IBM SPSS Modeler Text Analytics 18.2.1 使用手冊

注意事項

使用此資訊和支援的產品之前，請先閱讀 第 207 頁的『注意事項』 中的資訊。

產品資訊

除非新版本另有指示，否則本版適用於 IBM SPSS Modeler Text Analytics 18.2.1 版以及所有後續版本和修訂版。

目錄

前言

IBM® SPSS® Modeler Text Analytics 提供強大的文字分析功能，其使用先進的語言技術及「自然語言處理程序 (NLP)」，可快速處理各式各樣的未結構化文字資料，並可從此文字擷取及歸納主要概念。此外，IBM SPSS Modeler Text Analytics 還可以將這些概念分門別類。

一個組織內所存放的資料大約有 80% 的格式為文字文件，例如，報告、網頁、電子郵件和客服中心註記。一個組織要能更深入瞭解其客戶的行為，文字為關鍵要素。納入 NLP 的系統可以用智慧的方式擷取概念，包括複合詞組。甚且，基礎語言知識可利用意義和上下文，將術語分類為相關的群組，如產品、組織或人員。由此，您可以快速判定資訊與您需求的相關性。這些擷取的概念和種類可以與現有的結構化資料（如個人背景資訊）結合，並套用到 IBM SPSS Modeler 的完整資料採礦工具套組中的建模，形塑出更完善且更專注的決策。

語言系統極度取決於知識，亦即其字典中所包含的資訊越多，結果的品質越高。IBM SPSS Modeler Text Analytics 隨附一組語言資源，如術語和同義字字典、檔案庫和範本。此產品進而可讓您開發及微調這些語言資源以符合您的環境定義。精煉語言資源經常是一項反覆的程序，對於精確擷取及分類概念實屬必要。另包含特定領域的自訂範本、檔案庫和字典，如 CRM 和基因體。

關於 IBM Business Analytics

IBM Business Analytics 軟體提供完整、一致且準確的資訊，決策者可信任此資訊，並藉以改善營運績效。包括商業智慧、預測分析、財務績效和策略管理，以及分析應用程式的整合型產品組合，為目前績效提供了清晰、即時且具行動性的前瞻眼界，以及預測未來成果的能力。結合了豐富的業界解決方案、有效實證和專業服務，每種規模的組織都能引爆最高效能，確實自動化執行決策，並且交付更棒的成果。

在這項產品組合中，IBM SPSS Predictive Analytics 軟體有助於組織預測未來事件，並且針對前瞻概念提前行動，創造更棒的營運成果。全球的商業、政府和學術客戶相當倚重 IBM SPSS 技術所帶來的競爭優勢，藉此做為吸引、保有和發展更多客戶，同時降低可能的不實詐欺風險。藉由將 IBM SPSS 軟體併入每天作業，這些組織成為預測型企業 – 足以駕馭決策並使決策自動化處理，以符合營運目標，並且達到可測知的競爭優勢。如需更多資訊，或是聯絡代表人員，請造訪 http://www.ibm.com/spss。

技術支援

技術支援人員可提供客戶維護的服務。客戶可以聯絡技術支援人員，尋求 IBM Corp. 產品使用協助，或尋求其中一個受支援硬體環境的安裝協助。若要聯絡技術支援人員，請參閱 IBM Corp. 網站，網址：http://www.ibm.com/support。請求協助時，請準備好識別您個人、組織和支援合約的相關資訊。

第 1 章 關於 IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics 提供強大的文字分析功能，其使用先進的語言技術及「自然語言處理程序 (NLP)」，可快速處理各式各樣的未結構化文字資料，並可從此文字擷取及歸納主要概念。此外，IBM SPSS Modeler Text Analytics 還可以將這些概念分門別類。

一個組織內所存放的資料大約有 80% 的格式為文字文件，例如，報告、網頁、電子郵件和客服中心註記。一個組織要能更深入瞭解其客戶的行為，文字為關鍵要素。納入 NLP 的系統可以用智慧的方式擷取概念，包括複合詞組。甚且，基礎語言知識可利用意義和上下文，將術語分類為相關的群組，如產品、組織或人員。由此，您可以快速判定資訊與您需求的相關性。這些擷取的概念和種類可以與現有的結構化資料（如個人背景資訊）結合，並套用到 IBM SPSS Modeler 的完整資料採礦工具套組中的建模，形塑出更完善且更專注的決策。

語言系統極度取決於知識，亦即其字典中所包含的資訊越多，結果的品質越高。IBM SPSS Modeler Text Analytics 隨附一組語言資源，如術語和同義字字典、檔案庫和範本。此產品進而可讓您開發及微調這些語言資源以符合您的環境定義。精煉語言資源經常是一項反覆的程序，對於精確擷取及分類概念實屬必要。另包含特定領域的自訂範本、檔案庫和字典，如 CRM 和基因體。

部署。 您可以使用 IBM SPSS Modeler Solution Publisher 來部署文字採礦串流以對非結構化資料進行即時評分。部署這些串流的能力可確保順利關閉迴圈文字採礦實作。例如，您的組織現在可以透過套用預測性模型來分析入埠或出埠呼叫程式的即時運算簿附註，以便即時提高您行銷訊息的精確度。

若要使用 IBM SPSS Modeler Solution Publisher 來執行 IBM SPSS Modeler Text Analytics，請將目錄 `<install_directory>/ext/bin/spss.TMWBServer` 新增至 `$LD_LIBRARY_PATH` 環境變數中。

註：從 18.1 版開始，已淘汰日文版的 IBM SPSS Modeler Text Analytics 配接器。

升級 IBM SPSS Modeler Text Analytics

在安裝 IBM SPSS Modeler Text Analytics 之前，您應從現行版本儲存並匯出要在新版本中使用的任何 TAP、範本和檔案庫。我們建議您將這些檔案儲存到在安裝最新版本時不會刪除或改寫的目錄。

在安裝 IBM SPSS Modeler Text Analytics 的最新版本之後，您可以載入已儲存的 TAP 檔案，新增任何已儲存的檔案庫，或是載入任何已儲存的範本以在最新版本中使用它們。

重要：如果您未先儲存及匯出要求的檔案即解除安裝目前版本，將會遺失在舊版執行的任何 TAP、範本與公用程式庫工作，且無法在 IBM SPSS Modeler Text Analytics 的最新版本中使用。

關於文字挖掘

現今，越來越多的資訊以非結構化及半結構化格式保留，例如客戶電子郵件、呼叫中心附註、開放式結尾意見調查回應、新聞資訊來源、Web 論壇等。資訊如此豐富為許多組織帶來一個問題，即詢問他們自己「我們如何可以收集、探索及利用此資訊？」

文字挖掘是一個處理程序，會分析文字資料集合以便擷取關鍵概念及佈景主題，以及揭露隱藏的關係與趨勢，而不需要您知道作者用來表達那些概念的精確單字或術語。雖然它們非常不同，但有時文字挖掘會與資訊擷取發生混淆。雖然準確的資訊擷取及儲存都是巨大的挑戰，但是資訊內品質內容、詞彙及關係的擷取及管理都是重要且關鍵的處理程序。

文字挖掘及資料挖掘

對於每一個文字文章，基於語言的文字挖掘都傳回概念索引，以及有關那些概念的資訊。這個經提取的結構化資訊可以與其他資料來源結合，以解決以下問題：

- 哪些概念一起發生？
- 它們還鏈結哪些內容？
- 可以從所擷取資訊建立哪些更高層次的種類？
- 概念或種類有何預測？
- 概念或種類如何預測行為？

將文字挖掘與資料挖掘結合，可比單獨使用結構化或非結構化資料獲得更深刻的見解。這個處理程序通常包括下列步驟：

1. **識別要挖掘的文字。**準備進行挖掘的文字。如果文字存在於多個檔案中，請將檔案儲存至單一位置。對於資料庫，判定包含文字的欄位。
2. **發掘文字並擷取結構化資料。**將文字挖掘演算法套用至來源文字。
3. **建置概念及種類模型。**識別主要概念及/或建立種類。從非結構化資料傳回的概念數目通常非常巨大。識別用於評分的最佳概念與種類。
4. **分析結構化資料。**採用傳統資料挖掘技術，例如形成叢集、分類及預測性建模，以探索概念之間的關係。合併所擷取概念與其他結構化資料，從而根據概念預測未來的行為。

文字分析與分類

文字分析是一種定量分析，可從文字中擷取有用的資訊，以便可以將此文字內包含的關鍵至適當數目的種類。可以對所有類型及長度的文字執行文字分析，雖然分析方法將有所不同。

記錄或文件越短，分類越簡單，因為它們的複雜度不同，並且通常包含較少的語意不明單字及回應。例如，在使用較短的開放式結尾問題時，如果我們要求人們命名三個最愛假期的活動，我們可能預期會看到許多較短的回答，例如去沙灘、參觀國家公園或什麼也不做。另一方面，較長的開放式結尾回答可能非常複雜且非常長，特別是當回覆者接受過良好教育、積極主動且時間充裕可完成問券時。如果我們要求人們在意見調查中告知我們有關政治信仰，或者提供有關政治的部落格資訊來源，我們可能預期收到有關各種問題及位置的部分較長評論。

能夠擷取主要概念並在極短時間內從這些較長的文字來源中建立有見解的種類是使用 IBM SPSS Modeler Text Analytics 的一個主要優點。此優點源自自動化語言與統計技術的組合，為每個文字分析處理階段產生更可靠的結果。

語言處理及 NLP

管理所有這個非結構化文字資料的主要問題在於沒有撰寫文字的標準規則以供電腦理解。每個文件及每部分文字的語言以及產生的意義均不相同。準確擷取及組織此類非結構化資料的唯一方法是分析語言，並揭露其意義。有數個不同的自動化方法，可從非結構化資訊中擷取概念。這些方法可以分為兩種，語言與非語言。

部分組織已嘗試根據統計與中性網路採用自動化非語言解決方案。這些解決方案使用電腦技術，能夠比人類閱讀者更快速地掃描及分類主要概念。不幸的是，此類解決方案的精確度相當低。大部分基於統計資料的系統只簡單計數單字發生的次數，並計算其與相關概念的統計近似性。它們產生許多不相關的結果或雜訊，並遺失應該找到的結果，這稱為無聲。

為了補償有限的精確度，部分解決方案採用複雜的非語言規則，協助區分相關與無關的結果。這稱為基於規則的文字挖掘。

另一方面，基於語言的文字挖掘將自然語言處理 (NLP) 原則（電腦協助的人類語言分析）套用至文字的單字、片語及語法或結構的分析。納入 NLP 的系統可以用智慧的方式擷取概念，包括複合詞組。甚且，基礎語言知識可利用意義和上下文，將概念分類為相關的群組，如產品、組織或人員。

基於語言的文字挖掘像人類一樣發現文字的意義，方法是辨識具有類似意義的各種單字形式，以及分析句子結構提供用來理解文字的架構。此方法提供基於語言的系統的速度與成本效率，但是它提供遠遠更高的精確度，且需要極少的人為介入。

若要說明擷取程序期間基於統計資料與基於語言的方法之間的差異，請考量每一種方法如何回應有關文件前置正式作業的查詢。基於統計資料與基於語言的解決方案都必須展開單字 reproduction 以包括同義字，例如 copy 及 duplication。否則，將忽略相關資訊。但是如果基於統計資料的解決方案嘗試執行此類型的同義字，搜尋具有相同意義的其他術語，則很可能也包括術語 birth，產生許多無關的結果。對語言的理解會解決文字的語義不明確，讓依定義的基於語言的文字挖掘稱為更可靠的方法。

理解擷取如何運作可以協助您在細部調整語言資源（檔案庫、類型同義字等）時進行作出主要決策。擷取程序中的步驟包括：

- 將來源資料轉換為標準格式
- 識別候選術語
- 識別相當的同義字類別及整合
- 指派類型
- 編製索引，並在要求時將型次要分析器相比對

步驟 1. 將來源資料轉換為標準格式

在這個首要步驟中，您匯入的資料會轉換為可用於進一步分析的唯一格式。這個轉換會在內部執行，且不會變更您的原始資料。

步驟 2. 識別候選術語

請務必瞭解語言擷取期間候選術語識別中語言資源的角色。每次執行擷取時，都使用語言資源。它們以範本、檔案庫及編譯的資源形式存在。檔案庫包括用於指定或調整擷取的單字、關係及其他資訊的清單。無法檢視或編輯編譯的資源。然而，可以在 範本編輯器 中，或者，如果您在互動式工作台階段作業中，則可以在資源編輯器中編輯剩餘資源。

編譯資源是 IBM SPSS Modeler Text Analytics 內擷取引擎的核心內部元件。這些資源包括含基本表單清單且具有部分語音代碼的一般定義檔（名詞、動詞、形容詞等）。

除了那些編譯的資源之外，還隨產品提供數個檔案庫，可用於補充所編譯資源中的類型及概念定義，以及提供同義字。這些檔案庫以及您建立的任何自訂檔案庫都由數個組成。這些包括類型定義檔、同義字定義檔及排除定義檔。

已匯入及轉換資料之後，擷取引擎將開始識別用於擷取的候選術語。候選術語是用於識別文字中概念的單字或單字組。處理文字期間，單字（單一術語）與複合字（多術語）會利用部分語音型樣擷取程式進行識別。然後，利用觀感文字鏈結分析識別候選觀感關鍵字。

註：前方提及的已編譯一般定義檔中的術語代表很可能作為單一術語無興趣或在語言上語意不明的所有單字清單。當您識別單一術語時，會從擷取排除這些單字。然而，當您判定部分語音或查看更長的候選複合字（多術語）時，會重新評估它們。

步驟 3. 識別相當的同義字類別及整合

識別候選單一術語及多術語之後，軟體使用正規化定義檔來識別相當的類別。相當的類別是片語的基本形式，或者同一片語的兩個變式的單一形式。若要確定哪個概念用於相當的類別，，則擷取引擎在列出的順序中套用下列規則：

- 檔案庫中的使用者指定的形式。
- 最常見的形式，如經過前置編譯的資源所定義。

步驟 4. 指派類型

接下來，將類型指派給所擷取的概念。類型是概念的語意分組。此步驟中同時使用所編譯的資源及檔案庫。類型包括此類項目，例如更高層次的概念、正面及負面單字、名字、地點、組織等。請參閱第 163 頁的『類型字典』主題，以取得更多資訊。

語言系統極度取決於知識，亦即其定義檔中所包含的資訊越多，結果的品質越高。修改定義檔內容（例如同義字定義）可以簡化產生的資訊。這通常是一項反覆的處理程序，對於精確擷取概念實屬必要。NLP 是 IBM SPSS Modeler Text Analytics 的核心元素。

擷取如何運作

從回應擷取主要概念及構想期間， IBM SPSS Modeler Text Analytics 依賴基於語言的文字分析。此方法提供基於統計資料的系統的速度及成本有效性。但是它的正確性更高，且需要更少的人為介入。基於語言的文字分析基於已知作為自然語言處理的研究領域，也稱為計算語言。

理解擷取如何運作可以協助您在細部調整語言資源（檔案庫、類型同義字等）時進行作出主要決策。擷取程序中的步驟包括：

- 將來源資料轉換為標準格式
- 識別候選術語
- 識別相當的同義字類別及整合
- 指派類型
- 編製索引
- 符合型樣及事件擷取

步驟 1. 將來源資料轉換為標準格式

在這個首要步驟中，您匯入的資料會轉換為可用於進一步分析的唯一格式。這個轉換會在內部執行，且不會變更您的原始資料。

步驟 2. 識別候選術語

請務必瞭解語言擷取期間候選術語識別中語言資源的角色。每次執行擷取時，都使用語言資源。它們以範本、檔案庫及編譯的資源形式存在。檔案庫包括用於指定或調整擷取的單字、關係及其他資訊的清單。無法檢視或編輯編譯的資源。然而，可以在 範本編輯器 中，或者，如果您在互動式工作台階段作業中，則可以在資源編輯器中編輯剩餘資源（範本）。

編譯資源是 IBM SPSS Modeler Text Analytics 內擷取引擎的核心內部元件。這些資源包括含基本表單清單且具有部分語音代碼的一般定義檔（名詞、動詞、形容詞、副詞、分詞、對等連接詞、限定詞或介詞）。資源還包括保留的內建類型，用來將許多擷取的術語指派給下列類型：<Location>、<Organization> 或 <Person>。如需相關資訊，請參閱主題 第 164 頁的『內建類型』。

除了那些編譯的資源之外，還隨產品提供數個檔案庫，可用於補充所編譯資源中的類型及概念定義，以及提供其他類型及同義字。這些檔案庫以及您建立的任何自訂檔案庫都由數個組成。這些包括類型定義檔、替代定義檔（同義字及選用元素）及排除定義檔。如需相關資訊，請參閱主題 第 155 頁的第 15 章, 『使用檔案庫』。

已匯入及轉換資料之後，擷取引擎將開始識別用於擷取的候選術語。候選術語是用於識別文字中概念的單字或單字組。處理文字期間，不在所編譯資源中的單字（單一術語）會視為候選術語擷取。候選複合字（多術語）使用部分語音型樣擷取程式進行識別。例如，多術語 sports car 後面是 "adjective noun" 部分語音型樣，具有兩個元件。多術語 fast sports car 後面是 "adjective adjective noun" 部分語音型樣，具有三個元件。

註：前方提及的已編譯一般定義檔中的術語代表很可能作為單一術語無興趣或在語言上語意不明的所有單字清單。當您識別單一術語時，會從擷取排除這些單字。然而，當您判定部分語音或查看更長的候選複合字（多術語）時，會重新評估它們。

最終，特殊演算法用來處理大寫字母字串，例如工作職稱，以便可以擷取這些特殊型樣。

步驟 3. 識別相當的同義字類別及整合

識別候選單一術語及多術語之後，軟體使用一組演算法來進行比較，以及識別相當的類別。相當的類別是片語的基本形式，或者同一片語的兩個變式的單一形式。將片語指派給相當類別的目的是確保，例如 president of the company 及 company president 不會視為個別概念。若要確定哪個概念用於相當的類別，即，將 president of the company 還是 company president 用作前導術語，則擷取引擎在列出的順序中套用下列規則：

- 檔案庫中的使用者指定的形式。
- 完整文字主體中最常見的形式。
- 完整文字主體中的最短形式（通常對應於基本形式）。

步驟 4. 指派類型

接下來，將類型指派給所擷取的概念。類型是概念的語意分組。此步驟中同時使用所編譯的資源及檔案庫。類型包括此類項目，例如更高層次的概念、正面及負面單字、名字、地點、組織等。使用者可以定義其他類型。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。

步驟 5. 編製索引

透過在文字位置與每一個相當類別的代表術語之間建立指標，為整個記錄或文件集編製索引。這會假設候選概念的所有受影響形式實例作為候選基本形式進行編製索引。為每一個基本形式計算廣域頻率。

步驟 6. 符合型樣及事件擷取

IBM SPSS Modeler Text Analytics 不僅能夠探索類型及概念，還能探索之間的關係。此產品提供數個演算法及檔案庫，能夠擷取類型與概念之間的關係型樣。嘗試探索特定意見（例如，產品反映）或人員或物件之間的關聯式鏈結（例如，政治群組或基因組之間的鏈結）時，它們特別有用。

分類如何運作

在 IBM SPSS Modeler Text Analytics 中建立種類模型時，您可以從數個不同的技術中進行選擇以建立種類。由於每個資料集都是唯一的，因此技術數目及您套用它們的順序可能變更。由於您對結果的解譯可能與他人不同，因此您可能需要體驗不同的技術，以查看哪個能夠為您的文字資料產生最佳結果。在 IBM SPSS Modeler Text Analytics 中，您可以在工作台階進一步段作業中建立種類模型，並可以在其中探索及細部調整種類。

在本手冊中，**種類建置**是指透過使用一個或多個內建的技術產生種類定義及分類，並且**分類**是指評分或加上標籤處理程序，其中將唯一的 ID（名稱/ID/值）指派給每一個記錄或文件的種類定義。

種類建置期間，擷取的概念及類型用作種類的建置取塊。當您建置種類時，如果記錄或文件包含符合種類定義之一個元素的文字，則會自動指派給種類。

IBM SPSS Modeler Text Analytics 為您提供數個自動化種類建置技術，可協助您快速分類文件或記錄。

分組技術

每一個可用的技術都適用於某些類型的資料和狀況，但是在同一個分析中結合技術來擷取完整範圍的文件或記錄往往非常有用。您可能會看到一個概念在多個種類中或是發現冗餘的種類。

概念根衍生。這項技術會透過取用某個概念，然後尋找與它相關的其他概念（藉由分析是否有任何概念元件在形態上相關或共用根）來建立種類。這項技術對於識別同義複合字概念非常有用，因為每一個產生的種類中的概念都是同義字或是在意義上緊密相關。它處理不同長度的資料，並產生數目較少的精簡種類。例如，概念 opportunities to advance 將會與概念 opportunity for advancement 和 advancement opportunity 群組。如需相關資訊，請參閱主題 第 94 頁的『概念根衍生』。

語意網路。這項技術會先從每一個概念的單字關係延伸索引識別其可能的觀念，然後藉由將相關概念分組來建立種類。這項技術最適用於概念為語意網路所知，且不會過於含糊不清時。當文字包含特殊化術語或是網路不知道專門術語時，它就比較不是那麼有用。在一個範例中，概念 granny smith apple 可以與 gala apple 和 winesap apple 群組，因為它們是 granny smith 的同層級。在另一個範例中，概念 animal 可能與 cat 和 kangaroo 群組，因為它們是 animal 的下義詞。在本版本中，這項技術僅提供英文文字。如需相關資訊，請參閱主題 第 95 頁的『語意網路』。

概念併入。這項技術會根據多詞彙概念（複合字）包含的單字是其他概念中單字的子集還是超集來將它們分組，藉此來建置種類。例如，概念 seat 將會與 safety seat、seat belt 和 seat belt buckle 群組。如需相關資訊，請參閱主題 第 95 頁的『概念併入』。

共現項目。這項技術會從在文字中發現的共現項目來建立種類。其理念是在文件或記錄中經常一起發現概念或概念型樣時，該共現項目反映出一種基礎關係，該關係可能是種類定義中的值。當單字顯著共現時，會建立共現項目規則，可用來作為新的子總類的種類描述子。比方說，如果許多記錄包含單字 price 和 availability（但是有幾個記錄包含其中一個單字而沒有另一個單字），則可以將這些概念群組成共現項目規則 (price & available)，並指派給種類 price 的子種類（舉例而言）。如需相關資訊，請參閱主題 第 96 頁的『共生規則』。

文件數目下限。 為幫助判定共現項目有多有趣，請定義必須包含給定的共現項目，才能用來作為種類中的描述子的文件或記錄數目下限。

IBM SPSS Modeler Text Analytics 節點

除了 IBM SPSS Modeler 隨附的許多標準節點之外，您還可以使用文字採礦節點將文字分析功能納入串流中。IBM SPSS Modeler Text Analytics 為您提供了數個文字採礦節點以執行此作業。這些節點儲存在節點選用區的 IBM SPSS Modeler Text Analytics 標籤中。

包括下列節點：

- **檔案清單來源節點**會產生文件名稱清單作為文字採礦程序的輸入。如果文字位於外部文件而不是資料庫或其他結構化檔案中，則此節點很有用。該節點會輸出一個針對每個所列文件或資料夾包含一筆記錄的欄位，可以選取該欄位作為後續「文字採礦」節點中的輸入。如需相關資訊，請參閱主題 第 9 頁的『檔案清單節點』。

- **「Web 資訊來源」來源節點**可以讀取 Web 資訊來源（例如採用 RSS 或 HTML 格式的部落格或新聞資訊來源）中的文字，並在文字採礦程序中使用此資料。該節點會針對資訊來源中發現的每筆記錄輸出一或多個欄位，可以選取這些欄位作為後續「文字採礦」節點中的輸入。如需相關資訊，請參閱主題 第 11 頁的『Web 資訊來源節點』。
- **語言 ID 節點**是一個程序節點，它會掃描來源文字以判定其撰寫語言，然後在新欄位中標示該語言。主要設計為用於大量資料中，如果您在資料來源中具有多個語言且只想處理一個語言，則此節點特別有用。如需相關資訊，請參閱主題 第 14 頁的『語言節點』。
- **文字採礦節點**使用語言方法來從文字中擷取主要概念，容許您使用這些概念及其他資料來建立種類，以及提供一種能力來根據已知型樣（稱為文字鏈結分析）識別概念之間的關係和關聯。可以使用該節點來探索文字資料內容或產生概念模型或種類模型。概念和種類可以與現有結構化資料（例如個人背景資訊）相結合並套用於建模中。如需相關資訊，請參閱主題 第 18 頁的『文字挖掘建模節點』。
- **文字鏈結分析節點**會擷取概念，還會根據文字內的已知型樣識別概念之間的關係。可以使用型樣擷取來探索概念之間的關係，以及附加至這些概念的任何觀點或限定元。「文字鏈結分析」節點提供了一個更直接的方法來識別及擷取文字中的型樣，然後將型樣結果新增至串流中的資料集。但您可以在「文字採礦」建模節點中使用互動式工作台階段作業來執行 TLA。如需相關資訊，請參閱主題 第 41 頁的『文字鏈結分析節點』。
- 對外部文件中的文字進行採礦時，可以使用**文字採礦輸出節點**來產生一個 HTML 頁面，其中包含指向從中擷取概念之文件的鏈結。如需相關資訊，請參閱主題 第 47 頁的『檔案檢視器節點』。

應用

一般來說，需要例行檢閱大量文件以識別要進一步探索之重要元素的人員可以從 IBM SPSS Modeler Text Analytics 中獲益。

部分特定應用包括：

- **科學和醫療研究。** 探索次要研究資料，例如專利報告、日誌登載文章及通訊協定出版品。識別先前不明的關聯（例如與特定產品相關聯的醫生），呈現街道以進一步探索。最大程度減少藥物探索程序所花費的時間。用作基因研究中的輔助。
- **投資研究。** 檢閱每日分析報告、新聞文章及公司新聞稿，以識別重要策略點或市場變化。此類資訊的趨勢分析顯示一個公司或產業在一段時間內的新興問題或機遇。
- **詐騙偵測。** 用於銀行業及醫療保健詐騙中以偵測異常以及在大量文字中探索紅色旗標。
- **市場調查。** 用於市場調查工作中以識別從開啟到結束之意見調查回應中的重要主題。
- **部落格及 Web 資訊來源分析。** 使用在新聞資訊來源、部落格等資訊中發現的重要構想來探索及建置模型。
- **CRM。** 使用所有客戶接觸點（例如電子郵件、交易及意見調查）中的資料來建置模型。

第 2 章 讀取原始檔文字

用於文字採礦的資料可以採用 IBM SPSS Modeler 使用的任何標準格式，其中包括資料庫或其他「矩形」格式，這些格式用列和欄或以文件格式（例如，不符合此結構的 Microsoft Word, Adobe PDF 或 HTML）來代表資料。

- 若要讀取不符合標準資料結構的文件（其中包括 Microsoft Word、Microsoft Excel 和 Microsoft PowerPoint，以及 Adobe PDF、XML、HTML 等）中的文字，則可以使用「檔案清單」節點來產生文件或資料夾的清單以作為文字採礦程序的輸入。如需相關資訊，請參閱『檔案清單節點』。
- 若要讀取 Web 資訊來源（例如採用 RSS 或 HTML 格式的部落格或新聞資訊來源）中的文字，則可以使用「Web 資訊來源」節點來格式化要輸入到文字採礦程序的 Web 資訊來源資料。如需相關資訊，請參閱第 11 頁的『Web 資訊來源節點』。
- 若要讀取 SPSS Modeler 使用的任何標準資料格式（例如具有一或多個客戶註解文字欄位的資料庫）中的文字，則可以使用任何 SPSS Modeler 來源節點。如需相關資訊，請參閱 SPSS Modeler 節點文件。
- 在處理可能包含數種不同語言的文字的大量資料時，請使用「語言」節點來識別特定欄位中使用的語言。如需相關資訊，請參閱第 14 頁的『語言節點』。

檔案清單節點

若要從以 Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF、XML、HTML 及其他格式儲存的非結構化文件中讀取文字，「檔案清單」節點可以用來產生文件或資料夾清單，作為文字挖掘處理程式的輸入。由於非結構化文字文件不能與 IBM SPSS Modeler 使用的其他資料一樣以相同的方式透過欄位與記錄（列和欄）代表，因此這是必要項。

「檔案清單」節點用作來源節點。

您可以在位於IBM SPSS Modeler視窗底端的節點選用區的IBM SPSS Modeler Text Analytics標籤上尋找這個節點。如需相關資訊，請參閱主題 第 6 頁的『IBM SPSS Modeler Text Analytics 節點』。

重要： 不支援任何含有在機器本端編碼中未包含的字元的目錄名稱和檔案名稱。當試圖執行含有「檔案清單」節點的串流時，任何包含這些字元的檔案或目錄名稱都會導致串流執行失敗。國外目錄名稱或檔案名稱可能會發生這種情況，例如法文語言環境上的德文檔案名稱。

本端資料支援。如果您要連接至遠端 IBM SPSS Modeler Text Analytics Server ，且具有含「檔案清單」節點的串流，則資料應該與 IBM SPSS Modeler Text Analytics Server 位於同一機器上，或者確保伺服器具有對儲存「檔案清單」節點中來源資料之資料夾的存取權。

註：您不能將「檔案清單」節點用於在 IBM SPSS Collaboration and Deployment Services - 評分 配置內進行評分。

檔案清單節點：設定標籤

在此標籤上，您可以定義此節點的目錄、副檔名及輸入。

註：文字挖掘擷取不能在非 Microsoft Windows 平台下處理 Microsoft Office 與 Adobe PDF 檔。然而，一律可以處理 XML、HTML 或文字檔。

不支援任何含有在機器本端編碼中未包含的字元的目錄名稱和檔案名稱。當試圖執行含有「檔案清單」節點的串流時，任何包含這些字元的檔案或目錄名稱都會導致串流執行失敗。國外目錄名稱或檔案名稱可能會發生這種情況，例如法文語言環境上的德文檔案名稱。

目錄 指定包含您要列出之文件的根資料夾。

- **包括子目錄** 指定還應該掃描的子目錄。

要在清單中包括的檔案類型： 您可以選取或取消選取要使用的檔案類型及副檔名。透過取消選取副檔名，忽略具有該副檔名的檔案。您可以按下列副檔名過濾：

表 *1.* 檔案類型按副檔名過濾

• *.rtf*、*.doc*、*.docx*、*.docm*	• *.xls*、*.xlsx*、*.xlsm*	• *.ppt*、*.pptx*、*.pptm*	• *.txt*、*.text*
• *.htm*、*.html*、*.shtml*	• *.xml*	• *.pdf*	• *.$*

註：如需相關資訊，請參閱第 9 頁的『檔案清單節點』。

如果您的檔案沒有副檔名，或者使用尾端句點副檔名（例如 `File01` 或 `File01.`），則使用**無副檔名**選項選取這些項。

輸入編碼 如果輸入欄位將包含確切文字，請從下列清單中選擇相關值：

- 自動（歐洲）
- UTF-8
- UTF-16
- ISO-8859-1
- ISO-8859-2
- Windows-1250
- US ascii

輸出顯示為 UTF-8 文件文字。

重要：從第 14 版開始，**目錄清單**選項不再可用，並且唯一的輸出是檔案清單。

檔案清單節點：其他標籤

「類型」標籤是 IBM SPSS Modeler 節點中的標準標籤，作為「註釋」標籤。

在文字挖掘中使用檔案清單節點

當文字資料位於諸如 Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF、XML、HTML 及其他格式的外部非結構化文件中時，使用「檔案清單」節點。

例如，假設我們已將「檔案清單」節點連接至「文字挖掘」節點，從而提供位於外部文件中的文字：

1. **「檔案清單」節點（「設定」標籤）。** 首先，我們將此節點新增至串流以指定儲存文字文件的位置。我們選取了目錄，其中包含要在其上執行文字採礦的所有文件。
2. **「文字採礦」節點（「欄位」標籤）。** 接下來，我們將「文字採礦」節點新增並連接至「檔案清單」節點。在此節點中，我們定義了輸入格式、資源範本和輸出格式。我們選取了從「檔案清單」節點產生的欄位名稱、文字欄位及其他設定。如需相關資訊，請參閱主題 第 25 頁的『在串流中使用文字挖掘節點』。

