歐盟 JRC 技術報告之 SeTA 語意文本分析工具介紹

一、前言

文本（text）指的是由語言符號或一段文字所組成的訊息結構體。文本的語意不可避免地會反映人的特定立場、觀點、價值和利益，因此，由文本內容分析，可以推斷文本提供者的意圖和目的。

世界上每天產生的文字與內容，數量遠超過我們的想像，不論是專利文件、研究論文或部落格文章等，如何將這些內容裡的文本「資料」轉換成對使用者有用的「資訊」，還需要有文本分析的技術取出文本中有價值的資訊；而其中主題提取的技術就是透過演算法自動找出這篇文章的主題，將大量的文字分類成重要的關鍵字供後續分析使用。

二、SeTA(Semantic Text Analysis tool)工具簡介

政策制定的過程中所有階段的資訊都需要及時且具關聯性，在歐盟法律文本資料庫(EUR-Lex，網址：https://eur-lex.europa.eu)每年有超過一萬五千個文本產生，而光一個科學論文資料庫(Scopus)，就有超過七千萬項，在沒有分析工具幫助的情況下，策略分析師是無法閱讀並消化如此大量的數據。

因此，為了克服這項難題，由歐盟委員會的 JRC 聯合研究中心(Joint Research Centre，European Commission)開發了一種新工具「SeTA (Semantic Text Analysis tool)語意文本分析工具」，該工具運用大數據、機器學習及自然語言處理等技術，並運用到一個知識探勘和推薦系統中，可以提供策略分析師：

(一) 理解概念：如立法時跨領域所使用的同義詞和背景。

(二) 理解詞彙發展：如過去 50 年不斷發展變化的意義和背景。

(三) 搜尋歐盟執行委員會公共知識資料庫，範圍包括：

(1) 歐盟法律文本資料庫(EUR-Lex)

https://eur-lex.europa.eu

(2) 歐盟出版品平台(EU Bookshop)

https://op.europa.eu/en/web/general-publications/publications

(3) 社會研究與發展資訊服務平台（Cordis）

https://cordis.europa.eu

(4) JRC 出版物管理系統(JRC PUBSY)

https://publications.jrc.ec.europa.eu/repository

(5) 歐盟開放資料平台(EU ODP)

https://data.europa.eu/euodp/en/home

(四) 通過內容相似性查詢文件：複製或查看時間內相關文件，並以其相似度最密切內容來探尋大多數相關文件。

(五) 各領域間轉換知識：例如「歐盟水資源框架指令(WATER FRAMEWORK DIRECTIVE)」是關於 WATER 水資源管理，系統如何分析出「廢棄物 (WASTE)」，經由系統處理後可以得出的結果包括「歐盟廢棄物框架指令 (Eu waste framework directive)」、「廢棄物框架指令 (waste framework directive)」、「修訂廢棄物框架指令 (revised waste framework directive)」。

(六) 直接查找內文最常見用詞，並解釋其用詞在兩者或多種用詞間的關聯性。

三、SeTA 語意文本分析工具分析流程

(一)語料庫準備(Corpus preparation)

語料庫來源與歐盟公共政策相關文件共有 50 多萬份，來源包括歐盟法律文本資料庫(EUR-LEX)、歐盟出版品平台(EU Bookshop)、社會研究與發展資訊服務平台（Cordis）、JRC 出版物管理系統(JRC PUBSY)和歐盟開放資料平台(EU ODP)。

(二)類神經網路訓練

類神經網路可以學習任何函數和可用資料函數定義。因此，類神經網路訓練過程基本要素包括資料整備(Data preparation)、特徵工程(Feature engineering)和範圍含蓋(Domain coverage)，成為訓練中獲得有意義和可分析的基本要素。

經過訓練的模型網路將語料庫中的數十億單詞和片語表示為簡短的數學向量(Mathematical Vectors)。這些向量來源於輸入語料庫中單詞的位置，不僅代表每個單詞，而且還能捕捉到每個單詞的意義。

