國立臺灣大學文學院圖書資訊學系

碩士論文

Department of Library and Information Science

College of Liberal Arts

National Taiwan University

Master's Thesis

使用共字網絡社群偵測與 LDA 主題建模技術對 TED Talks 進行主題萃取

Using co-word network community detection and LDA topic modeling to extract topics on TED Talks.

洪莉婷

Li-Ting Hung

指導教授：唐牧群 博士、林頌堅 博士

Advisors: Muh-Chyun, Tang, Ph.D., Sung-Chien, Lin, Ph.D.

中華民國 111 年 6 月

June, 2022

doi:10.6342/NTU202201269

中文摘要

一、研究目的與問題

本研究藉由文字探勘技術對 TED Talks 上的影片進行探勘，並根據不同欄位的資料類型，在社群偵測和主題建模兩種自動化方法之間選擇合適的方法探索 TED Talks 的主題，以進一步觀察呈現影片不同面向資訊的資料欄位，在詮釋同一份資料集的主題上有何同與不同之處。TED Talks 平台的資料除了有影片基本描述資訊外還有影片內容的逐字稿，因此提供了以量化方式研究影片內容的可能。

由於不同欄位的資料類型所適用的主題萃取方法不同，針對資料類型選用最能夠有效探勘其內容的方式，才能夠確保結果具有解讀意義。此外，不同欄位的資料描述的是影片不同面向的資訊，因此就算是針對同一份資料集進行分析，也可能因為分析對象來自不同欄位以及其搭配的分析方法，而得到不同數量或詞彙組合的分群結果。然而，也正因探索的對象是同一份資料集，所以不同方法的結果所呈現出來的主要主題不應相去太遠，甚至可受惠於探勘的資料面向多元而相互補充，使得最終對於資料集中蘊含之主題的詮釋更加完整。

具體而言，本研究切入分析的角度與欲回答的研究問題為：（一）對 TED Talks 資料集進行探索性資料分析，以認識資料集概貌，並初探資料集中可能包含的主題；（二）使用不同方法對 TED Talks 的不同資料欄位進行主題萃取，以完整對資料集的主題探索；（三）以質化和量化的方式，觀察不同主題萃取結果中的主題並計算主題之間的相似性。

二、研究資料與方法

本研究使用的 TED Talks 影片資料是在 2021 年 4 月 8 號使用 Gupta（2020）在 Github 平台公開提供之程式碼，以 Python 進行爬取並在當天完成。本次研究選用來萃取 TED Talks 資料集主題的欄位包含「相關標籤」、「影片描述」（由「影片簡介」和「講者序言」合併而成）以及「逐字稿」。

doi:10.6342/NTU202201269

本研究首先對 TED Talks 資料集的多個欄位進行探索性資料分析，以認識資料集概貌。接著，則是根據三個欄位的資料格式，使用不同方法進行主題萃取，以完整對資料集的主題探索——針對相關標籤，本研究透過 Gephi 建立共字網絡以進行社群偵測，並在嘗試不同過濾策略後選擇模組化分數與可解讀性較高的網絡作為主題萃取結果。屬於自然語言的影片描述和逐字稿欄位則需先進行資料預處理，再使用 Gensim 套件進行主題建模，並考量模型的一致性分數選擇合適的主題模型。最後，再根據主題中的關鍵字以及關鍵字組合所反映出的主題意涵，並以典型文本為輔助，給與各主題命名。在三種主題萃取結果的比較，則是分別以質化和量化的方式，觀察不同主題萃取結果中的主題，並以相關標籤為依據建立主題的向量座標，以計算主題之間的相似性。

三、研究結果

資料集中的熱門影片通常是與日常生活較為相關的經驗與議題，而這些影片的演講調性較為輕鬆，講者傾向用與觀眾生活經驗連結的方式，與觀眾分享知識，而這些知識也多是落在與思想、行為與感受等範疇。在影片的標籤使用上，2000 年以前的常見標籤為科技（technology）、設計（design）與科學（science），與 TED 創立時欲傳播的三大主題的思想（TED, technology, entertainment, design）大致相符；2000 以後至今，代表不同議題的標籤變得愈來愈常見，於各地獨立策劃的 TEDx 演講和 TED-Ed 系列的影片也愈來愈多。

在三種主題萃取方法所得到的結果中，相關標籤共字網絡的社群偵測到 13 個主題，對影片描述欄位和逐字稿欄位進行主題建模，則分別得到了 25 與 40 個主題。相關標籤中的主題反映出 TED 初期創辦時關注的三個核心思想相關的主題，包含在熱門影片觀察到的生活經驗相關知識，以及醫學、環境、生物技術等主題，並也偵測到 TED-Ed 動畫與音樂表演這兩種較為特別的主題。在影片描述欄位的主題模型中，則萃取出比相關標籤欄位萃取出更多可以描述影片內容的主題，逐字稿則因篇幅較大且直接反映文本內容，因此無論是在數量或是對於內容深度的

doi:10.6342/NTU202201269

探索，都比從相關標籤欄位或是影片描述欄位要萃取出主題種類多元且意涵更為聚焦的主題。此外，影片描述的主題模型顯示「D8 心理學」與「D12 商業」同為近兩年佔比最高的主題；逐字稿主題模型中近年來較常見的主題則為「T6 希臘羅馬神話」、「T27 環境污染與碳排放」、「T11 個人與職場」、「T37 人工智慧與機器學習」和「T20 社群媒體與網路」等。

