十九與二十世紀各國經典文學作品的高頻字與主題標記分析

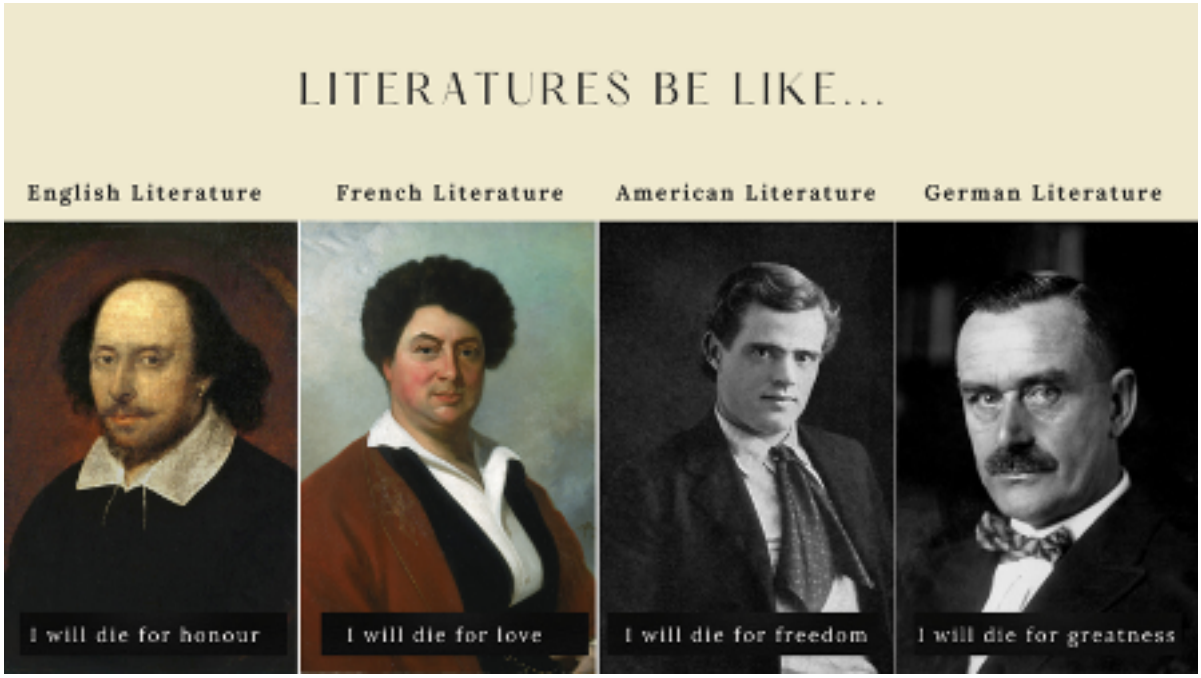
組名：G02 Sherry Don’t Go
組員：陳聯輝 周昕妤 黃彙茹 吳錤

一、主題

畢業、離開校園的時刻即將來臨，在外文系的四年裡，我們修過了各種的文學課，讀了也分析了來自各個年代、風格各異的文本。文學讓我們能一窺那個時代的思想與生活樣貌。咀嚼文字的芬芳之間，帶我們坐著時光機，回到作者筆下美好，又或不美好的年代。

在還沒讀過各個國家的文學作品前，或許對某些國家的文學會存有既定印象。英國封建制度下的貴族傳統，崇尚以榮譽、責任、勇氣等一系列核心精神，那英國的文學作品是否同樣會圍繞這些概念？法國常常被跟浪漫連結在一起，而讓我們對法國有種種美麗憧憬及遙遠印象，那法國的文學作品是否也同樣如此？我們想藉由分析多個國家的文學作品，來了解這些文本是否多半圍繞在特定主題，並且與某些刻板印象又有甚麼差異。

