相較於這類的集群演算法之中最為知名的Density Based Spatial Clustering of Applications with Noise（DBSCAN）（Ester, Kriegel, Sander, & Xu, 1996）會將密度較低的部分也聚集成集群，HDBSCAN只注重在聚集

密度較高的主題，更適合於這個應用。因此，為了確認彼此相似而聚集成群的穩定主題，並且減少選出只有少數相似的不確定主題，本研究選擇HDBSCAN作為這個應用的集群演算法。並且在進行主題集群之前，本研究先使用McInnes、Healy與Melville（2018）提出的Uniform Manifold Approximation and Projection（以下簡稱UMAP）演算法將主題進行降維（dimensionality reduction）。在資料維度相當高的情形下，運用UMAP進行降維可以避免資料在高維度空間上的稀疏性，有助於提高集群的效果。利用UMAP進行降維然後利用HDBSCAN進行集群分析的運算模式，已經被應用於NTM的Top2Vec（Angelov, 2020）和BERTopic（Grootendorst, 2022）測量文件之間的相似性與群組關係。

UMAP降維演算法可以根據設定鄰近主題數目的大小控制主題空間中局部結構與全體結構的平衡，較小的鄰近主題數目將著重於精細的細節結構，而較大的鄰近主題數目將可以獲得更廣泛的空間資訊。而HDBSCAN集群演算法只需指定最少成群所需的鄰近主題數目，輸出空間中密度較高且符合主題數目的主題群組。指定的鄰近主題數目較小，HDBSCAN演算法會輸出較多的主題群組；反之，若是指定較大的鄰近主題數目，則會讓較多無法成群的主題都被視為雜訊，並且輸出較少的群組。由於原本的LDA主題建模演算法並不穩定，使得第一階段輸出的有效主題群組有時多於原先設定的主題數目，有時則少於原先設定的主題數目。在計算主題建模的穩定性時，每個模型應具有相同的主題數目，才能比較它們之間的相似性。因此，本研究需要解決產生的主題群組數目不一致的問題。本研究利用調整UMAP和HDBSCAN兩個演算法中的鄰近主題數目，使得找出的主題群組數目（$K'$）需要大於或等於原本預定主題數目的90%並且小於原本預定主題數目（$K$），也就是$0.9K \leq K' < K$，並在第二階段引導式主題建模中以背景主題（background topics）補足模型的預設主題數目。如果產生的主題群組數目不在上述範圍內，便調整成鄰近主題數目重新計算UMAP和HDBSCAN演算法，一直到輸出的主題群組數目符合條件為止。

最後，以圖3的虛擬碼（pseudo-code）說明在第一階段「有效主題選取」的演算法：

「有效主題選取」演算法的說明如下：

Step 1. 將要處理的語料庫（$C$）及主題數目（$K$）輸入一般LDA主題建模程式，固定α、β和推斷循環次數等參數（可使用程式內定的數值），重複進行$M$次建模，建立$M$個主題模型。這些主題模型以下稱為基本模型。

有效主題選取

Input: 語料庫 $C$, 主題數目 $K$, 模型數目 $M$

Output: 主題群組集合 $G$

Begin

Step 1: 將 $C$ 與 $K$ 輸入 LDA，進行 $M$ 次主題建模，產生 $M$ 個基本模型。

Step 2: 產生各主題的代表特徵。

Step 2.1: 自 $M$ 個基本模型中，取出所有主題的詞語機率分布$\phi_k$。

Step 2.2: 計算每對$\phi_k$之間的於餘弦值，測量每對主題的相似性。

Step 2.3: 根據主題之間的相似性，利用 UMAP 降維，產生主題特徵。

Step 3: 輸入主題特徵，利用 HDBSCAN 進行主題集群，產生主題群組集合 $G$。

Step 4: 如果 $G$ 的群組數量$K'$，$0.9K \leq K' < K$，則結束有效主題選取，輸出 $G$；否則調整鄰近主題數目，重新計算 Step 2 與 Step 3。

**圖3　穩定性改善技術第一階段「有效主題選取」的演算法虛擬碼**

Step 2. 從每個基本模型中，取出各模型上代表主題的詞語機率分布參數$\phi_k$（所有模型共計$M \times K$個詞語機率分布），利用餘弦測量計算$\phi_k$之間的相似性。將每對$\phi_k$之間的餘弦相似性輸入UMAP進行降維，產生代表主題的特徵。

Step 3. 輸入降維後的主題特徵，利用HDBSCAN集群演算法進行主題集群，將所有的主題區分為若干個群組以及雜訊，產生主題群組的集合**G**。每一個群組中即包含彼此相似而較穩定的有效主題。

Step 4. 如果產生的主題群組集合**G**內的群組數目（$K'$）大於或等於原先設定的主題數目（$K$），或是小於原先設定主題數目的90%（$0.9 \times K$），便調整UMAP和HDBSCAN演算法的鄰近主題數目，重新計算步驟2與步驟3，一直到輸出的主題群組數目符合條件（$0.9K \leq K' < K$）為止。

## 三、引導式主題建模

第二階段「引導式主題建模」的演算法虛擬碼如圖4。

引導式主題建模

Input: 主題群組集合 $G$

Output: 引導式主題建模產生的改良主題模型

Begin

Step 1: 對於主題群組集合 $G$ 上的每一個群組$g$，計算主題特徵$\phi_g = \sum_{k \in g} p_{gk} \times \phi_k / \sum_{k \in g} p_{gk}$。