在比較三個主題萃取結果後則可以發現，1) 以社群偵測作為方法，在相關標籤共字網絡中萃取出來之主題的數量較少且意涵較為廣泛，原因在於相關標籤在平台上有分類與檢索的功能，故可以將其視為一種如控制詞彙之較為固定的檢索點，而經過相似性的計算可以發現標籤欄位所偵測出來的主題範疇不出另外兩個欄位的主題萃取結果，並確實可以反映影片資訊與內容。2) 分別在影片描述欄位和逐字稿欄位建模後得到的主題，數量較多、主題內的意涵較為狹義，且多具有較高的專指性，原因在於其分析的對象是包含多元詞彙的自然語言文本，又因資料來源的欄位不同，而分別反映出影片的背景資訊和影片的實際內容。3) 三個主題萃取結果中的主題，也分別以不同角度形塑本研究對資料集的認識：相關標籤的主題顯示 TED Talks 主要談論的議題（discussed issues），而影片描述的主題展示出影片所關懷的是議題中的哪些面向（concerned aspects），逐字稿的主題則可以代表影片內容真正的核心主題（core topics）。

四、結論與討論

透過同時對三個欄位進行文字探勘，並選擇使用適合該欄位之資料型態的主題萃取方法，最後進行結果之間的比較，本研究不僅是從影片觀看者的角度發掘TED Talks 演講中所蘊含的主題，亦透過文字探勘掌握資料集的主題範疇，使得對於 TED 網站的內容探索也更為全面與完整。

關鍵詞：TED 演講、共字網絡、社群偵測、LDA 主題建模、主題萃取

doi:10.6342/NTU202201269

Abstract

1. Purpose and Question

This study applies text mining techniques to explore topics in TED Talks videos. Two automated clustering methods - community detection and LDA topic model – were used to perform topic extraction based on the type of text data in various fields of TED Talks videos. The purpose of topic extraction is to find the topic structure that implies semantic connections between the data from the content characteristics of the input. Automated topic extraction can be realized by a variety of unsupervised learning methods, such as cluster analysis, community detection, and topic models. The results of topic extraction can help users understand the coverage of topics of the data collection, identify similarities and differences between data, and improve the efficiency and comprehensibility of information retrieval.

TED is a non-profit organization dedicated to spreading ideas in short and powerful talks. The early emphasis of the talks was on technology, entertainment, and design, but now they cover topics from all subjects and fields. The reason for choosing TED Talks as our data collection for topic extraction is that, in addition to the wide range of the topics they cover, the video content contains many fields that can be used for topic extraction. Despite different aspects of the topic information various fields can provide, fields of different data types require different topic extraction methods to ensure that the results are of good quality and that effective interpretations can be made from them.

As a result, it remains an empirical question whether different methods applied in different fields might generate relevant or similar results, and the purpose of the study is to investigate whether the results define equivalent topic ranges of the data set and yet yield complementary topics that make the interpretation of the topic structure more complete.

doi:10.6342/NTU202201269

2. Dataset

All 5050 TED Talks data was collected using a Python program on April 08, 2021. Topic extraction was performed on three data fields of the videos: "related tags", "video descriptions" (combined with text from the fields "talk description" and "speaker: why listen"), and "transcript".

3. Method

For topic extraction from “related tags,” a co-word network based on which was constructed using Gephi for community detection. To find communities that best represent topics implied in “related tags,” different filtering strategies were used to generate a community detection result with a higher modularity score and better interpretability. As for “talk description” and “transcript,” since texts in which are both in natural language forms, data preprocessing including removing punctuations, deleting stop words, and lemmatization were required before applying Gensim for topic modeling. Coherence scores of different models trained under different topic numbers would be considered when selecting the model that best represents the topic extraction result of the data field. The result of a topic extraction is some topics with their own sets of keywords; thus, the topics would be named by inferring the meaning behind the combination of their keywords.

Lastly, the relevance and similarities between topics from three extraction results would be discussed both qualitatively and quantitatively. The qualitative analysis involves the comparison of topics with similar namings, and the quantitative analysis would be performed by generating a vector coordinate for each topic based on related tags and then calculating the cosine similarities between them.

4. Results

The result of community detection on the related tag network reveals 13 topics, and the best topic models for video description and transcript are of 25 and 40 topics, respectively. Most topics in LDA topic models can correspond to a certain tag group in the community detection result. Relevance or similarities between topics can be found either from topic names or by calculating their cosine similarities, and the latter shows that topics extracted from two natural text data fields by LDA topic models share more similar topics and also higher similarity scores. Overall, the results suggest that all three topic extraction methods can unveil equivalent topic range implied in the content of TED Talks videos.

5. Conclusion

Among the three results, however, groups generated by community detection express the broader meaning of topics as related tags, the data source, carry not only labeling but retrieval and classification functions on the TED Talks website, and visualization has the advantage of giving a readily accessible overview of the topics covered. Topics of video description and transcript in videos, which were discovered by the LDA topic model, are more delicate and precise as the source data are in natural language form. Keywords in topics are found to be able to show nuances in between topics and thus help give a more complete interpretation, and videos with the highest probabilities to appear in topics also help understand the core meaning of the topics. Nevertheless, there is still a little difference between the two LDA model results. Because transcripts are text data directly reflecting video contents and therefore can be seen as their surrogates, topics generated from them are more intuitive, making it easier to interpret the result.

Kewords: TED Talks; Co-word Network; Community Detection; LDA Topic Modeling;

Topic Extraction

doi:10.6342/NTU202201269