二、假設



三、資料取得

我們選擇英、美、德、法，四個國家，每個國家各選5篇長篇文學進行文本分析。書單如下表所示：

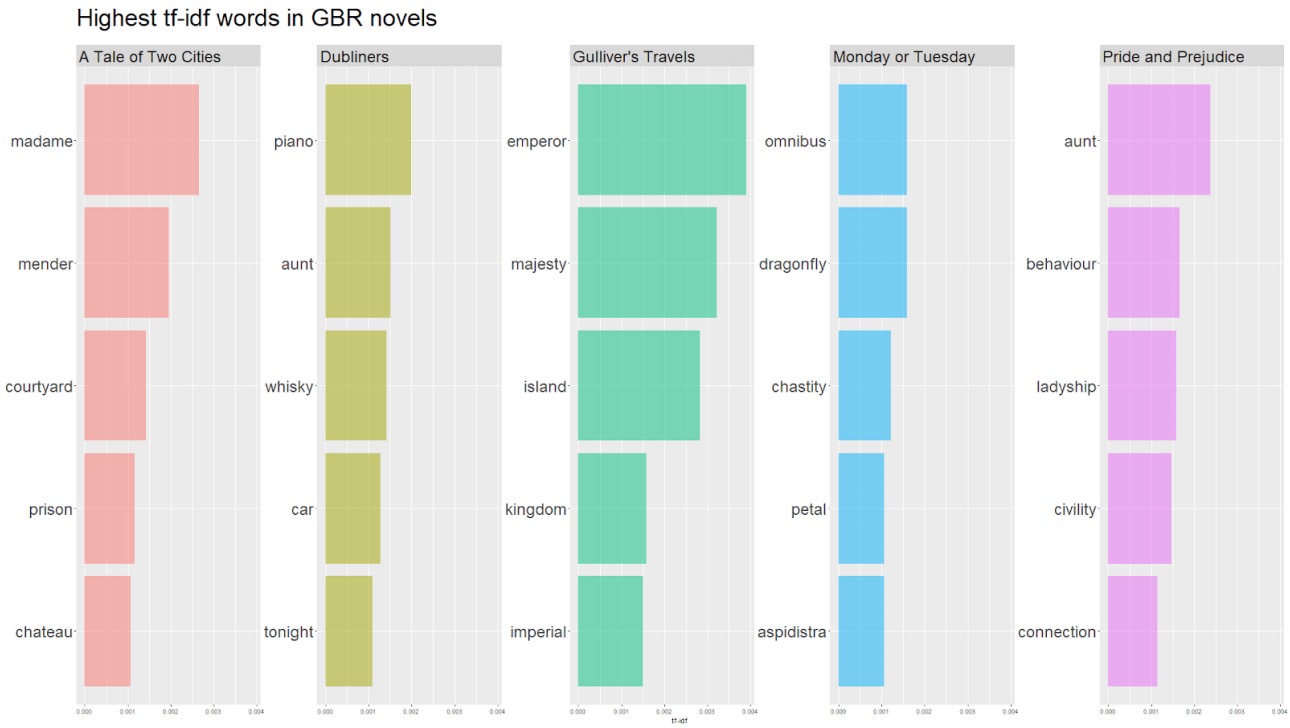
英國	美國
• A Tale of Two Cities • Dubliners • Gulliver's Travels into Several Remote Nations of the World • Monday or Tuesday • Pride and Prejudice	• Adventures of Huckleberry Finn • Little Women • Moby Dick • The Great Gatsby • The Scarlet Letter
德國	**法國**
• Metamorphosis • Royal Highness • Siddhartha • The Sorrows of Young Werther • The Trial	• Candide • Madame Bovary • Notre-Dame de Paris • The Flowers of Evil • The Phantom of the Opera

這些文本是我們於大學期間曾經閱讀過，或是在19、20世紀期間，具代表性的作品。我們利用Gutnerbergr套件將其中幾本書從Project Gutenberg: Free eBooks(https://www.gutenberg.org/)下載下來，而未收錄於此套件中的書籍，我們則直接使用httr和rvest套件，直接從網站中爬下來。其中德國及法國的文本，我們則是採用英文譯本作為我們的分析文本。