Step 2: 以詞語均勻分布做為背景主題特徵，將主題特徵的總數補足到模型預定之主題數目 $K$。

Step 3: 根據主題群組特徵及背景主題特徵，利用引導式主題建模演算法產生最終的改良模型。

End

**圖4　穩定性改善技術第二階段「引導式主題建模」的演算法虛擬碼**

以下詳細說明各步驟：

Step 1. 雖然在第一階段已經區分主題建模結果中較穩定的有效主題與不穩定的雜訊，然而在每一個群組中的主題雖然都是來自同一個典型主題，但有些主題相當接近典型主題，有些主題則可能僅具有典型主題的一部分，或是合併了其他主題，而使得它們與典型主題較不相似。HDBSCAN演算法輸出的主題集群結果，除了賦予每個主題所屬的群組編號之外，還包括一個群組機率值，表示該主題在所屬群組上的核心程度。群組機率值較大的主題位於群組的核心位置，也就是這些主題與典型主題較相似，它們的詞語出現機率對於群組主題特徵的貢獻應該比不屬於核心的主題來得大。所以本研究利用群組機率值作為加權，計算群組內主題的平均詞語出現機率值，作為每一個群組的代表主題特徵。以某一個主題群組g為例，其代表主題特徵如 $\phi_g = \frac{\sum_{k \in g} p_{gk} \times \phi_k}{\sum_{k \in g} p_{gk}}$。在此，$\phi_k$

是屬於$g$的一個主題$k$上的詞語出現機率分布，$p_{gk}$是這個主題在群組$g$上的機率值。因此，主題特徵$\phi_g$是一個維度為$V$的向量，在特徵向量上數值較大的詞語對群組上的主題較重要。

Step 2. Vorontsov與Potapenko（2015）認為語料庫中出現的主題可分為領域特定主題（domain-specific topics）和背景主題。領域特定主題的詞語出現機率集中在語料庫的少數詞語上，並且這些詞語僅分布在一部分文本當中，例如在討論職場生涯議題的語料庫中，某些有關勞資契約和勞基法方面的詞語便經常會共同出現在相同文本上，例如「勞工」、「契約」、「雇主」、「資遣費」、「合約」等等。而背景主題則可能會廣泛的出現在語料庫的各個文本上，並且主題上的詞語出現機率較分散。本研究以第一階段有效主題選取產生的主題群組作為領域特定主題，並以AlSumait、Barbará、Gentle與Domeniconi（2009）建議的詞語均勻分布（uniform distribution）作為背景主題的特徵，將主題特徵補足到原本主題模型預定之$K$個主題。

Step 3. 以主題特徵進行引導式主題建模。在引導式主題建模過程當中，主題特徵將使主題模型上的文本主題組成比例和主題的詞語出現機率等參數朝向預先設定的主題（Zhou, Kan, Huang, & Silbernagel, 2023）。本研究中針對語料庫上出現的每一個詞語，利用前面兩個步驟產生的主題群組特徵$\phi_g$以及背景主題特徵，進行引導式主題建模。如此一來，可以保留第一階段獲得的有效主題資訊，並且得知文件上包含的主題組成比例。

## 四、其他主題建模技術的適用性

雖然本研究提出的主題建模穩定性改善方法主要以LDA演算法為應用對象，然而也可以同樣適用於其他主題建模演算法的穩定性改善。

這個穩定性改善方法的第一階段運用主題相似性比對、維度縮減、主題集群等技巧，從重複建模產生的多個基本模型中選擇經常出現的有效主題。為此，本方法所使用的主題模型，其主題形式必須是長度為詞彙量的向量，向量上的每一個元素代表某一個詞語在這個主題上的重要程度。以LDA演算法而言，即是主題模型中的主題詞語出現機率分布$\phi_k$。事實上，目前幾乎所有的主題建模演算法都能夠產生由重要詞語構成的主題，並且以表現其重要性的向量形式呈現。

第二階段則是應用引導式主題建模技術，利用第一階段選取有效主

題上的重要詞語引導主題建模。目前的引導式主題建模技術，除了先前所提到的引導式LDA主題建模以外，NMF（Vendrow, Haddock, Rebrova, & Needell, 2021）以及BERTopic也都有研究者發展出引導式主題建模。

另外，特別提出來說明的是本研究所提出改善方法的兩個階段可以應用不同類型的主題建模演算法。例如在第一階段利用Belford等人（2018）提出的NMF技術產生有效主題，然後在第二階段進行引導式LDA主題建模；或是在第一階段產生多個LDA主題模型並選擇有效主題後，第二階段利用BERTopic的引導式主題建模。

### 五、節省時間的技巧

過去的LDA主題建模方法通常只執行一次，產生一個主題模型。但為了使得LDA主題建模產生較穩定的結果，本研究提出的兩階段改善方法在第一階段時需要執行多次主題建模來獲得多個基本模型，因此將要花費更多的時間。

為了提高這個方法的效率，本研究建議在第一階段重複進行LDA主題建模，建立多個基本模型時，每次只隨機取出語料庫中的部分文件進行建模，藉此減少每次建模所需的時間。由於建立基本模型的目的是從中選取穩定出現的主題，因此雖然每次只從語料庫中隨機取出部分文件進行建模，但本研究假設每部分的文件中都很有可能包含大部分的有效主題，所以從累積多次建模產生的模型應該也能夠選取出其中多數穩定出現的有效主題。為了驗證上述的推論，本研究將進行實驗，改變建模時使用的文件數量以及基本模型數量，瞭解所提出改善方法與節省時間技巧的成效與限制。

## 肆、研究過程與結果

本研究將使用兩個語言與文類（genres）完全不同的語料庫進行一系列主題建模穩定性實驗，瞭解所提出改善方法的有效性。這些實驗比較主題建模穩定性改善方法與傳統LDA對於不同性質的語料庫的穩定性改善程度，以及第一階段中基本模型數量對改善方法的影響。本研究也將探討藉由隨機選取部分語料庫，減少建立基本模型使用的文本資料比例達到節省時間的可行性。本節將先說明本研究中使用的語料庫蒐集和預處理（preprocessing）的方法，然後報告主題建模穩定性的測量方法，最後是穩定性的改善程度與減少時間花費的實驗結果。