如需使用「文字挖掘」節點的相關資訊，請參閱第 18 頁的『文字挖掘建模節點』。

Web 資訊來源節點

「Web 資訊來源」節點可用來從「Web 資訊來源」準備文字資料，以進行文字採礦程序。這個節點接受採用兩種格式的 Web 資訊來源：

- RSS 格式。RSS 是一種用於 Web 內容的簡式 XML 型標準化格式。此格式的 URL 會指向一個頁面，該頁面具有一組鏈結的文章，如企業聯盟新聞來源和部落格。由於 RSS 是一種標準化格式，因此會自動將每一個鏈結的文章識別並視為產生的資料串流中的各別記錄處理。除非您要將過濾技術套用到文字，否則不需要進一步的輸入就能夠從資訊來源識別重要的文字資料和記錄。
- HTML 格式。您可以在「輸入」標籤中定義一或多個到 HTML 頁面的 URL。然後，在「記錄」標籤中，定義記錄起始標籤，以及識別界定目標內容的標籤及指派那些標籤給您選擇的輸出欄位（說明、標題、修改日期等等）。如需相關資訊，請參閱主題 第 12 頁的『Web 資訊來源節點：記錄標籤』。

重要事項！ 如果您在嘗試透過 Proxy 伺服器從網路擷取資訊，則必須在 IBM SPSS Modeler Text Analytics 用戶端及伺服器的 `net.properties` 檔中啟用 Proxy 伺服器。請遵循本檔案內詳述的指示。由於這些連線通過 Java™，因此當透過「Web 資訊來源」節點存取網路或擷取「軟體即服務 (SaaS)」授權時，此項適用。依預設，此檔案位於 *C:\Program Files\IBM\SPSS\Modeler\18.2.1\jre\lib\net.properties* 中。

此節點的輸出是一組用來說明記錄的欄位。**說明**欄位是最常用的欄位，因為它包含文字內容的主體，不過，您可能也會對其他欄位感興趣，例如記錄的簡要說明（**簡要說明**欄位）或記錄的標題（**標題**欄位）。可以選取任何輸出欄位作為後續的「文字採礦」節點的輸入。

註：要在IBM SPSS Collaboration and Deployment Services - 評分配置內進行評分，不能使用「Web 資訊來源」節點。

您可以在位於IBM SPSS Modeler視窗底端的節點選用區的IBM SPSS Modeler Text Analytics標籤上尋找這個節點。如需相關資訊，請參閱主題 第 6 頁的『IBM SPSS Modeler Text Analytics 節點』。

Web 資訊來源節點：輸入標籤

「輸入」標籤用來指定一或多個網址或 URL，以便擷取文字資料。在文字採礦的環境定義中，您可以指定包含文字資料之資訊來源的 URL。

重要：使用非 RSS 資料時，您可能偏好使用 Web 立即運算工具（如 WebQL®）來自動化內容收集，然後從該工具使用不同的來源節點參照輸出。

您可以設定下列參數：

輸入或貼上 URL。 在此欄位中，您可以鍵入或貼上一或多個 URL。如果您要輸入多個 URL，請每行僅輸入一個 URL，並使用 **Enter/Return** 鍵來分隔各行。輸入檔案的完整 URL 路徑。這些 URL 可以用於兩種格式之一的資訊來源：

- `RSS` 格式。 RSS 是一種用於 Web 內容的簡式 XML 型標準化格式。此格式的 URL 會指向一個頁面，該頁面具有一組鏈結的文章，如企業聯盟新聞來源和部落格。由於 RSS 是一種標準化格式，因此會自動將每一個鏈結的文章識別並視為產生的資料串流中的各別記錄處理。除非您要將過濾技術套用到文字，否則不需要進一步的輸入就能夠從資訊來源識別重要的文字資料和記錄。
- `HTML` 格式。 您可以在「輸入」標籤中定義一或多個到 HTML 頁面的 URL。然後，在「記錄」標籤中，定義記錄起始標籤，以及識別界定目標內容的標籤及指派那些標籤給您選擇的輸出欄位（說明、標題、修改日期等等）。使用非 RSS 資料時，您可能偏好使用 Web 立即運算工具（如 WebQL®）來自動化內容收集，然後從該工具使用不同的來源節點參照輸出。如需相關資訊，請參閱主題 第 12 頁的『Web 資訊來源節點：記錄標籤』。

每個 URL 要讀取的最新項目數。 此欄位指定要針對欄位中列出的每一個 URL 讀取的記錄數目上限（從在資訊來源中發現的第一筆記錄開始）。文字的數量會影響在「文字採礦」節點或「文字鏈結分析」節點中擷取下游期間的處理速度。

盡可能儲存並重複使用先前的 Web 資訊來源。 使用這個選項時，會掃描 Web 資訊來源並快取處理的結果。然後，在執行後續的串流後，如果給定的資訊來源的內容未變更，或是無法存取資訊來源（例如網際網路中斷），則會使用快取的版本來加速處理時間。也會快取在這些資訊來源中探索到的任何新內容，以供下次執行節點時使用。

- **標籤。**如果您選取**盡可能儲存並重複使用先前的 Web 資訊來源**，則必須指定結果的標籤名稱。這個標籤用來說明在伺服器上快取的資訊來源。如果未指定任何標籤或是標籤無法辨識，則無法進行任何重複使用。您可以在 IBM SPSS Deployment Manager 中所包含的 IBM SPSS Text Analytics Administration Console 的階段作業表格中管理這些 Web 資訊來源快取。如需相關資訊，請參閱 Deployment Manager 使用手冊。

Web 資訊來源節點：記錄標籤

「記錄」標籤用來指定非 RSS 資訊來源的文字內容，作法為識別每一個新記錄開始的位置，以及關於每一個記錄的相關資訊。如果您知道非 RSS 資訊來源 (HTML) 包含位於多個記錄中的文字，則您必須在這裡確認記錄起始標籤，否則該文字將被視為一個記錄處理。縱使 RSS 資訊來源已標準化，不需要此標籤上的任何標籤規格，您還是可以在「預覽」標籤中預覽內容。

重要： 使用非 RSS 資料時，您可能偏好使用 Web 立即運算工具（如 WebQL®）來自動化內容收集，然後從該工具使用不同的來源節點參照輸出。

URL。 這個下拉清單包含在「輸入」標籤上輸入的 URL 清單。會呈現經 HTML 和 RSS 格式化的資訊來源。如果 URL 位址對於此下拉清單過長，則會使用省略符號取代裁剪的文字，在中間自動裁剪它，例如 *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*。

- 使用 **HTML 格式化資訊來源**時，如果資訊來源包含多個記錄（或項目），您可以定義哪些 HTML 標籤包含對應於表格中所顯示欄位的資料。例如，您可以定義起始標籤，以指出新記錄已開始、修改的日期標籤或作者名稱。
- 使用 **RSS 格式化資訊來源**時，會提示您輸入任何標籤，因為 RSS 是標準化格式。不過，若有需要，您可以在「預覽」標籤上檢視範例結果。所有已辨識的 RSS 資訊來源前面都會有 RSS 標誌影像。

「原始碼」標籤。 在這個標籤上，您可以檢視任何 HTML 資訊來源的原始碼。此程式碼不可編輯。您可以使用「尋找」欄位來尋找此頁面上的特定標籤或資訊，然後就可複製並貼入下方表格。「尋找」欄位不區分大小寫，並會比對部分字串。

「預覽」標籤。 在這個標籤上，您可以預覽「Web 資訊來源」節點將如何讀取記錄。這對於 HTML 資訊來源特別有用，因為您可以透過在「預覽」標籤下方的表格中定義 HTML 標籤來變更讀取記錄的方式。

非 RSS 記錄起始標籤。 這個選項僅適用於非 RSS 資訊來源。如果您的 HTML 資訊來源包含多個要拆開成多個記錄的文字，請在這裡指定 HTML 標籤，以標明記錄的開頭（如文章或部落格文章）。如果您不對非 RSS 資訊來源定義 HTML 標籤，Modeler 將會嘗試猜測 XML 格式並傳回對應的記錄。如果 Modeler 無法猜測 XML 格式，則不會傳回任何記錄。如果您的目標是要匯入一個頁面的整個內容，並在稍後處理它，我們建議使用具有更強大功能的各別 XML 讀取器，然後將結果匯入 Modeler Text Analytics 中。

欄位表格。 這個選項僅適用於非 RSS 資訊來源。在此表格中，您可以輸入任何預先定義輸出欄位的起始標籤，藉此將文字內容拆開成特定的輸出欄位。請只限輸入起始標籤。所有的比對是透過剖析 HTML 並將表格內容與在 HTML 中發現的標籤名稱和屬性比對來完成。您可以使用位於底端的按鈕來複製您已定義的標籤，並在其他資訊來源中重複使用它們。

表 2. 非 *RSS* 資訊來源可能的輸出欄位（*HTML* 格式）

輸出欄位名稱	預期的標籤內容
標題	此標籤界定記錄標題（選用項目）。
簡要說明	此標籤界定簡要說明或標籤（選用項目）。
說明	此標籤界定主要文字。如果保留空白，此欄位將會包含 <body> 標籤（如果有單筆記錄）中的所有其他內容，或是在現行記錄內發現的內容（當已指定記錄定界字元時）。
作者	此標籤界定文字的作者（選用項目）。
貢獻者	此標籤界定貢獻者的名稱（選用項目）。
發佈日期	此標籤界定發佈文字時的日期。如果留白，此欄位將會包含節點讀取資料時的日期。
修改日期	此標籤界定修改文字時的日期。如果留白，此欄位將會包含節點讀取資料時的日期。

當您在表格中輸入標籤時，會使用此標籤作為要符合的最小標籤而非完全相符，來掃描資訊來源。也就是說，如果您在「標題」欄位中輸入 ，這會符合資訊來源中的任何 標籤，其中包括那些具有指定屬性的標籤（如 ），使得 等於主要標籤 ( ) 以及任何包含屬性的衍生項，並在「標題」輸出欄位中使用該內容。如果您輸入主要標記，則也會包含任何更深層的屬性。

表 3. 識別輸出欄位文字所用 *HTML* 標籤的範例

如果您輸入：	它將會符合：	也會符合：	但是不符合：
			任何其他標籤
<p class="auth">	<p class="auth">	<p color="black" class="auth" id="85643">	<p color="black">

Web 資訊來源節點：內容過濾器標籤

「內容過濾器」標籤用來套用過濾器技術至 RSS 資訊來源內容。這個標籤不適用於 HTML 資訊來源。您可能會想要過濾資訊來源是否在標頭、標底、功能表、廣告等等的表單中包含大量的文字。您可以使用此標籤來從內容中去除掉不要的 HTML 標籤、JavaScript 和較短的單字或行。

內容過濾：如果您不要套用清除技術，請選取**無**。否則，請選取 **RSS 內容清除程式**。

RSS 內容清除程式選項。如果您選取 **RSS 內容清除程式**，可以選擇根據特定準則來捨棄行。行是由 HTML 標籤（如 <p> 和 ）定界，但是不包括行內標籤（如 、<b> 和 <font>）。請注意，
 標籤會被視為換行處理。

- **捨棄短行**。這個選項會忽略不包含這裡定義的**單字數目下限**的行。
- **捨棄含有短單字的行**。這個選項會忽略具有超過這裡定義的**平均單字長度下限**的行。
- **捨棄含有許多單一字元單字的行**。這個選項會忽略包含超過特定**單一字元單字比例**的行。
- **捨棄包含特定標籤的行**。這個選項會忽略行中包含欄位中指定的任何標籤的文字。
- **捨棄包含特定文字的行**。這個選項會忽略包含欄位中指定的任何文字的行。

在文字採礦中使用 Web 資訊來源節點

「Web 資訊來源」節點可用來從「網際網路 Web 資訊來源」準備文字資料，以進行文字採礦程序。這個節點接受採用 HTML 或 RSS 格式的 Web 資訊來源。這些資訊來源用來作為文字採礦程序（後續的「文字採礦」或「文字鏈結分析」節點）的輸入。

如果您使用「Web 資訊來源」節點，則必須確保指定「文字」欄位代表「文字採礦」或「文字鏈結分析」節點中的**實際文字**，以指出這些資訊來源直接鏈結至每一篇文章或部落格文章。

重要事項！ 如果您在嘗試透過 Proxy 伺服器從網路擷取資訊，則必須在 IBM SPSS Modeler Text Analytics 用戶端及伺服器的 net.properties 檔中啟用 Proxy 伺服器。請遵循本檔案內詳述的指示。由於這些連線通過 Java，因此當透過「Web 資訊來源」節點存取網路或擷取「軟體即服務 (SaaS)」授權時，此項適用。依預設，此檔案位於 *C:\Program Files\IBM\SPSS\Modeler\18.2.1\jre\lib\net.properties* 中。

範例：「Web 資訊來源」節點（RSS 資訊來源）與「文字採礦」建模節點

舉例而言，假設我們將「Web 資訊來源」節點連接至「文字採礦」節點，以便將 RSS 資訊來源中的文字資料提供到文字採礦程序中。

1. **「Web 資訊來源」節點（「輸入」標籤）。** 首先，我們將此節點新增至串流，以指定資訊來源內容所在的位置及驗證內容結構。在第一個標籤上，我們提供了 URL 給 RSS 資訊來源。由於我們的範例是針對 RSS 資訊來源，已經定義格式化，因此無需在「記錄」標籤上進行任何變更。有可供用於 RSS 資訊來源的選用內容過濾演算法，不過在此情況下，並未套用它。
2. **「文字採礦」節點（「欄位」標籤）。** 接下來，我們將「文字採礦」節點新增並連接至「Web 資訊來源」節點。在此標籤上，我們依「Web 資訊來源」節點定義了文字欄位輸出。在此情況下，我們想要使用**說明**欄位。我們也選取了選項「文字」欄位代表**實際文字**以及其他設定。
3. **「文字採礦」節點（「模型」標籤）。** 接下來，在「模型」標籤上，我們選擇建置模式和資源。在此範例中，我們選擇直接從此節點使用預設資源範本來建置概念模型。

如需使用「文字採礦」節點的相關資訊，請參閱第 18 頁的『文字挖掘建模節點』。

語言節點

您可以使用「語言」節點，以識別來源資料內文字欄位的自然語言。

此節點的輸出是包含所偵測到語言碼得衍生欄位。

註：您不能將「語言」節點用於在 IBM SPSS Collaboration and Deployment Services - 評分 配置內評分。

您可以在位於IBM SPSS Modeler視窗底端的節點選用區的IBM SPSS Modeler Text Analytics標籤上尋找這個節點。如需相關資訊，請參閱主題 第 6 頁的『IBM SPSS Modeler Text Analytics 節點』。

語言節點：設定標籤

在此標籤上，您可以指定如何輸出所選取文字欄位的語言詳細資料。

文字欄位 選取您要識別其語言的文字欄位。

衍生欄位名稱 輸入將包含所偵測到語言碼得衍生欄位的名稱。預設值為語言。

無法識別語言時的預設值 指定無法識別語言時要建立的欄位名稱。可用的選項為：

- **未定義** 如果選取，則衍生的欄位包含空值。
- **受支援** 如果選取，您可以從下列其中一個受支援的 ISO 語言中進行選擇：
  - 英文 (EN)
  - 德文 (DE)
  - 西班牙文 (ES)
  - 法文 (FR)
  - 義大利文 (IT)
  - 荷蘭文 (NL)

  – 葡萄牙文 (PT)
- **自訂** 如果沒有受支援的語言適用，請使用此選項以指定應該使用自訂值。通常，這可能是含 2 個字母的 ISO 語言碼，但可能是您需要的任何字串。

第 3 章 概念和種類的挖掘

「文字挖掘」建模節點用於產生兩種文字挖掘模型塊的其中一種：

- 概念模型塊可從您的結構化或非結構化文字資料發現並擷取突出的概念。
- 種類模型塊可對文件和記錄進行評分，並為其指派種類，種類由擷取的概念（和型樣）組成。

從您的模型塊擷取的概念、型樣及種類皆可以與現有的結構化資料（例如，個人背景資訊）相整合，並使用 IBM SPSS Modeler 提供的完整工具套組套用以生產更多更好的聚焦決策。例如，如果客戶頻繁列出登入問題作為阻礙完成線上帳戶管理作業的主要原因，您可能想要將「登入問題」合併至您的模型中。

此外，「文字挖掘」建模節點已完全整合在 IBM SPSS Modeler 內，因此您可以透過 IBM SPSS Modeler Solution Publisher 部署文字挖掘串流，以便對應用程式內的非結構化資料（例如，PredictiveCallCenter）進行即時評分。部署這些串流的能力可確保順利關閉迴圈文字挖掘實作。例如，您的組織現在可以透過套用預測性模型來分析入埠或出埠呼叫程式的即時運算簿附註，以便即時提高您行銷訊息的精確度。使用文字挖掘模型會導致顯示串流，從而改進預測性資料模型的精確度。

若要使用 IBM SPSS Modeler Solution Publisher 來執行 IBM SPSS Modeler Text Analytics，請將目錄 `<install_directory>/ext/bin/spss.TMWBServer` 新增至 `$LD_LIBRARY_PATH` 環境變數中。

在 IBM SPSS Modeler Text Analytics 中，我們一般參照擷取的概念與種類。瞭解概念和種類的意義極為重要，因為它們可以幫助您在實驗性工作和模型建置期間作出更明智的決策。

概念和概念模型塊

擷取程序期間，系統會掃描並分析文字資料，以便識別有興趣或相關的單字，例如 `election` 或 `peace`，以及單字片語，例如 `presidential election`、`election of the president` 或 `peace treaties`。這些單字和片語統一稱作術語。使用語言類資源擷取相關的術語，並將相似的術語一併分組在一個前導術語下叫作**概念**。

透過此方法，概念可以根據您的文字及所使用的一組語意資源，代表多個基礎術語。例如，讓我們假設我們具有員工滿意度意見調查，且已擷取概念 `salary`。再讓我們假設當您查看與 `salary` 相關聯的記錄時，我們注意到 `salary` 未一律呈現在文字中，而是包含某些類似的記錄，例如術語 `wage`、`wages` 及 `salaries`。由於擷取引擎將這些術語視為類似，或者根據處理規則或語意資源將它們判定為同義字，因此根據 `salary` 對它們進行分組。在此情況下，任何包含這些其術語中任何術語的文件或記錄都被視為它們包含單字 `salary`。

如果要查看某個概念下分組的術語，您可以在互動式工作台內探索概念，或是查看概念模型中顯示的同義字。如需相關資訊，請參閱主題 第 28 頁的『概念模型中的基礎術語』。

概念模型塊包含一組概念，可用於識別同樣包含概念（包括其任意同義字或分組術語）的記錄或文件。可以兩種方式概念模型。第一種是探索和分析在原始文字內探索到的概念，或是快速識別感興趣的文件。第二種是將此模型套用至新的文字記錄或文件，以便快速識別新文字/記錄中相同的主要概念，例如，從呼叫中心即時探索即時運算簿資料內的主要概念。

如需相關資訊，請參閱主題 第 26 頁的『文字挖掘塊：概念模型』。

種類和種類模型塊

您可以建立**種類**，以從本質上代表高層次概念或主題，以擷取主要構想、知識和文字中表達的本質。種類由一組描述子（例如，概念、類型和規則）組成。配合使用這些描述子可以用於識別記錄或文件是否屬於給定的種類。可以掃描文件或記錄來查看其文字是否符合描述子。如果發現相符項目，則會將文件/記錄指派給該種類。此過程叫作**分類**。

可以使用產品豐富的自動化技術來自動建置種類，也可以使用您可能擁有的其他資料相關功能或資料技術組合來手動建置。您還可以透過本節點的「模型」標籤來從文字分析套件加載一組預置種類。手動建立種類或優化種類只能透過互動式工作台來執行。如需相關資訊，請參閱主題 第 21 頁的『文字採礦節點：模型標籤』。

種類模型塊包含一組種類及其描述子。模型可以用來根據每一個文件/記錄中的文字來對一組文件或記錄進行分類。讀取每一個文件或記錄，然後將其指派給具備相符描述子的每一個種類。在此方式下，可以將文件或記錄指派給多個種類。例如，您可以使用種類模型塊來查看開放式意見調查回應或一組部落格文章內的基本構想。

如需相關資訊，請參閱主題 第 34 頁的『文字挖掘塊：種類模型』。

文字挖掘建模節點

「文字挖掘」節點使用語言與頻率技術，從文字擷取主要概念，並利用這些概念及其他資料建立種類。節點可以用來探索文字資料內容，或者用來提供概念模型塊或種類模型塊。當我們執行這個建模節點時，內部語言擷取引擎會使用自然語言處理程方法，擷取及組織概念、型樣及/或種類。

您可以執行「文字挖掘」節點，以及使用**直接產生**選項自動產生概念或種類模型。或者，您可以採用**以互動方式建置**模式使用更具實踐性的上機方式，您不僅可以擷取概念、建立種類及精簡語言資源，還可以執行文字鏈結分析及探索叢集。如需相關資訊，請參閱主題 第 21 頁的『文字採礦節點：模型標籤』。

您可以在位於IBM SPSS Modeler視窗底端的節點選用區的IBM SPSS Modeler Text Analytics標籤上尋找這個節點。如需相關資訊，請參閱主題 第 6 頁的『IBM SPSS Modeler Text Analytics 節點』。

需求。「文字挖掘」建模節點接受「Web 資訊來源」節點、「檔案清單」節點或任何標準來源節點的文字資料。此節點隨 IBM SPSS Modeler Text Analytics 一起安裝，且可以在 IBM SPSS Modeler Text Analytics 選用區上進行存取。

註：此節點會取代「文字擷取」節點，這在舊版產品中進行提供。如果您有使用舊節點或模型塊的較舊串流，則必須使用「文字挖掘」節點重建串流。

文字挖掘節點：欄位標籤

使用「欄位」標籤，可指定您將從中擷取概念之資料的欄位設定。使用較大的資料集減少處理次數時，請考量使用此節點的「樣本」節點上游。如需相關資訊，請參閱主題 第 25 頁的『取樣上游以節省時間』。

您可以設定下列參數：

ID 欄位 選取包含文字記錄 ID 的欄位。ID 必須是整數。ID 欄位用作個別文字記錄的索引。如果文字欄位代表要挖掘的文字，則使用 ID 欄位。

「文字」欄位。 選取含有要發掘之文字的欄位。這個欄位視資料來源而定。

語言欄位 選取包含兩個字母 IOS 語言 ID 的欄位。如果您未選取欄位，則假設每一個文件的語言即是所提供範本的語言。

文件類型。 文件類型指定文字的結構。請選取下列一種類型：

- **全文。** 用於大部分的文件或文字來源。會掃描整組文字以進行擷取。有別於其他選項，這個選項沒有其他設定。
- **結構化文字。** 用於書目表單、專利，以及任何包含可以識別及分析之一般結構的檔案。這種文件類型用來跳過全部或部分擷取程序。它可讓您定義術語分隔字元、指派類型，以及強制最小頻率值。如果您選取這個選項，則必須按一下**設定**按鈕，然後在「文件設定」對話框的**結構化文字格式化**區域中輸入文字分隔字元。如需相關資訊，請參閱主題『欄位標籤的文件設定』。

文字個體。 從下列選項中選取擷取模式：

- **文件模式。** 用於較為簡短且在語意上同質的文件，例如來自新聞社的文章。
- **段落模式。** 用於網頁及非帶標記文件。擷取程序會利用內部標籤和語法之類的性質，依語意分割文件。如果選取了這種模式，會逐個段落套用評分。因此，舉例而言，只有在相同的段落中發現 `apple` 和 `orange`，規則 `apple & orange` 才會為真。

  註：由於從 PDF 文件擷取文字的方式，因此**段落模式**不適用於這些文件。這是因為擷取會抑制換行標記。

段落模式設定。 只有在您將文字個體選項設定為**段落模式**時，才有這個選項可用。請指定要在任何擷取中使用的字元臨界值。實際大小會向上或向下捨入到最接近的期間。如果要確保從文件集合的文字所產生的單字關聯為代表，請避免指定過小的擷取大小。

- **最小值。** 指定要在任何擷取中使用的字元數目下限。
- **最大值。** 指定要在任何擷取中使用的字元數目上限。

分割區模式 使用分割區模式，以選擇是根據類型節點設定分割，還是選取另一個分割區。分割會將資料分為訓練與測試樣本。

欄位標籤的文件設定

結構化文字格式

如果您因為具有結構化資料，或是想要強制執行有關如何處理文字的規則，而要跳過全部或部分的擷取程序，請使用**結構化文字**文件類型選項，並在「文件設定」對話框的**結構化文字格式**區段中宣告包含該文字的欄位或標籤。系統只會從宣告的欄位或標籤（以及子標籤）內所含的文字衍生擷取的詞彙。將忽略任何未宣告的欄位或標籤。

在某些環境定義中，並不需要語言處理程序，而且明確宣告可以取代語言擷取引擎。在其中關鍵字欄位是以分號 (;) 或逗點 (,) 之類的分隔字元區隔的參考書目檔案中，擷取兩個分隔字元之間的字串已足夠。基於此原因，您可以暫停完整擷取程序，改為定義特殊處理規則來宣告詞彙分隔字元、將類型指派給擷取的文字，或是強制執行擷取頻率計數下限。

宣告結構化文字元素時，請使用下列規則：

- 每一行只能宣告一個欄位、標籤或元素。它們不必存在於資料中。
- 宣告區分大小寫。
- 如果宣告具有屬性的標籤（如 `<title id="1234">`），並且想要包含所有變式或（在本案例中）所有 ID，請新增沒有屬性或結尾角括弧的標籤 (>)，如 `<title`
- 在欄位或標籤名稱後面加入冒號來表示這是結構化文字。請緊接在欄位或標籤後面，但是在任何分隔字元、類型或頻率值前面加入這個冒號，例如 `author:` 或 `<place>:`。
- 如果要指出欄位或標籤中包含多個詞彙，並指出要使用分隔字元來指定個別詞彙，請在冒號後面宣告分隔字元，例如 `author:,` 或 `<section>:;`。
- 如果要將類型指派給在標籤中發現的內容，請在冒號和分隔字元後面宣告類型名稱，例如 `author:,Person` 或 `<place>:;Location`。使用出現在「資源編輯器」中的名稱來宣告類型。

- 如果要為欄位或標籤定義頻率計數下限，請在行尾處宣告一個數字，如 `author:,Person1` 或 `<place>:;Location5`。其中 `n` 是您定義的頻率計數，在欄位中發現的詞彙在要擷取的整組文件或記錄中，必須至少要出現 `n` 次。這也需要您定義分隔字元。
- 如果您有標籤包含冒號，則在冒號前面必須有一個反斜線字元，以便宣告不會被忽略。比方說，如果您有一個稱為 `<topic:source>` 的欄位，請將它輸入為 `<topic\:source>`。

為說明語法，讓我們假設您具有下列重複出現的書目欄位：

```
author:Morel, Kawashima
abstract:This article describes how fields are declared.
publication:Text Mining Documentation
datepub:March 2010
```

比方說，如果我們要擷取程序聚焦在作者和摘要上，而忽略其餘內容，則我們只會宣告下列欄位：

```
author:,Person1
abstract:
```

在本例中，`author:,Person1` 欄位宣告指出在欄位內容上已暫停語言處理程序。反之，它指出作者欄位包含多個名稱（以逗點分隔字元與下一個名稱區隔），且這些名稱應指派給「人員」類型，並指出如果該名稱在整組文件或記錄中至少出現一次，則應該擷取它。由於列出的欄位 `abstract:` 沒有任何其他宣告，因此在擷取期間將會掃描該欄位，並會套用標準語言處理程序和類型設定。

XML 文字格式

如果您要將擷取程序僅限於特定 XML 標籤內的文字，請使用 **XML 文字**文件類型選項，並在「文件設定」對話框的 **XML 文字格式**區段中宣告包含該文字的標籤。系統只會從這些標籤或其子標籤內所含的文字衍生擷取的詞彙。

重要事項！ 如果您要跳過擷取程序，並對詞彙分隔字元強制施行規則、將類型指派給擷取的文字，或是對擷取的詞彙強制執行頻率計數，請使用接下來說明的**結構化文字**選項。

宣告 XML 文字格式的標籤時，請使用下列規則：

- 每一行只能宣告一個 XML 標籤。
- 標籤元素區分大小寫。
- 如果標籤具有屬性（如 `<title id="1234">`），並且想要包含所有變式或（在本案例中）所有 ID，請新增沒有屬性或結尾角括弧的標籤 (>)，如 `<title`

為說明語法，讓我們假設您具有下列 XML 文件：

```
<section>Rules of the Road
     <title id="01234">Traffic Signals</title>
     <p>Road signs are helpful.</p>
</section>
<p>Learning the rules is important.</p>
```

針對這個範例，我們將宣告下列標籤：

```
<section>
<title
```

在本例中，由於您已宣告標籤 `<section>`，因此在擷取程序期間會掃描此標籤及其巢套標籤中的文字 `Traffic Signals` 和 `Road signs are helpful`。不過，會忽略 `Learning the rules is important`，因為未明確宣告標籤 `<p>`，也未明確宣告在宣告的標籤內巢套的標籤。

文字採礦節點：模型標籤

使用「模型」標籤來為節點輸出指定建置方法和一般模型設定。

您可以設定下列參數：

模型名稱。您可以根據目標或 ID 欄位（或模型類型，若未指定此類欄位）自動產生模型名稱，或者指定自訂名稱。

使用分割的資料。如果定義分割區欄位，則此選項會確保僅將訓練分割區的資料用於建置模型。

建置模式。指定執行具有此「文字採礦」節點的串流時，將如何產生模型塊。或者，您可以採用**以互動方式建置**模式使用更具實踐性的上機方式，您不僅可以擷取概念、建立種類及精簡語言資源，還可以執行文字鏈結分析及探索叢集。

- **以互動方式建置。**當執行串流時，此選項會啟動互動式介面，您可以從中擷取概念及型樣、探索及細部調整所擷取結果、建置及精簡種類、細部調整語言資源（範本、同義字、類型、檔案庫等），以及建置種類模型塊。如需相關資訊，請參閱 『以互動方式建置』。
- **直接產生。**此選項指出執行串流時，應該會自動建立模型並新增至「模型」選用區。與互動式工作台不同，執行時，除了節點中定義的設定之外，不需要您提供任何其他操作。如果您選取此選項，則會顯示模型特定的選項，您可以利用這些選項定義要產生之模型的類型。如需相關資訊，請參閱 第 22 頁的『直接產生』。

將大型模型儲存在 AS 中。如果您具有 IBM SPSS Analytic Server 的連線，請選取此選項，以從遠端在伺服器上儲存您的模型。

註：任何在伺服器上建置及儲存的模型只能在該伺服器上進行評分。若要回復包含此類模型的互動式工作台，您需要用來建立階段作業之原始伺服器的連線。

複製資源來源。 對文字進行採礦時，擷取程序不僅是基於「專家」標籤中的設定，還基於語言資源。這些資源在擷取期間用作如何處理文字的基礎，從而取得概念、類型，並有時取得型樣。您可以將資源從資源範本、文字分析套件 (.tap) 或 SPSS Text Analytics for Surveys 專案檔 (.tas) 複製到此節點。進行選擇然後按一下**載入**以定義將從中複製資源的範本、套件或專案。當您載入時，會將資源的副本儲存在節點中。因此，如果您要使用更新的資源，則必須在這裡或在互動式工作台階段作業中重新載入它。為了便於使用，資源的複製及載入日期和時間顯示在節點中。如需相關資訊，請參閱 第 23 頁的『從範本及 TAP 複製資源』。

文字語言。 識別所要發掘之文字的語言。在節點中所複製的資源控制存在的語言選項。請選取已調整其資源的語言。

以互動方式建置

在文字挖掘建模節點的「模型」標籤中，您可以選擇模型塊的建置模式。如果您選擇**以互動方式建置**，則當您執行串流時，會開啟互動式介面。在這個互動式工作台中，您可以：

- 擷取及探索擷取結果，包括概念並鍵入以探索文字資料中顯著的構想。
- 使用各種方法從概念、類型、TLA 型樣及規則建置及延伸種類，以便您可以將文件及記錄評分至這些種類。
- 精簡您的語言資源（資源範本、檔案庫、定義檔、同義字等），以便您可以透過反覆運算處理程序改良您的結果，在該處理程序期間會擷取、檢查及精簡您的概念。
- 執行文字鏈結分析 (TLA) 並使用探索到的 TLA 型樣，建置更好的種類模型塊。「文字鏈結分析」節點不提供相同的探索選項或建模功能。
- 產生叢集以探索新的關係，並在「視覺化」窗格中探索概念、類型、型樣與種類之間的關係。
- 為 IBM SPSS Modeler 中的「模型」選用區產生精簡的種類模型塊，並在其他串流中使用它們。

註：如果您要建立 IBM SPSS Collaboration and Deployment Services 工作，則不能建置互動式模型。

使用前次節點更新的階段作業工作（種類、TLA、資源等）。當您在互動式工作台階段作業中工作時，可以利用階段作業資料（擷取參數、資源、種類定義等）更新節點。**使用階段作業工作**選項可讓您利用儲存的階段作業資料重新啟動互動式工作台。由於無法儲存任何階段作業資料，因此在您第一次使用此節點時，此選項已停用。若要瞭解如何使用階段作業資料更新節點，以便您可以使用此選項，請參閱第 66 頁的『更新建模節點及儲存』。

