本文摘自Lisa M. Rhody的〈[Topic Modeling and Figurative Language](http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/)〉，由邱永泰統整後撰文。

本文旨在介紹主題模型是什麼、是如何運作的、該如何運用它、用主題模型工具來分析一般的文本與象徵性語言有何差異、會有哪些難處等。而本文所運用的主題模型為LDA（隱含狄利克雷分布）。

LDA能非常分析大量的文本或語料庫，藉由分析文本的構造、每個字的上下文、LDA能將相似概念的字，或者是上下文組成類似的字找出來、將他們歸類。結果就是、一群擁有共同概念或特性的字會被歸類到一塊，這一塊字，就屬於同一個「主題」。

這邊以一個簡單的例子來描述主題模型究竟是如何分析文章，從而產出主題：你聽說了隔壁的村子開了花市，你很想知道那邊的花市究竟賣了哪些花，然而你沒有空親自去一趟，於是拜託了媽媽每天去一趟花市買一些花回來。媽媽每天都帶回來很多不同的花，經過每天的觀察與計算，便可以得知這個花市大致上有賣「哪幾種」花，而各種花的「數量」又相差多少。這便是LDA的概念，LDA可以從大量的文本中找出各種主題的大致分佈，主題的數量對應的是花市中花的種類。要注意的是，LDA做的事僅僅是將一群有類似概念的字集合在一起，主題代表的只是一個「集合」，不同集合的字就是不同的主題，但LDA並不會給這個主題（這個集合）取名，只是我們人類為了方便理解，根據各個主題內含字的共通概念取名。

主題模型最初是創造來分析非象徵性語言的，譬如科學期刊的文章或百科全書的內容，看過文章中的字便可以得知這篇文章的意義；相對的，我們很難單由字面上的字去了解象徵性語言（譬如詩作）所隱含的意義。本文便是以此為出發點，使用主題模型嘗試分析象徵性語言，觀察會得出何種不一樣的結果。

另外，對於一個主題模型而言，有兩個很重要的要素：第一：分析出來的結果必須是前後一致且可重複的，也就是說，做一百次分析要得到一百次相近的結果，套用到不同的文本上也要能得到可預期的結果；第二：分析出來的結果必須是人類所能理解的，以一般的文本舉例來說，便是要能得到此份文本的文意。

根據原文作者描述，他嘗試使用LDA分析大量詩作得到了四種不一樣的結果，且這樣結果都是會一直重複出現。第一種是掃描文件沒有掃好所出現的錯別字，通常是將原本的單字拼錯或是換成沒有意義的符號。顯而易見的，這類型的主題並沒有意義。第二種是因為某一文本規模太大，或是單一文本內出現太多重複的字，導致主題模型直接將這些大量重複字視作為一個主題。第三種是，到看似富有語意的主題。這類型的主題看起來就像是以LDA分析非象徵性語言會得到的主題，但我們不能以同樣的思維去看待這種主題。這類型的主題通常只是描繪出一份詩作最淺層的表象，例如詩作中的風景等等。第四種是，到看似語意不明的主題。當單看這種主題內的字，會發現這些字似乎好像沒有太大的共通點，然而當我們將這種主題套用回各個文本，卻發現了有些文本存在著共通點。這類型的主題，能為學者帶來新的發現，得出文本背後潛在的模式，使學者能從不同的角度對文本提出新的疑問。這便是使用主題模型分析象徵性語言所希望能得到的成果。