## 一、語料庫與文本資料處理

本研究所使用的語料庫一為自行蒐集的中文BBS發文語料庫，另一為網路上公開的英文研討會論文語料庫。

中文BBS發文語料庫（以下簡稱PTT-Salary）由本研究利用自行撰寫的程式，從批踢踢實業坊電子布告欄系統（PTT bulletin board system，以下簡稱PTT BBS）的職場板（https://www.ptt.cc/bbs/Salary/index.html）上蒐集網友發布的文章作為語料庫。蒐集文章的發布時間範圍自2019年6月起到2023年5月，共獲得來自6,465位作者的14,319筆發文。這些發文先利用中央研究院中文詞知識庫小組開發的ckiptagger套件（Li, Fu, & Ma, 2020）進行斷詞處理（word segmentation）與詞類標示（part-of-speech tagging），將輸入的每一筆發文內容切分為詞語和其對應詞類的序列。本研究依據詞語的詞類和在整個語料庫上出現的總次數和發文數選擇詞語序列中的詞語，保留普通名詞、專有名詞、地方詞、名物化動詞、動詞與外文標記等詞類的詞語，但刪除出現發文數少於20次或是出現發文數在總發文數1/10以上的詞語。最後，使用的發文資料共有2,121種不同詞語。選擇內容中包含10個詞語以上的發文，建立建模用的語料庫，共計7,488筆發文，每一筆發文選擇詞語數目的平均值為34.17，標準差為24.65。

英文研討會論文語料庫採用Neural Information Processing Systems（以下簡稱NIPS）資料集的論文全文（Hamner, 2017），此資料集包括1987年到2017年間各屆NIPS研討會的論文資料，資料中包含題名、作者、摘要及由PDF檔轉出的全文，共計7,241筆論文資料。本研究首先將全文資料進行小寫化處理和去除標點與數學運算符號，然後利用自然語言工具包（natural language toolkit, NLTK）套件將論文全文進行分詞（tokenization）與詞形還原（lemmatization）等處理，然後進行統計，刪除出現發文數少於100次或是出現發文數在總發文數1/10以上的詞語，最後再以nltk套件的英文停用詞（stopwords）進行過濾，最後使用的論文資料共有3,418種不同詞語。選擇內容中包含50個詞語以上的全文資料，建立建模用的語料庫，共計7,221筆論文全文資料，每一筆論文資料選擇詞語數目的平均值為298.14，標準差為116.45。

為了進行以下的實驗，本研究先決定主題模型內的主題數目$K$。依據先前的研究經驗，在選用模型的主題數目時，以具有最高平均主題協調性（average topic coherence）（Röder, Both, & Hinneburg, 2015）的主題數目決定。本研究針對語料庫建立多種不同主題數目（$K$ = 5, 10, ...,

50）的主題模型，其中以PTT-Salary語料庫以30個主題時的平均主題協調性最高，而NIPS語料庫則以55個主題時最高。因此，在以下的主題建模穩定性實驗，兩個語料庫便分別採用30和55個主題。

表3總結兩個語料庫的各項統計特性。PTT-Salary語料庫文本內容較短，詞語數目也較少，因此使得平均主題協調性最高的主題模型不需太多的主題數目，而NIPS語料庫為學術研討會的論文全文資料，文本內容較長並且有較多種的詞語，使得平均主題協調性最高的主題模型需要較多的主題數目。

## 二、主題建模穩定性的測量

本研究分別測量一般LDA演算法以及所提出方法的穩定性，藉以驗證改善方法的有效性。圖5是每一實驗組的主題建模穩定性測量過程示意圖，詳細的計算方式可參考林頌堅（2022）。

首先，產生多組主題模型，這些主題模型依據實驗需求可以是採用傳統建立的模型，或是由本研究為改善主題建模穩定性所提出方法建立的改良模型。然後進行每一對模型之間的主題相似性估計，以餘弦測量詞語出現機率分布（$\phi_k$）的相似性，計算兩個模型中每一對主題的相似性分數。因為主題上的詞語出現機率都是大於或等於0，因此主題的相似性分數在0與1之間，而且主題愈相似，分數愈接近1。

接下來，藉由這兩個模型上彼此相似主題計算每一對模型之間的一致性分數。根據前面對於主題建模穩定性測量研究的分析，本研究利用

**表3**

**PTT-Salary與NIPS語料庫的各項統計特性**

| 語料庫名稱 | PTT-Salary | NIPS |
|---|---|---|
| 蒐集文本數 | 14,319 | 7,241 |
| 詞語出現發文數下限 | 20 | 100 |
| 詞語出現發文數上限 | 1/10蒐集文本數 | 1/10蒐集文本數 |
| 研究選用詞語數 | 2,121 | 3,418 |
| 文本選擇詞語數下限 | 10 | 50 |
| 研究選用文本數 | 7,488 | 7,221 |
| 文本選擇詞語平均數 | 34.17 | 298.14 |
| 文本選擇詞語標準差 | 24.65 | 116.45 |
| 模型主題數 | 30 | 55 |

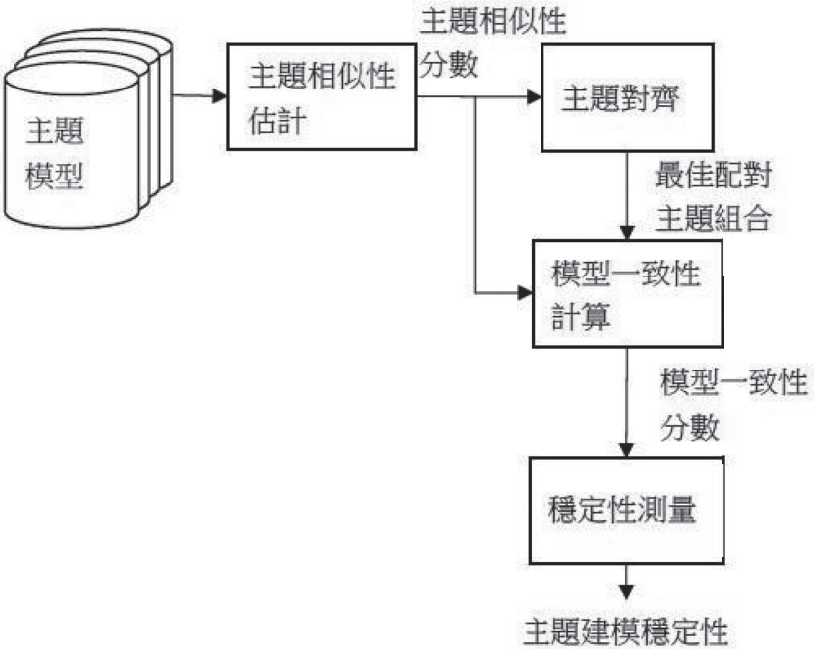


**圖5　主題建模穩定性的測量過程示意圖**

匈牙利演算法進行主題對齊，計算兩個模型之間最佳配對主題組合中的相似性分數平均值，做為這對模型之間的一致性分數。一致性分數介於0與1之間，如果獲得較大的平均相似性分數，則表示這兩個主題模型具有許多相似的主題，兩個主題模型之間相當一致。