如果您使用此選項啟動階段作業，則當您下一次啟動階段作業時，前次從互動式工作台階段作業執行節點更新時的擷取設定、種類、資源及任何其他工作可用。由於儲存的階段作業資料與此選項搭配使用，因此會停用並忽略某些內容（例如從下方範本中複製的資源）及其他標籤。但是如果您不使用此選項啟動階段作業，則只會使用目前定義的節點內容，表示您在工作台中執行的所有先前的工作將無法使用。

附註：如果您在利用**使用階段作業工作...** 選項快取擷取結果之後變更串流的來源節點，並且您想要取得更新的擷取結果，則在啟動互動式工作台階段作業之後將需要執行新的擷取。

跳過擷取並重複使用所快取資料及結果。您可以在互動式工作台階段作業中重複使用任何快取的擷取結果及資料。如果啟動階段作業時，您想要節省時間並重複使用擷取結果，而不是等待執行全新的擷取，此選項特別有用。為了使用此選項，您必須先前已從互動式工作台階段作業內更新此節點，並選擇選項以**保持階段作業工作並利用擷取結果快取文字資料以重複使用**。若要瞭解如何使用階段作業資料更新節點，以便您可以使用此選項，請參閱第 66 頁的『更新建模節點及儲存』。

開始階段作業時間。選取該選項指出啟動互動式工作台階段作業時您希望首先啟用的視圖與動作。無論您從哪個視圖開始，只要您在階段作業中，就可以切換至任何視圖。

- **使用擷取結果建置種類。**此選項會在「種類與概念」視圖中啟動互動式工作台，並在適用時執行擷取。在此視圖中，您可以建立種類並產生種類模型。您還可以切換至另一個視圖。如需相關資訊，請參閱主題 第 57 頁的第 7 章, 『互動式工作台模式』。
- **探索文字鏈結分析 (TLA) 結果。**此選項透過擷取並識別文字中概念之間的關係啟動並開始，例如「文字鏈結分析」視圖中的意見或其他鏈結。您必須選取包含 TLA 型樣規則的範本或文字分析套件，從而使用此選項及取得結果。如果您要使用更大的資料集，則 TLA 擷取可能需要一些時間。在此情況下，您可能想要考量使用「樣本」節點上游。如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。
- **分析共存單字叢集。**此選項在「叢集」視圖中啟動，並更新所有過期的擷取結果。在此視圖中，您可以執行共存單字叢集分析，這會產生一組叢集。共存單字形成叢集是一個處理程序，透過根據給定記錄或文件中兩個概念的共生，評量兩個概念之間的鏈結值強度開始，並以將高度鏈結的概念分組到叢集結束。如需相關資訊，請參閱主題 第 57 頁的第 7 章, 『互動式工作台模式』。

直接產生

在文字挖掘建模節點的「模型」標籤中，您可以選擇模型塊的建置模式。如果您選擇**直接產生**，則可以在節點中設定選項，然後只需執行串流。輸出是概念模型塊，直接放置在「模型」選用區下方。與互動式工作台不同，除了在節點中為此選項定義的頻率設定之外，執行時不需要您進行任何其他操作。

要在模型中包括的概念數目上限。僅當您自動建置模型（非互動式）時，此選項才適用，指出您想要建立概念模型。它還說明此模型包含的概念不應該超過指定數目。

- **根據最高頻率勾選概念。概念數目上限。**從具有最高頻率的概念開始，這是將勾選的概念數目。這裡，頻率是指概念（及其所有基礎術語）在整個文件/記錄集中出現的次數。此數目可能會高於記錄計數，因為一個概念可能會在一筆記錄中出現多次。
- **取消勾選在太多記錄中出現的概念。記錄百分比。**取消勾選記錄計數百分比高於您指定之數目的概念。此選項對於排除文字或每筆記錄中頻繁出現，但是對分析無意義的概念非常有用。

為評分速度最佳化。依預設選取此選項，可確保建立的模型壓縮，且在高速度評分。取消選取此選項會建立更大的模型，且評分速度更慢。然而，模型較大會確保在所產生概念模型中最初顯示的分數，與評分與模型塊相同的文字時取得的分數相同。

從範本及 TAP 複製資源

對文字進行採礦時，擷取不僅基於「匯出」標籤中的設定，還基於語言資源。這些資源在擷取期間用作如何處理文字的基礎，從而取得概念、類型，並有時取得型樣。您可以從資源範本將資源複製到此節點，並且如果您在「文字採礦」節點中，還可以選取文字分析套件 (TAP) 或 SPSS Text Analytics for Surveys 專案 (.tas)。

依預設，將節點新增至畫布時，會將資源從產品授權語言的基本範本中複製到節點。如果您已授權多種語言，則選取的第一個語言用來判定自動載入的範本。

當您載入時，會將所選取資源的副本儲存在節點中。如果範本、TAP 或 SPSS Text Analytics for Surveys 本身未鏈結至節點，則只會複製範本、TAP 或 SPSS Text Analytics for Surveys 專案資源的內容。這表示如果稍後更新資源，則這些更新不會在節點中自動可用。簡言之，一律使用載入節點的資源，除非您重新載入新的資源副本，或者除非您更新「文字採礦」節點並選取**使用階段作業工作**選項。如需**使用階段作業工作**的相關資訊，請進一步參閱本節。

當您選取資源時，選擇與您的文字資料使用相同語言的資源。您只能使用獲授權之語言的資源。如果您想要執行文字鏈結分析，惡必須選取包含 TLA 型樣的範本。如果範本包含 TLA 型樣，則將在「載入資源範本」對話框的 TLA直欄中顯示一個圖示。

註：您不能將 TAP 或 SPSS Text Analytics for Surveys 專案載入「文字鏈結分析」節點。

資源範本

資源範本是一組預先定義的檔案庫及進階語言與非語言資源，已針對特定網域或使用進行細部調整。在文字採礦建模節點中，當您將節點新增至串流時，已在節點中載入基本範本的資源副本，但是您可以透過選取**資源範本**或**文字分析套件**，然後按一下**載入**，變更範本或載入文字分析套件。對於範本，然後您可以在「載入資源範本」對話框中選取範本。

註：如果您未在清單中看到想要的範本，但是在機器上具有已匯出的副本，則可以立即匯入。您還可以從此對話框中匯出，以與其他使用者共用。如需相關資訊，請參閱 第 151 頁的『匯入及匯出範本』。

文字分析套件 (TAP) 和用於意見調查的文字分析專案 (TAS)

文字分析套件 (TAP) 是一組預先定義的檔案庫集及進階語言與非語言資源，並組合一組或多組預先定義的種類。IBM SPSS Modeler Text Analytics 提供數個針對特定網域進行細部調整的預先建置 TAP。您可以編輯這些 TAP 並將其儲存至其他目錄以供用來立即開始建置種類模型。您還可以在互動式階段作業中建立您自己的 TAP。如需相關資訊，請參閱 第 114 頁的『載入文字分析套件』。

如果您選擇匯入 SPSS Text Analytics for Surveys 專案 (.tas)，則它將轉換為 TAP。

註：您不能將 TAP 或 SPSS Text Analytics for Surveys 專案載入「文字鏈結分析」節點。

使用「使用階段作業工作」選項（模型標籤）

由於資源複製到「模型」標籤中的節點，您還可以稍後在互動式階段作業中對資源進行變更，以及想要利用這些最新的變更來更新文字採礦建模節點。在此情況下，您可選取文字採礦建模節點中「模型」標籤內的**使用階段作業工作**選項。

如果您選取**使用階段作業工作**，並且在節點中停用**載入**按鈕以指出將使用來自互動式工作台的那些資源，而不是先前在這裡載入的資源。

若要在您選取**使用階段作業工作**選項後對資源進行變更，您可以直接在互動式工作台階段作業內部，透過資源編輯器視圖直接編輯或切換資源。如需相關資訊，請參閱 第 150 頁的『載入之後更新節點資源』。

文字挖掘節點：專家標籤

「專家」標籤包含某些進階參數，影響文字的擷取及處理方式。此對話框中的參數會控制擷取程序的基本行為，以及少數進階行為。然而，它們僅代表為您提供的部分選項。還有許多語言資源與選項會影響擷取結果，這由您在「模型」標籤上選取的資源範本進行控制。如需相關資訊，請參閱主題 第 21 頁的『文字採礦節點：模型標籤』。

註：如果您已使用在「模型」標籤上儲存的互動式工作台資訊選取**以互動方式建置**模式，則在此情況下，會從前次儲存的工作台階段作業中取得擷取設定。

擷取時，您可以設定下列參數：

將擷取限制為廣域頻率至少為 [n] 的概念。 指定單字或詞組要能夠被擷取所必須在文字中發生的次數下限。如此一來，值 5 會將擷取限制為那些在整個記錄或文件集中發生至少五次的單字或詞組。

在某些情況下，變更此限制可能會在產生的擷取結果中造成巨大的差異，並因此造成種類上的差異。假設您正在處理某餐廳資料，並在此選項中不將限制增加超過 1。在此情況下，擷取結果中可能會出現 *pizza (1)*、*thin pizza (2)*、*spinach pizza (2)* 和 *favorite pizza (2)*。不過，如果您要將擷取限制為廣域頻率 5 或以上，然後重新擷取，則您不會再取得這其中的三個概念。而是會取得 *pizza (7)*，因為 *pizza* 是最簡單的格式，而且這個字已經存在為可能的候選字。視其餘的文字而定，您實際的頻率可能超過七，這取決於是否仍然有其他詞組在文字中有 pizza。此外，如果 *spinach pizza* 已經是種類描述子，則您可能需要將 *pizza* 新增為描述子而不是擷取所有記錄。基於此原因，每當已經建立種類時，都要小心變更此限制。

請注意，這是僅限擷取的特性；如果您的範本包含術語，並且在文字中發現範本的術語，則無論其頻率為何，都會檢索該術語。

例如，假設您使用「基本資源」範本，此範本在 Core 檔案庫中的 `<Location>` 類型下包含 "los angeles"；如果您的文件僅包含 Los Angeles 一次，則 Los Angeles 將會是概念清單的一部分。如果要防止此情況，您需要將過濾器設定為顯示發生的次數至少與在**將擷取限制為廣域頻率至少為 [n] 的概念**欄位中輸入的值相同的概念。

容納標點符號錯誤。此選項會在擷取期間暫時正規化包含標點符號錯誤的文字（例如，用法不當），以改進概念的擷取能力。當文字較短且品質較差時（例如，在開放式意見調查回應、電子郵件及 CRM 資料中時），或者當文字包含許多縮寫時，此選項極其有用。

容納單字字元長度下限 [n] 的拼字 此選項套用模糊分組技術，可協助將通常拼字錯誤的單字或根據一個概念拼字接近的單字分組在一起。模糊分組演算法暫時去掉所有母音（除了第一個），並從擷取的單字中去掉雙/三重輔音，然後比較它們以查看它們是否相同，以便將 `modeling` 與 `modelling` 分組在一起。然而，如果將每一個術語指派給不同的類型（排除 `<Unknown>` 類型），則將不會套用模糊分組技術。

您也可以在使用模糊分組之前，定義需要的根字元數目下限。術語中的根字元數目計算方式為所有字元總數減去形成字形變化字尾的字元，若為複合字術語，則再減去限定詞與介詞。例如，術語 `exercises` 將以 "exercise" 形式計為 8 個根字元，位於單字末尾的字母 *s* 是字形變化（複數形式）。類似地，`apple sauce` 計為 10 個根字元 ("apple sauce")，而 `manufacturing of cars` 計為 16 個根字元 ("manufacturing car")。此計數方法僅用於檢查是否應該套用模糊分組，但不會影響單字的相符程度。

註：如果您稍後發現某些單字未正確地分組，則可以透過在「進階資源」標籤的**模糊分組：異常狀況**區段中明確地宣告，從此技術中排除單字配對。如需相關資訊，請參閱主題 第 177 頁的『模糊分組』。

擷取單一術語 只要單字尚且不是複合字的一部分，或者如果它是名詞或無法辨識的語音的一部分，則此選項會擷取單字（單一屬於）。

擷取非語言實體 此選項擷取非語言實體，例如電話號碼、社會安全號碼、時間、日期、貨幣、數位、百分比、電子郵件位址及 HTTP 位址。您可以在「進階資源」標籤的**非語言實體：配置**區段中包括或排除某些類型的非語言實體。透過停用任何不需要的實體，擷取引擎不會浪費處理時間。如需相關資訊，請參閱主題第 181 頁的『配置』。

大寫演算法 只要術語的第一個字母為大寫形式，此選項就擷取不在內建目錄中的簡式及複合術語。此選項提供良好的方法以擷取最適當的名詞。

可能時將部分及全部人員名稱分組在一起 此選項將文字中看起來不同的名稱分組在一起。由於通常在文字開頭以名稱的完整形式對名稱以縮寫進行參照，之後僅使用縮寫版本，因此本功能有用。此選項嘗試將任何類型為 <Unknown> 的單一術語與鍵入為 <Person> 的任何複合術語的最後一個單字進行比對。例如，如果發現 *doe*，且最初鍵入為 <Unknown>，則擷取引擎會檢查以查看 <Person> 類型中是否有任何複合術語包括 *doe* 作為最後一個單字，例如 *john doe*。由於大部分名稱從不作為單一術語擷取，因此本選項不適用於第一個名稱。

非功能單字排列上限 此選項指定套用排列技術時可以呈現的非功能單字數目上限。此排列技術僅依照所包含的非功能單字，將彼此不同的類似片語分組在一起（例如，of 及 the），而不考量字形變化。例如，讓我們假設將此值設為最多兩個單字，並擷取 company officials 與 officials of the company。在此情況下，由於當忽略 of the 時，兩個術語被視為相同，因此兩個擷取的術語將在最終概念清單中分組在一起。

分組多術語時使用衍生 處理海量資料時，選取此選項以透過使用衍生規則分組多術語。

註：若要啟用「文字鏈結分析」結果的擷取，您必須利用**探索文字鏈結分析結果**選項開始階段作業，以及選擇包含 TLA 定義的資源。在互動式工作台階段作業期間，您一律可以稍後透過「擷取設定」對話框擷取 TLA 結果。如需相關資訊，請參閱主題 第 70 頁的『擷取資料』。

取樣上游以節省時間

當您有大量資料時，處理時間可能從數分鐘到數小時，特別是當使用互動式工作台階段作業時。資料大小越大，擷取與分類處理程序所花費的時間越長。若要更有效地工作，您可以從「文字挖掘」節點新增 IBM SPSS Modeler 的「樣本」節點上游。使用此「樣本」節點可取得隨機樣本，利用較小的文件或記錄子集執行前幾次傳遞。

較小的樣本通常特別足以用來決定如何編輯資源，甚至是建立大部分（如果不是全部）種類。當您在較小的資料集上執行，並對結果滿意後，可以將同一種類建立技術套用至整個資料集。然後，您可以尋找不適合所建立種類的文件或記錄，並根據需要進行調整。

註：「樣本」節點是標準 IBM SPSS Modeler 節點。

在串流中使用文字挖掘節點

「文字挖掘」建模節點用來在串流中存取資料及擷取概念。您可以使用任何來源節點存取資料，例如「資料庫」節點、「變數檔案」節點、「Web 資訊來源」節點或「修正檔案」節點。對於位於外部文件中的文字， 可以使用「檔案清單」節點。

範例 1：用來直接建置概念模型塊的「檔案清單」節點及「文字挖掘」節點

下列範例顯示如何使用「檔案清單」節點以及「文字挖掘」建模節點來產生概念模型塊。如需使用「檔案清單」節點的相關資訊，請參閱第 9 頁的『檔案清單節點』。

1. **檔案清單節點（設定標籤）**。首先，我們將此節點新增至串流，以指定將儲存文字文件的位置。我們選取包含您要執行文字挖掘之所有文件的目錄。
2. **文字挖掘節點（欄位標籤）**。接下來，我們新增「文字挖掘」節點，並連接至「檔案清單」節點。在此節點中，我們定義輸入格式、資源範本及輸出格式。我們選取從「檔案清單」節點產生的欄位名稱，並選取文字欄位，以及其他設定。如需相關資訊，請參閱主題第 25 頁的『在串流中使用文字挖掘節點』。
3. **文字挖掘節點（模型標籤）**。接下來，在「模型」標籤上，我們選取建置模式以直接從此節點產生概念模型塊。您可以選取不同的資源範本，或者保持基本資源。

範例 2：用來以互動方式建置種類模型的 Excel 檔案及文字挖掘節點

此範例顯示「文字挖掘」節點如何還可以啟動互動式工作台階段作業。如需互動式工作台的相關資訊，請參閱第 57 頁的第 7 章, 『互動式工作台模式』。

1. **Excel 來源節點（資料標籤）**。首先，我們將此節點新增至串流，以指定將儲存文字的位置。
2. **文字挖掘節點（欄位標籤）**。接下來，我們新增「文字挖掘」節點，並進行連接。在這個第一個標籤上，我們定義輸入格式。我們從來源節點中選取欄位名稱。
3. **文字挖掘節點（模型標籤）**。接下來，在「模型」標籤上，我們選取以互動方式建置種類模型塊，以及使用擷取結果自動建置種類。在此範例中，我們從文字分析套件載入資源副本及一組種類。
4. **互動式工作台階段作業**。接下來，我們執行串流，即會開啟互動式工作台介面。執行擷取之後，我們開始探索資料，並改良我們的種類。

文字挖掘塊：概念模型

只要您順序執行「文字挖掘」模型節點（您已在其中選取選項以在「模型」標籤中**直接產生模型**），就會建立「文字挖掘」概念模型塊。文字挖掘概念模型塊用於其他文字資料中主要概念的即時探索，例如呼叫中心的即時運算簿資料。

概念模型塊本身包含概念清單，這已指派給類型。您可以選取該模型中的任何或全部概念，以針對其他資料進行評分。當您執行包含「文字挖掘」模型塊的串流時，會根據建置模型之前在「文字挖掘」建模節點的「模型」標籤上W選取的建置模型，將新的欄位新增至資料。如需相關資訊，請參閱主題 第 27 頁的『概念模型：模型標籤』。

如果是使用轉換的文件產生模型塊，則會在轉換的語言中執行評分。同樣地，如果是使用英文作為語言產生模型塊，則可以在模型塊中指定轉換語言，因為文件就會轉換成英文。

當產生「文字採礦」模型塊時，它們會被放在模型塊選用區中（位於IBM SPSS Modeler視窗右上側的「模型」標籤上）。

檢視結果

若要查看模型塊的相關資訊，請用滑鼠右鍵按一下模型塊選用區中的節點，然後從快速功能表中選擇**瀏覽**（如果是串流中的節點則選擇**編輯**）。

新增模型至串流

如果要將模型塊新增至串流，請按一下模型塊選用區中的圖示，然後按一下要在其中放置節點的串流畫布。或是用滑鼠右鍵按一下圖示，並從快速功能表中選擇**新增至串流**。然後將串流連接至節點，您就準備好傳遞資料以產生預測。

警告：如果您想要使用評分塊來重新產生包含種類模型與所用範本的建模節點，我們建議您建立 TAP，並將它用在互動式階段作業中，取代建模節點，然後再產生評分塊。

概念模型：模型標籤

在概念模型中，「模型」標籤顯示擷取的一組概念。概念以表格格式呈現，一列代表每一個概念。此標籤上的目標是選取哪個概念將用於評分。

附註：如果您改為產生種類模型塊，則此標籤將呈現不同的資訊。如需相關資訊，請參閱主題 第 34 頁的『種類模型塊：模型標籤』。

依預設，會選取所有概念以進行評分，如最左側窗格中勾選框所顯示。已勾選的方框表示將概念用於評分。未勾選的方框表示將從評分中排除概念。您可以透過選取多列，並按一下選擇中的其中一個勾選框來勾選多列。

若要進一步瞭解每一個概念，我們可以查看下列每一個直欄中提供的其他資訊：

概念。這是擷取的前導單字或片語。在部分情況下，此概念代表概念名稱，以及與此概念相關聯的部分其他基礎術語。若要查看哪個基礎術語是概念的一部分，請顯示此標籤內部的「基礎術語」窗格，並選取概念以查看位於對話框底端的對應術語。如需相關資訊，請參閱主題 第 28 頁的『概念模型中的基礎術語』。

廣域。這裡，廣域（頻率）是指概念（及其所有基礎術語）在整個文件/記錄集中出現的次數。

- **長條圖。**此概念在文字資料中的廣域頻率呈現為長條圖。長條圖採用指派概念的目標類型的顏色，從而以視覺化方式識別類型。
- **%.** 此概念在文字資料中的廣域頻率呈現為百分比。
- **N.** 此概念在文字資料中的實際出現次數。

文件。這裡，文件是指文件計數，即出現概念（及其所有基礎術語）的文件/記錄平均數。

- **長條圖。**此概念的文件計數，呈現為長條圖。長條圖採用指派概念的目標類型的顏色，從而以視覺化方式識別類型。
- **%.** 此概念的文件計數，呈現為百分比。
- **N.** 包含此概念的文件或記錄的實際數目。

類型。指派概念的目標類型。對於每一個概念，「廣域」與「文件」直欄以顏色顯示，表示指派此概念的目標類型。**類型**是概念的語意分組。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。

使用概念

透過用滑鼠右鍵按一下表格中的資料格，您可以顯示一個快速功能表，您可以從中：

- **全選。**將選取表格中的所有列。
- **複製。**所選取的概念會複製到剪貼簿。
- **使用欄位複製** 所選取的概念複製到剪貼簿，以及直欄標題。
- **勾選所選取項。**勾選表格中所選取列的所有勾選框，從而包括那些概念以進行評分。
- **取消勾選所選取項。**取消勾選表格中所選取列的所有勾選框。
- **全選。**勾選表格中的所有勾選框。這會造成在最終輸出中使用所有概念。
- **取消全選。**取消勾選表格中的所有勾選框。取消勾選概念表示它將不會用在最終輸出中。

- **包括概念。**顯示「包括概念」對話框。如需相關資訊，請參閱主題 『用於包括評分概念的選項』。

用於包括評分概念的選項

若要快速勾選或取消勾選那些將用於評分的概念，請按一下**包括概念**的工具列按鈕。

圖 *1.* 包括概念工具列按鈕

按一下此工具列按鈕將會開啟「包括概念」對話框，以讓您根據規則選取概念。將包括在「模型」標籤上具有勾號的所有概念以用於評分。在此子對話框中套用規則以變更將用於評分的概念。

您可以從下列選項中選擇：

根據最高頻率勾選概念。前幾個概念數目。 從具有最高廣域頻率的概念開始勾選，這是將勾選的概念數目。這裡的頻率指的是概念（及其所有基礎術語）在整個文件/記錄集中的出現次數。此數目可能會高於記錄計數，因為一個概念可能會在一筆記錄中出現多次。

根據文件計數勾選概念。最大計數。 這是要勾選概念所必須達到的最低文件計數。這裡的文件計數指的是概念（及其所有基礎術語）在其中出現的文件/記錄數。

勾選指派給某個類型的概念。 從下拉清單中選取類型，以勾選指派給此類型的所有概念。在擷取程序期間會自動將概念指派給類型。**類型**是語義上的概念分組。類型包括較高層次概念、肯定字與否定字及限定元、環境定義限定元、名字、位置、組織，等等。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。

取消勾選在太多記錄中出現的概念。記錄百分比。 取消勾選記錄計數百分比高於所指定數目的概念。如果要排除在文字或每筆記錄中頻繁出現但對分析不重要的概念，則此選項很有用。

取消勾選指派給某個類型的概念。 取消勾選符合您從下拉清單中所選類型的概念。

概念模型中的基礎術語

您可以查看針對表格中所選概念而定義的基礎術語。按一下工具列上的基礎術語切換按鈕，就會在對話框底端的分割窗格中顯示基礎術語表格。

這些基礎術語包括語言資源中定義的同義字（無論它們是否出現於文字中）、在文字中找到的用來產生模型塊的任何已擷取的複數/單數形式、排列詞、智慧型分組中的術語，等等。

圖 *2.* 顯示基礎術語工具列按鈕

附註：您無法編輯基礎術語清單。此清單是透過替代、同義字定義（在替代字典中）、智慧型分組等產生的 - 所有這些項目都在語言資源中。若要變更術語在概念下的分組方式或術語的處理方式，則必須直接在資源中進行變更（可在互動式工作台的 資源編輯器 或在 範本編輯器 中編輯，然後在節點中重新載入），然後重新執行串流以取得具有更新結果的新模型塊。

用滑鼠右鍵按一下包含基礎術語或概念的資料格，就會顯示一個快速功能表，您可以在其中執行下列動作：

- **複製。** 會將選取的資料格複製到剪貼簿。
- **連同欄位一起複製。** 會將選取的資料格連同直欄標題一起複製到剪貼簿。

- **全選**。將會選取表格中的所有資料格。

概念模型：設定標籤

「設定」標籤用於定義新輸入資料的文字欄位值（必要的話）。還可以在該處定義輸出（評分模式）的資料模型。

註：僅當將模型塊放置到畫布上時，才會顯示此標籤。當您直接在「模型」選用區中存取此對話框時，它不存在。

評分模式：概念作為記錄

使用此評分模式，會為每一個 concept/document 配對建立新的記錄。通常，輸出中的記錄數多於輸入中的記錄數。

除了輸入欄位之外，會將以下新欄位新增至資料：

表 *4.* 「概念作為記錄」的輸出欄位

欄位	**說明**
概念	包含在文字資料欄位中找到的已擷取概念名稱。
類型	儲存概念的類型作為完整類型名稱，例如位置或人員。**類型**是概念的語意分組。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。
計數	顯示文字主題（記錄/文件）中該概念（及其基礎術語）的出現數目。

當您選取此選項時，會停用除了**容納標點符號錯誤**之外的所有其他選項。

評分模式：概念作為欄位

在概念模型中，對於每一筆輸入記錄，為在給定文件中找到的每個概念建立一筆新記錄。因此，輸出記錄數與輸入記錄數相同。然而，現在每一筆記錄（列）都針對在「模型」標籤上選取（使用勾號）的每個概念包含一個新的欄位（直欄）。每一個概念欄位的值都取決於您是選取**旗標**還是**計數**作為此標籤上的欄位值。

註：如果您要使用非常大型的資料集，例如使用 DB2 資料庫，則使用**概念作為欄位**可能由於資料量而發生處理問題。在此情況下，我們建議改為使用**概念作為記錄**。

欄位值。選擇每一個概念的新欄位是否將包含計數或旗標值。

- **旗標。**此選項用來取得輸出中具有兩個不同值的旗標，例如 *Yes/No*、*True/False*、*T/F* 或 *1* 及 *2*。系統會自動設定儲存體類型以反映所選擇值。例如，如果您為旗標輸入數值，則它們將自動作為整數值處理。旗標的儲存體類型可以是字串、整數、實數或日期/時間。輸入 **True** 及 **False** 的旗標值。
- **個數。** 用來取得給定記錄中發生概念次數的計數。

欄位副檔名。指定欄位名稱的副檔名。透過使用概念名稱加上此副檔命，可產生欄位名稱。

- **新增為。**指定哪裡應該將副檔名新增至欄位名稱。選擇**字首**以將副檔名新增至字串的開頭。選擇**字尾**以將副檔名新增至字串的結尾。

容納標點符號錯誤。此選項會在擷取期間暫時正規化包含標點符號錯誤的文字（例如，用法不當），以改進概念的擷取能力。當文字較短且品質較差時（例如，在開放式意見調查回應、電子郵件及 CRM 資料中時），或者當文字包含許多縮寫時，此選項極其有用。

概念模型：欄位標籤

必要的話，「欄位」標籤會定義新輸入資料的文字欄位值。

註：僅當模型塊放置在串流中時，才會顯示此標籤。如果您直接在「模型」選用區中存取此輸出，則不會顯示此標籤。

「文字」欄位。 選取含有要發掘之文字的欄位。這個欄位視資料來源而定。

文件類型。 文件類型指定文字的結構。請選取下列一種類型：

- **全文。** 用於大部分的文件或文字來源。會掃描整組文字以進行擷取。有別於其他選項，這個選項沒有其他設定。
- **結構化文字。** 用於書目表單、專利，以及任何包含可以識別及分析之一般結構的檔案。這種文件類型用來跳過全部或部分擷取程序。它可讓您定義術語分隔字元、指派類型，以及強制最小頻率值。如果您選取這個選項，則必須按一下**設定**按鈕，然後在「文件設定」對話框的**結構化文字格式化**區域中輸入文字分隔字元。如需相關資訊，請參閱主題第 19 頁的『欄位標籤的文件設定』。

輸入編碼。 只有在您指出文字欄位代表**要記載的路徑名稱**時，才有這個選項可用。它指定預設文字編碼。 會執行從指定的或辨識的編碼到 ISO-8859-1 的轉換。因此即使您指定另一個編碼，擷取引擎都會在處理它之前將它轉換為 ISO-8859-1。任何不合配 ISO-8859-1 編碼定義的字元都會轉換為空格。

文字語言。 識別要發掘的文字的語言；這是在擷取期間偵測到的主要語言。如果您有興趣購買您目前無權存取之受支援語言的授權，請聯絡您的業務代表。

概念模型：摘要標籤

「摘要」標籤會呈現模型自身（分析資料夾）、模型中所用欄位（欄位資料夾）、建置模型時所用設定（建置設定資料夾）及模型訓練（訓練摘要資料夾）的相關資訊。

第一次瀏覽建模節點時，「摘要」標籤上的資料夾是收合的。若要查看相關結果，請使用資料夾左側的展開控制項來顯示結果，或按一下**全部展開**按鈕以顯示全部結果。若要在檢視結果之後隱藏結果，請使用展開控制項來收合您要隱藏的特定資料夾，或按一下**全部收合**按鈕以收合所有資料夾。

在串流中使用概念模型塊

當使用「文字挖掘」建模節點時，您可以產生概念模型塊或種類模型塊（透過互動式工作台階段作業）。下列範例顯示如何在簡式串流中使用概念模型。

範例：具有概念模型塊的統計資料檔案節點

下列範例顯示如何使用「文字挖掘」概念模型塊。

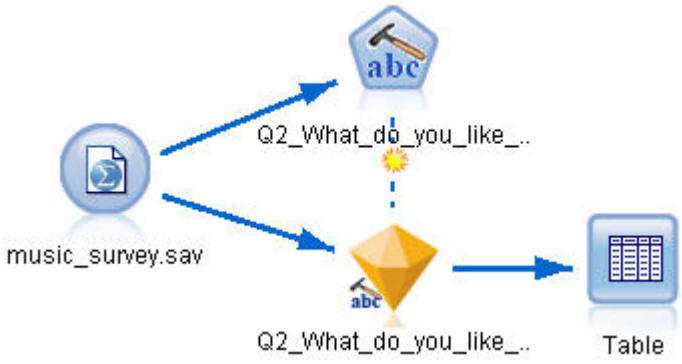


圖 3. 範例串流：具有文字挖掘概念模型塊的統計資料檔案節點

1. **統計資料檔案節點（資料標籤）**。首先，我們將此節點新增至串流，以指定將儲存文字文件的位置。

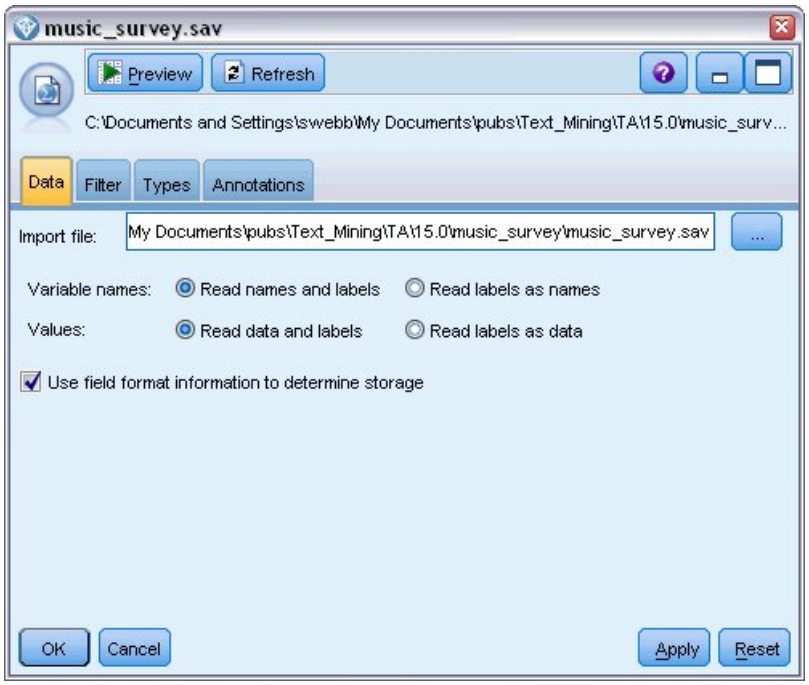


圖 4. 統計資料檔案節點對話框：資料標籤

2. **文字挖掘概念模型塊（模型標籤）**。接下來，我們新增概念模型塊，並連接至「統計資料檔案」節點。我們選取想要用來評分資料的概念。

圖 *5.* 文字挖掘模型塊對話框：模型標籤

3. **文字挖掘概念模型塊（設定標籤）**。接下來，我們定義輸出格式，並選取概念作為欄位。將在輸入中為「模型」標籤中選取的每一個概念建立一個新的欄位。每一個欄位名稱將由概念名稱與字首 "*Concept_*" 組成