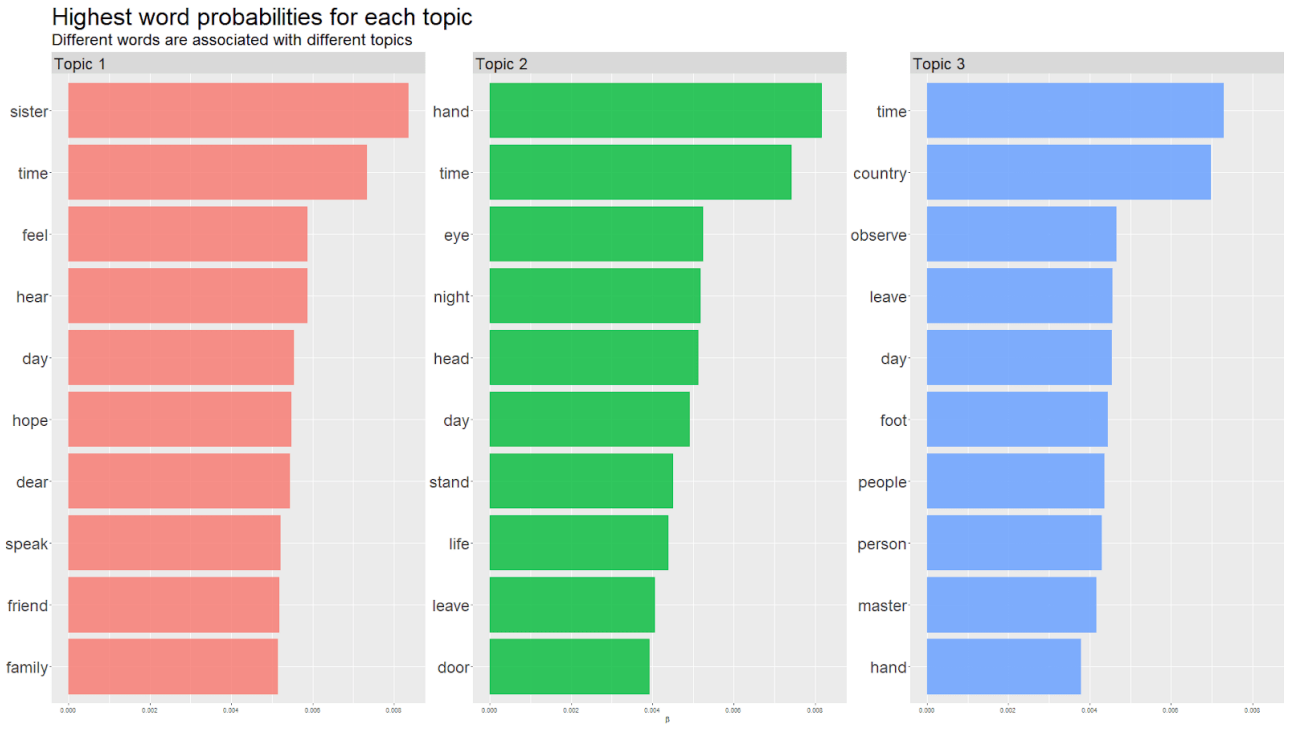
四、分析方式

在將文章存下來後，我們將文本進行斷詞(Tokenization)和詞型還原(Lemmatization)，並移除文本中的停用詞(Stop Words)，完成文本前處理。