最後，穩定性的測量值為所有模型之間一致性分數的平均值。

## 三、兩階段集成式改善方法與傳統LDA在穩定性上的比較

為了瞭解本研究提出的兩階段集成式改善方法是否比傳統LDA產生模型更穩定以及基本模型數量對改善方法的影響，本研究將首先以全部語料庫建立多個基本模型。在有效主題選取部分，本研究分別從10、20、30和40個基本模型進行主題集群，選取穩定出現的有效主題，然後進行引導式主題建模，產生改良模型，與傳統LDA演算法產生的模型比較。每一組改良模型和傳統LDA模型都是10個模型，對每個實驗組內的每一對模型計算它們的一致性分數，再將45個一致性分數進行平均，做為各組的穩定性分數。表4是兩種語料庫在各實驗組所獲得的穩定性分數。

從表4的穩定性測量結果可以很明顯地看出本研究所提出的方法確

表4

PTT-Salary與NIPS語料庫在各實驗組所獲得的穩定性分數

| | | 改善方法使用的基本模型數目 | | | |
|---|---|---|---|---|---|
| 語料庫 | 傳統LDA | 10個 | 20個 | 30個 | 40個 |
| PTT-Salary | 0.70 | 0.82 | 0.81 | 0.85 | 0.82 |
| NIPS | 0.64 | 0.77 | 0.80 | 0.82 | 0.82 |

實能夠改善LDA主題建模的穩定性。在兩個語料庫的實驗當中，不論是使用多少個基本模型來選取有效主題，使用改善方法的實驗組都比傳統LDA演算法明顯地得到更好的穩定性結果。其中，NIPS語料庫由於具有較多的文本內容、詞語數量較多，因此產生的主題模型具有較多的主題數目，傳統LDA的穩定性並不理想。然而本研究所提出的改善方法不僅能夠有效地達到較好的穩定性，而且若是在第一階段使用較多的基本模型選取有效主題，改良模型的穩定性也能夠達到相當好的穩定性。因此，本研究的改善方法具有相當程度的穩定性改善能力，特別是對於詞語數較大、主題數較多的主題模型。

為了進一步比較本研究所提出的穩定性改良方法與傳統LDA，下面以圖6和圖7分別表示PTT-Salary與NIPS各實驗組與傳統LDA演算法的一致性和穩定性測量結果。圖上灰色小點表示組內每一對模型的一致性分數，黑色方塊則是一致性分數的平均值，也就是穩定性分數。

從兩個語料庫的主題建模穩定性實驗的圖形上可以觀察到，不論使用多少基本模型，每對改良模型之間的一致性分數幾乎都比傳統LDA建立的每對模型之間一致性分數更高，而且在NIPS語料庫的實驗當中，隨著基本模型的數目增加，模型之間愈發一致，穩定性有小幅度的增加。至於在PTT-Salary語料庫的實驗結果，由於產生的主題模型較不複雜，使用10個基本模型選取有效主題，便已經達到相當好的改善程度。在30個基本模型結果，改良模型之間都相當一致，甚至有一些模型之間的一致性分數達到0.9以上。雖然在20個基本模型結果，一致性分數的平均值明顯比傳統LDA建模來得好，但仍有一部分的主題模型之間不一致，使得這組的穩定性分數是各實驗組中最低。