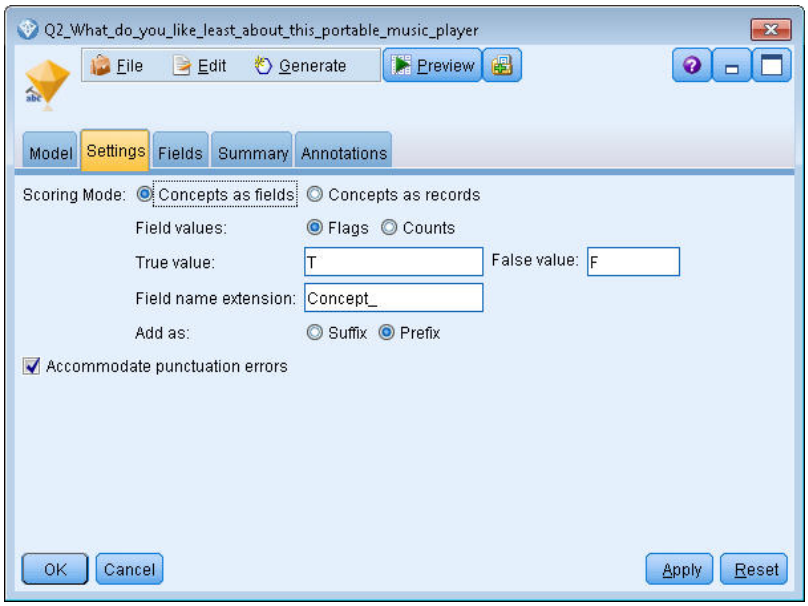


圖 *6.* 文字挖掘概念模型塊對話框：設定標籤

4. **文字挖掘概念模型塊（欄位標籤）**。接下來，我們選取文字欄位 **Q2_What_do_you_like_least_about_this_portable_music_player**，這時來自「統計資料檔案」節點的欄位名稱。我們還選取選項**文字欄位代表：實際文字**。

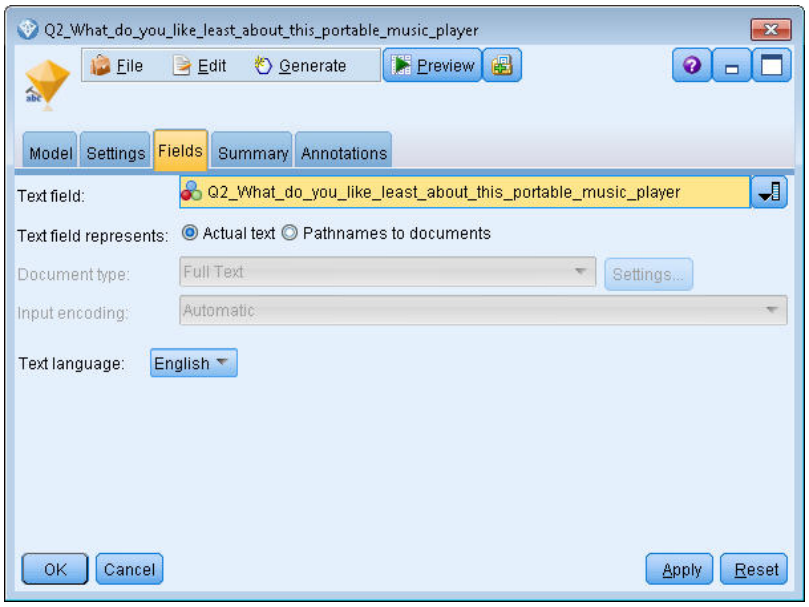
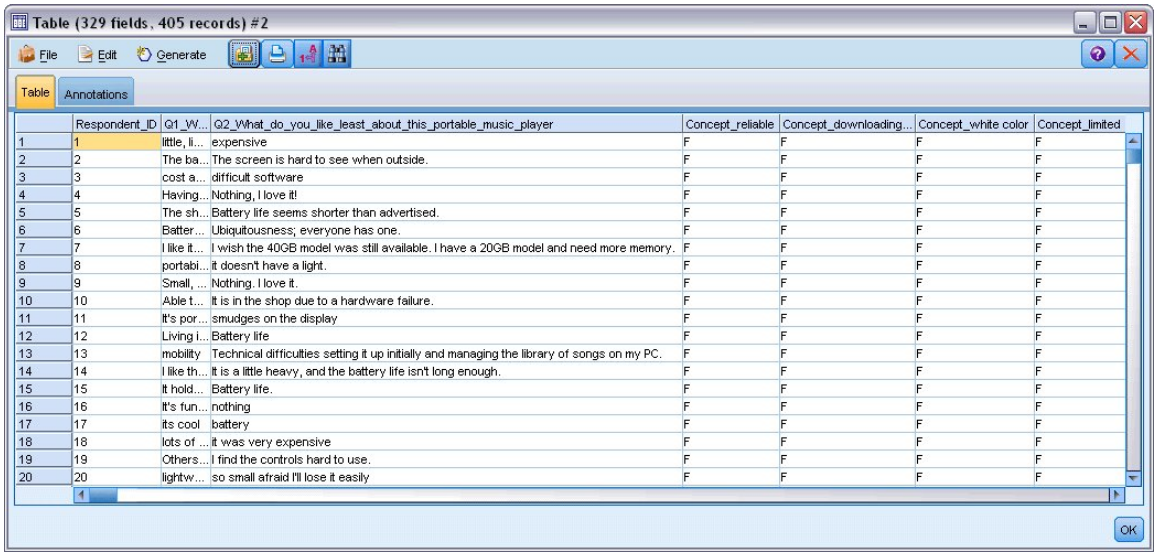


圖 *7.* 文字挖掘概念模型塊對話框：欄位標籤

5. **表格節點**。接下來，我們附加表格節點以查看結果，並執行串流。即會在畫面上開啟表格輸出。



	Respondent_ID	Q1_W...	Q2_What_do_you_like_least_about_this_portable_music_player	Concept_reliable	Concept_downloading...	Concept_white color	Concept_limited
1	1	little, li...	expensive	F	F	F	F
2	2	The ba...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Having...	Nothing, I love it!	F	F	F	F
5	5	The sh...	Battery life seems shorter than advertised.	F	F	F	F
6	6	Batter...	Ubiquitousness; everyone has one.	F	F	F	F
7	7	I like it...	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F
8	8	portabi...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing. I love it.	F	F	F	F
10	10	Able t...	It is in the shop due to a hardware failure.	F	F	F	F
11	11	It's por...	smudges on the display	F	F	F	F
12	12	Living i...	Battery life	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F
14	14	I like th...	It is a little heavy, and the battery life isn't long enough.	F	F	F	F
15	15	It hold...	Battery life.	F	F	F	F
16	16	It's fun...	nothing	F	F	F	F
17	17	its cool	battery	F	F	F	F
18	18	lots of ...	it was very expensive	F	F	F	F
19	19	Others...	I find the controls hard to use.	F	F	F	F
20	20	lightw...	so small afraid I'll lose it easily	F	F	F	F

圖 *8.* 表格輸出捲動以顯示概念旗標

文字挖掘塊：種類模型

只要您從互動式工作台內產生種類模型，就會建立「文字挖掘」種類模型塊。此建模塊包含一組種類，其定義由概念、類型、TLA 型樣及/或種類規則組成。該塊用於分類意見調查回應、部落格文章、其他 Web 資訊來源及任何其他文字資料。

如果您在建模節點中啟動互動式工作台階段作業，則可以在產生種類模型之前探索擷取結果、精簡資源、細部調整您的種類。當您執行包含「文字挖掘」模型塊的串流時，會根據建置模型之前在「文字挖掘」建模節點的「模型」標籤上W選取的建置模型，將新的欄位新增至資料。如需相關資訊，請參閱主題 『種類模型塊：模型標籤』。

如果是使用轉換的文件產生模型塊，則會在轉換的語言中執行評分。同樣地，如果是使用英文作為語言產生模型塊，則可以在模型塊中指定轉換語言，因為文件就會轉換成英文。

當產生「文字採礦」模型塊時，它們會被放在模型塊選用區中（位於IBM SPSS Modeler視窗右上側的「模型」標籤上）。

檢視結果

若要查看模型塊的相關資訊，請用滑鼠右鍵按一下模型塊選用區中的節點，然後從快速功能表中選擇**瀏覽**（如果是串流中的節點則選擇**編輯**）。

新增模型至串流

如果要將模型塊新增至串流，請按一下模型塊選用區中的圖示，然後按一下要在其中放置節點的串流畫布。或是用滑鼠右鍵按一下圖示，並從快速功能表中選擇**新增至串流**。然後將串流連接至節點，您就準備好傳遞資料以產生預測。

警告：如果您想要使用評分塊來重新產生包含種類模型與所用範本的建模節點，我們建議您建立 TAP，並將它用在互動式階段作業中，取代建模節點，然後再產生評分塊。

種類模型塊：模型標籤

對於種類模型，模型標籤在左側的種類模型中顯示種類清單，並在右側顯示所選取種類的描述子。每一個種類都由許多描述子組成。對於您選取的每一個種類，相關聯的描述顯示在表格中。這些描述子可以包括概念、種類規則、類型及 TLA 型樣。還會顯示每一個描述子的類型，以及每一個描述子代表的部分範例。

在標籤上，目標是選取您要用於評分的種類。對於種類模型，將文件及記錄評分至種類。如果文件或記錄在其文字或任何基礎術語中包含一個或多個描述子，則會將文件或記錄指派給描述子屬於的種類。這些基礎術語包括在語言資源中定義的同義字（而無論是否在文字中找到它們），以及在用於產生模型塊、已排列術語、模糊分組中的術語等項目的文字中找到的任何擷取的複數/單數術語。

附註：如果您產生概念模型塊，則此標籤將包含不同的結果。如需相關資訊，請參閱主題 第 27 頁的『概念模型：模型標籤』。

種類樹狀結構

若要進一步瞭解每一個種類，請選取該種類，並檢閱為該種類中描述子出現的資訊。對於每一個描述子，您可以檢閱下列資訊：

- **描述子**名稱。此欄位包含一個圖示，代表描述子的種類，以及描述子名稱。

表 5. 描述子圖示

	概念		TLA 型樣
	類型		種類規則

- **類型。**此欄位包含描述子的類型名稱。類型是類似概念（語意分組）的集合，例如組織名稱、產品或正面意見。規則未指派給類型。
- **詳細資料。**此欄位包含該描述子中包括的項目清單。根據相符項的數目，由於對話框中的大小限制，您可能不會看到每一個描述子的整個清單。

選取及複製種類

依預設，會選取所有最上層種類以進行評分，如左窗格中勾選框所顯示。已勾選的方框表示將種類用於評分。未勾選的方框表示將從評分中排除種類。您可以透過選取多列，並按一下選擇中的其中一個勾選框來勾選多列。此外，如果選取種類或子種類，但是未選取其中一個子種類，則指出在所選取種類的子項中只有部分選擇。

透過用滑鼠右鍵按一下樹狀結構中的種類，您可以顯示一個快速功能表，您可以從中：

- **勾選所選取項。**勾選表格中所選取列的所有勾選框。
- **取消勾選所選取項。**取消勾選表格中所選取列的所有勾選框。
- **全選。**勾選表格中的所有勾選框。這會造成在最終輸出中使用所有種類。您還可以使用工具列上的對應勾選框圖示。
- **取消全選。**取消勾選表格中的所有勾選框。取消勾選種類表示它將不會用在最終輸出中。您還可以使用工具列上對應的空勾選框圖示。

透過用滑鼠右鍵按一下描述子表格中的資料格，您可以顯示一個快速功能表，您可以從中：

- **複製。**所選取的概念會複製到剪貼簿。
- **利用欄位複製。**所選取的描述子會隨直欄標題一起複製到剪貼簿。
- **全選。**將選取表格中的所有列。

種類模型塊：設定標籤

「設定」標籤用於定義新輸入資料的文字欄位值（必要的話）。還可以在該處定義輸出（評分模式）的資料模型。

註：僅當模型塊放置在畫布上或串流中時，此標籤才顯示在節點對話框中。當您直接在「模型」選用區中存取此塊時，它不存在。

評分模式：種類作為欄位

使用此選項，輸出記錄數與輸入記錄數相同。然而，現在每一筆記錄都針對在「模型」標籤上選取（使用勾號）的每個種類包含一個新的欄位。對於每一個欄位，為 **True** 及 **False** 輸入一個旗標值，例如 *Yes/No*、*True/False*、*T/F* 或 *1* 與 *2*。系統會自動設定儲存體類型以反映所選擇的值。例如，如果您為旗標輸入數值，則它們將自動作為整數值處理。旗標的儲存體類型可以是字串、整數、實數或日期/時間。

註：如果您要使用非常大型的資料集，例如使用 DB2 資料庫，則使用**種類作為欄位**可能由於資料量而發生處理問題。在此情況下，我們建議改為使用**種類作為記錄**。

欄位副檔名。您可以選擇為欄位名稱指定副檔名字首/字尾，或者您可以選擇使用種類代碼。透過使用種類名稱加上此副檔名，產生欄位名稱。

- **新增為。**指定哪裡應該將副檔名新增至欄位名稱。選擇**字首**以將副檔名新增至字串的開頭。選擇**字尾**以將副檔名新增至字串的結尾。

如果未選取子種類。 這個選項可讓您指定將會如何處理未選取其所屬子種類以進行評分的描述子。有兩個選項。

- **從評分完全排除其描述子**選項將會導致在評分期間忽略及未使用沒有勾號（未選取）之子種類的描述子。
- **聚集那些在母種類中之種類的描述子**選項將會導致沒有勾號（未選取）之子種類的描述子被用來作為母種類（在此子種類之上的種類）的描述子。如果未選取數個層次的子種類，則會在第一個可用的母種類下累積描述子。

僅評分最低層的相符種類。 使用這個選項，可將種類僅輸出在單一行上（比方說，如果種類是 GeneralSatisfaction/Pos，則選取這個選項會導致 GeneralSatisfaction/Pos。如果沒有這個選項，則會出現兩行： GeneralSatisfaction 和 GeneralSatisfaction/Pos）。

容納標點符號錯誤。此選項會在擷取期間暫時正規化包含標點符號錯誤的文字（例如，用法不當），以改進概念的擷取能力。當文字較短且品質較差時（例如，在開放式意見調查回應、電子郵件及 CRM 資料中時），或者當文字包含許多縮寫時，此選項極其有用。

評分模式：種類作為記錄

使用此選項，會為每一個 category、document 配對建立新的記錄。通常，輸出中的記錄數多於輸入中的記錄數。除了輸入欄位之外，還會根據模型的種類，將新的欄位新增至資料。

表 *6.* 「種類作為記錄」的輸出欄位

新的輸出欄位	說明
Category	包含指派文字文件的目標種類名稱。進行如果種類是另一個種類的子種類，則種類名稱的完整路徑會由您在此對話框中選擇的值進行控制。

階層式種類的值。此選項會控制如何在輸出中顯示子種類的名稱。

- **完整種類路徑。**此選項將輸出種類的名稱及母項種類的完整路徑，如果適用，使用斜線分隔種類名稱與子種類名稱。
- **簡短種類名稱。**此選項將僅輸出種類的名稱，但是使用省略符號顯示問題中種類的母項種類數目。
- **底端層次種類。**此選項將僅輸出種類的名稱，但不顯示完整路徑或母項種類。

如果未選取子種類。 這個選項可讓您指定將會如何處理未選取其所屬子種類以進行評分的描述子。有兩個選項。

- **從評分完全排除其描述子**選項將會導致在評分期間忽略及未使用沒有勾號（未選取）之子種類的描述子。
- **聚集那些在母種類中之種類的描述子**選項將會導致沒有勾號（未選取）之子種類的描述子被用來作為母種類（在此子種類之上的種類）的描述子。如果未選取數個層次的子種類，則會在第一個可用的母種類下累積描述子。

容納標點符號錯誤。此選項會在擷取期間暫時正規化包含標點符號錯誤的文字（例如，用法不當），以改進概念的擷取能力。當文字較短且品質較差時（例如，在開放式意見調查回應、電子郵件及 CRM 資料中時），或者當文字包含許多縮寫時，此選項極其有用。

種類模型塊：其他標籤

種類模型塊的「欄位」標籤及「設定」標籤與概念模型塊相同。

- 「欄位」標籤。如需相關資訊，請參閱主題 第 30 頁的『概念模型：欄位標籤』。
- 「摘要」標籤。如需相關資訊，請參閱主題 第 30 頁的『概念模型：摘要標籤』。

在串流中使用類別模型塊

「文字挖掘」種類模型塊產生自互動式工作台階段作業。您可以在串流中使用此模型塊。

範例：具有種類模型塊的統計資料檔案節點

下列範例顯示如何使用「文字挖掘」模型塊。

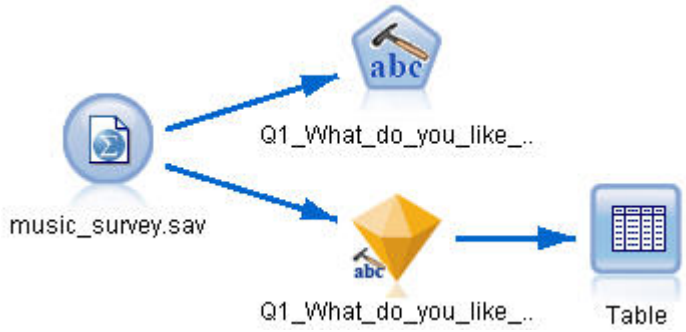


圖 *9.* 範例串流：具有文字挖掘種類模型塊的統計資料檔案節點

1. **統計資料檔案節點（資料標籤）**。首先，我們將此節點新增至串流，以指定將儲存文字文件的位置。

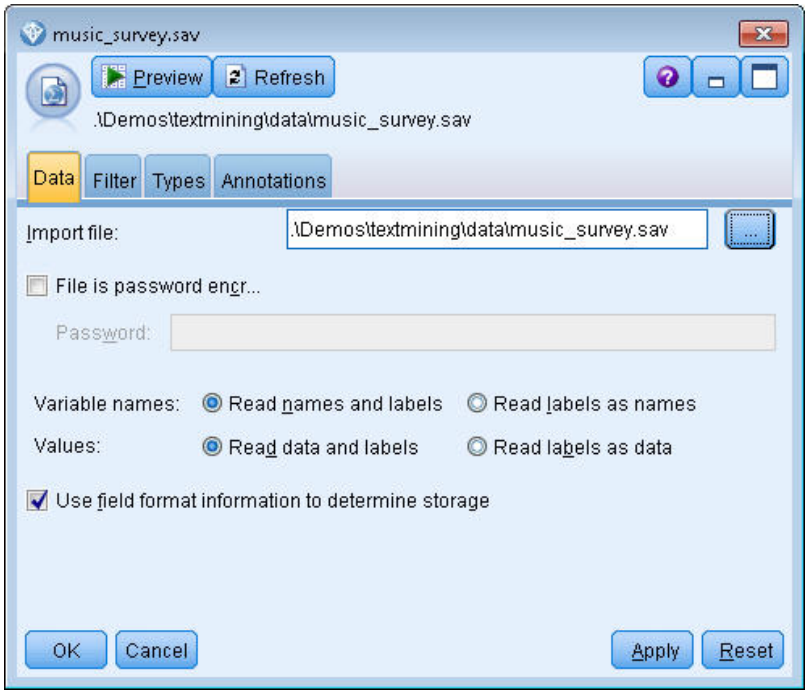


圖 *10.* 統計資料檔案節點對話框：資料標籤

2. **文字挖掘種類模型塊（模型標籤）**。接下來，我們新增種類模型塊，並連接至「統計資料檔案」節點。我們選取想要用來評分資料的種類。

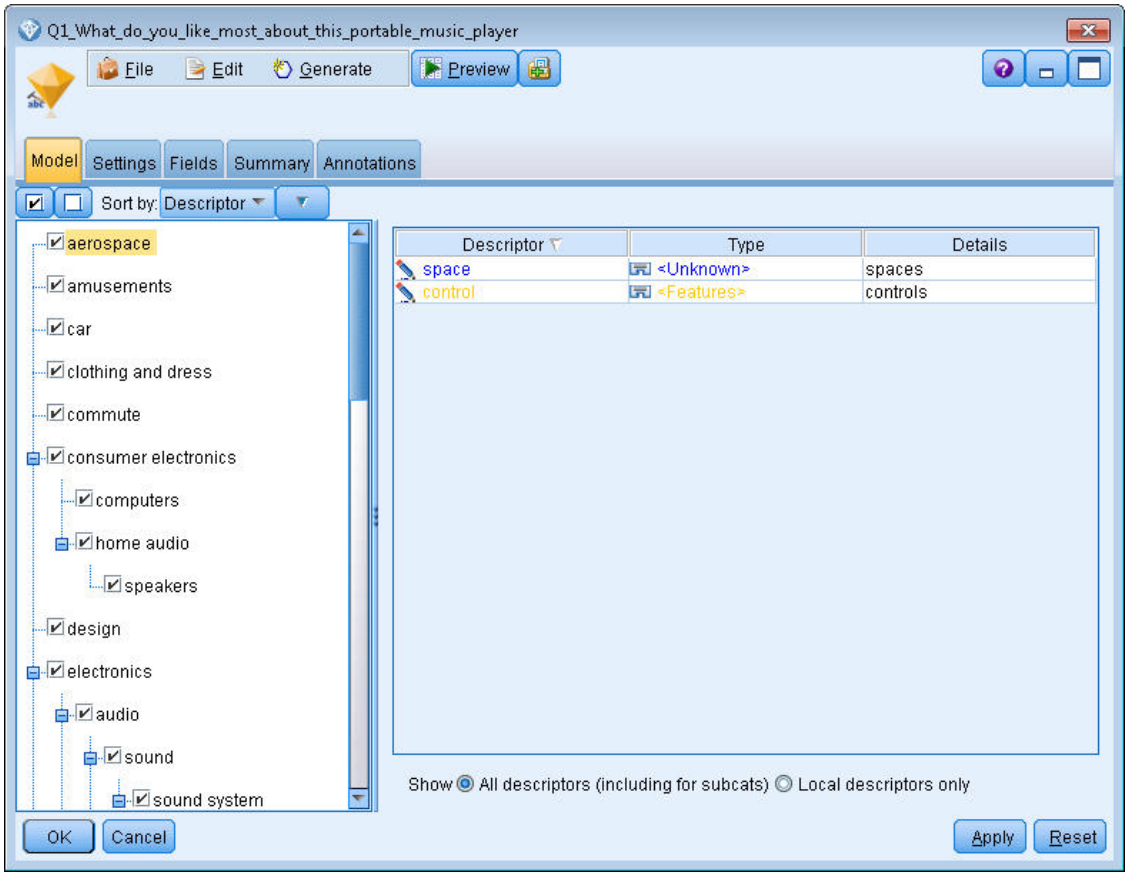


圖 *11.* 文字挖掘模型塊對話框：模型標籤

3. **文字挖掘模型塊（設定標籤）**。接下來，我們定義輸出格式**種類作為欄位**。

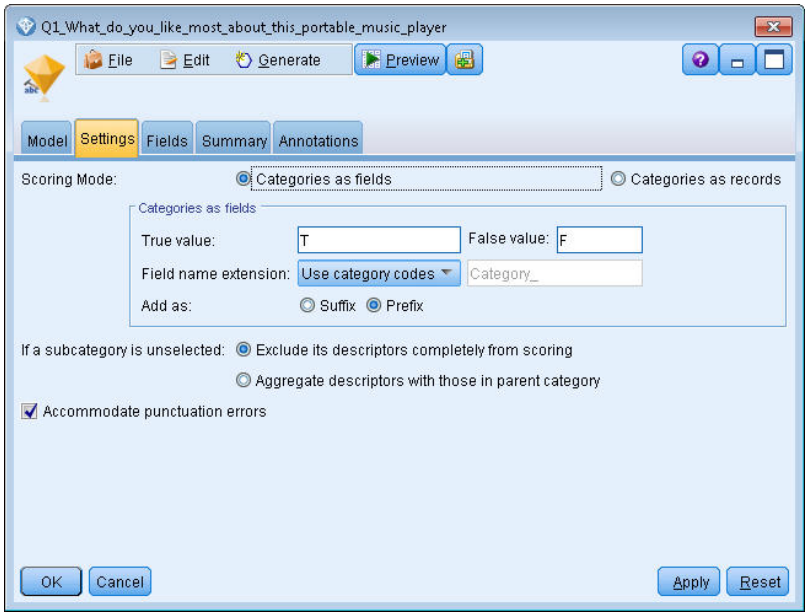


圖 *12.* 種類模型塊對話框：設定標籤

4. **文字挖掘種類模型塊（欄位標籤）**。接下來，我們選取文字欄位變數，這是來自「統計資料欄位」節點的欄位名稱，並且選取的選項「文字」欄位代表**實際文字**，以及其他設定。

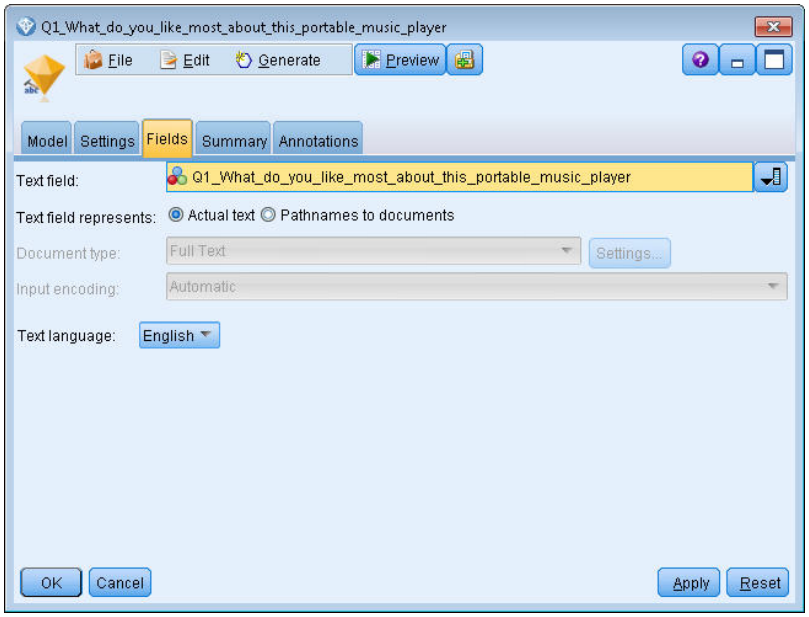
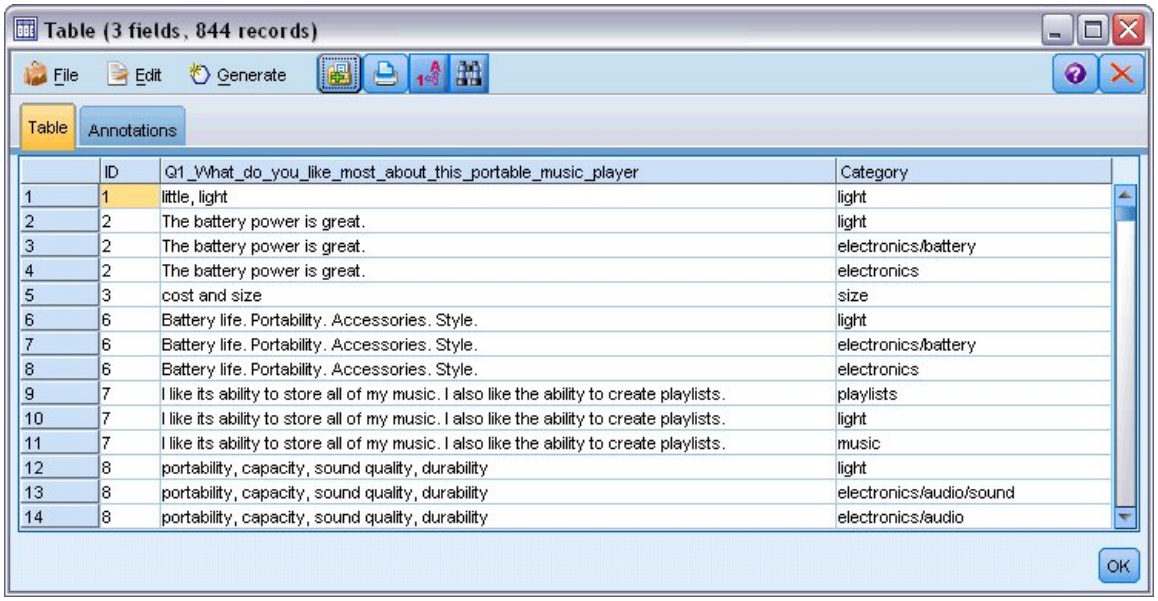


圖 *13.* 文字挖掘模型塊對話框：欄位標籤

5. **表格節點**。接下來，我們附加表格節點以查看結果，並執行串流。



	ID	Q1_What_do_you_like_most_about_this_portable_music_player	Category
1	1	little, light	light
2	2	The battery power is great.	light
3	2	The battery power is great.	electronics/battery
4	2	The battery power is great.	electronics
5	3	cost and size	size
6	6	Battery life. Portability. Accessories. Style.	light
7	6	Battery life. Portability. Accessories. Style.	electronics/battery
8	6	Battery life. Portability. Accessories. Style.	electronics
9	7	I like its ability to store all of my music. I also like the ability to create playlists.	playlists
10	7	I like its ability to store all of my music. I also like the ability to create playlists.	light
11	7	I like its ability to store all of my music. I also like the ability to create playlists.	music
12	8	portability, capacity, sound quality, durability	light
13	8	portability, capacity, sound quality, durability	electronics/audio/sound
14	8	portability, capacity, sound quality, durability	electronics/audio

圖 *14.* 表格輸出

第 4 章 文字鏈結採礦

文字鏈結分析節點

「文字鏈結分析 (TLA)」節點將一個型樣相符技術新增至文字採礦的概念擷取中，以根據已知型樣識別文字資料中概念之間的關係。這些關係可以說明客戶關於產品的使用感受、哪些公司有商業合作，甚至可以說明基因機構或醫藥機構之間的關係。

例如，您可能對擷取競爭者的產品名稱並不是很感興趣。但是您還可以使用此節點來瞭解人們對此產品的使用感受（如果資料中存在此類觀點的話）。這些關係和關聯是透過將已知型樣與您的文字資料進行比對來識別及擷取的。

您可以使用 IBM SPSS Modeler Text Analytics 隨附的某些資源範本內的 TLA 型樣規則，也可以建立/編輯您專屬的型樣規則。型樣規則由巨集、單字清單及字隙組成以構成與輸入文字進行比較的布林查詢或規則。使用 TLA 型樣規則來比對文字時，可以將此文字作為 TLA 結果並重組為輸出資料。如需相關資訊，請參閱主題 第 187 頁的第 18 章, 『關於文字鏈結規則』。

「文字鏈結分析」節點提供了一個更直接的方法來識別及擷取文字中的 TLA 型樣結果，然後將結果新增至串流中的資料集。但「文字鏈結分析」節點並不是您可以執行文字鏈結分析的唯一方法。您還可以在「文字採礦」建模節點中使用互動式工作台階段作業。

在互動式工作台中，您可以探索 TLA 型樣結果並將其用作種類描述子，以及/或使用往下探查圖形來進一步瞭解結果。如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。實際上，使用「文字採礦」節點來擷取 TLA 結果非常適合用於探索及精細調整範本資料以供稍後直接用於 TLA 節點中。

輸出最多可以由 6 個屬性或組件來代表。如需相關資訊，請參閱主題 第 44 頁的『TLA 節點輸出』。

您可以在 IBM SPSS Modeler 視窗底端的節點選用區的 IBM SPSS Modeler Text Analytics 標籤中找到此節點。如需相關資訊，請參閱主題 第 6 頁的『IBM SPSS Modeler Text Analytics 節點』。

需求。「文字鏈結分析」節點可以接受使用任何標準來源節點（「資料庫」節點、「純文字檔」節點等）讀取到欄位中的文字資料，或者讀取到列出外部文件（由「檔案清單」節點或「Web 資訊來源」節點產生）之欄位中的文字資料。