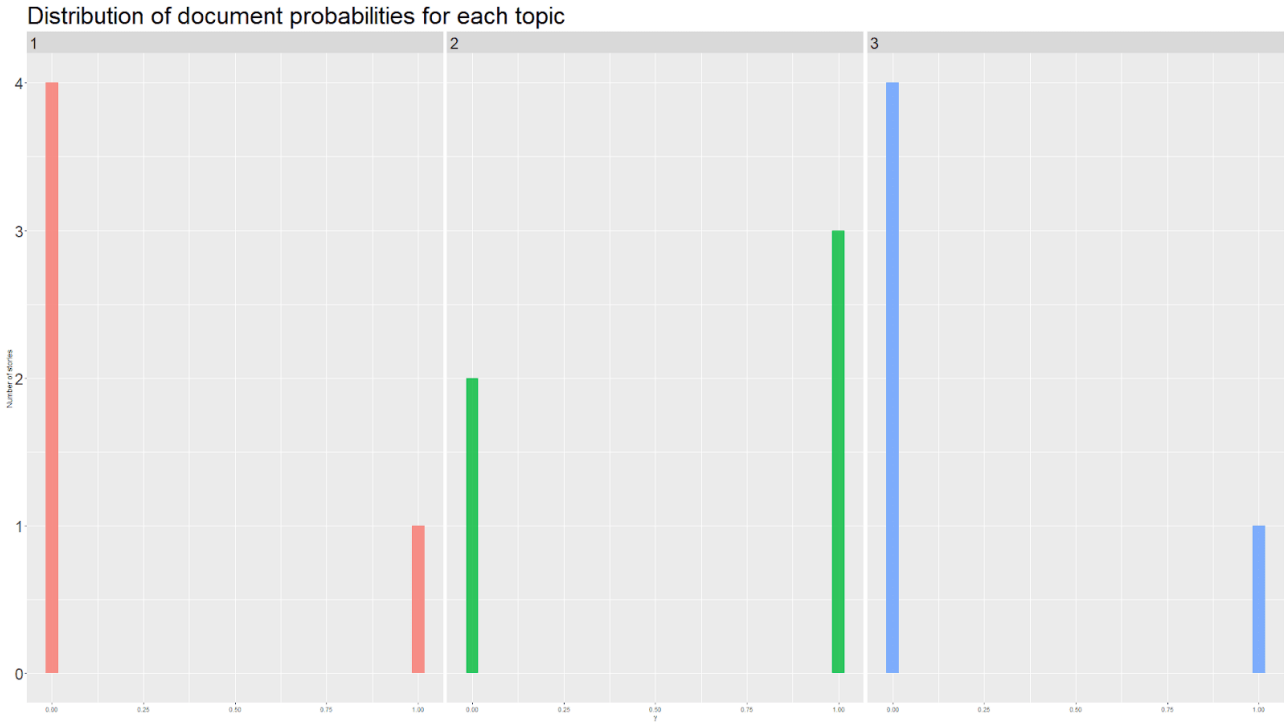
經過文本處理後，我們接著利用TF-IDF找出各國書籍的高頻字，並建立各國的詞頻表，接著再以ggplot2 套件，將詞頻表進行視覺化（如下圖）。



此外，我們也使用stm和quanteda這兩個套件建立主題模型，將文本進行主題模型分析，找出各國文學常出現的主題，並將分析結果視覺化（如下圖）。



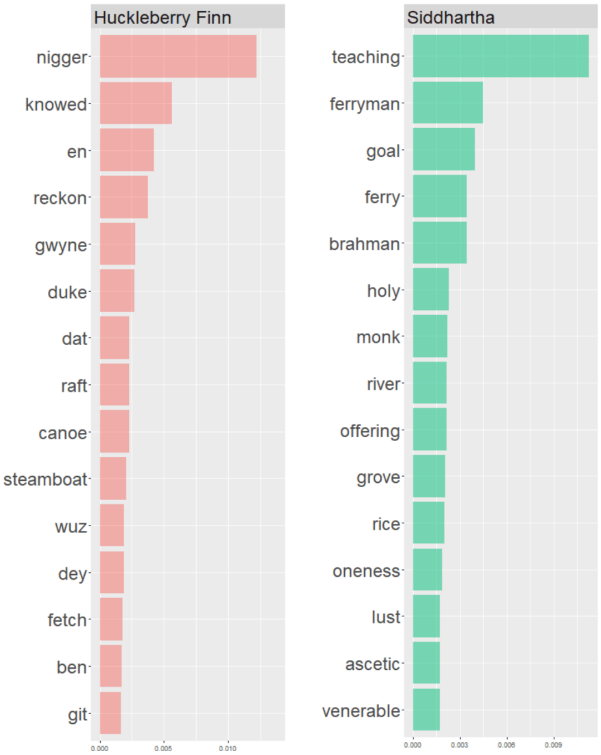
最後，我們也將主題模型分析過後的結果，繪製出主題分布圖（如下圖）。利用該圖，我們就能得知各個國家有那些較為熱門的主題。