## 四、以部分語料庫文本節省改善方法時間的研究

為了減少基本模型所需的建模時間，藉以達成加速本研究提出之改善方法的目的，我們在建立每個基本模型時只從語料庫中隨機選取部分

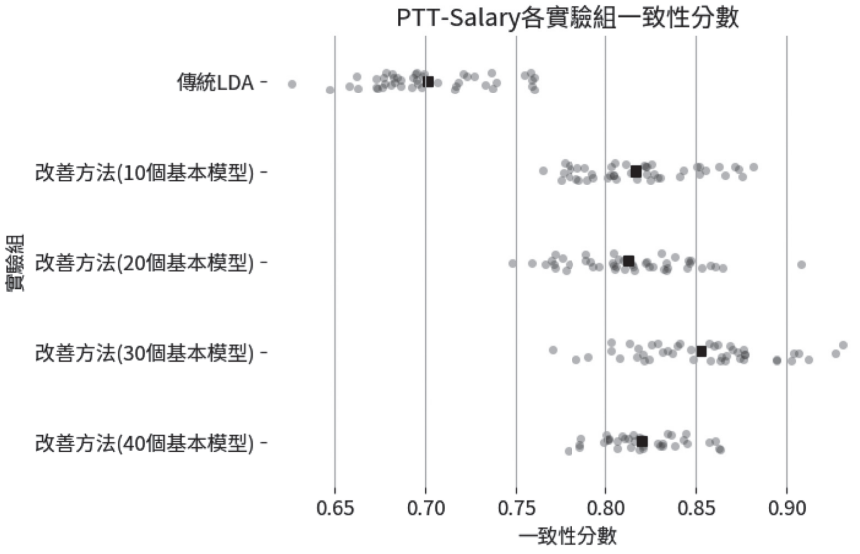


圖6　PTT-Salary語料庫各實驗組的主題模型一致性和穩定性測量結果

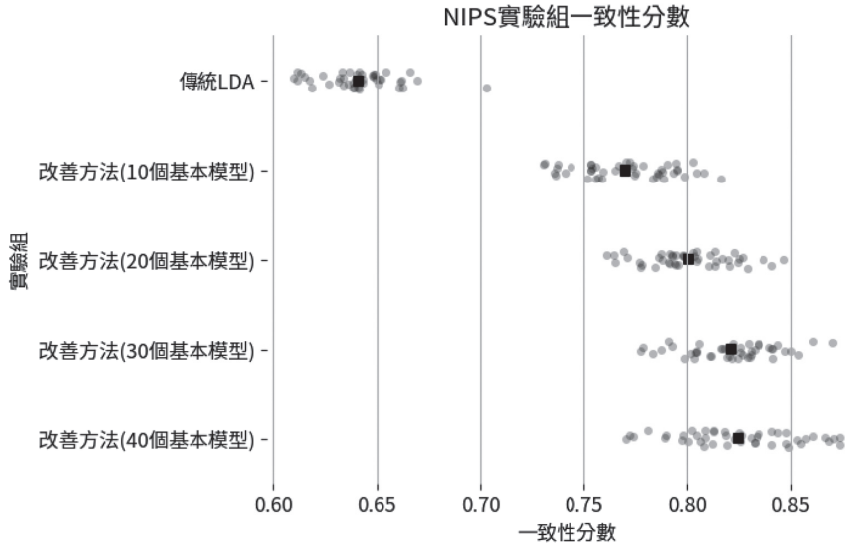


圖7　NIPS語料庫各實驗組的主題模型一致性和穩定性測量結果

文本。表5與表6分別是以本研究使用的PTT-Salary與NIPS語料庫為例，依據不同比例隨機選取文本進行100次基本模型訓練，測量每次建模時間的平均值與其標準差。結果顯示較少的文本數量確實可以達到減少平均建模時間。

以下分別以20%到100%的語料庫文本建立從10、20、30和40個基本模型選取穩定出現的有效主題，然後產生10個改良模型計算各組的穩定性，藉以瞭解第一階段中基本模型數目對改善方法的影響。表7與表8分別是兩個語料庫在各種基本模型下使用不同數量文本的穩定性測量結果。

從表7與表8的實驗結果，本研究提出的主題建模穩定性改善方法，在某些情形下，可以在第一階段產生基本模型時，藉由減少文本資料以達到節省時間的目的，並且還能相當程度維持較好的穩定性分數。比較傳統LDA主題建模的穩定性分數與各組改良模型所得到的穩定性分數，

**表5**

**對不同數量的PTT-Salary語料庫文本進行100次基本模型訓練的平均建模時間與其標準差**

| 使用語料庫文本比例 | 文本數 | 平均建模時間 | 標準差 |
|---|---|---|---|
| 100% | 7,488 | 19.96 sec. | 1.32 sec. |
| 80% | 5,990 | 16.36 sec. | 1.19 sec. |
| 60% | 4,492 | 12.13 sec. | 0.99 sec. |
| 40% | 2,995 | 7.74 sec. | 0.86 sec. |
| 20% | 1,497 | 3.74 sec. | 0.70 sec. |

**表6**

**對不同數量的NIPS語料庫文本進行100次基本模型訓練的平均建模時間與其標準差**

| 使用語料庫文本比例 | 文本數 | 平均建模時間 | 標準差 |
|---|---|---|---|
| 100% | 7,221 | 125.48 sec. | 2.13 sec. |
| 80% | 5,776 | 106.15 sec. | 2.32 sec. |
| 60% | 4,332 | 72.02 sec. | 2.48 sec. |
| 40% | 2,888 | 47.85 sec. | 2.89 sec. |
| 20% | 1,444 | 23.14 sec. | 1.62 sec. |

表7

PTT-Salary語料庫使用不同數量文本的穩定性測量結果

| 文本比例 | 10個模型 | 20個模型 | 30個模型 | 40個模型 |
|---|---|---|---|---|
| 20% | 0.77 | 0.77 | 0.77 | 0.78 |
| 40% | 0.78 | 0.83 | 0.82 | 0.83 |
| 60% | 0.79 | 0.82 | 0.83 | 0.83 |
| 80% | 0.80 | 0.81 | 0.81 | 0.84 |
| 100% | 0.82 | 0.81 | 0.85 | 0.82 |

表8

NIPS語料庫使用不同數量文本的穩定性測量結果

| 文本比例 | 10個模型 | 20個模型 | 30個模型 | 40個模型 |
|---|---|---|---|---|
| 20% | 0.71 | 0.75 | 0.77 | 0.78 |
| 40% | 0.72 | 0.79 | 0.79 | 0.78 |
| 60% | 0.76 | 0.79 | 0.80 | 0.81 |
| 80% | 0.76 | 0.79 | 0.82 | 0.84 |
| 100% | 0.77 | 0.80 | 0.82 | 0.83 |

不論是表7 PTT-Salary語料庫或是表8 NIPS語料庫，不同數量的文本產生各種數量的基本模型所得到的穩定性分數都比原先傳統LDA主題建模（職場板：0.70，NIPS：0.64）要好。

並且如同直觀的想法，在本次實驗當中，使用愈完整的文本資料會得到愈高的穩定性分數。在PTT-Salary語料庫的實驗當中，除了使用20%文本資料的各個實驗組都沒有達到0.80以外，其他各種數量的各個實驗組，幾乎每一組都獲得接近0.80或以上的穩定性分數。80%或60%的文本資料在第一階段只要產生10個基本模型進行主題集群便可以獲得0.80以上的穩定性分數，如果是使用40%或60%的文本資料則只要產生20個基本模型便可以有0.80以上的穩定性分數。至於模型上詞語以及主題數目都較多的NIPS語料庫，60%或80%的文本資料可以獲得稍遜於完整的文本資料的穩定性，同時也可以發現文本資料數量以及基本模型數目增加都能夠使得最後建立的主題模型獲得比較好的穩定性。