最重要的發現是 Word2Vec 中的 Skip-Gram 類神經網路模型，能進一步的歸納出相似的群集，從下圖中可以清楚看出自動分類後導出的結果。

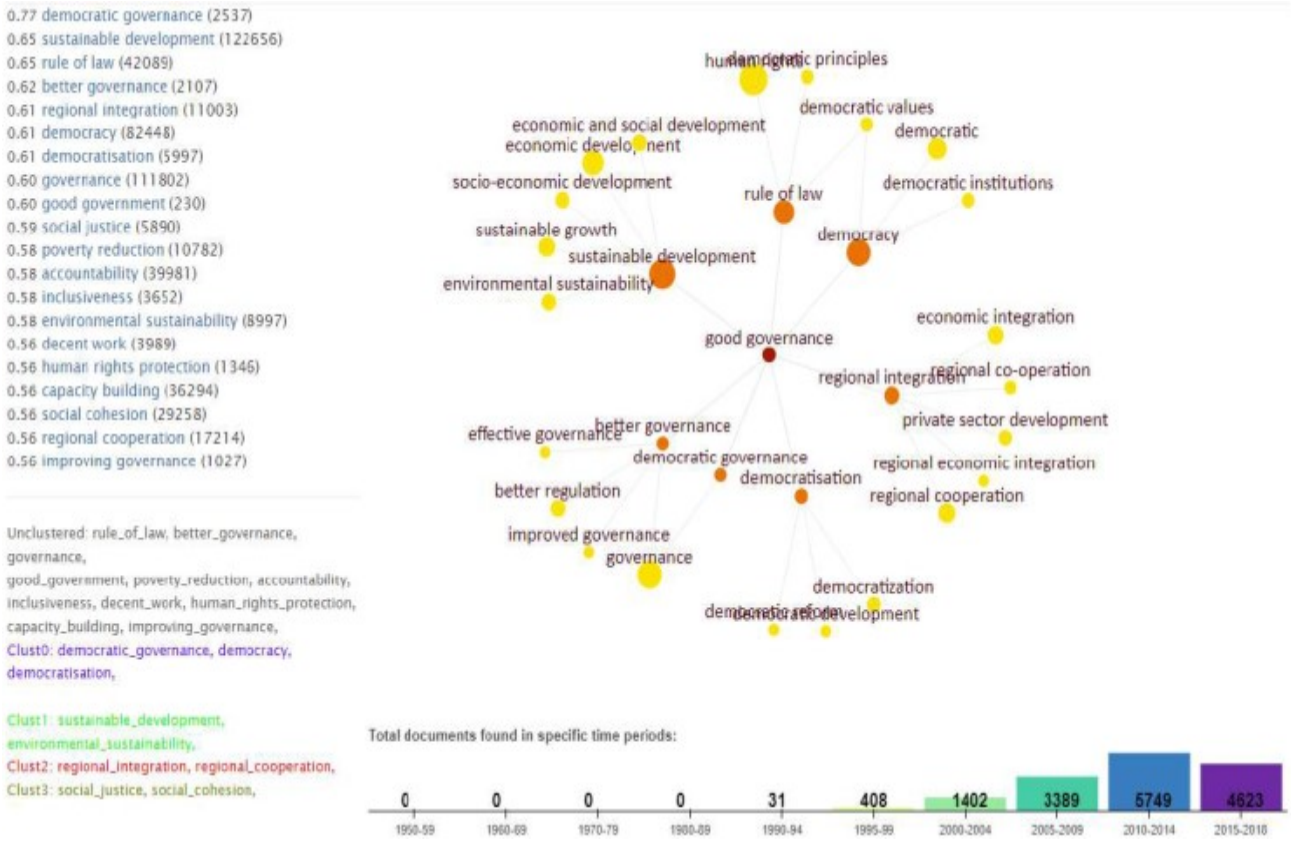

圖1 Word2Vec 中的 Skip-Gram 類神經網路模型圖

右下角顯示了這些術語在過去 50 年中出現的情況。在右上角，我們可以看到由詞相似性構造的網路圖。

(三)研究分析、驗證和結果分析。

每個經過訓練的類神經網路都會產生一個包含 300 行的表格，並儲存類神經網路比重的狀態，以及每一行都有 1~2 百萬列的單詞及片語。

四、未來研究方向

針對目前的 SeTA 工具，仍有許多未來研究方向是值得探討研究的，包括：

(一) 事實查核：將歐洲統計局資料庫(EUROSTAT)與語意句分析集成在一起。

(二) 豐富數據：提取的資訊可附有來自其他來源的相關資訊。目前，團

隊以維基百科網站(Wikipedia web)作為一個外域知識擷取的來源，目前也正在考慮其他資料擷取來源。

(三)文件偏見與意圖解析：分析文件應該是公正的，而從句子提示分析中可以有效幫助分析人員辨識公正性的部分。

(四)使用 API 介面：動態產生完整記錄文件和演算法，包括連結存取知識庫。

五、結論

最近幾年，機器學習 (Machine Learning)受到各界極大的關注，也在許多應用上獲得令人驚艷的成果。機器學習可以應用的範疇非常廣泛，技術途徑與工具也是五花八門，常常讓人無法適從。

相較於其他領域所處理的數據是以影像、聲音、數字為主，檔案管理則是以文字 (text)為主要處理對象。今天我們要介紹的就是一種由歐盟的「歐洲委員會(European Commission)」下屬的「聯合研究中心 (Joint Research Centre)」所開發的文字語意分析器(Semantic Text Analyser，SeTA)，適用於分析大量的文件，由文字中萃取有用的資訊，並且轉化為以視覺方式呈現的知識圖 (Knowledge graphs)，可以協助管理者與政策制定者輕易地獲得更廣泛，更完整，原本被隱藏在文件中的資訊與知識。

SeTA 融合了在大數據、機器學習、自然語言處理(Natural Language Processing, NLP)方面最新的技術與發展，建構了知識探索(Knowledge exploration)