強度。 「文字鏈結分析」節點除了基本概念擷取之外，還會提供概念之間關係的相關資訊，以及資料中可能揭露的相關觀點或限定元資訊。

文字鏈結分析節點：欄位標籤

使用「欄位」標籤來指定將從中擷取概念之資料的欄位設定。您可以設定下列參數：

ID 欄位。 選取包含文字記錄 ID 的欄位。 ID 必須是整數。ID 欄位充當個別文字記錄的索引。如果文字欄位代表要發掘的文字，請使用 ID 欄位。

「文字」欄位。 選取含有要發掘之文字的欄位。這個欄位視資料來源而定。

語言欄位。 選取包含兩個字母之 ISO 語言 ID 的欄位。如果不選取欄位，則會假定每個文件的語言都是所提供範本的語言。

文件類型。 文件類型指定文字的結構。請選取下列一種類型：

- **全文。** 用於大部分的文件或文字來源。會掃描整組文字以進行擷取。有別於其他選項，這個選項沒有其他設定。
- **結構化文字。** 用於書目表單、專利，以及任何包含可以識別及分析之一般結構的檔案。這種文件類型用來跳過全部或部分擷取程序。它可讓您定義術語分隔字元、指派類型，以及強制最小頻率值。如果您選取這個選項，則必須按一下**設定**按鈕，然後在「文件設定」對話框的**結構化文字格式化**區域中輸入文字分隔字元。如需相關資訊，請參閱主題第 19 頁的『欄位標籤的文件設定』。

文字個體。 從下列選項中選取擷取模式：

- **文件模式。** 用於較為簡短且在語意上同質的文件，例如來自新聞社的文章。
- **段落模式。** 用於網頁及非帶標記文件。擷取程序會利用內部標籤和語法之類的性質，依語意分割文件。如果選取了這種模式，會逐個段落套用評分。因此，舉例而言，只有在相同的段落中發現 `apple` 和 `orange`，規則 `apple & orange` 才會為真。

  註：由於從 PDF 文件擷取文字的方式，因此**段落模式**不適用於這些文件。這是因為擷取會抑制換行標記。

段落模式設定。 只有在您將文字個體選項設定為**段落模式**時，才有這個選項可用。請指定要在任何擷取中使用的字元臨界值。實際大小會向上或向下捨入到最接近的期間。如果要確保從文件集合的文字所產生的單字關聯為代表，請避免指定過小的擷取大小。

- **最小值。** 指定要在任何擷取中使用的字元數目下限。
- **最大值。** 指定要在任何擷取中使用的字元數目上限。

複製資源來源。 挖掘文字時，擷取程序不僅是基於「專家」標籤中的設定，還基於語言資源。這些資源充當擷取期間如何處理文字以取得概念、類型及 TLA 型樣的基礎。您可以從資源範本將資源複製到此節點。

資源範本是已針對特定網域或用途進行精細調整的檔案庫及進階語言和非語言資源的預先定義集。這些資源充當擷取期間如何處理資料的基礎。按一下**載入**並選取要從中複製資源的範本。

範本是在您選取範本時而不是在執行串流時載入的。載入範本時，會在節點中儲存資源副本。因此，如果您想要使用更新的範本，則需要在這裡重新載入範本。如需相關資訊，請參閱主題 第 23 頁的『從範本及 TAP 複製資源』。

文字語言。 識別所要發掘之文字的語言。在節點中所複製的資源控制存在的語言選項。請選取已調整其資源的語言。

文字鏈結分析節點：專家標籤

在此節點中，會自動啟用擷取文字鏈結分析 (TLA) 型樣結果。「專家」標籤包含一些其他參數，這些參數會影響擷取及處理文字的方式。此對話框中的參數會控制擷取程序的基本行為以及一些進階行為。還有一些語言資源及選項，它們也會影響受所選資源範本控制的擷取結果。

將擷取限制為廣域頻率至少為 [n] 的概念。 指定單字或詞組要能夠被擷取所必須在文字中發生的次數下限。如此一來，值 5 會將擷取限制為那些在整個記錄或文件集中發生至少五次的單字或詞組。

在某些情況下，變更此限制可能會在產生的擷取結果中造成巨大的差異，並因此造成種類上的差異。假設您正在處理某餐廳資料，並在此選項中不將限制增加超過 1。在此情況下，擷取結果中可能會出現 *pizza (1)*、*thin pizza (2)*、*spinach pizza (2)* 和 *favorite pizza (2)*。不過，如果您要將擷取限制為廣域頻率 5 或以上，然後重新擷取，則您不會再取得這其中的三個概念。而是會取得 *pizza (7)*，因為 *pizza* 是最簡單的格式，而且這個字已經存在為可能的候選字。視其餘的文字而定，您實際的頻率可能超過七，這取決於是否仍然有其他詞組在文字中有 pizza。此外，如果 *spinach pizza* 已經是種類描述子，則您可能需要將 *pizza* 新增為描述子而不是擷取所有記錄。基於此原因，每當已經建立種類時，都要小心變更此限制。

請注意，這是僅限擷取的特性；如果您的範本包含術語，並且在文字中發現範本的術語，則無論其頻率為何，都會檢索該術語。

例如，假設您使用「基本資源」範本，此範本在 Core 檔案庫中的 `<Location>` 類型下包含 "los angeles"；如果您的文件僅包含 Los Angeles 一次，則 Los Angeles 將會是概念清單的一部分。如果要防止此情況，您需要將過濾器設定為顯示發生的次數至少與在**將擷取限制為廣域頻率至少為 [n] 的概念**欄位中輸入的值相同的概念。

容納標點符號錯誤。此選項會在擷取期間暫時正規化包含標點符號錯誤的文字（例如，用法不當），以改進概念的擷取能力。當文字較短且品質較差時（例如，在開放式意見調查回應、電子郵件及 CRM 資料中時），或者當文字包含許多縮寫時，此選項極其有用。

容納單字字元長度下限 [n] 的拼字 此選項套用模糊分組技術，可協助將通常拼字錯誤的單字或根據一個概念拼字接近的單字分組在一起。模糊分組演算法暫時去掉所有母音（除了第一個），並從擷取的單字中去掉雙/三重輔音，然後比較它們以查看它們是否相同，以便將 `modeling` 與 `modelling` 分組在一起。然而，如果將每一個術語指派給不同的類型（排除 `<Unknown>` 類型），則將不會套用模糊分組技術。

您也可以在使用模糊分組之前，定義需要的根字元數目下限。術語中的根字元數目計算方式為所有字元總數減去形成字形變化字尾的字元，若為複合字術語，則再減去限定詞與介詞。例如，術語 `exercises` 將以 "exercise" 形式計為 8 個根字元，位於單字末尾的字母 *s* 是字形變化（複數形式）。類似地，`apple sauce` 計為 10 個根字元 ("apple sauce")，而 `manufacturing of cars` 計為 16 個根字元 ("manufacturing car")。此計數方法僅用於檢查是否應該套用模糊分組，但不會影響單字的相符程度。

註：如果您稍後發現某些單字未正確地分組，則可以透過在「進階資源」標籤的**模糊分組：異常狀況**區段中明確地宣告，從此技術中排除單字配對。如需相關資訊，請參閱主題 第 177 頁的『模糊分組』。

擷取單一術語 只要單字尚且不是複合字的一部分，或者如果它是名詞或無法辨識的語音的一部分，則此選項會擷取單字（單一屬於）。

擷取非語言實體 此選項擷取非語言實體，例如電話號碼、社會安全號碼、時間、日期、貨幣、數位、百分比、電子郵件位址及 HTTP 位址。您可以在「進階資源」標籤的**非語言實體：配置**區段中包括或排除某些類型的非語言實體。透過停用任何不需要的實體，擷取引擎不會浪費處理時間。如需相關資訊，請參閱主題第 181 頁的『配置』。

大寫演算法 只要術語的第一個字母為大寫形式，此選項就擷取不在內建目錄中的簡式及複合術語。此選項提供良好的方法以擷取最適當的名詞。

可能時將部分及全部人員名稱分組在一起 此選項將文字中看起來不同的名稱分組在一起。由於通常在文字開頭以名稱的完整形式對名稱以縮寫進行參照，之後僅使用縮寫版本，因此本功能有用。此選項嘗試將任何類型為 <Unknown> 的單一術語與鍵入為 <Person> 的任何複合術語的最後一個單字進行比對。例如，如果發現 *doe*，且最初鍵入為 <Unknown>，則擷取引擎會檢查以查看 <Person> 類型中是否有任何複合術語包括 *doe* 作為最後一個單字，例如 *john doe*。由於大部分名稱從不作為單一術語擷取，因此本選項不適用於第一個名稱。

非功能單字排列上限 此選項指定套用排列技術時可以呈現的非功能單字數目上限。此排列技術僅依照所包含的非功能單字，將彼此不同的類似片語分組在一起（例如，of 及 the），而不考量字形變化。例如，讓我們假設將此值設為最多兩個單字，並擷取 company officials 與 officials of the company。在此情況下，由於當忽略 of the 時，兩個術語被視為相同，因此兩個擷取的術語將在最終概念清單中分組在一起。

分組多術語時使用衍生 處理海量資料時，選取此選項以透過使用衍生規則分組多術語。

TLA 節點輸出

執行「文字鏈結分析」節點之後，會重組資料。請務必要瞭解文字採礦重組資料的方式。如果您想要將不同結構用於資料採礦，則可以使用「欄位作業」選用區上的節點來達成此目的。例如，如果在您所使用的資料中，每一列皆代表一筆文字記錄，則會為來源文字資料中未涵蓋的每個型樣建立一列。輸出中的每一列都有 15 個欄位：

- 6 個欄位（**概念編號**，例如 **Concept1**、**Concept2**... 及 **Concept6**）代表型樣比對中找到的所有概念。
- 6 個欄位（**類型編號**，例如 **Type1**、**Type2**... 及 **Type6**）代表每個概念的類型。
- **規則名稱**代表用來比對文字及產生輸出的文字鏈結規則的名稱。
- 使用您在節點中所指定 ID 欄位之名稱的欄位，其代表記錄或文件在輸入資料中的 ID
- **相符文字**代表原始記錄或文件中符合 TLA 型樣的文字資料部分。

註：任何預先存在的串流（包含 5.0 之前版次中的「文字鏈結分析」節點）可能無法完全可執行，直到您更新這些節點為止。 IBM SPSS Modeler 較新版本中的某些改良可能需要將較舊節點取代為較新版本，後者更易於部署且功能更強大。

還可以對某些語言執行自動翻譯。此功能可讓您能夠發掘您可能不會讀寫的語言版本的文件。如果您要使用翻譯功能，您必須具有 SDL 軟體即服務 (SaaS) 的存取權。如需相關資訊，請參閱主題 翻譯設定。

快取 TLA 結果

如果進行快取，則文字鏈結分析結果位於串流中。若要避免每次執行串流時都會重複擷取文字鏈結分析結果，請選取「文字鏈結分析」節點，並從功能表中選擇**編輯 > 節點 > 快取 > 啟用**。下次執行串流時，輸出就會快取在節點中。節點圖示會顯示一個小「文件」圖形，填入快取時，該圖形會從白色變更為綠色。在階段作業期間會保留快取。若要延長保留快取一天（在關閉並重新開啟串流之後），請選取節點並從功能表中選擇 **編輯 > 節點 > 快取 > 儲存快取**。下次開啟串流時，您可以重新載入已儲存的快取而不是再次執行翻譯。

或者，您也可以儲存或啟用節點快取，方法是用滑鼠右鍵按一下節點並從快速功能表中選擇**快取**。

在串流中使用文字鏈結分析節點

「文字鏈結分析」節點用來在串流中存取資料及擷取概念。您可以使用任何來源節點來存取資料。

範例：具有文字鏈結分析節點的統計資料檔案節點

下列範例顯示如何使用「文字鏈結分析」節點。

圖 *15.* 範例：具有文字鏈結分析節點的統計資料檔案節點

1. **統計資料檔案節點（資料標籤）**。先將此節點新增至串流以指定文字的儲存位置。

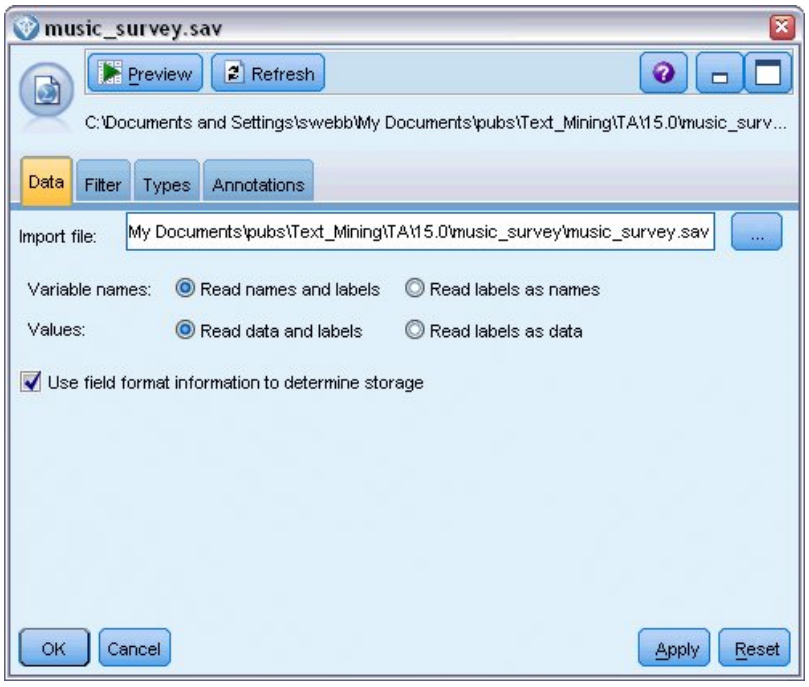


圖 *16.* 「統計資料檔案」節點對話框：「資料」標籤

2. **文字鏈結分析節點（「資料」標籤）**。 接著將此節點附加至串流以擷取概念用於下游建模或檢視。已指定 ID 欄位及包含資料的文字欄位名稱，以及其他設定。

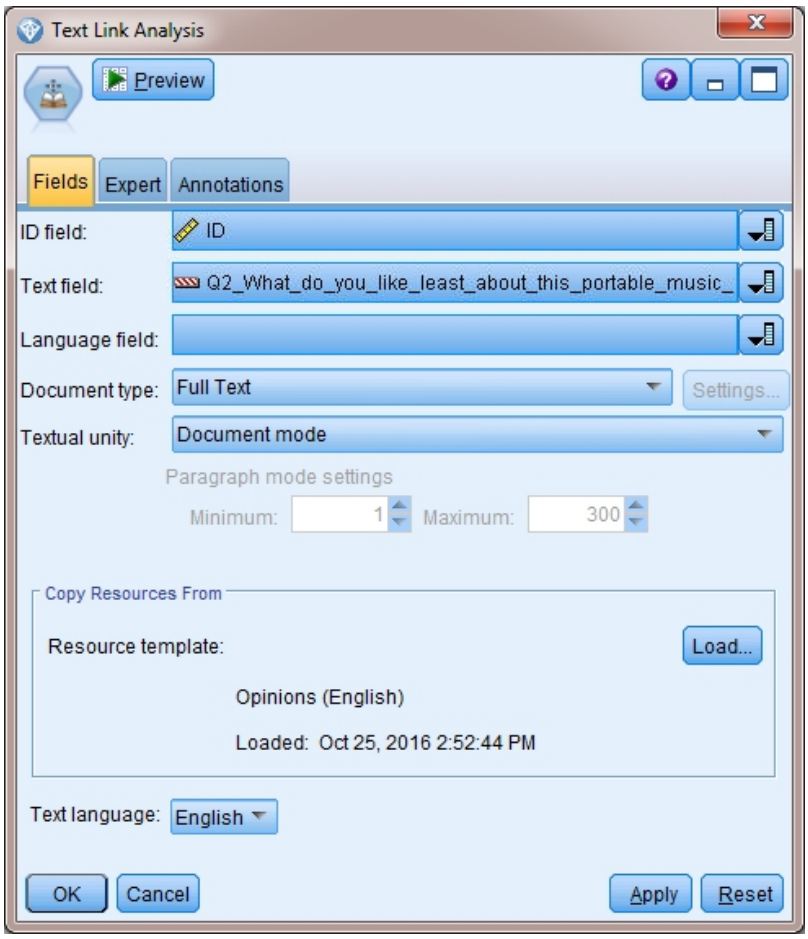
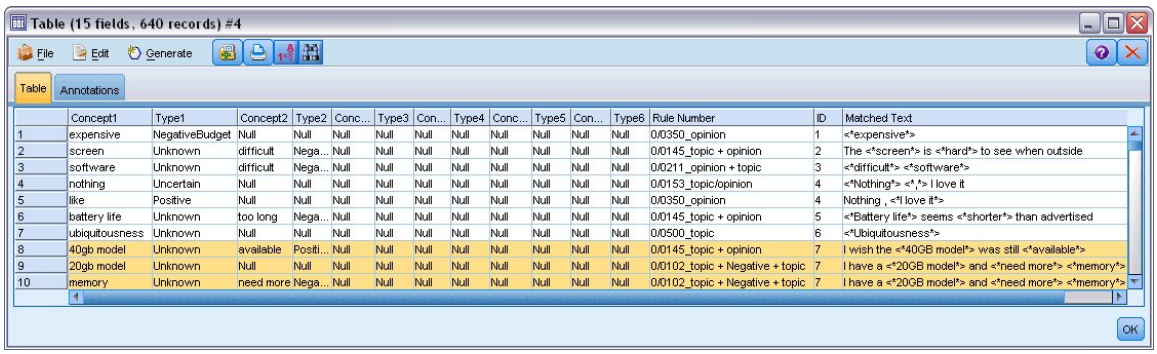


圖 *17.* 文字鏈結分析節點對話框：欄位標籤

3. **表格節點**。最後，附加「表格」節點以檢視從文字文件中擷取的概念。在顯示的表格輸出中，使用「鏈結分析」節點執行此串流之後，您可以看到在資料中找到的 TLA 型樣結果。部分結果顯示僅符合一個概念/類型。而在其他節點中，結果則更為複雜且包含數個類型和概念。此外，透過「文字鏈結分析」節點執行資料以及擷取概念之後，也會導致變更資料的數個方面。範例中的原始資料包含 8 個欄位及 405 筆記錄。執行「文字鏈結分析」節點之後，現在有 15 個欄位及 640 筆記錄。現在，找到的每個 TLA 型樣結果都有一個對應列。例如，ID 7 從原始變成了 3 列，因為擷取了 3 個 TLA 型樣結果。如果您想要將此輸出資料合併到原始資料中，則可以使用「合併」節點。



	Concept1	Type1	Concept2	Type2	Conc...	Type3	Con...	Type4	Conc...	Type5	Con...	Type6	Rule Number	ID	Matched Text
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*expensive*>
2	screen	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	2	The <*screen*> is <*hard*> to see when outside
3	software	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0211_opinion + topic	3	<*difficult*> <*software*>
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0153_topic/opinion	4	<*Nothing*> <*,*> I love it
5	like	Positive	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	4	Nothing , <*I love it*>
6	battery life	Unknown	too long	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	5	<*Battery life*> seems <*shorter*> than advertised
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0500_topic	6	<*Ubiquitousness*>
8	40gb model	Unknown	available	Positi...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	7	I wish the <*40GB model*> was still <*available*>
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>
10	memory	Unknown	need more	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>

圖 *18.* 表格輸出節點

第 5 章 瀏覽外部來源文字

檔案檢視器節點

對一系列文件進行採礦時，您可以直接在「文字採礦」建模節點中指定檔案的完整路徑名稱。但是，如果輸出到「表格」節點，則您將只能看到文件的完整路徑名稱而不是文件內的文字。「檔案檢視器」節點可以當作「表格」節點使用，它可讓您能夠存取每一個文件內的實際文字，而無需將它們全都合併到單一檔案中。

「檔案檢視器」節點可以協助您更充分的瞭解文字擷取結果，方法是它可以讓您存取從中擷取概念的來源文字或未翻譯文字，而在串流中是無法存取這些文字的。此節點會新增至串流中的「檔案清單」節點之後，以取得所有檔案的鏈結清單。

此節點的結果是一個視窗，其中顯示已讀取並用來擷取概念的所有文件元素。從此視窗中，您可以按一下工具列圖示以在外部瀏覽器中啟動報告，其中會以超鏈結的形式列出文件名稱。您可以按一下鏈結以開啟集合中的對應文件。如需相關資訊，請參閱主題 『使用檔案檢視器節點』。

您可以在位於IBM SPSS Modeler視窗底端的節點選用區的IBM SPSS Modeler Text Analytics標籤上尋找這個節點。如需相關資訊，請參閱主題 第 6 頁的『IBM SPSS Modeler Text Analytics 節點』。

註：當您在主從架構模式下工作，且「檔案檢視器」節點是串流的一部分時，必須將文件集合儲存在伺服器上的 Web 伺服器目錄中。因為「文字採礦」輸出節點會產生 Web 伺服器目錄中儲存的文件清單，所以 Web 伺服器的安全設定會管理這些文件的許可權。

檔案檢視器節點設定

您可以為「表檔案檢視器」節點指定下列設定。

文件欄位。 從資料中選取包含要顯示之文件的完整名稱和路徑的欄位。

所產生 HTML 頁面的標題。 建立要顯示在包含文件清單之頁面頂端的標題。

使用檔案檢視器節點

下列範例顯示如何使用「檔案檢視器」節點。

範例：「檔案清單」節點及「檔案檢視器」節點

圖 *19.* 說明如何使用檔案檢視器節點的串流

1. **檔案清單節點（「設定」標籤）。** 先新增此節點以指定文件所在的位置。

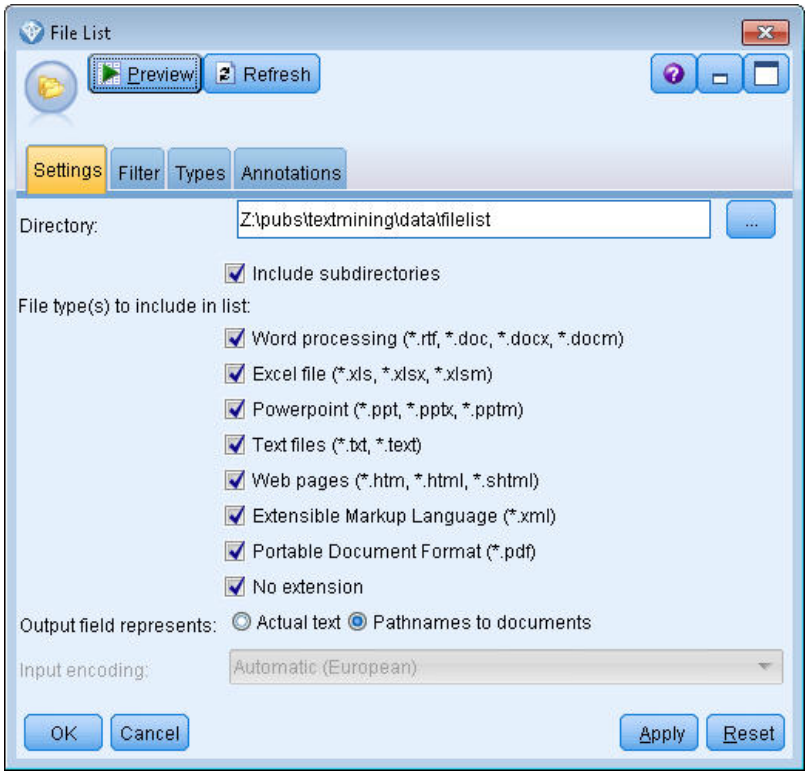


圖 *20.* 檔案清單節點對話框：設定標籤

2. **檔案檢視器節點（「設定」標籤）。** 接著附加「檔案檢視器」節點以產生文件的 HTML 清單。

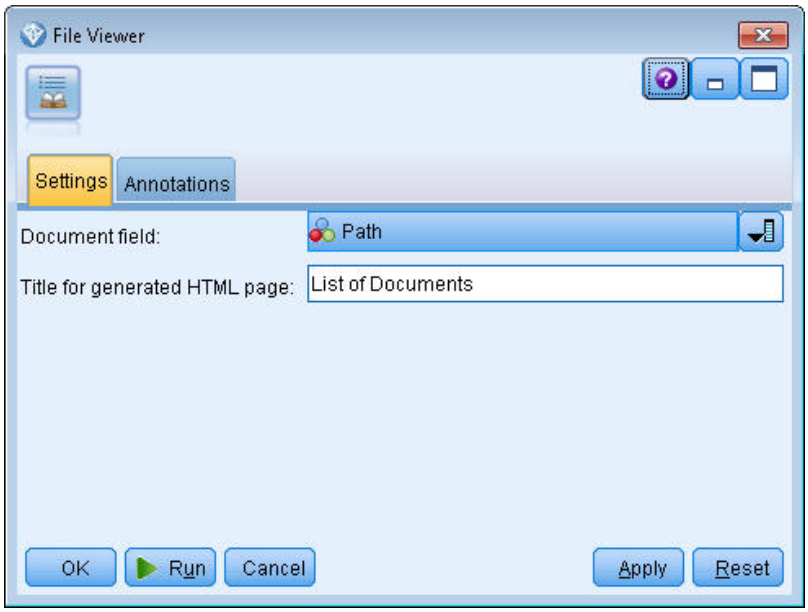


圖 *21.* 檔案檢視器節點對話框：設定標籤

3. **檔案檢視器輸出對話框。** 然後執行串流以在新視窗中輸出文件清單。

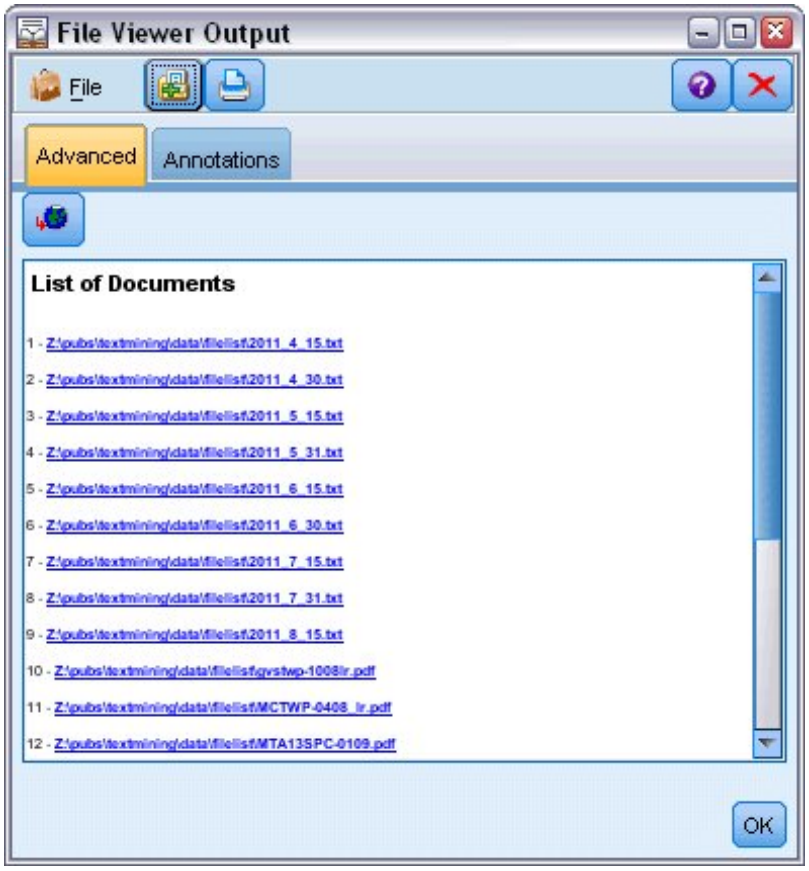


圖 22. 檔案檢視器輸出

4. 若要查看文件，可以按一下工具列按鈕，該按鈕顯示帶有紅色箭頭的小球。這會在瀏覽器中開啟文件超鏈結清單。

第 6 章 用於 Scripting 的節點內容

IBM SPSS Modeler 具有 Scripting 語言，可讓您從指令行執行串流。在這裡，您可以瞭解特定於 IBM SPSS Modeler Text Analytics 所交付之每個節點的節點內容。如需 IBM SPSS Modeler 所交付之標準節點集的相關資訊，請參閱 Scripting and Automation Guide。

檔案清單節點：filelistnode

您可以在下表中使用內容進行 scripting。稱為 `filelistnode`。

表 *7.* 檔案清單節點 *scripting* 內容

Scripting 內容	資料類型
`path`	字串
`recurse`	旗標
`word_processing`	旗標
`excel_file`	旗標
`powerpoint_file`	旗標
`text_file`	旗標
`web_page`	旗標
`xml_file`	旗標
`pdf_file`	旗標
`no_extension`	旗標

附註：'Create list' 參數不再可用，並且包含該選項的任何 Script 都將自動轉換為 'Files' 輸出。

Web 資訊來源節點：webfeednode

您可以使用下表中的內容進行 Scripting。節點本身稱為 `webfeednode`。

表 *8. Web* 資訊來源節點 *Scripting* 內容

Scripting 內容	資料類型	內容說明
`URL`	*string1 string2 ...stringn*	每一個 URL 都在清單結構中指定。以"\n"區隔的 URL 清單
`recent_entries`	旗標	
`limit_entries`	整數	每個 URL 要讀取的最新項目數。
`use_previous`	旗標	儲存及重複使用 Web 資訊來源快取。
`use_previous_label`	*string*	已儲存的 Web 快取的名稱。
`start_record`	*string*	非 RSS 起始標籤。
`url` *n* `.title`	*string*	針對清單中的每一個 URL，您也必須在這裡定義一個 URL。第一個 URL 將會是 `url1.title`，其中的數字符合它在 URL 清單中的位置。這是包含內容標題的起始標籤。
`url` *n* `.short_description`	*string*	對於 `url` *n* `.title` 相同。
`url` *n* `.description`	*string*	對於 `url` *n* `.title` 相同。

表 *8. Web* 資訊來源節點 *Scripting* 內容 *(*繼續*)*

Scripting 內容	資料類型	內容說明
url *n* .authors	*string*	對於 url *n* .title 相同。
url *n* .contributors	*string*	對於 url *n* .title 相同。
url *n* .published_date	*string*	對於 url *n* .title 相同。
url *n* .modified_date	*string*	對於 url *n* .title 相同。
html_alg	無 HTMLCleaner	內容過濾方法。
discard_lines	旗標	捨棄短行。與 min_words 搭配使用
min_words	整數	單字數目下限。
discard_words	旗標	捨棄短行。與 min_avg_len 搭配使用
min_avg_len	整數	
discard_scw	旗標	捨棄含有許多單一字元單字的行。與 max_scw 搭配使用
max_scw	整數	一行中的單一字元單字的比例 0-100 百分比上限
discard_tags	旗標	捨棄包含特定標籤的行。
標籤	*string*	特殊字元必須以反斜線 \ 跳出。
discard_spec_words	旗標	捨棄包含特定字串的行
單字	*string*	特殊字元必須以反斜線 \ 跳出。

語言節點：languageidentifier

您可以在下表中使用內容進行 scripting。節點本身稱為 languageidentifier。

表 *9.* 語言節點 *scripting* 內容

Scripting 內容	資料類型	內容說明
text	欄位	
language_field_name	字串	作為輸出產生的欄位名稱。
unidentified_language_value	Undefined Supported Custom	無法識別語言時要使用的預設值。
unidentified_language_supported	en de es fr it ja nl pt	ISO 代碼。僅當 unidentified_language_value 是 Supported 時可用。
unidentified_language_custom	字串	僅當 unidentified_language_value 是 Custom 時可用。

文字挖掘節點：TextMiningWorkbench

您可以使用下列參數，透過 Scripting 定義或更新節點。節點本身稱為 TextMiningWorkbench。

重要：可以透過 Scripting 指定不同的資源範本。如果您認為需要範本，則必須在節點對話框中進行選取。

表 *10.* 文字採礦建模節點 *Scripting* 內容

Scripting 內容	**資料類型**	**內容說明**
文字	欄位	
方法	ReadText ReadPath	
docType	整數	在可能的值 (0,1,2) 中，0 = 全文，1 = 結構化文字，2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	請注意，應該要用引號括住具有特殊字元的值（如 "UTF-8"）以避免與數學運算子混淆。
個體	整數	在可能的值 (0,1) 中，0 = 段落，1 = 文件
para_min	整數	
para_max	整數	
mtag	*string*	包含所有 mtag 設定（從 XML 檔案的「設定」對話框）
mclef	*string*	包含所有 mclef 設定（從「結構化文字」檔案的「設定」對話框）
partition	欄位	
custom_field	旗標	指出是否將指定分割區欄位。
use_model_name	旗標	
model_name	字串	
use_partitioned_data	旗標	如果定義分割區欄位，則僅將訓練資料用於模型建置。
model_output_type	互動式 Model	Interactive 產生種類模型。Model 產生概念模型。
use_interactive_info	旗標	僅適用於在工作台階段作業中以互動方式建置。
reuse_extraction_results	旗標	僅適用於在工作台階段作業中以互動方式建置。
interactive_view	Categories TLA Clusters	僅適用於在工作台階段作業中以互動方式建置。
extract_top	整數	當 model_type = Concept 時，使用此參數
use_check_top	旗標	
check_top	整數	