## 伍、結論

應用主題建模方法可以分析文本資料的主題結構，以便瞭解文本資料的包含的主題，並且進一步使用於各種後續的自然語言處理應用，如資訊檢索、情感分析。雖然主題建模出現愈來愈多的新技術，包括統計推論、非負矩陣分解和神經主題建模等等各種類型的演算法，效能也愈來愈強大，但是在應用到文本分析時，主題建模的穩定性仍然是一個相當值得關注的問題。主題建模的不穩定是由於大多數的主題建模技術運用隨機性策略來推導模型參數，因為每次使用的隨機種子不一樣，使得產生的模型會有所不同。如果產生的主題模型相當不一致，將會影響文本分析時的研究信度以及自然語言處理應用的效用。

本研究針對LDA主題建模的穩定性改善進行研究。前人提出的主題建模穩定性改善方法所使用的策略，不論是改變模型參數的推斷方法以及選取較穩定的主題或模型，仍然有許多需要克服的問題。本研究提出一個兩階段主題建模方法整合這些策略的優點，第一階段將從多個基本模型中選取穩定出現的有效主題，第二階段則根據這些有效主題的詞語出現資訊進行引導式主題建模產生穩定性較高的改良模型。為了節省整體的建模時間，本研究並且建議第一階段建立基本模型時可以只選用一部分的語料庫。

為了證明所建議之改善方法和節省時間策略的成效，本研究並以兩個不同的語料庫進行測試。結果發現本研究比傳統LDA主題建模，在不同語料庫及各種基本模型之下都明顯地得到更好的穩定性分數。並且對於原本而穩定性較不理想的NIPS語料庫，本研究所提出的方法對穩定性有相當大幅度的改善。在利用語料庫的部分文本資料建立基本模型以減少改善方法所需時間的測試當中，即便只使用相當少部分的文本資料（20%），所得到的穩定性分數也比傳統LDA主題建模更高。對於PTT-Salary語料庫這種主題模型較簡單的語料庫，使用較少的文本資料（40%）並選用較多的基本模型（20個或以上）所得到的穩定性分數，便相當接近於完整的文本資料所得到的分數。如NIPS語料庫這種需要較複雜主題模型的語料庫，則需要使用稍多的文本資料（60%或80%）才可以獲得稍遜於完整的文本資料的穩定性。

未來的主題建模穩定性研究除了繼續提出更有效、更節省時間與資源的改善方法以及擴大實驗的語料庫範圍、增加一致性測量的組數之外，更重要的是不同主題建模穩定性改善方法之間的比較。目前雖然已經有相當多改善方法的研究提出。但由於使用的策略不同，應用於不同

語料庫時需要調整的參數與程序相當多，主要的幾種方法並沒有程式釋出，也缺乏各種參數的設定資料作為調整各種方法最佳成效的依據。因此，目前尚未有比較不同方法成效的研究。未來將可以針對此研究缺口，進行深入的比較研究。另一個相當值得深入探討的是對於其他主題建模演算法的穩定性改善研究。雖然本研究中主要將所提出的改善方法應用在LDA主題建模，然而也討論了其他演算法的適用性，未來相當值得進行深入的研究。

## 參考文獻


邱志洲、吳忠敏、簡德年、高凌菁、邱德生（2023）。加護病房患者臨床結果預測 —— 機器學習與主題模型法之應用。*醫務管理期刊*，*24*(3)，221-248。doi:10.6174/JHM.202309_24(3).221【Chiu, C.-C., Wu, C.-M., Chien, T.-N., Kao, L.-J., & Qiu, J. T. (2023). Predicting clinical outcomes of the patients in intensive care units with machine-learning and topic model techniques. *Journal of Healthcare Management*, *24*(3), 221-248. doi:10.6174/JHM.202309_24(3).221 (in Chinese)】

林頌堅（2022）。主題相似性估計與其在主題建模穩定性測量之應用。*教育資料與圖書館學*，*59*(2)，201-231。doi:10.6120/JoEMLS.202207_59(2).0014.RS.AM【Lin, S.-C. (2022). Estimation of topic similarity and its application to measuring stability of topic modeling. *Journal of Educational Media & Library Sciences*, *59*(2), 201-231. doi:10.6120/JoEMLS.202207_59(2).0014.RS.AM (in Chinese)】

郝沛毅、歐仁彬、黃天受、林振穎、吳建生（2018）。透過新聞文章預測股價漲跌趨勢 —— 結合情緒分析、主題模型與模糊支持向量機。*資訊管理學報*，*25*(4)，363-395。【Hao, P.-Y., Ou, J.-B., Huang, T.-S., Lin, Z.-Y., & Wu, J.-S. (2018). Sentiment and topic analysis on financial news for stock movement prediction by using fuzzy support vector machine. *Journal of Information Management*, *25*(4), 363-396. (in Chinese)】

陳怡璇、劉桂君（2023）。以瘟疫之名：探討公眾風險感知的COVID-19網路論述與詮釋，以批踢踢八卦板與政黑板為例。*新聞學研究*，*156*， 47-104。doi:10.30386/MCR.202307.0013【Chen, V. Y., & Liu, A. K. C. (2023). In the name of the pandemic: Exploring COVID-19 online discourse. *Mass Communication Research*, *156*, 47-104. doi:10.30386/

MCR.202307.0013 (in Chinese)】

傅文成、黃瑎戩、顏瑞宏（2021）。以資料科學方法輔助民意趨勢分析：戰略及戰爭風險感知的網路民意研究。*新聞學研究*，*149*，1-49。doi:10.30386/MCR.202109.0020【Fu, W.-C., Huang, W.-J., & Yen, J.-H. (2021). Using data science approach to explore online public opinion: Strategic and war risk perceptions on the Internet. *Mass Communication Research*, *149*, 1-49. doi:10.30386/MCR.202109.0020 (in Chinese)】

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*. doi:10.1016/j.is.2022.102131

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74-88. doi:10.1016/j.infsof.2018.02.005