與推薦引擎(Recommendation Engine)，使得政策制定者能夠由大量的文件中找出特定的術語、術語的同義字及術語的發展歷程，並且透過內容相似度 (Content Similarity)找出所有與這個術語相關的文件。換句話說，SeTA 可以從數以萬計的文件中，依據使用者輸入的關鍵字，例如輸入全球暖化，可以找出的相關術語有化石燃料、北極冰山、亞馬遜雨林等，並且呈現這些相似或相關術語的文件，按照語意相關度或相似度來排序，並以圖形視覺方式呈現所推薦的文件。

因此，SeTA 可以被視為一種高階的文件過濾器，使得使用者能夠將精力集中於相關的文件，大幅提高工作效率與品質。

綜觀現今國內語意文本分析進展，中央研究院於 2011 年開發了一個中文斷詞系統，此系統目前建立約 10 萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料，並且在 2019 年決定將中文斷詞工具開源釋出，此舉對於國內致力於開發自然語言處理的研究人員，無疑是一大福音。未來中研院的詞庫小組計畫也會建立一個具有百萬中文詞庫的知識庫，並且持續發展相關的自然語言處理工具。

未來政府部分或許可以應用相關技術來找尋相關資訊，像是應用在機關業務系統、公文檔管相關系統或檔案系統等進行相關搜尋，來提高工作效率與品質。

六、參考資料

(一) https://ec.europa.eu/jrc/en/publication/semantic-text-analysis-tool-seta

(二) https://eur-lex.europa.eu

(三) https://publications.europa.eu/en/web/general-publications/publications

(四) https://cordis.europa.eu/

(五) http://data.europa.eu/euodp/en/home

(六) http://ckipsvr.iis.sinica.edu.tw/

(七) https://panx.asia/archives/50161

(八) https://www.ithome.com.tw/news/132838