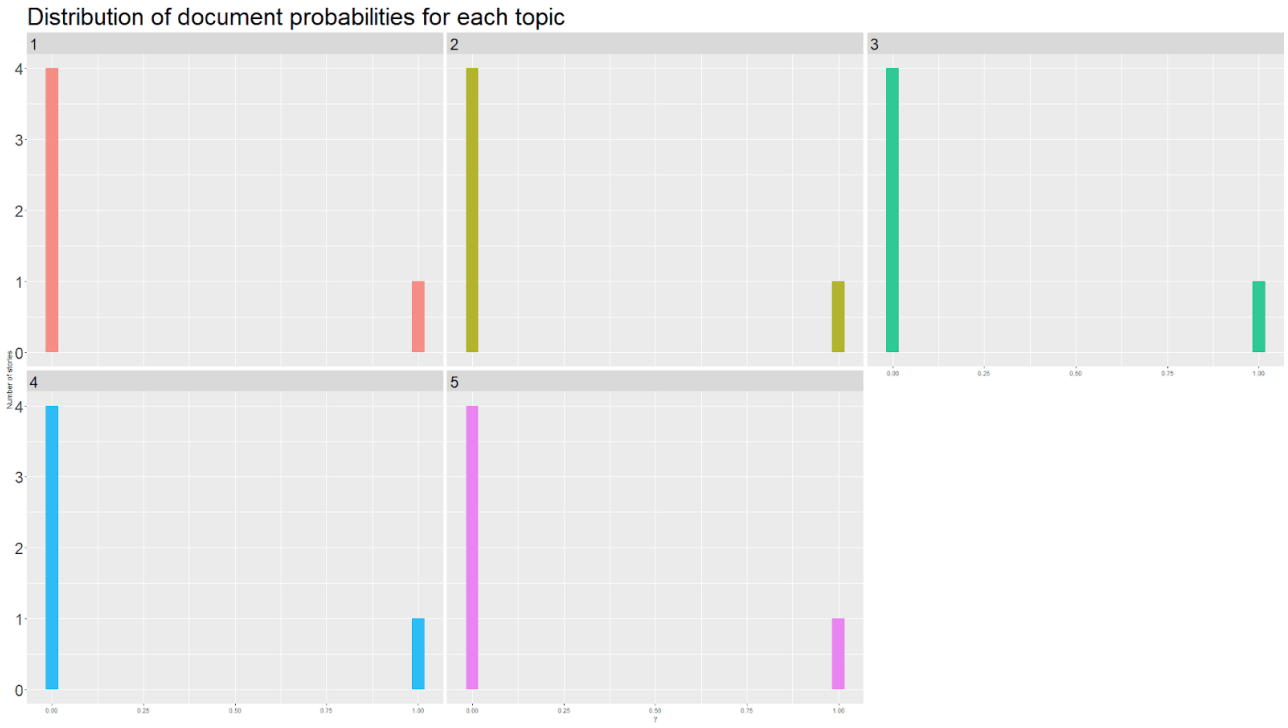


五、結論

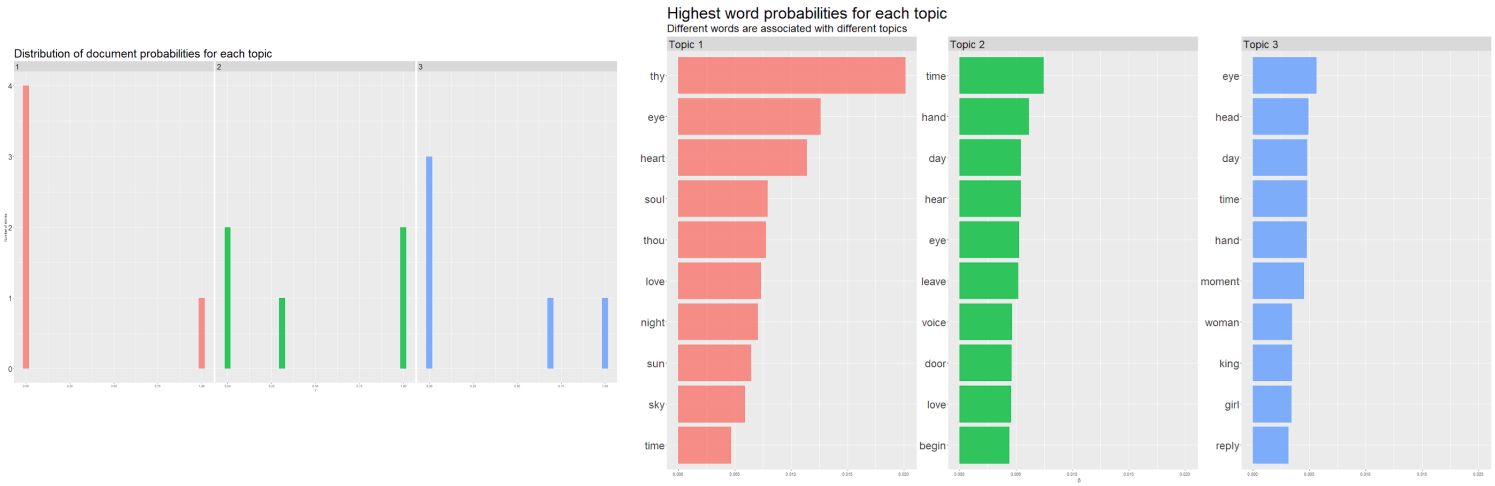
根據分析結果，可以從tf-idf的結果快速的窺見每本書大致的重點，且許多書有其獨特的寫作手法，tf-idf也某種程度上有顯示出來。