Alexander, E., & Gleicher, M. (2016). Task-driven comparison of topic models. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 320-329. doi:10.1109/TVCG.2015.2467618

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD 2009)* (pp. 67-82). Bled, Slovenia: Springer Berlin Heidelberg. doi:10.1007/978-3-642-04180-8_22

Angelov, D. (2020). *Top2vec: Distributed representations of topics*. arXiv. doi:10.48550/arXiv.2008.09470

Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, *16*(1), 101224. doi:10.1016/j.joi.2021.101224

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications: An International Journal*, *91*, 159-169. doi:10.1016/j.eswa.2017.08.047

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Chauhan, U., & Shah, A. (2022). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*, *54*(7), 1-35.

doi:10.1145/3462478

Chen, S., Andrienko, N., Andrienko, G., Adilova, L., Barlet, J, Kindermann, J., ...Turkay, C. (2020). LDA ensembles for interactive exploration and categorization of behaviors. *IEEE Transactions on Visualization and Computer Graphics*, *26*(9), 2775-2792. doi:10.1109/TVCG.2019.2904069

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In M. Rada, C. Joyce, & S. Anoop (Eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 175-184). Denver, Colorado: Association for Computational Linguistics. doi:10.3115/v1/N15-1018

de Waal, A., & Barnard, E. (2008). Evaluating topic models with stability. In F. Nicolls (Ed.), *Proceedings of the 19th annual symposium of the Pattern Recognition Association of South Africa* (pp. 79-84). Cape Town, South Africa: Pattern Recognition Association of South Africa.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Elgesem, D., Steskal, L., & Diakopoulos, N. (2015). Structure and content of the discourse on climate change in the blogosphere: The big picture. *Environmental Communication*, *9*(2), 169-188. doi:10.1080/17524032.2014.983536

Ester, M., Kriegel, H.-P., Sander, J., & Xu. X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)* (pp. 226-231). Portland, Oregon: AAAI Press.

Greene, D., O'Callagha, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Proceedings of the Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science (ECML PKDD 2014)* (pp. 498-513). Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-44848-9_32

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228-5235. doi:10.1073/pnas.0307752101

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. doi:10.48550/arXiv.2203.05794

Hamner, B. (2017). *NIPS Papers—Titles, authors, abstracts, and extracted text for all NIPS papers (1987-2017)*, [Kaggle Dataset]. Retrieved from https://www.kaggle.com/datasets/benhamner/nips-papers

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*(pp. 50-57). New York, NY: Association for Computing Machinery. doi:10.1145/312624.312649

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89-106. doi:10.1080/21670811.2015.1093271

Jagarlamudi, J., Daumé H. III, & Udupa, R. (2012). Incorporating lexical priors into topic models. In W. Daelemans (Ed.), *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204-213). Avignon, France: Association for Computational Linguistics.

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus. In A. F. Gelbukh (Ed.), *Proceedings of the Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science (CICLing 2011)* (pp. 163-176). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-19437-5_13

Koltcov, S., Nikolenko, S. I., Koltsova, O., Filippov, V., & Bodrunova. S. (2016). Stable topic modeling with local density regularization. In F. Bagnoli, A. Satsiou, I. Stavrakakis, P. Nesi, G. Pacini, Y. Welp, T. Tiropanis, & D. DiFranzo (Eds.), *Internet Science. Lecture Notes in Computer Science (INSCI 2016)* (pp. 176-188). Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-319-45982-0_16

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83-97. doi:10.1002/nav.3800020109

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral,

L. A. N. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, *5*(1), 011007. doi:10.1103/PhysRevX.5.011007

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788-791. doi:10.1038/44565

Li, P.-H., Fu, T.-J., & Ma, W.-Y. (2020). Why attention? Analyze BiLSTM deficiency and its remedies in the case of NER. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8236-8244. doi:10.1609/aaai.v34i05.6338

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93-118. doi:10.1080/19312458.2018.1430754

Mantyla, M. V., Claes, M., & Farooq, U. (2018). Measuring LDA topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'18)* ( pp. 1-4). New York, NY: Association for Computing Machinery. doi:10.1145/3239235.3267435

McInnes, L., & Healy, J. (2017). Accelerated hierarchical density based clustering. In *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 33-42). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ICDMW.2017.12

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: *Uniform manifold approximation and projection for dimension reduction*. arXiv. doi:10.48550/arXiv.1802.03426

Rieger, J., Rahnenführer, J., & Jentsch, C. (2020). Improving Latent Dirichlet Allocation: On reliability of the novel mthod LDAPrototype. In E. Métais, F. Meziane, H. Horacek, & P. Cimiano (Eds.), *Natural Language Processing and Information Systems. Lecture Notes in Computer Science (NLDB 2020)* (pp 118-125). Cham, Switzerland: Springer International Publishing. doi:10.1007/978-3-030-51310-8_11

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In X. Cheng (Ed.), *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15)*

(pp. 399-408). New York, NY: Association for Computing Machinery. doi:10.1145/2684822.2685324

Srivastava, A., & Sutton, C. (2017, April). *Autoencoding variational inference for topic models*. Paper presented at the 5th International Conference on Learning Representations (ICLR 2017). Hyēres, France.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583-617. doi:10.1162/153244303321897735

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463-479. doi:10.1509/jmr.12.0106

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided Latent Dirichlet Allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, *56*(1), 18-36. doi:10.1177/0022243718820559

Vendrow, J., Haddock, J., Rebrova, E., & Needell, D. (2021). On a guided nonnegative matrix factorization. In *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)* (pp. 3265-3269). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ICASSP39728.2021.9413656

Vorontsov, K., & Potapenko, A. (2015). Additive regularization of topic models. *Machine Learning*, *101*(1-3), 303-323. doi:10.1007/s10994-014-5476-6

Yang, Y., Pan, S., Lu, J., Topkara, M., & Song, Y. (2016). The stability and usability of statistical topic models. *ACM Transactions on Interactive Intelligent Systems*, *6*(2), 1-23. doi:10.1145/2954002