表 *10.* 文字採礦建模節點 *Scripting* 內容 *(*繼續*)*

Scripting 內容	資料類型	內容說明
use_uncheck_top	旗標	
uncheck_top	整數	
語言	de en es fr it ja nl pt	
frequency_limit	整數	14.0 中已淘汰。
concept_count_limit	整數	將擷取限制為廣域頻率至少為此值的概念。
fix_punctuation	旗標	
fix_spelling	旗標	
spelling_limit	整數	
extract_uniterm	旗標	
extract_nonlinguistic	旗標	
upper_case	旗標	
group_names	旗標	
排列	整數	非函數單字排列上限（預設值是 3）。

文字採礦模型塊：TMWBModelApplier

您可以使用下表中的內容來進行 Scripting。模型塊自身稱為 TMWBModelApplier。

表 *11.* 文字採礦模型塊內容

Scripting 內容	資料類型	內容說明
scoring_mode	欄位 記錄	
field_values	旗標 計數	此選項在「種類」模型塊中不可用。對於旗標，請設定為 TRUE 或 FALSE
true_value	字串	對於 Flags，將值定義為 true。
false_value	字串	對於 Flags，將值定義為 false。
extension_concept	字串	指定欄位名稱的副檔名。欄位名稱是使用概念名稱加上此副檔名而產生。使用 add_as 值來指定放置此副檔名的位置。
extension_category	字串	欄位名稱副檔名。您可以選擇指定欄位名稱的副檔名字首/字尾，也可以選擇使用種類代碼。欄位名稱是使用種類名稱加上此副檔名而產生。使用 add_as 值來指定放置此副檔名的位置。
add_as	字尾 字首	
fix_punctuation	旗標	

表 *11.* 文字採礦模型塊內容 *(*繼續*)*

Scripting 內容	資料類型	內容說明
excluded_subcategories_descriptors	RollUpToParent 忽略	僅用於種類模型。如果已取消選取子種類。此選項可讓您指定將如何處理屬於未選取為用於評分之子種類的描述子。有兩個選項。 • 忽略。「從評分中完全排除其描述子」選項將會導致在評分期間忽略及不使用沒有勾號（已取消選取）之子種類的描述子。 • RollUpToParent。「將描述子與母項種類中的描述子聚合」選項會導致將沒有勾號（已取消選取）之子種類的描述子，用作母項種類（此子種類之上的種類）的描述子。如果子種類有數個層次且已取消選取，則會在第一個可用的母項種類下向上捲動描述子
check_model	旗標	已在第 14 版中淘汰
text	欄位	
method	ReadText ReadPath	
docType	整數	可能的值 (0,1,2)，其中 0 = Full Text、1 = Structured Text 及 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	請注意，應該要用引號括住具有特殊字元的值（如 "UTF-8"）以避免與數學運算子混淆。
語言	de en es fr it ja nl pt	

文字鏈結分析節點：textlinkanalysis

您可以使用下表中的參數，透過 Scripting 來定義或更新節點。節點自身稱為 textlinkanalysis。

重要：不能透過 Scripting 來指定資源範本。若要選取範本，您必須從節點對話框內選取。

表 *12.* 文字鏈結分析 *(TLA)* 節點 *Scripting* 內容

Scripting 內容	資料類型	內容說明
id_field	欄位	
文字	欄位	

表 12. 文字鏈結分析 (TLA) 節點 Scripting 內容 (繼續)

Scripting 內容	資料類型	內容說明
方法	ReadText ReadPath	
docType	整數	在可能的值 (0,1,2) 中，0 = 全文，1 = 結構化文字，2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	請注意，應該要用引號括住具有特殊字元的值（如 "UTF-8"）以避免與數學運算子混淆。
個體	整數	在可能的值 (0,1) 中，0 = 段落，1 = 文件
para_min	整數	
para_max	整數	
mtag	*string*	包含所有 mtag 設定（從 XML 檔案的「設定」對話框）
mclef	*string*	包含所有 mclef 設定（從「結構化文字」檔案的「設定」對話框）
語言	de en es fr it ja nl pt	
concept_count_limit	整數	將擷取限制為廣域頻率至少為此值的概念。
fix_punctuation	旗標	
fix_spelling	旗標	
spelling_limit	整數	
extract_uniterm	旗標	
extract_nonlinguistic	旗標	
upper_case	旗標	
group_names	旗標	
排列	整數	非函數單字排列上限（預設值是 3）。

第 7 章 互動式工作台模式

從文字採礦建模節點中，您可以選擇在串流執行期間啟動互動式工作台階段作業。在此工作台中，您可以從文字資料中擷取主要概念、建置種類、探索文字鏈結分析型樣及叢集，並產生種類模型。在本章中，我們會從高階視景討論工作台介面以及使用的主要元素，包括：

- **擷取結果。**在執行擷取之後，這些是從文字資料中識別和擷取的關鍵字及詞組，也稱為概念。這些概念會分到類型群組中。使用這些概念和類型，您可以探索資料以及建立種類。這些會在**種類和概念**視圖中管理。
- **種類。** 使用描述子（例如擷取結果、型樣及規則）作為定義，您可以根據是否包含種類定義的一部分手動或自動建立一組為其指派文件或記錄的種類。這些會在**種類和概念**視圖中管理。
- **叢集。** 叢集是一種概念分組，這些概念之間的鏈結已經探索到，可指出它們之間的關係。概念分組使用的是複式演算法，除了其他因素以外，還會使用兩個概念一起出現的頻率與它們分別出現的頻率之比。這些會在**叢集**視圖中管理。您還可以將組成叢集的概念新增至種類。
- **文字鏈結分析型樣。** 如果您在語言資源中有「文字鏈結分析 (TLA)」型樣規則或正在使用已經有部分 TLA 規則的資源範本，則可以從文字資料中擷取型樣。這些型樣可以協助您發現資料中概念之間的有趣關係。您還可以將這些型樣作為描述子在種類中使用。這些會在**文字鏈結分析**視圖中管理。
- **語言資源。**擷取程序依賴一組參數及語言定義來控管如何擷取和處理文字。這些會在**資源編輯器**視圖中以範本和程式庫的形式管理。

潛在的互動式工作台問題

- 多個互動式工作台階段作業可能會導致行為遲緩。啟動互動式工作台階段作業時，SPSS Modeler Text Analytics 和 SPSS Modeler 共用一般 Java 執行時期引擎。視您在 SPSS Modeler 階段作業期間呼叫的互動式工作台階段作業數而定，系統記憶體可能會導致應用程式變得遲緩，即使開啟和關閉相同的階段作業也是如此。如果您使用的是大型資料或機器的 RAM 設定小於建議的 4GB，則此效果可能特別明顯。如果您發現機器回應很慢，建議您儲存所有工作，關閉 SPSS Modeler，然後重新啟動該應用程式。在記憶體小於建議大小的機器上執行 SPSS Modeler Text Analytics ，特別是在使用的資料集較大或長時間工作時，可能會導致 Java 記憶體不足並關閉。如果您使用的是大型資料，強烈建議您升級至建議的記憶體設定或更大（或使用 SPSS Modeler Text Analytics Server）。
- 在執行多個 SPSS Modeler Text Analytics 互動式工作台階段作業而不重新啟動應用程式之後，SPSS Modeler Client 可能會記憶體不足。在狀態行中監視記憶體用量，如果不足，請關閉並重新開啟 SPSS Modeler Client。

種類和概念視圖

應用程式介面由數個視圖組成。「種類和概念」視圖是您可以在其中建立和探索種類以及探索和調整擷取結果的視窗。種類是指透過評分程序為其指派文件及記錄的一組密切相關的構想及型樣。而概念是指可用作種類建置區塊（稱為描述子）的最基本層次的擷取結果。

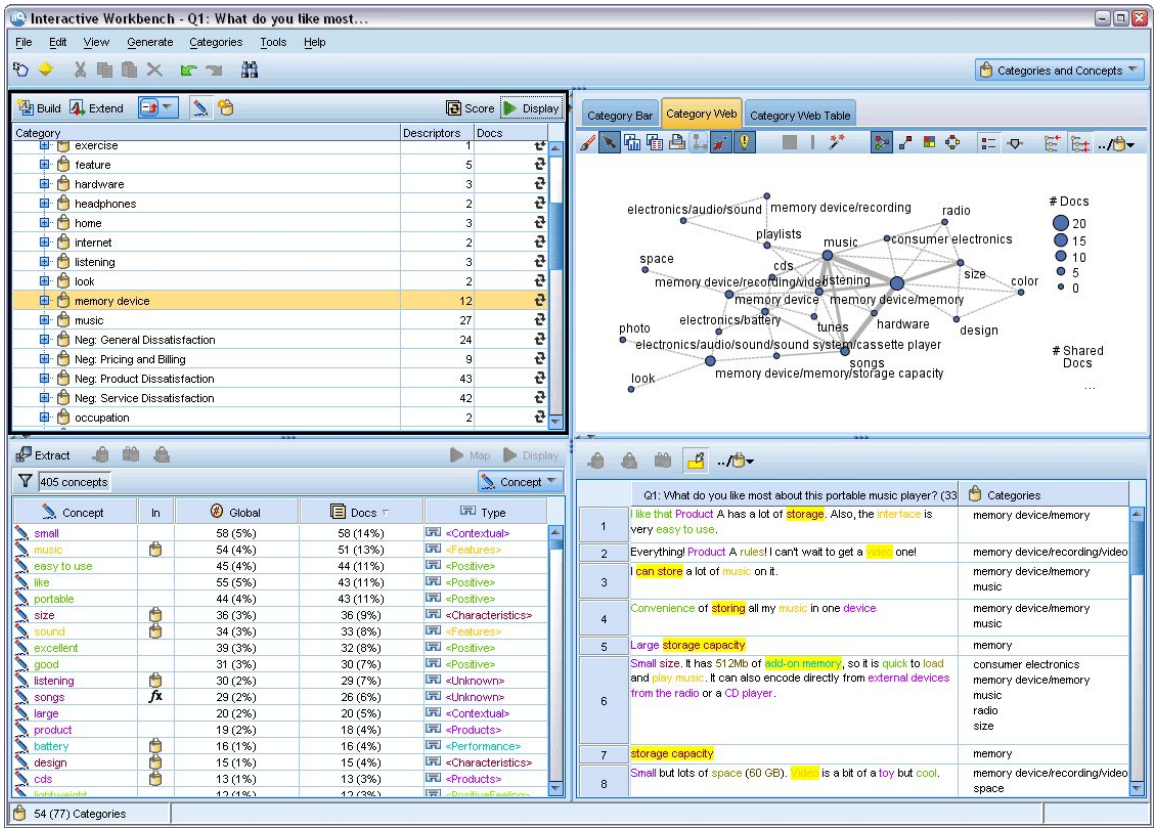


圖 *23.* 種類和概念視圖

「種類和概念」視圖組織為四個窗格，每一個都可以透過從「檢視」功能表中選取其名稱來隱藏或顯示。如需相關資訊，請參閱主題 第 81 頁的第 9 章, 『分類文字資料』 。

種類窗格

此區域位於左上角，會呈現您可以在其中管理任何建置種類的表格。從文字資料中擷取概念和類型之後，您可以透過使用語意網路及概念併入等技術或透過手動建立它們來開始建置種類。如果您按兩下種類名稱，則「種類定義」對話框會開啟並顯示組成其定義的所有描述子，例如，概念、類型及規則。如需相關資訊，請參閱主題 第 81 頁的第 9 章, 『分類文字資料』 。並非所有自動技術都適用於所有語言。

當您在窗格中選取一列時，您可以顯示「資料」及「視覺化」窗格中對應文件/記錄或描述子的相關資訊。

擷取結果窗格

此區域位於左下角，會呈現擷取結果。當您執行擷取時，擷取引擎會通讀文字資料，識別相關概念，並將類型指派給每一個概念。概念是從文字資料中擷取的單字或詞組。類型是以類型字典形式儲存的概念語意分組。擷取完成時，概念和類型會在「擷取結果」窗格中與顏色編碼一起出現。如需相關資訊，請參閱主題 第 69 頁的『擷取結果：概念和類型』 。

您可以透過將滑鼠移至概念名稱上方，將一組基礎術語用於概念。這樣做將顯示一個工具提示，顯示概念名稱，以及根據該概念分組的最多數行術語。這些基礎術語包括在語言資源中定義的同義字（而無論是否在文字中找到它們），以及任何擷取的複數/單數術語、已排列術語、模糊分組中的術語等。您可以透過用滑鼠右鍵按一下概念名稱，並選擇環境定義功能表選項，以複製這些術語或查看完整的基礎術語集。

文字採礦是一個疊代過程，在這個過程中，根據文字資料的環境定義檢閱擷取結果，進行細部調整來產生新結果，然後重新評估。可以透過修改語言資源精簡擷取結果。可以直接從「擷取結果」或「資料」窗格中，但也可以直接在「資源編輯器」視圖中執行部分細部調整。如需相關資訊，請參閱主題 第 63 頁的『資源編輯器視圖』。

註：如果可見窗格能夠容納更多的結果，則您可以使用窗格底部的控制項在結果中前後移動，或輸入要跳至的頁碼。

視覺化窗格

此區域位於右上角，會呈現文件/記錄分類中的多個共同性視景。每一個圖形或圖表提供的資訊非常相似，但會以不同的方式或不同的詳細程度來呈現。這些圖表及圖形可用來分析分類結果，並協助對種類進行細部調整或產生報告。例如，在某個圖形中您可能會發現太過相似（例如，它們共用 75% 以上的記錄）或差異太大的種類。圖形或圖表中的內容對應於其他窗格中的選取內容。如需相關資訊，請參閱主題第 135 頁的『種類圖形與圖表』。

資料窗格

「資料」窗格位於右下角。此窗格呈現的表格包含對應於視圖中另一區域中選取內容的文件或記錄。視選取的內容而定，只有對應的文字會在「資料」窗格中出現。選取後，按一下**顯示**按鈕可將對應文字移入「資料」窗格。

如果您在另一個窗格中進行選取，則對應文件或記錄會顯示以顏色強調顯示的概念，來協助您在文字中輕鬆識別它們。您還可以將滑鼠移至顏色編碼項目上方來顯示工具提示，顯示在其下面擷取的概念的名稱及為其指派的類型的名稱。如需相關資訊，請參閱主題 第 88 頁的『資料窗格』。

在種類和概念視圖中搜尋並尋找

在某些情況下，您可能需要在特定章節中快速找到資訊。使用「尋找」工具列，您可以輸入想要搜尋的字串並定義其他搜尋準則，例如，區分大小寫或搜尋方向。然後您可以選擇想要在其中搜尋的窗格。

使用尋找功能

1. 在「種類和概念」視圖中，從功能表中選擇**編輯 > 尋找**。「尋找」工具列在「種類」窗格及「視覺化」窗格上方出現。
2. 在文字框中輸入您想要搜尋的單字字串。您可以使用工具列按鈕來控制區分大小寫、部分相符及搜尋方向。
3. 在工具列中，按一下您想要在其中搜尋的窗格的名稱。如果找到相符項，則會在視窗中強調顯示該文字。
4. 若要尋找下一相符項，請再按一下窗格的名稱。

叢集視圖

在「叢集」視圖中，您可以建置並探索在文字資料中找到的叢集結果。叢集是由叢集演算法根據概念出現的頻率及它們一起出現的頻率產生的概念分組。叢集的目標是將一起共現的概念分組，而種類的目標是根據文件或記錄包含的文字與每一個種類的描述子（概念、規則、型樣）的相符程度將它們分組。

叢集內的概念一起出現的頻率越高，再加上它們與其他概念出現的頻率越低，叢集越擅長識別有趣的概念關係。當兩個概念都在相同的文件或記錄中出現（或其中一個同義字或術語出現）時，它們共現。如需相關資訊，請參閱主題 第 119 頁的第 10 章, 『分析叢集』。

您可以建置叢集並在一組圖表及圖形中探索它們，這可協助您發現概念之間的關係，否則尋找它們會非常耗費時間。雖然您無法將整個叢集新增至種類，但可以透過「叢集定義」對話框將叢集中的概念新增至種類。如需相關資訊，請參閱主題 第 122 頁的『叢集定義』。

您可以對叢集的設定進行變更來影響結果。如需相關資訊，請參閱主題 第 120 頁的『建置叢集』。

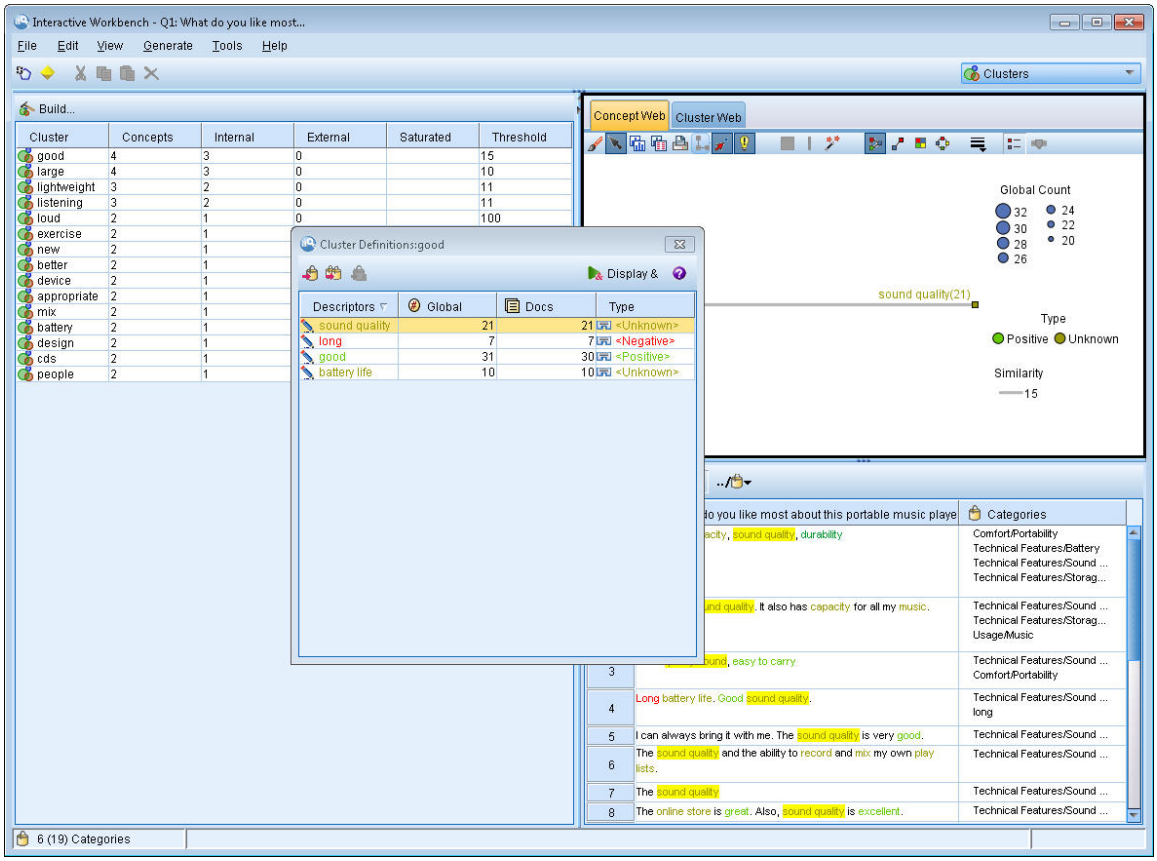


圖 *24.* 叢集視圖

「叢集」視圖組織為三個窗格，每一個都可以透過從「檢視」功能表中選取其名稱來隱藏或顯示。一般只有「叢集」窗格及「視覺化」窗格可見。

叢集窗格

此窗格位於左側，會呈現在文字資料中探索到的叢集。您可以透過按一下**建置**按鈕來建立叢集結果。叢集由叢集演算法組成，該演算法會嘗試識別經常一起出現的概念。

每當發生新的擷取，都會清除叢集結果，您必須重建叢集來取得最新結果。建置叢集時，您可以變更部分設定，例如要建立的叢集數上限、它可包含的概念數上限或它可具有的包含外部概念的鏈結數上限。如需相關資訊，請參閱主題 第 122 頁的『探索叢集』。

視覺化窗格

此窗格位於右上角，會提供兩個叢集視景：概念 Web 圖形及叢集 Web 圖形。如果不可見，則可以從「檢視」功能表（**檢視 > 視覺化**）中存取此窗格。視在叢集窗格中選取的內容而定，您可以檢視叢集之間或叢集內的對應互動。結果會以多種格式呈現：

- **概念 Web**。Web 圖形，其中顯示所選取叢集內的所有概念以及該叢集外部的鏈結概念。

- **叢集 Web**。Web 圖形，其中顯示從所選取叢集到其他叢集的鏈結以及其他那些叢集之間的任何鏈結。

註：為了顯示叢集 Web 圖形，您必須已使用外部鏈結建置了叢集。外部鏈結是不同叢集中概念配對之間的鏈結（一個叢集內的概念和另一個叢集中的概念）。如需相關資訊，請參閱主題 第 136 頁的『叢集圖形』。

資料窗格

「資料」窗格位於右下角，依預設已隱藏。您無法從「叢集」窗格顯示任何「資料」窗格結果，因為這些叢集跨越多個文件/記錄，使資料結果非常無趣。但您可以查看對應於「叢集定義」對話框內選取內容的資料。視該對話框中的選取內容而定，只有對應的文字會在「資料」窗格中出現。選取後，按一下**顯示 &** 按鈕可將一起包含所有概念的文件或記錄移入「資料」窗格。

對應文件或記錄會顯示以顏色強調顯示的概念，來協助您在文字中輕鬆識別它們。您還可以將滑鼠移至顏色編碼項目上方來顯示在其下面擷取的概念及為其指派的類型。「資料」窗格可以包含多個直欄，但一律顯示文字欄位直欄。它會攜帶擷取期間使用的文字欄位的名稱，如果文字資料在多個不同的檔案中，則會攜帶文件名稱。其他直欄是可用的。如需相關資訊，請參閱主題 第 88 頁的『資料窗格』。

文字鏈結分析視圖

在「文字鏈結分析」視圖中，您可以建置並探索在文字資料中找到的文字鏈結分析型樣。文字鏈結分析 (TLA) 是型樣比對技術，可讓您定義 TLA 規則並將其與實際擷取的概念及在文字中找到的關係進行比較。

嘗試探索特定主題的相關概念或意見之間的關係時，型樣極其有用部分範例包括想要從意見調查中擷取產品的相關意見、從醫學研究論文內擷取基因組關係，或從情報資料中擷取人員或地點之間的關係。

擷取部分 TLA 型樣之後，您可以在「資料」或「視覺化」窗格中探索它們，甚至可將它們新增至「種類和概念」視圖中的種類。必須在您使用的資源範本或程式庫中定義部分 TLA 規則，以便擷取 TLA 結果。如需相關資訊，請參閱主題 第 187 頁的第 18 章, 『關於文字鏈結規則』。

如果您選擇擷取 TLA 型樣結果，則會在此視圖中呈現結果。如果您未選擇這麼做，則不得不使用**擷取**按鈕並選擇選項來啟用型樣擷取。

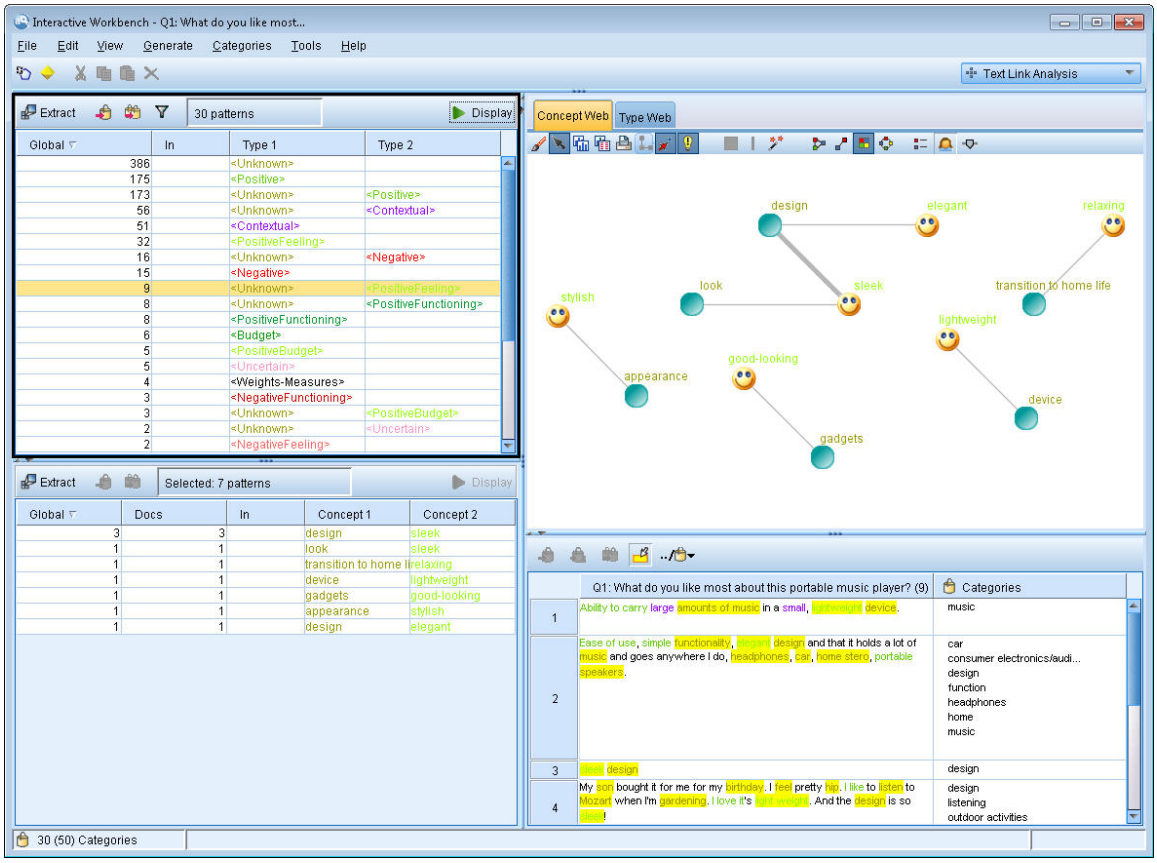


圖 25. 文字鏈結分析視圖

「文字鏈結分析」視圖組織為四個窗格，每一個都可以透過從「檢視」功能表中選取其名稱來隱藏或顯示。如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。

類型型樣和概念型樣窗格

「類型型樣」和「概念型樣」窗格位於左側，是兩個交互連接的窗格，您可以在其中探索及選取 TLA 型樣結果。型樣由一系列類型（最多六種）或概念（最多六個）組成。就像在語言資源中定義的一樣，TLA 型樣規則會指定型樣結果的複雜性。如需相關資訊，請參閱主題 第 187 頁的第 18 章, 『關於文字鏈結規則』。

型樣結果先在類型層次分組，然後分成概念型樣。基於此原因，有兩個不同的結果窗格：類型型樣（左上方）和概念型樣（左下方）。

- **類型型樣。**「類型型樣」窗格會呈現擷取的型樣，由兩個或多個符合 TLA 型樣規則的相關類型組成。類型型樣顯示為 `<Organization> + <Location> + <Positive>`，可提供特定位置中組織的相關正面意見。
- **概念型樣。**「概念型樣」窗格會針對在其上的「類型型樣」窗格中目前選取的所有類型型樣，在概念層次呈現擷取的型樣。概念型樣接在建築物之後，例如，`hotel + paris + wonderful`。

就像「種類和概念」視圖中的擷取結果一樣，您可以在這裡檢閱結果。如果您查看希望對組成這些型樣的類型及概念進行的精簡，您可在「種類和概念」視圖中的「擷取結果」窗格中進行，或直接在「資源」編輯器中進行，然後重新擷取型樣。

視覺化窗格

此窗格位於「文字鏈結分析」視圖的右上角，將所選取型樣的 Web 圖形呈現為類型型樣或概念型樣。如果不可見，則可以從「檢視」功能表（**檢視** > **視覺化**）中存取此窗格。視在其他窗格中選取的內容而定，您可以檢視文件/記錄與型樣之間的對應互動。

結果會以多種格式呈現：

- **概念圖形**。此圖形會呈現所選取型樣中的所有概念。概念圖形中的線寬及節點大小（如果類型圖示不顯示）會顯示所選取表格中的廣域出現次數。
- **類型圖形**。此圖形會呈現所選取型樣中的所有類型。該圖形中的線寬及節點大小（如果類型圖示不顯示）會顯示所選取表格中的廣域出現次數。節點由類型顏色或圖示代表。

如需相關資訊，請參閱主題第 138 頁的『文字鏈結分析圖形』。

資料窗格

「資料」窗格位於右下角。此窗格呈現的表格包含對應於視圖中另一區域中選取內容的文件或記錄。視選取的內容而定，只有對應的文字會在「資料」窗格中出現。選取後，按一下**顯示**按鈕可將對應文字移入「資料」窗格。

如果您在另一個窗格中進行選取，則對應文件或記錄會顯示以顏色強調顯示的概念，來協助您在文字中輕鬆識別它們。您還可以將滑鼠移至顏色編碼項目上方來顯示工具提示，顯示在其下面擷取的概念的名稱及為其指派的類型的名稱。如需相關資訊，請參閱主題 第 88 頁的『資料窗格』。

資源編輯器視圖

IBM SPSS Modeler Text Analytics 使用強大的擷取引擎，快速且準確地從文字資料擷取關鍵概念。此引擎主要依賴於語言資源，以決定應該分析及解譯多大量的未結構化文字資料。

您可以在 資源編輯器 視圖中檢視及細部調整用於擷取概念、根據類型分組、在文字資料中探索型樣等的語言資源。IBM SPSS Modeler Text Analytics 提供數個預先配置的資源範本。此外，在部分語音中，您也可以在文字分析套件中使用資源。如需相關資訊，請參閱主題 第 113 頁的『使用文字分析套件』。

由於這些資源可能不會一律完美適合資料的環境定義，因此您可以在資源編輯器中為特定環境定義或網域建立、編輯及管理您自己的資源。請參閱第 155 頁的第 15 章, 『使用檔案庫』主題，以取得更多資訊。

若要簡化細部調整語言資源的處理程序，您可以透過「擷取結果」與「資料」窗格的環境定義功能表，直接從種類與概念視圖中執行共用定義檔作業。如需更多資訊，請參閱主題第 76 頁的『精簡擷取結果』。

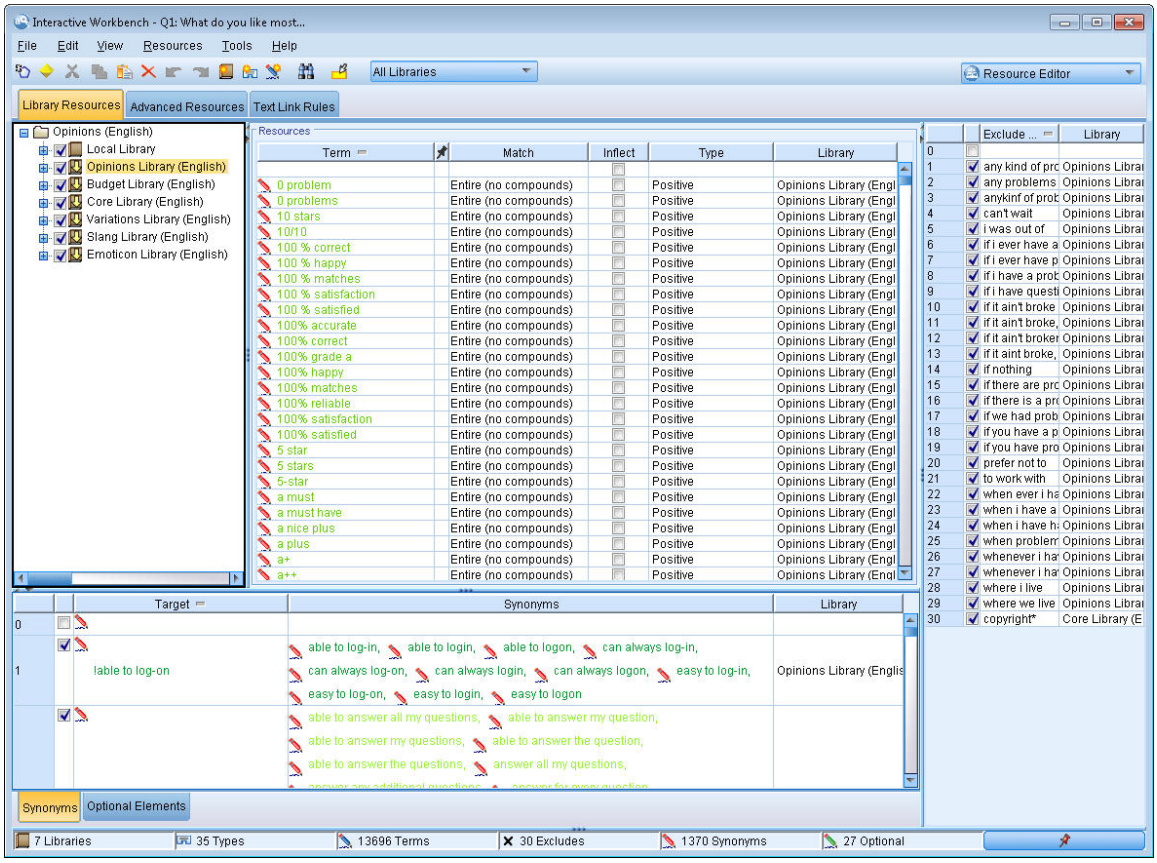


圖 *26.* 資源編輯器視圖

您在資源編輯器視圖中執行的作業會解決有關語言資源的管理及細部調整。這些資源以範本及檔案庫形式儲存。資源編輯器 視圖組織為四個部分：「檔案庫樹狀結構」窗格、「類型定義檔」窗格、「替代定義檔」窗格及「排除定義檔」窗格。