以美國文學中的馬克吐溫的頑童歷險記來說（左圖左），作者刻意誤拼了很多字，以表達美國南方的黑人口音，像是gwyne代表going、dat、wuz等等；再舉德國的流浪者之歌為例，從monk、oneness、ascetic這些字，我們可以推測或許與修行、對真理的頓悟有關。





以主題模型的結果來說，我們選的不同國家的書各自內容主題差異蠻大，所以可看到在主題的分佈上（上圖），以法國文學為例，一旦分類成五個主題，每本書就只會各自對應到一個不同的主題。



但若分類成三個主題（上圖左），就會有兩本以上的書被分配到第二和第三個主題。然而再看到三個主題下的高頻字（上圖右），仍舊很難辨識出每個主題大概的內容，我們推測是因為每本書的內容實在差異太大，若硬分三個主題，會造成模型的結果不佳。

這樣的差異，或許可以告訴我們，每個作家的文風其實大大不同，以德國為例，有卡夫卡悖謬的寓言式故事，以隱喻的手法勾勒出一個既冷漠陌生又熟悉的社會，主角逃不過暗處強大力量的操縱、在荒誕的真實中絕望的煎熬；又有歌德筆下的少年維特為愛殉道的理想主義者，內心狂放的感情因得不到宣洩，而將生命陪葬愛情。雖然我們只取了某些作者的單一本書，可能難以具代表性，但我們能大略推測出即便處在相同時代或國家，不同作者表達的手法與內容有其獨特性，無法如同先前的迷因圖，以一句話做概括。

在此次專案中，我們雖然初步分析了來自不同國家文學作品的高頻字與主題比較，資料的取得與分析還是受到了限制。資料分析上，遇到如文本長度不一與Gutenberg上難以找到非英文文學的英文翻譯等問題，未來期許透過更嚴謹的設計來解決。文本分析上也能透過限縮文本主題類似、而作者來自國家不同的分類方法，以增加文本的可比性，或許能解決主題模型中主題間關聯不高的問題。

在以往的文學課中，我們多透過一頁頁精讀文本，並深入探討的分析手法來了解作者作者們留給我們的文字。經過這次專案，透過程式進行文本分析，能讓我們對整個長篇文本有更宏觀的認識，並且更迅速的做文本間的比較，相信對於未來文學賞析上，會是一個新穎特別的視角。

附錄：

分工表		
姓名	**工作**	**投入程度**
陳聯輝 (code組)	資料蒐集、資料處理、TF-IDF、Topic Modeling、結果視覺化	6
	簡報內容發想、圖片製作	4
周昕妤 (code組)	資料蒐集、資料處理、Shiny介面製作、結果視覺化	6
	簡報內容發想、圖片製作	4
黃彙茹 (報告組)	資料處理、Shiny介面製作	4
	簡報製作、結果分析、報告專案成果、分析與限制	6
吳錤 (報告組)	資料處理、TF-IDF、Topic Modeling	4
	簡報製作、報告專案動機、過程與方法	6