Zhou, S., Kan, P., Huang, Q., & Silbernagel, J. (2023). A guided Latent Dirichlet Allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura. *Journal of Information Science*, *49*(2), 465-479. doi:10.1177/01655515211007724

# Improving the Stability of LDA Topic Modeling with Ensemble Learning

**Sung-Chien Lin**

Associate Professor
Department of Information and Communications
Shih Hsin University

## Introduction

Topic modeling enables the rapid discovery of latent thematic structures within large amounts of unstructured textual data. This has led to various applications, including helping researchers analyze the themes across all texts in a corpus or providing textual features for subsequent natural language processing tasks such as information retrieval and sentiment analysis. However, in practice, many topic modeling algorithms, including Latent Dirichlet Allocation (LDA), suffer from stability issues due to stochastic strategies that derive model parameters from the corpus (e.g., Abdelrazek, Eid, Gawish, Medhat, & Hassan, 2023; Mantyla, Class & Farooq., 2018). This means that even with identical parameter configurations and repeated modeling on the same corpus, the results may still vary.

Current methods for improving the stability of topic modeling have focused on two strategies: altering the inference of model parameters (Koltcov, Nikolenko, Koltsova, Filippov, & Bodrunova, 2016; Lancichinetti et al., 2015; Maier et al., 2018) and selecting more stable topics (Chuang et al., 2015; Mantyla, Claes, & Farooq, 2018). The former strategy involves word clustering from the corpus and uses these word clusters to guide the inference of topic model parameters. However, this approach tends to assign each word primarily to a single topic, contradicting the fundamental assumption in topic modeling that allows words to exist across different topics.

The latter strategy employs cluster algorithms to identify more stable topics from multiple models using identical parameter settings on the

same corpus. Nevertheless, this method fails to determine the proportional composition of topics within a text, posing inconvenience and challenges for subsequent textual analysis tasks.

The present study introduces a two-stage ensemble learning method designed to improve the stability of LDA in topic modeling, addressing the limitations of the strategies mentioned above. Furthermore, this paper aims to empirically validate the feasibility of this proposed approach. To mitigate the higher computational demands associated with this improvement method, our research explores the initial modeling of a randomly sampled subset of documents from the corpus in the proposed method's first stage.

## Method

The two-stage integrated method proposed in this study harnesses the strengths of the previously mentioned strategies. The first stage, "Effective Topic Selection," establishes multiple topic models. Subsequently, it employs the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm by McInnes and Healy (2017) in conjunction with the Uniform Manifold Approximation and Projection (UMAP) algorithm by McInnes, Healy, and Melville (2018) to cluster similar topic groups. This process effectively distinguishes stable, meaningful topics from sporadic noise using only a few models. Notably, words exist across multiple significant topics based on their significance, aligning with the core principles of topic modeling.

The second stage, "Guided Topic Modeling," uses the topic groups generated in the first stage to establish feature vectors representing each effective topic. The influence of each topic in constructing effective topic feature vectors is determined by its cluster membership score. These feature vectors, and the several vectors derived from a uniform distribution of words to represent background topics, serve as the initial set of topic feature vectors. As such, this method employs the initial feature vectors for guided topic modeling, resulting in comprehensive and more stable topic models.

To expedite the establishment of multiple base models in the first stage, this study suggests randomly using only a subset of documents from the corpus each time. The assumption, based on the possibility that each subset of

documents may contain a significant portion of the effective topics, reduces the required time while enhancing this method's efficiency.

## Results

This study uses two different corpora of different languages and text genres: PTT-Salary, collected posts from the Salary Board on the PTT Bulletin Board System (BBS), and NIPS, comprising full-text papers from the Neural Information Processing Systems conference. Each corpus is preprocessed using tokenization, part-of-speech tagging or lemmatization, stop-word removal, and by selecting appropriately sized texts. The preprocessed data is input to the LDA topic modeling algorithm as a bag-of-words model to establish the base models.

In the experiments conducted on these two corpora, the experimental groups using the improvement method consistently demonstrated significantly better stability results than traditional LDA topic modeling, regardless of the number of base models used for selecting effective topics. For the PTT-Salary genre, employing just 10 or more base models in the first stage of the "Effective Topic Selection" elevated stability from 0.70 to above 0.82. The NIPS data included more content and a higher word count. This resulted in a greater number of topics in the generated models, where traditional LDA topic modeling showed a stability of only 0.64. However, this study's proposed improvement method achieved a stability score of 0.77 with 10 base models and could reach 0.82 with over 30 base models.

The study experiments also confirmed that reducing amounts of textual data can save time while maintaining a relatively good level of stability scores. In the experiments conducted on the PTT-Salary corpus, using only 40% of the textual data resulted in stability scores close to or above 0.80. Regarding the NIPS corpus, using 60% or 80% of the textual data yielded slightly lower stability compared to using the complete set.

## Conclusions and Future Works

This study draws the following conclusions:

1. Compared to traditional LDA topic modeling, this research consistently achieved significantly better stability scores across various corpora and different numbers of base models. This was particularly evident in corpora with larger text quantities and topic counts.
2. Even when using a relatively small portion of textual data to establish base models, the resulting stability scores surpassed those obtained through traditional LDA topic modeling. Moreover, increasing the proportion of the textual data used and augmenting the number of base models contributed to better stability in the final developed topic models.

Future research can further validate the feasibility of this method in enhancing the stability of topic modeling. This includes expanding the scope of experimental corpora, increasing the number of topic models used to compute the agreement score between models, comparing it with different approaches for improving topic modeling stability, and proposing more effective and resource-efficient methods for enhancing topic modeling stability. Additionally, while the method proposed in this study is primarily designed for LDA topic modeling, it can be applied to other topic modeling techniques, such as NMF (Non-Negative Matrix Factorization) by Lee and Seung (1999) and the recently prominent BERTopic by Grootendorst (2022). Therefore, exploring our proposed model's adaptability and adjustments for other techniques will be a valuable direction for further investigation.