註：如需相關資訊，請參閱主題 第 146 頁的『編輯器介面』。

設定選項

您可以在「選項」對話框中設定 IBM SPSS Modeler Text Analytics 的一般選項。此對話框包含下列標籤：

- **階段作業。**此標籤包含一般選項和定界字元。
- **顯示。** 此標籤包含介面中所使用顏色的選項。
- **音效。**此標籤包含音效提示的選項。

編輯選項

1. 從功能表中，選擇**工具 > 選項**。「選項」對話框即會開啟。
2. 選取包含您要變更的資訊的標籤。
3. 變更任何選項。
4. 按一下**確定**以儲存變更。

選項：階段作業標籤

在此標籤上，您可以定義某些基本設定。

資料窗格與種類圖形顯示。 這些選項會影響資料在「資料」窗格以及在「種類與概念」視圖中的「視覺化」窗格中的呈現方式。

- **「資料」窗格和「種類 Web」的顯示限制。** 這個選項設定要顯示或用來移入「資料」窗格或「種類與概念」視圖中的圖形和圖表的文件數目上限。
- **在顯示時間顯示文件/記錄的種類。** 如果選取此選項，則每當您按一下「顯示」時，就會將文件或記錄予以評分，以便可以在「資料」窗格中的「種類」直欄以及種類圖形中顯示它們所屬的任何種類。在某些情況，特別是較大型的資料集，您可能需要關閉這個選項，以便能資料和圖形能遠遠更快顯示。

從資料窗格新增至種類。 這些選項會影響從「資料」窗格新增文件和記錄時，什麼內容會新增至種類。

- **在「種類與概念」視圖中複製。** 從此視圖中的「資料」窗格新增文件或記錄將會複製在**僅概念**或是**概念與型樣**兩者之上。
- **在「文字鏈結分析」視圖中複製。** 從此視圖中的「資料」窗格新增文件或記錄將會複製在**僅型樣**或是**概念與型樣**兩者之上。

「資源編輯器」定界字元。 選取在「資源編輯器」視圖中輸入元素（如概念、同義字）和選用元素時，要用來作為定界字元的字元。

選項：顯示標籤

在此標籤上，您可以編輯影響應用程式整體外觀與操作方式的選項，以及用來識別元素的顏色。

附註：若要將產品的外觀與操作方式切換到典型外觀或舊版的外觀，請在 IBM SPSS Modeler 主視窗中的「工具」功能表中開啟「使用者選項」對話框。

自訂顏色。編輯畫面上出現的元素的顏色。針對表格中的每一個元素，您可以變更顏色。若要指定自訂顏色，請按一下您想要變更的元素右側的顏色區域，並從下拉顏色清單中選擇顏色。

- **非擷取的文字。**未擷取的文字資料在「資料」窗格可見。
- **強調顯示背景。**在窗格中選取元素或在「資料」窗格中選取文字時的文字選取背景顏色。
- **需要擷取的背景。**「擷取結果」、「型樣」及「叢集」窗格的背景顏色，指出已對程式庫進行了變更並需要擷取。
- **種類意見背景。**在作業之後出現的種類背景顏色。
- **預設類型。**在「資料」窗格及「擷取結果」窗格中出現的類型和概念的預設顏色。此顏色會套用至您在「資源編輯器」中建立的任何自訂類型。您可以透過在資源編輯器中編輯自訂類型字典的內容，置換這些類型字典的這個預設顏色。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。
- **帶狀線表 1.** 兩個顏色的第一個，在「編輯強制概念」對話框的表格中以交互方式使用兩個顏色以區分每一組線條。
- **帶狀線表 2.** 兩個顏色的第二個，在「編輯強制概念」對話框的表格中以交互方式使用兩個顏色以區分每一組線條。

附註：如果您按一下**重設為預設值**按鈕，則此對話框中的所有選項都會重設為第一次安裝此產品時它們具有的值。

選項：音效標籤

在此標籤上，您可以編輯影響音效的選項。在「音效事件」下面，您可以指定音效，用來在發生事件時通知您。可用的音效有很多。使用省略符號按鈕 (**...**) 來瀏覽並選取音效。用來為 IBM SPSS Modeler Text Analytics 建立音效的 *.wav* 檔案會儲存在安裝目錄的 *media* 子目錄中。如果不想播放音效，請選取**忽略所有音效**。依預設會忽略音效。

附註：如果您按一下**重設為預設值**按鈕，則此對話框中的所有選項都會重設為第一次安裝此產品時它們具有的值。

說明的 Microsoft Internet Explorer 設定

Microsoft Internet Explorer 設定

此應用程式中的大部分「說明」功能使用基於 Microsoft Internet Explorer 的技術。部分版本的 Internet Explorer（包括 Microsoft Windows XP Service Pack 2 提供的版本）依預設會封鎖其認為是本端電腦上 Internet Explorer 視窗中「主動式內容」的內容。此預設值可能會導致封鎖「說明」功能中的部分內容。若要查看所有「說明」內容，您可以變更 Internet Explorer 的預設行為。

1. 從 Internet Explorer 功能表中選擇：

   工具 > **網際網路選項...**
2. 按一下**進階**標籤。
3. 向下捲動至**安全性**區段。
4. 選取**允許主動式內容在我電腦上的檔案中執行**。

產生模型塊和建模節點

當您處於互動式階段作業中時，可能想要使用已完成的工作來產生：

- **文字採礦建模節點。**從互動式工作台階段作業中產生的建模節點是「文字採礦」節點，其設定及選項會反映在開啟的互動式階段作業中儲存的那些設定及選項。當您不再有原始「文字採礦」節點或想要製作新版本時，這可能非常有幫助。如需相關資訊，請參閱主題 第 17 頁的第 3 章, 『概念和種類的挖掘』。
- **種類模型塊。**從互動式工作台階段作業中產生的模型塊是種類模型塊。您必須在「種類和概念」視圖中至少有一個種類才能產生種類模型塊。如需相關資訊，請參閱主題 第 34 頁的『文字挖掘塊：種類模型』。

產生文字採礦建模節點

1. 從功能表中，選擇**產生** > **產生建模節點**。會使用工作台階段作業中目前的所有設定將「文字採礦」建模節點新增至工作畫布。會根據文字欄位命名該節點。

產生種類模型塊

1. 從功能表中，選擇**產生** > **產生模型**。會使用預設名稱直接在「模型」選用區上產生模型塊。

更新建模節點及儲存

當您在互動式階段作業中工作時，我們建議您經常更新建模節點來儲存變更。您還應該每次在互動式工作台階段作業中完成工作和想要儲存工作時更新建模節點。更新建模節點時，工作台階段作業內容會儲存回互動式工作台階段作業起源的「文字採礦」節點。這不會關閉輸出視窗。

重要事項！ 此更新不會儲存串流。若要儲存串流，請在更新建模節點之後在 IBM SPSS Modeler 主視窗中這麼做。

更新建模節點

1. 從功能表中，選擇**檔案** > **更新建模節點**。會使用建置和擷取設定以及您有的任何選項和種類更新建模節點。

關閉及結束階段作業

當您在階段作業中完成工作時，可以用三種不同的方式離開階段作業：

- **儲存**。這個選項可讓您先將工作儲存回起源的建模節點中以用於未來的階段作業，以及發佈任何檔案庫以供在其他階段作業中重複使用。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。在儲存之後，階段作業視窗即會關閉，並且會從IBM SPSS Modeler視窗中的「輸出」管理程式刪除階段作業。
- **結束**。這個選項會捨棄任何未儲存的工作、關閉階段作業視窗，以及從IBM SPSS Modeler視窗中的「輸出」管理程式刪除階段作業。如果要釋出記憶體，我們建議儲存任何重要工作，然後結束階段作業。
- **關閉**。這個選項不會儲存或捨棄任何工作。這個選項會關閉階段作業視窗，但是階段作業將會繼續執行。您可以在IBM SPSS Modeler視窗中的「輸出」管理程式中選取此階段作業來再度開啟階段作業視窗。

關閉工作台階段作業

1. 從功能表中，選擇**檔案 > 關閉**。

鍵盤協助工具

互動式工作台介面提供鍵盤快速鍵，使產品的功能更易於存取。在最基本層次上，您可以按 Alt 鍵加上適當的按鍵來啟動視窗功能表（例如，按 Alt+F 鍵可存取「檔案」功能表），或是按 Tab 鍵來捲動對話框控制項。本節將涵蓋用於替代導覽的鍵盤快速鍵。IBM SPSS Modeler 介面有其他的鍵盤快速鍵。

表 *13.* 一般鍵盤快速鍵

快速鍵	功能
Ctrl+1	在具有許多標籤的窗格中顯示第一個標籤。
Ctrl+2	在具有許多標籤的窗格中顯示第二個標籤。
Ctrl+A	選取具有焦點之窗格的所有元素。
Ctrl+C	將所選的文字複製到剪貼簿中。
Ctrl+E	在「種類與概念」和「文字鏈結分析」視圖中啟動擷取。
Ctrl+F	在 資源編輯器/範本編輯器 中顯示「尋找」工具列（如果尚未看到），並在那裡放置焦點。
Ctrl+I	在「種類與概念」視圖中，針對選取的種類啟動「種類定義」對話框。在「叢集」視圖中，針對選取的叢集啟動「叢集定義」對話框。
Ctrl+R	在 資源編輯器/範本編輯器 中開啟「新增詞彙」對話框。
Ctrl+T	在 資源編輯器/範本編輯器 中開啟「類型內容」對話框以建立新類型。
Ctrl+V	貼上剪貼簿內容。
Ctrl+X	從 資源編輯器/範本編輯器 剪下選取的項目。
Ctrl+Y	重做視圖中的最後一個動作。
Ctrl+Z	復原視圖中的最後一個動作。
F1	顯示「說明」，或是在對話框中時，顯示項目的上下文說明。
F2	在表格資料格中切換編輯模式。
F6	在作用中視圖中的主要窗格之間移動焦點。
F8	將焦點移至窗格分割線列以重新調整大小。
F10	展開主要「檔案」功能表。
上移鍵、下移鍵	選取了分割線列時，垂直調整窗格大小。
左移鍵、右移鍵	選取了分割線列時，水平調整窗格大小。
Home、End 鍵	選取了分割線列時，將窗格調整為最小或最大大小。
Tab 鍵	在視窗、窗格或對話框中的項目之間往前移動。
Shift+F10	顯示項目的快速功能表。
Shift+Tab 鍵	在視窗或對話框中的項目之間往回移動。

表 *13.* 一般鍵盤快速鍵 *(*繼續*)*

快速鍵	功能
Shift+箭頭	在處於編輯模式 (F2) 時，選取編輯欄位中的字元。
Ctrl+Tab 鍵	將焦點往前移至視窗中的下一個主要區域。
Shift+Ctrl+Tab 鍵	將焦點往回移至視窗中的前一個主要區域。

對話框的快速鍵

當您在使用對話框時，有數個快速鍵和螢幕閱讀器按鍵非常有用。在進入對話框後，您可能需要馬上按下 Tab 鍵，以將焦點置於第一個控制項上及起始螢幕閱讀器。下表中提供完整的特殊鍵盤和螢幕閱讀器快速鍵清單。

表 *14.* 對話框快速鍵

快速鍵	功能
Tab 鍵	在視窗或對話框中的項目之間往前移動。
Ctrl+Tab 鍵	在文字框中向前移動至下一個項目。
Shift+Tab 鍵	在視窗或對話框中的項目之間往回移動。
Shift+Ctrl+Tab 鍵	從文字框往回移動至前一個項目。
空白鍵	選取具有焦點的控制項或按鈕。
Esc 鍵	取消變更並關閉對話框。
Enter 鍵	驗證變更並關閉對話框（相等於「確定」按鈕）。如果您在文字框中，則必須先按下 Ctrl+Tab 鍵才能結束文字框。

第 8 章 擷取概念和類型

每當您執行啟動互動式工作台的串流時，便會對串流中的文字資料自動執行擷取。這種擷取的最終結果是一組概念、類型及型樣（在這種情況下，語言資源中存在 TLA 型樣）。您可以在「擷取結果」窗格中檢視並使用概念和類型。如需相關資訊，請參閱 第 4 頁的『擷取如何運作』。

如果您想要對擷取結果進行細部調整，則可以修改語言資源並重新擷取。如需相關資訊，請參閱 第 76 頁的『精簡擷取結果』。擷取程序依賴「擷取」對話框中的資源及任何參數來指定如何擷取和組織結果。您可以使用擷取結果來定義大部分（如果不是全部）的種類定義。

註：從 18.2 版開始，已改良擷取的概念結果（現在，它們類似於 IBM SPSS Text Analytics for Surveys 中擷取的概念結果）。

擷取結果：概念和類型

擷取程序期間，系統會掃描所有文字資料，識別、擷取相關概念並將它們指派給類型。擷取完成後，結果會顯示在位於「種類和概念」視圖左下角的「擷取結果」窗格中。首次啟動階段作業時，您在節點中選取的語言資源範本用來擷取及組織這些概念及類型。

註：如果可見窗格能夠容納更多的結果，則您可以使用窗格底部的控制項在結果中前後移動，或輸入要跳至的頁碼。

擷取的概念、類型與 TLA 型樣統一稱為**擷取結果**，並且它們用作您的種類的描述子或建置區塊。您也可以在種類規則中使用概念、類型及型樣。此外，自動技術使用概念及類型以建置種類。

文字挖掘是反覆運算處理程序，在該程序中，根據文字資料的環境定義檢閱擷取結果，並進行細部調整以產生新的結果，然後進行重新評估。擷取後，您應該檢閱結果，並透過修改語言資源進行您認為必要的變更。您可以直接從「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框部分細部調整資源。請參閱第 76 頁的『精簡擷取結果』主題，以取得更多資訊。您也可以直接在「資源編輯器」視圖中執行此動作。如需更多資訊，請參閱主題第 63 頁的『資源編輯器視圖』。

細部調整後，您可以重新擷取以查看新的結果。透過從開始細部調整擷取結果，您可以假設每次重新擷取時，您都將在種類定義中取得相同的結果，完美適合資料的環境定義。透過此方法，可以更準確且可重複的方式將文件/記錄指派給您的種類定義。

概念

擷取程序期間，系統會掃描及分析文字資料，從而識別文字中有興趣或相關的單字（例如 `election` 或 `peace`），以及單字片語（例如 `presidential election`、`election of the president` 或 `peace treaties`）。這些單字和片語統一稱作術語。使用語言類資源擷取相關的術語，然後將相似的術語一併分組在一個前導術語下叫作**概念**。

您可以透過將滑鼠移至概念名稱上方，將一組基礎術語用於概念。這樣做將顯示一個工具提示，顯示概念名稱，以及根據該概念分組的最多數行術語。這些基礎術語包括在語言資源中定義的同義字（而無論是否在文字中找到它們），以及任何擷取的複數/單數術語、已排列術語、模糊分組中的術語等。您可以透過用滑鼠右鍵按一下概念名稱，並選擇環境定義功能表選項，以複製這些術語或查看完整的基礎術語集。

依預設，概念以小寫形式顯示，並根據文件計數（文件直欄）按遞減順序排序。擷取概念後，會為它們指派一個類型，以協助將類似的概念分組在一起。它們根據此類型進行顏色編碼。顏色在資源編輯器內的類型內容中進行定義。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。

在種類定義中使用概念、類型或型樣時，圖示顯示在可排序的**中**直欄內。

類型

類型 是概念的語意分組。擷取概念後，會為它們指派一個類型，以協助將類似的概念分組在一起。IBM SPSS Modeler Text Analytics 提供數個內建類型，例如 <Location>、<Organization>、<Person>、<Positive>、<Negative> 等。例如，<Location> 類型分組地理關鍵字與工作區。此類型將指派給諸如 chicago、paris 及 tokyo 等概念。對於大部分語音，未在任何類型定義檔中找到但擷取自文字的概念會自動鍵入為 <Unknown>如需更多資訊，請參閱主題第 164 頁的『內建類型』。

當您選取「類型」視圖時，依預設，擷取的類型按廣域頻率以遞減順序顯示。您還可以看到類型按顏色編碼，以協助進行識別。顏色屬於類型內容。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。您還可以建立您自己的類型。

型樣

型樣也可以擷取自文字資料。然而，您必須具有一個檔案庫，包含資源編輯器中的部分「文字鏈結分析 (TLA)」型樣規則。您還必須使用選項**啟用文字鏈結分析型樣擷取**，在 IBM SPSS Modeler Text Analytics 節點設定或在「擷取」對話框中擷取這些型樣。如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。

擷取資料

需要擷取時，「擷取結果」窗格會變成黃色，並在此窗格的工具列下方顯示訊息**按下擷取按鈕以擷取概念**。

如果您尚無任何擷取結果，已對語言資源進行變更且需要更新擷取結果，或者已重新開啟您未儲存擷取結果的階段作業（**工具** > **選項**），則可能需要擷取。

註：如果您在利用**使用階段作業工作...** 選項快取擷取結果之後變更串流的來源節點，並且您想要取得更新的擷取結果，則在啟動互動式工作台階段作業之後將需要執行新的擷取。

執行擷取之後，會顯示進度指示器，以提供有關擷取狀態的意見。在此期間，擷取引擎會讀取所有文字資料，識別相關術語及型樣，擷取它們，並將它們指派給某個類型。然後，引擎會嘗試根據一個前導術語（稱為概念）分組同義字術語。處理程序完成後，產生的概念、類型及型樣會顯示在「擷取結果」窗格中。

擷取程序會產生一組概念及類型，以及「文字鏈結分析 (TLA)」型樣（如果已啟用）。您可以在「種類與概念」視圖的「擷取結果」窗格中檢視及使用這些概念及類型。如果您已擷取 TLA 型樣，則可以在「文字鏈結分析」視圖中看到這些型樣。

註：資料集大小與其完成擷取程序所花費的時間之間存在關係。您一律可以考量插入「樣本」節點上游，或者最佳化您機器的配置。

要擷取資料

1. 從功能表中，選擇**工具** > **擷取**。或者，按一下**擷取**工具列按鈕。
2. 如果您選擇一律顯示「擷取設定」對話框，則會顯示以便您可以進行任何變更。如需每一個設定的描述子，請進一步參閱此主題。

3. 按一下**擷取**以開始擷取程序。擷取開始後，將開啟進度對話框。擷取後，結果會顯示在「擷取結果」窗格中。依預設，概念以小寫形式顯示，並根據文件計數（Doc. 直欄）按遞減順序排序。

您可以使用工具列選項檢閱結果，以不同的方式排序結果，過濾結果，以及切換至不同的視圖（概念或類型）。您還可以透過使用語言資源來精簡擷取結果。如需更多資訊，請參閱主題第 76 頁的『精簡擷取結果』。

潛在擷取問題

多個「互動式工作台」階段作業可能導致緩慢行為。啟動互動式工作台階段作業時，SPSS Modeler Text Analytics 與 SPSS Modeler 共用一般 Java 執行時期引擎。根據您在 SPSS Modeler 階段作業期間呼叫的「互動式工作台」階段作業數目，即使開啟及關閉同一階段作業，系統記憶體也可能導致應用程式變得緩慢。當您使用大量資料，或者機器低於建議的 RAM 設定 4 GB 時，這個效果可能尤其顯著。如果您注意到機器回應緩慢，則建議您儲存所有工作，關閉 SPSS Modeler，並重新啟動應用程式。在低於建議的記憶體的機器上執行 SPSS Modeler Text Analytics ，特別是在使用大量資料集或延遲使用時，可能導致 Java 用完記憶體或關閉。如果您使用大量資料，強烈建議您升級至建議的記憶體大小或更大（或者使用 SPSS Modeler Text Analytics 伺服器）。

使用於英文、法文、德文、義大利文、葡萄牙文及西班牙文文字

「擷取設定」對話框包含部分基本擷取選項。

啟用文字鏈結分析型樣擷取。指定您想要從文字資料擷取 TLA 型樣。還假設您在「資源編輯器」的其中一個檔案庫中具有 TLA 型樣規則。此選項可能會顯著延長擷取時間。如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。

容納標點符號錯誤。此選項會在擷取期間暫時正規化包含標點符號錯誤的文字（例如，用法不當），以改進概念的擷取能力。當文字較短且品質較差時（例如，在開放式意見調查回應、電子郵件及 CRM 資料中時），或者當文字包含許多縮寫時，此選項極其有用。

容納單字字元長度下限 [n] 的拼字 此選項套用模糊分組技術，可協助將通常拼字錯誤的單字或根據一個概念拼字接近的單字分組在一起。模糊分組演算法暫時去掉所有母音（除了第一個），並從擷取的單字中去掉雙/三重輔音，然後比較它們以查看它們是否相同，以便將 `modeling` 與 `modelling` 分組在一起。然而，如果將每一個術語指派給不同的類型（排除 `<Unknown>` 類型），則將不會套用模糊分組技術。

您也可以在使用模糊分組之前，定義需要的根字元數目下限。術語中的根字元數目計算方式為所有字元總數減去形成字形變化字尾的字元，若為複合字術語，則再減去限定詞與介詞。例如，術語 `exercises` 將以 "exercise" 形式計為 8 個根字元，位於單字末尾的字母 *s* 是字形變化（複數形式）。類似地，`apple sauce` 計為 10 個根字元 ("apple sauce")，而 `manufacturing of cars` 計為 16 個根字元 ("manufacturing car")。此計數方法僅用於檢查是否應該套用模糊分組，但不會影響單字的相符程度。

註：如果您稍後發現某些單字未正確地分組，則可以透過在「進階資源」標籤的**模糊分組：異常狀況**區段中明確地宣告，從此技術中排除單字配對。如需相關資訊，請參閱主題 第 177 頁的『模糊分組』。

擷取單一術語 只要單字尚且不是複合字的一部分，或者如果它是名詞或無法辨識的語音的一部分，則此選項會擷取單字（單一屬於）。

擷取非語言實體 此選項擷取非語言實體，例如電話號碼、社會安全號碼、時間、日期、貨幣、數位、百分比、電子郵件位址及 HTTP 位址。您可以在「進階資源」標籤的**非語言實體：配置**區段中包括或排除某些類型的非語言實體。透過停用任何不需要的實體，擷取引擎不會浪費處理時間。如需相關資訊，請參閱主題第 181 頁的『配置』。

大寫演算法 只要術語的第一個字母為大寫形式，此選項就擷取不在內建目錄中的簡式及複合術語。此選項提供良好的方法以擷取最適當的名詞。

可能時將部分及全部人員名稱分組在一起 此選項將文字中看起來不同的名稱分組在一起。由於通常在文字開頭以名稱的完整形式對名稱以縮寫進行參照，之後僅使用縮寫版本，因此本功能有用。此選項嘗試將任何類型為 <Unknown> 的單一術語與鍵入為 <Person> 的任何複合術語的最後一個單字進行比對。例如，如果發現 *doe*，且最初鍵入為 <Unknown>，則擷取引擎會檢查以查看 <Person> 類型中是否有任何複合術語包括 *doe* 作為最後一個單字，例如 *john doe*。由於大部分名稱從不作為單一術語擷取，因此本選項不適用於第一個名稱。

非功能單字排列上限 此選項指定套用排列技術時可以呈現的非功能單字數目上限。此排列技術僅依照所包含的非功能單字，將彼此不同的類似片語分組在一起（例如，of 及 the），而不考量字形變化。例如，讓我們假設將此值設為最多兩個單字，並擷取 company officials 與 officials of the company。在此情況下，由於當忽略 of the 時，兩個術語被視為相同，因此兩個擷取的術語將在最終概念清單中分組在一起。

分組多術語時使用衍生 處理海量資料時，選取此選項以透過使用衍生規則分組多術語。

概念對映的索引選項 指定您想要在擷取時建置的對映索引，以便稍後可以快速繪製概念對映。若要編輯索引設定，請按一下**設定**。如需相關資訊，請參閱主題 第 75 頁的『建置概念地圖索引』。

一律在啟動擷取之前顯示此對話框 指定您是否想要在每一次擷取之前查看「擷取設定」對話框，您是否除非前往「工具」功能表否則永遠不想要查看，或者您是否想要在每一次擷取時被詢問是否想要編輯任何擷取設定。

過濾擷取結果

當您使用的資料集非常大時，擷取程序可能會產生上百萬的結果。對於許多使用者來說，這個數量會使有效檢閱結果的難度增加。因此，為了聚焦於那些最有趣的結果，您可以透過「擷取結果」窗格中提供的「過濾器」對話框過濾這些結果。

請記住，此「過濾器」對話框中的所有設定會一起使用，來過濾可用於種類的擷取結果。

依頻率過濾 您可以過濾以僅顯示具有某個廣域或文件頻率值的那些結果。

- **廣域頻率**是概念在整個文件或記錄集中出現的總次數，會在**廣域**直欄中顯示。
- **文件頻率**是概念在其中出現的文件或記錄總數，會在**文件數**直欄中顯示。

例如，如果概念 nato在 500 個記錄中出現了 800 次，我們會說，此概念的廣域頻率為 800，文件頻率為 500。

依據類型 您可以過濾以僅顯示屬於某些類型的那些結果。您可以選擇所有類型或僅選擇特定類型。

依據比對文字 您還可以過濾來僅顯示符合您在這裡所定義規則的那些結果。在**比對文字**欄位中輸入要比對的字元集，然後選取在其中套用比對的條件。

表 *15.* 比對文字條件

條件	說明
包含	如果字串在某個位置出現則符合文字。（預設選項）
開始於	僅在概念或類型以指定文字開頭時符合文字。
結尾是	僅在概念或類型以指定文字結尾時符合文字。
完全相符	整個字串必須符合概念或類型名稱。

在擷取結果窗格中顯示的結果

下面是一些範例，說明如何能夠根據過濾器在「擷取結果」窗格工具列中以英文顯示結果。

表 *16.* 過濾器意見範例

過濾器意見	說明
	該工具列會顯示結果數。因為沒有文字符合過濾器且未符合上限，所以不會顯示其他圖示。
	該工具列顯示結果限制為過濾器中指定的上限，在此情況下為 300。如果紫色圖示存在，這表示符合概念數上限。如需相關資訊，請移至圖示上方。
	該工具列顯示使用比對文字過濾器限制了結果。這透過放大鏡圖示顯示。

過濾結果

1. 從功能表中，選擇**工具** > **過濾器**。「過濾器」對話框即會開啟。
2. 選取並精簡您要使用的過濾器。
3. 按一下**確定**以套用過濾器，並在「擷取結果」窗格中查看新結果。

探索概念地圖

您可以建立概念地圖來探索如何相互關聯概念。透過選取單一概念並按一下**地圖**，會開啟概念地圖視窗，這樣您就可以探索與所選取概念相關的概念集。您可以透過編輯設定（例如，要包括哪些類型、要尋找哪些類型的關係等）來過濾出顯示哪些概念。

重要：在建立地圖之前，必須先產生索引。這可能需要花費數分鐘的時間。但產生索引之後，在重新擷取之前不必再次重新產生索引。如果您想在每次擷取時自動產生索引，請在擷取設定中選取該選項。如需相關資訊，請參閱主題 第 70 頁的『擷取資料』。

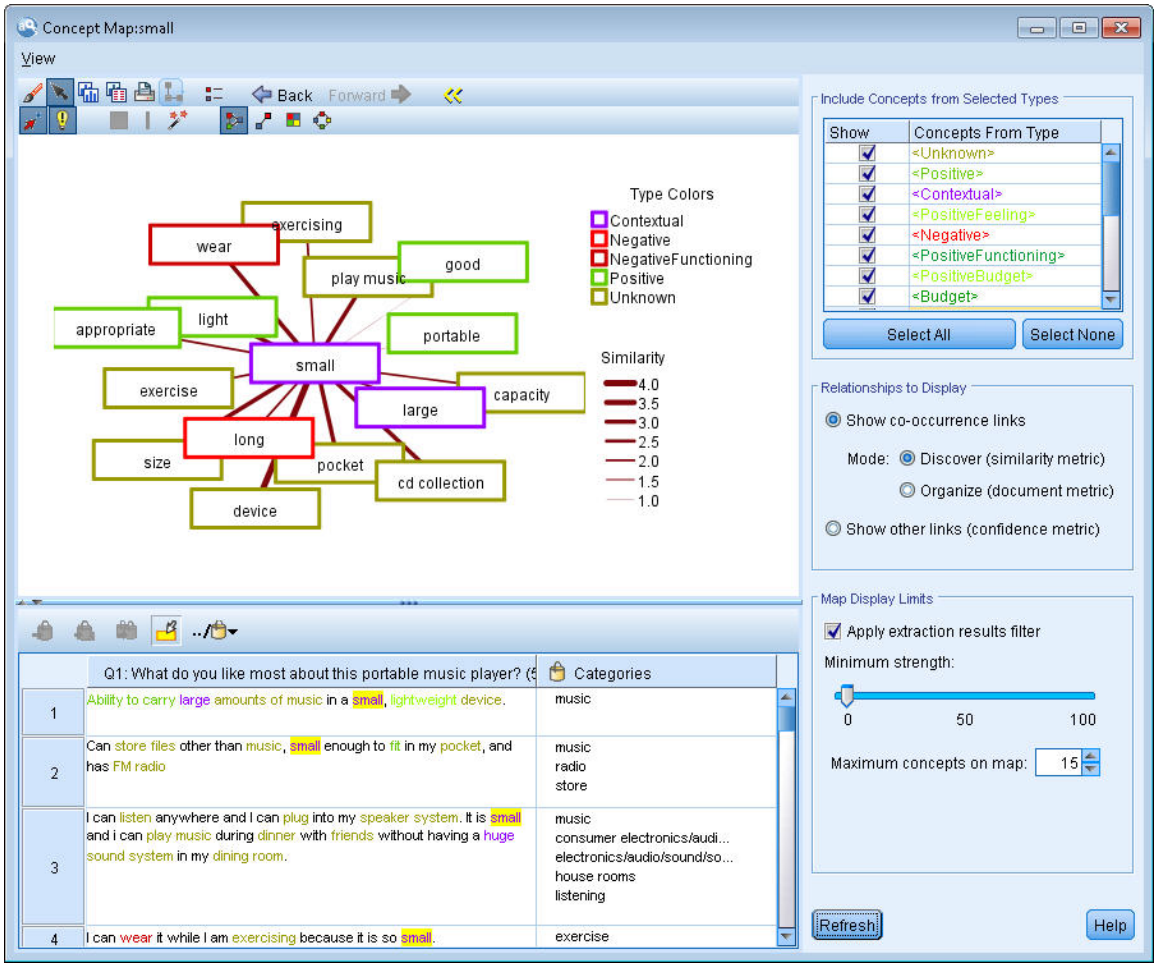


圖 27. 所選取概念的概念地圖

檢視概念地圖

1. 在「擷取結果」窗格中，選取單一概念。
2. 在此窗格的工具列中，按一下**地圖**按鈕。如果地圖索引已產生，則概念地圖會在不同的對話框中開啟。如果地圖索引未產生或已過期，則必須重建索引。此處理程序可能需要花費數分鐘的時間。
3. 在地圖周圍按一下以進行探索。如果您按兩下鏈結的概念，則地圖會重新繪製本身，並向您顯示剛剛按兩下的概念的鏈結概念。
4. 頂端工具列提供一些基本的地圖工具，例如，後移至前一個地圖、根據關係強度過濾鏈結，還有開啟過濾器對話框來控制出現的概念類型以及要代表的關係類型。另一個工具列行包含圖形編輯工具。請參閱第 138 頁的『使用圖形工具列及選用區』主題，以取得更多資訊。
5. 如果您對正在尋找的鏈結類型不滿意，請檢查此地圖右側顯示的地圖設定。

地圖設定：包括所選取類型中的概念

僅屬於表格中所選取類型的那些概念會在地圖中顯示。若要對某個類型隱藏概念，請在表格中取消選取該類型。

地圖設定：要顯示的關係

顯示共現鏈結 如果您要顯示共現鏈結，請選擇該模式。該模式會影響鏈結強度的計算方式。

- 探索（相似性度量值）。透過此度量值，會使用更複雜的計算來計算鏈結強度，計算會考量到兩個概念分開出現的頻率以及它們一起出現的頻率。較高的強度值表示一對概念一起出現的頻率往往比分開出現的頻率更高。使用下列公式，任何浮點數值都會轉換為整數。

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

圖 *28.* 相似性係數公式

在此公式中，$_I$ 是其中出現概念 I 的文件或記錄數。

C_J 是其中出現概念 J 的文件或記錄數。

C_{IJ} 是文件集中其中同時出現概念配對 I 和 J 的文件或記錄數。

- 組織（文件度量值）。使用此度量值的鏈結強度由共現的原始計數判定。一般兩個概念出現得越頻繁，它們一起出現的可能性就越大。較高的強度值表示一對概念經常一起出現。

顯示其他鏈結（信任度量值）。您可以選擇其他鏈結來顯示；這些鏈結可能是語意、衍生（形態）或併入（語法），並與從所鏈結概念移除概念的步驟數有關。這些鏈結可以協助您調整資源，特別是同義或釐清。如需上述每一個分組技術的簡要說明，請參閱 第 92 頁的『進階語言設定』

註：請記住，如果建置索引時未選取這些鏈結或者如果找不到任何關係，則一個都不會顯示。如需相關資訊，請參閱主題 『建置概念地圖索引』。

地圖設定：地圖顯示限制

套用擷取結果過濾器。如果您不想使用所有概念，則可以在擷取結果窗格中使用過濾器來限制顯示的內容。然後選取此選項，IBM SPSS Modeler Text Analytics 會使用此過濾集來尋找相關概念。如需相關資訊，請參閱主題 第 72 頁的『過濾擷取結果』。

強度下限。在這裡設定鏈結強度下限。關係強度低於此限制的任何相關概念都會對地中隱藏。

地圖上的概念數上限。指定要在地圖上顯示的關係數上限。

建置概念地圖索引

在建立地圖之前，必須先產生概念關係的索引。每當您建立概念地圖時，IBM SPSS Modeler Text Analytics 都會參照此索引。您可以透過在此對話框中選取技術來選擇要編制索引的關係。

分組技術。選擇一或多種技術。如需上述每一種技術的簡要說明，請參閱 第 93 頁的『關於語言技術』 並非所有技術都適用於所有文字語言。

防止特定概念配對。選取此勾選框可停止處理程序在輸出中將兩個概念一併分組或配對。若要建立或管理概念配對，請按一下**管理配對**。如需相關資訊，請參閱主題 第 93 頁的『管理鏈結異常狀況配對』。

建置索引可能需要花費數分鐘的時間。但產生索引之後，不必再次重新產生，直到重新擷取為止或除非您想要變更設定來包括更多的關係。如果您想要在每次擷取時都產生索引，則可以在擷取設定中選取該選項。如需相關資訊，請參閱主題 第 70 頁的『擷取資料』。

精簡擷取結果

擷取是您可以擷取、檢閱結果、進行變更然後重新擷取以更新結果的反覆運算處理程序。由於精確度及連續性對於成功文字挖掘及分類至關重要，從頭細部調整擷取結果會確保您每一次重新擷取時，都會從種類定義中取得完全相同的結果。透過此方法，可以更準確且可重複的方式將記錄及文件指派給您的種類。

擷取結果用作種類的建置區塊。當您使用這些擷取結果建立種類時，如果記錄及文件包含符合一個或多個種類描述子的文字，則會自動將它們指派給種類。雖然您可以在對語言資源進行精簡之前開始分類，則在開始之前至少檢閱擷取結果一次會非常有用。

檢閱結果時，您可能發現想要擷取引擎以不同方式處理的元素。請考量下列範例：

- **無法辨識的同義字。**假設您找到數個認為同義的概念，例如 smart、intelligent、bright 及 knowledgeable，並且它們顯示為擷取結果中的個別概念。您可以建立同義字定義，其中 intelligent、bright 及 knowledgeable 全部在目標概念 smart 下進行分組。這樣做會將所有這些項目分與 smart 分組在一起，並且廣域頻率計數也將更高。如需相關資訊，請參閱主題 『新增同義字』。
- **錯誤區分類型的概念。**假設您的擷取結果中的概念是一個類型，並且您希望將它們指派給另一個。在另一個範例中，假設您在擷取結果中找到 15 個蔬菜概念，並且您想要將它們全部新增至名為 <Vegetable> 的新類型。對於大部分語音，未在任何類型定義檔中找到但擷取自文字的概念會自動鍵入為 <Unknown>您可以將概念新增至類型。如需相關資訊，請參閱主題 第 77 頁的『將概念新增至類型』。
- **不重要的概念。**假設您找到已擷取且具有極高頻率計數的概念，即，在許多記錄或文件中找到它。然而，您將此概念視為對於您的分析不重要。您可以從擷取中排除它。如需相關資訊，請參閱主題 第 78 頁的『從擷取中排除概念』。
- **不正確的相符項。**假設在檢閱包含某個特定概念的記錄或文件時，您發現兩個單字未正確地分組在一起，例如 faculty 及 facility。此相符可能是由於內部演算法，稱為模糊分組，即暫時地忽略雙重或三重輔音與母音，將一般拼錯分組。您可以將這些單字新增至應該分組的單字配對清單。如需相關資訊，請參閱主題 第 177 頁的『模糊分組』。
- **未擷取的概念。**假設您預期找到擷取的某些概念，但是在檢閱記錄或文件文字時注意到未擷取少數單字或片語。通常這些單字是您無興趣的動詞或形容詞。然而，有時您想要使用未作為種類定義一部分進行擷取的單字或片語。若要擷取概念，您可以強制術語進入類型定義檔。如需相關資訊，請參閱主題 第 79 頁的『強制單字執行擷取』。

可以透過選取一個或多個元素，並按一下滑鼠右鍵以存取快速功能表，從而直接從「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框中執行其中許多變更。

進行變更之後，窗格背景顏色會變更，以表明您需要重新擷取以檢視您的變更。請參閱第 70 頁的『擷取資料』主題，以取得更多資訊。如果您在使用較大的資料集，則在進行數個變更之後而不是在進行每一個變更之後進行重新擷取可能更為有效。

註：您可以在資源編輯器視圖（檢視 > 資源編輯器）中檢視用來產生擷取結果的完整可編輯語言資源集。這些資源在此視圖中檔案庫及定義檔的形式以呈現。您可以直接在檔案庫及定義檔內自訂概念及類型。如需相關資訊，請參閱主題 第 155 頁的第 15 章, 『使用檔案庫』。

新增同義字

同義字建立具有相同意義的兩個或多個單字的關聯。同義字通常用來將術語與其縮寫分組在一起，或者將通常錯誤拼字的單字與正確的拼字分組在一起。透過使用同義字，目標概念的頻率較大，這讓探索文字資料中以不同方式呈現的類似資訊更為容易。

語言資源範本及產品提供的檔案庫包含許多預先定義的同義字。然而，如果您探索到未辨識的同義字，則可以定義它們，以便您在下一次擷取時會辨識它們。

第一個步驟是決定目標、或前導、概念為何。目標概念是您要據以在最終結果中分組所有同義字術語的單字或片語。擷取期間，同義字在此目標概念下分組。第二個步驟是識別此概念的所有同義字。系統會替換最終擷取中所有同義字的目標概念。必須將術語擷取至同義字。然而，不需要擷取目標概念即可發生替代。例如，如果您要將 intelligent 取代為 smart，則 intelligent 是同義字，且 smart 是目標概念。

如果您建立新的同義字定義，則會將新的目標概念新增至定義檔。然後，您必須將同義字新增至該目標概念。只要您建立或編輯同義字，就會在資源編輯器中的同義字定義檔內記錄這些變更。如果您要檢視這些同義字定義檔的整個內容，或者如果您要進行大量變更，則可能偏好直接在資源編輯器中工作。如需相關資訊，請參閱主題 第 169 頁的『替代/同義字字典』。

所有新的同義字將自動儲存在資源編輯器視圖中檔案庫樹狀結構內列出的第一個檔案庫中，依預設，這是**本端檔案庫**。

註：如果您尋找同義字定義，並且無法透過快速功能表或直接在資源編輯器中找到它，則可能從內部模糊分組技術產生相符項。如需相關資訊，請參閱主題 第 177 頁的『模糊分組』。

若要建立新的同義字

1. 在「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框中，選取您要為其建立新同義字的概念。
2. 從功能表中，選擇**編輯** > **新增至同義字** > **新建**。即會開啟「建立同義字」對話框。
3. 在「目標」文字框中輸入目標概念。這是將據以分組所有同義字的概念。
4. 如果您要新增更多同義字，請在「同義字」清單框中輸入它們。使用廣域分隔字元來分隔每一個同義字術語。如需相關資訊，請參閱主題 第 64 頁的『選項：階段作業標籤』。
5. 按一下**確定**，套用您的變更。即會關閉該對話框，變更「擷取結果」窗格的背景顏色，指出您需要重新擷取以查看變更。如果您有數個變更，請在重新擷取之前進行變更。

若要新增同義字

1. 在「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框中，選取您要新增至現有同義字定義的概念。
2. 從功能表中，選擇**編輯** > **新增至同義字**。該功能表會顯示一組同義字，並且最新建立的同義字會顯示在清單頂端。選取您要新增所選取概念的目標同義字名稱。如果您看到所尋找的同義字，請進行選取，並且選取的概念會新增至該同義字定義。如果您看不到，請選取**更多**以顯示「所有同義字」對話框。
3. 在「所有同義字」對話框中，您可以按自然順序（建立順序）或者以遞增或遞減順序排序清單。選取您要新增所選取概念的目標同義字的名稱，然後按一下**確定**。即會關閉對話框，並且概念會新增至同義字定義。

將概念新增至類型

只要執行擷取，就會將擷取的概念指派給類型，從而將共用內容的術語分組在一起。 IBM SPSS Modeler Text Analytics 同時提供許多內建類型。如需相關資訊，請參閱主題 第 164 頁的『內建類型』。對於大部分語音，未在任何類型定義檔中找到但擷取自文字的概念會自動鍵入為 <Unknown>

檢閱結果時，您可能發現部分概念出現在一個類型中，但是您想要指派給另一個概念，或者您可能發現一組單字實際上自己屬於一個新的類型。在這些情況下，您可能想要將概念重新指派給另一個類型，或者建立一個全新的類型。

例如，假設您要使車用關於輛的意見調查資料，並且您對透過聚焦不同領域的車輛進行分類有興趣。您應該建立稱為 <Dashboard> 的類型，以將在車輛儀表板上找到之量規及旋鈕相關的所有概念分組在一起。然後，您可以將概念（例如 gas gauge、heater、radio 及 odometer）指派給該新類型。

在另一個範例中，假設您正在使用與大學及學院相關的意見調查資料，並且擷取鍵入 Johns Hopkins（大學）作為 <Person> 類型，而不是作為 <Organization> 類型。在此情況下，您可以將此概念新增至 <Organization> 類型。

只要您建立類型，或者將概念新增至類型的術語清單，就會在資源編輯器中語言資源檔案庫內的類型定義檔中記錄這些變更。如果您要檢視這些檔案庫的內容，或者進行大量變更，則可能偏好直接在資源編輯器中工作。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。

若要將概念新增至類型

1. 在「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框中，選取您要新增至現有類型的概念。
2. 按一下滑鼠右鍵以開啟快速功能表。
3. 從功能表中，選擇**編輯** > **新增至類型**。該功能表會顯示一組類型，並且最新建立的類型會顯示在清單頂端。選取您要新增所選取概念的目標類型名稱。如果您看到所尋找的類型名稱，請進行選取，並且選取的概念會新增至該類型。如果您看不到，請選取**更多**以顯示「所有類型」對話框。
4. 在「所有類型」對話框中，您可以按自然順序（建立順序）或者以遞增或遞減順序排序清單。選取您要新增所選取概念的目標類型的名稱，然後按一下**確定**。即會關閉對話框，並且它們作為術語新增至類型。

若要建立新的類型

1. 在「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框中，選取您要建立新類型的概念。
2. 從功能表中，選擇**編輯** > **新增至類型** > **新建**。即會開啟「類型內容」對話框。
3. 在「名稱」欄位中輸入此類型的新名稱，並對其他欄位進行任何變更。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。
4. 按一下**確定**，套用您的變更。即會關閉該對話框，變更「擷取結果」窗格的背景顏色，指出您需要重新擷取以查看變更。如果您有數個變更，請在重新擷取之前進行變更。

從擷取中排除概念

檢閱您的結果時，您有時可能發現不想要擷取的概念，或者由任何自動化種類建置技術使用的概念。在某些情況下，這些概念具有非常高的頻率計數，對您的分析完全不重要。在此情況下，您可以將概念標記為從最終擷取中排除。通常，您新增至此清單的概念是文字中使用以取得連續性的填寫單字或片語，但是那不會新增任何重要事項，並且可能讓擷取結果雜亂。透過將概念新增至排除定義檔，您可以確保永不擷取它們。

透過排除概念，您下一次擷取時，所排除概念的所有變化都會從擷取結果中消失。如果此概念已在種類中顯示為描述子，重新擷取後，它將保留在具有零計數的種類中。

排除後，這些變更會記錄在資源編輯器中的排除定義檔內。如果您要檢視所有排除定義並直接編輯它們，您可以偏好直接在資源編輯器中工作。如需相關資訊，請參閱主題 第 172 頁的『排除字典』。

若要排除概念

1. 在「擷取結果」窗格、「資料」窗格、「種類定義」對話框或「叢集定義」對話框中，選取您要從擷取中排除的概念。
2. 按一下滑鼠右鍵以開啟快速功能表。
3. 選取**從擷取中排除**。概念即會新增至資源編輯器中的排除定義檔，並且「擷取結果」窗格的背景顏色會變更，指出您需要重新擷取以查看變更。如果您有數個變更，請在重新擷取之前進行變更。

註：您排除的所有單字將自動儲存在資源編輯器中檔案庫樹狀結構內列出的第一個檔案庫中，依預設，這是**本端檔案庫**。

強制單字執行擷取

擷取後在「資料」窗格中檢閱文字資料時，您可能發現未擷取部分單字或片語。通常，這些單字是您無興趣的動詞或形容詞。然而，有時您想要使用未作為種類定義一部分進行擷取的單字或片語。

如果您要擷取這些單字及片語，則可以強制術語進入類型檔案庫。如需相關資訊，請參閱主題 第 168 頁的『強制術語』。

重要事項！ 將定義檔中的術語標記為強制並不十分簡單。透過此方法，我們明確表明即使您已明確將術語新增至定義檔，有時在您重新擷取後，它可能不會呈現在「擷取結果」窗格中，或者它未完全按照您宣告的方式呈現。雖然很少發生此情況，但是當已擷取某個單字或片語作為較長片語的一部分時，可能發生此情況。為了防止此情況，將**整個（無複合字）**符合選項套用至類型定義檔中的這個術語。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。

第 9 章 分類文字資料

在「種類與概念」視圖中，您可以建立種類，以從本質上代表高層次概念或主題，以擷取主要構想、知識與文字中表達的本質。

自 IBM SPSS Modeler Text Analytics 14 版開始，種類也可以具有階層式結構，這表示它們可以包含子種類，並且那些子種類也可以具有其自己的子種類，依此類推。您可以匯入具有階層式種類的預先定義種類結構（先前稱為代碼訊框），以及在產品內建置這些階層式種類。

實際上，階層式種類可讓您建置具有一個或多個子種類的樹狀結構，從而更準確地分組項目，例如不同的概念或主題區域。簡式範例可以與休閒活動相關；回答問題，例如如果您有更多時間會希望執行什麼活動？，您可能具有最上層種類，例如 *sports*、*art and craft*、*fishing* 等；向下一個層次，在 *sports* 下方，您可能具有子種類以查看這是否為 *ball games*、*water-related* 等。

種類由一組描述子（例如，概念、類型、型樣及種類規則）組成。這些描述子一起用於識別文件或記錄是否屬於給定的種類。可以掃描文件或記錄內的文字，以查看是否有任何文字符合描述子。如果發現相符項目，則會將文件/記錄指派給該種類。此過程叫作分類。

您可以利用在「種類與概念」視圖的四個窗格中呈現的資料，使用、建置及視覺化探索種類，例如可以透過從「視圖」功能表中選取其名稱已隱藏或顯示的每一個種類。

- **「種類」窗格。**在此窗格中建置及管理您的種類。如需相關資訊，請參閱主題 第 82 頁的『種類窗格』。
- **「擷取結果」窗格。**在此窗格中探索及使用所擷取概念及類型。如需相關資訊，請參閱主題 第 69 頁的『擷取結果：概念和類型』。
- **視覺化窗格。**在此窗格中以視覺化方式探索您的種類及它們的互動方式。如需相關資訊，請參閱主題 第 135 頁的『種類圖形與圖表』。
- **資料窗格。**在此窗格中探索及檢閱對應於選擇之文件及記錄內包含的文字。如需相關資訊，請參閱主題 第 88 頁的『資料窗格』。

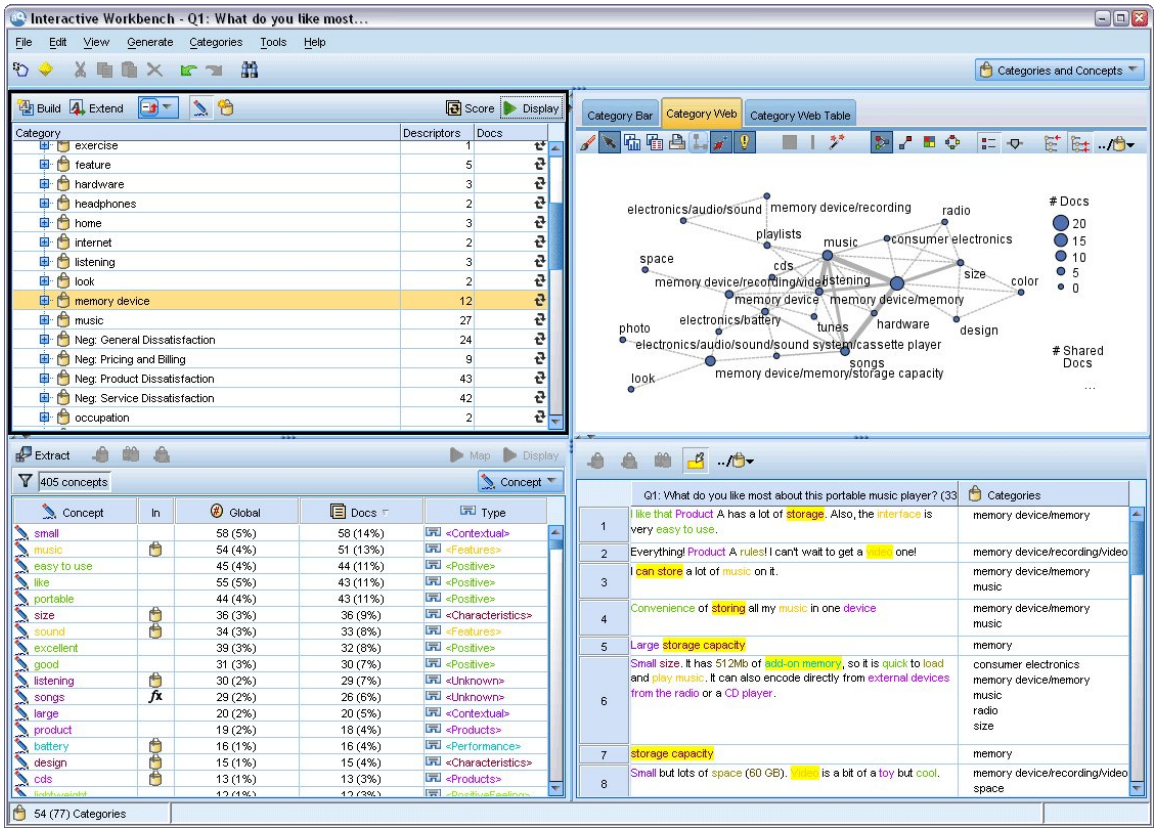


圖 29. 種類與概念視圖

由於您可能從文字分析套件 (TAP) 中的一組種類開始，或者從預先定義的種類檔案匯入，您還可能需要建立您自己的種類。可以使用產品豐富的自動化技術來自動建置種類，這會使用擷取結果（概念、類型及型樣）產生種類及其描述子。也可以使用您可能擁有的其他資料相關見解來手動建立。然而，您只能手動建立種類，或者透過互動式工作台細部調整。請參閱第 21 頁的『文字採礦節點：模型標籤』主題，以取得更多資訊。您可以透過將擷取結果拖放至種類，手動建立種類定義。您可以透過將種類規則新增至種類，使用您自己的預先定義的種類，或者組合，豐富這些種類或任何空的種類。

每一個技術及方法都完美適合某些類型的資料及狀況，但是它通常有助於將技術結合在同一分析中，以擷取完整範圍的文件或記錄。並且在分類的過程中，您可能看到要對語言資源進行的其他變更。

種類窗格

「種類」窗格是您可以建置及管理您的種類的區域。此窗格位於「種類與概念」視圖的左上角。從文字資料擷取概念及類型之後，您可以開始手動或利用概念併入、共生等技術自動建置種類。請參閱第 90 頁的『建置種類』主題，以取得更多資訊。

每一次建立或更新種類時，可以透過按一下**評分**按鈕評分文件或記錄，以查看是否有任何文字符合給定種類中的某個描述子。如果發現相符項目，則會將文件或記錄指派給該種類。最終結果是根據種類中的描述子，大部分（如果不是全部）文件或記錄指派給種類。

註：如果有多個種類可以適合可見的窗格，您可以使用位於窗格底端的控制項，在種類中向前或向後移動，或者輸入要跳至的頁碼。

種類樹狀結構表格

此窗格中德樹狀結構表格代表種類、子種類及描述子集。樹狀結構還具有數個直欄，呈現每一個樹狀結構項目的資訊。下列直欄可能可供顯示：

- **代碼** 列出每一個種類的代碼值。依預設會隱藏此直欄。您可以透過功能表顯示此直欄：**檢視** > **種類窗格**。
- **種類**。包含顯示種類及子種類名稱的種類樹狀結構。此外，如果按一下描述子工具列圖示，還將顯示描述子集。
- **描述子**。提供組成其定義的描述子數目。此技術不包括子種類中的描述子數目。在**種類**直欄中顯示描述子名稱時，不會給定任何計數。 您可以透過功能表在樹狀結構中顯示或隱藏描述子本身：**檢視** > **種類窗格** > **所有描述子**。
- **文件** 評分之後，此直欄提供分類至種類及其所有子種類的文件或記錄。因此如果根據描述子 5 筆記錄符合您的最上層種類，並且根據描述子 7 筆不同的記錄符合子種類，最上層種類的文件總數是兩項的總和，在此案例中為 12。然而，如果同一記錄符合最上層種類及其子種類，則計數將為 11。

不存在任何種類時，表格仍包含兩列。最上層列（稱為**所有文件**）是文件或記錄的總數。第二列（稱為**未分類**）顯示尚未分類的文件/記錄數。

對於窗格中的每一個種類，會在種類名稱前面顯示一個小的黃色儲存區圖示。如果您按兩下種類，或者在功能表中選擇**視圖 > 種類定義**，即會開啟「種類定義」對話框，並呈現組成其定義的所有元素（稱為描述子），例如概念、類型、型樣及種類規則。如需相關資訊，請參閱主題第 87 頁的『關於種類』。依預設，種類樹狀結構表格不會顯示種類中的描述子。如果您要直接在樹狀結構中（而不是在「種類定義」對話框中）查看描述子，請按一下工具列中具有鉛筆圖示的切換按鈕。選取此切換按鈕時，您也可以展開樹狀結構以查看描述子。

評分種類

種類樹狀結構表格中的**文件**直欄顯示分類至該特定種類的文件或記錄。如果數字過期或未計算，則會在該直欄中顯示一個圖示。 您可以在窗格工具列上按一下**評分**，以重新計算文件的數目。請記住，當您使用更大的資料集時，評分處理程序可能需要一些時間。

在樹狀結構中選取種類

在樹狀結構中進行選擇時，您只能選取同層級種類，即，如果您選取最上層種類，則還可以選取子種類。或者，如果您選取給定種類的 2 個子種類，則不能同步選取另一個種類的子種類。選取不連續的種類將導致遺失先前的選擇。

在資料及視覺化窗格中顯示

當您選取表格中的某個列時，您可以按一下**顯示**按鈕，利用對應於您的選擇的資訊，重新整理「視覺化」及「資料」窗格。如果窗格不可見，則按一下**顯示**將顯示窗格。

精簡您的種類

首次嘗試時，分類可能不會針對資料產生結果，可能有您要刪除或與其他種類結合的種類。透過檢閱擷取結果，您還可能發現有部分未建立的種類非常有用。如果這樣，您可以對結果進行手動變更，以針對特定環境定義細部調整它們。如需更多資訊，請參閱主題第 115 頁的『編輯及精簡種類』。

用來建立種類的方法及策略

如果您尚未擷取，或者您的擷取結果已過期，則使用其中一個種類建置或延伸技術將提示您自動延伸。套用技術之後，分組至種類的概念及類型仍可用於利用其他技術進行種類建置。這表示您可能在多個種類中看到一個概念，除非您選擇不重複多個使用。

為了協助您建立最佳種類，請檢閱下列項：

- **用來建立種類的方法**
- **用來建立種類的策略**
- **用來建立種類的提示**

用來建立種類的方法

由於每個資料集都是唯一的，因此一段時間之後，種類建立方法數目及您套用它們的順序可能發生變更。此外，由於各個資料集的文字挖掘目標可能不同，因此您可能需要體驗不同的方法，以查看哪個為給定的文字資料產生最佳結果。沒有任何自動技術將完美分類您的資料；因此我們建議尋找並套用適合您資料的一個或多個自動技術。

除了搭配使用文字分析套件（TAP、*.tap*）與預先建置的種類集之外，您還可以利用下列方法的組合分類回應：

- **自動建置技術。**數個基於語言及基於頻率的種類選項可為您自動建置種類。如需相關資訊，請參閱主題 第 90 頁的『建置種類』。
- **自動延伸技術。**數個語言技術可新增及加強描述子，以便擷取更多記錄，從而延伸現有種類。如需相關資訊，請參閱主題 第 98 頁的『延伸種類』。
- **手動技術。**存在數種手動方法，例如拖放。如需相關資訊，請參閱主題 第 100 頁的『手動建立種類』。

用於建立種類的策略

下列策略清單並不詳盡，但是可為您提供一些有關如何建置種類的構想。

- 當您定義文字挖掘節點時，從文字分析套件 (TAP) 中選取一個種類集，以便您利用部分預先建置的種類開始您的分析。這些種類可能足以從頭分類您的文字。然而，如果您要新增更多種類，則可以編輯「建置種類」設定（**種類 > 建置設定**）。開啟**進階設定：語言**對話框，然後選擇「種類」輸入選項**未用的擷取結果**，並建置其他種類。
- 當您定義節點時，從「互動式工作台」中「種類與概念」視圖內的 TAP 選取一個種類集。接下來，將您認為適當的未用概念或型樣拖放至種類。然後，展開您剛剛編輯的現有種類（**種類 > 展開種類**），以取得與現有種類描述子相關的更多描述子。
- 使用進階語言設定（**種類 > 建置種類**）自動建置種類。然後，透過刪除描述子、刪除種類或合併類似的種類，直到您對產生的種類滿意為止，從而手動精簡種類。此外，如果您最初在**不**使用**可能時利用萬用字元一般化**選項的情況下建置種類，則還可以嘗試利用**一般化**選項使用「延伸種類」自動簡化種類。
- 使用極具描述性的種類名稱及/或註釋匯入預先定義的種類檔案。此外，如果您最初已在**未**選擇選項從種類名稱匯入或產生描述子的情況下匯入，則可以稍後使用「延伸種類」對話框，並選擇**利用從種類名稱產生的描述子延伸空的種類**。選項。然後，第二次展開那些種類，但是此次使用分組技術。
- 按頻率排序概念或概念型樣，然後將最有興趣的項目拖放至「種類」窗格，從而手動建立第一組種類。具有起始種類集之後，使用「延伸」特性（**種類 > 延伸種類**）延伸並精簡所有選取的種類，以便它們包括其他相關描述子，從而符合更多記錄。

套用這些技術之後，我們建議您檢閱產生的種類，並使用手動技術進行次要調整，移除任何錯誤分類，或者新增可能遺漏的記錄或單字。此外，由於使用不同的技術可能產生冗餘種類，您還可以根據需要合併或刪除種類。如需更多資訊，請參閱主題第 115 頁的『編輯及精簡種類』。

建立種類的提示

為了協助您建立更好的種類，您可以檢閱部分提示，可協助您作出有關方法的決策。

有關種類與文件比例的提示

由於至少兩個原因，在定量文字分析中，指派文件及記錄的目標種類通常不會互斥。

- 首先，一般經驗法則表明文字文件文件或記錄越長，表達的構想及意見越具體。因此，可以為文件或記錄指派多個種類的機會大大增加。
- 第二，通常有各種方法可分組及解譯在邏輯上分隔的文字文件或記錄。如果意見調查包含有關回覆者政治信仰的開放式結尾問題，我們可以建立種類，例如 *Liberal* 及 *Conservative* 或 *Republican* 及 *Democrat*，以及更具體的種類，例如 *Socially Liberal*、*Fiscally Conservative* 等。這些種類無需徹底互斥。

有關要建立之種類數目的提示

種類建立應該直接從資料進行，當您看到有關資料的有興趣內容時，您可以建立種類以代表該資訊。通常，您建立的種類數目沒有建議上限。然而，當然可能建立太多種類而無法管理。兩個原則適用：

- **種類頻率**。要讓種類有用，它必須包含文件或記錄數目下限。一個或兩個文件可能包括一些非常有趣的內容，但如果它們是 1,000 個文件中的一個或兩個，則它們包含的資訊可能不夠頻繁，實際上無用。
- **複雜性**。您建立越多種類，則在完成分析後要檢閱及彙總的資訊越多。然而，種類太多會增加複雜性，可能不會新增有用的明細。

不幸的是，沒有規則用來判定多少種類太多或用來判定每個種類的記錄數目下限。您將必須根據特定狀況下的需求進行此類判定。

然而，我們可以提供有關從何處開始的建議。雖然種類數目不應過多，但是在分析早期，種類太多比種類太少要好。將相對類似的種類分組在一起要比將案例分割為新種類簡單，因此從較多種類減少為較少種類這一策略通常是最佳做法。此軟體計劃提供文字挖掘的反覆運算本質，並且可以簡單達成，因此開始時接受建置較多種類。

選擇最佳描述子

下列資訊包含有關選擇或建立最佳種類描述子（概念、類型、TLA 型樣及種類規則）的部分準則。描述子是種類的建置區塊。當文件或記錄中的部分或全部文字符合描述子時，文件或記錄會符合種類。

除非描述子包含或對應於所擷取的概念或型樣，否則它不會符合任何文件或記錄。因此，如下列段落中所述使用概念、類型、型樣及種類規則。

由於概念不僅代表其自己，還代表一組基礎術語，因此範圍從複數/單數形式到同義字，到拼字作變化，只有概念本身應該用作描述子，或者用部分描述子。若要進一步瞭解所有給定概念的基礎術語，按一下「種類與概念」視圖之「擷取結果」窗格中的概念名稱。當您將游標移至概念名稱上方時，會顯示一個工具提示，顯示前次擷取期間在文字中找到的所有基礎術語。並非所有概念都具有基礎術語。例如，如果 `car` 與 `vehicle` 是同義字，但是 `car` 擷取作為概念，而 `vehicle` 擷取作為基礎術語，則您只想要在描述子中使用 `car`，因為它自動將文件或記錄與 `vehicle` 相符。

概念及類型作為描述子

當您想要尋找包含該概念（或者其任何基礎術語）的所有文件或記錄時，可以將概念用作描述子。在此情況下，由於確切的概念名稱足夠，不需要使用更複雜的種類規則。請記住，當您使用擷取意見的資源時，有時概念可以在 TLA 型樣擷取期間變更，以擷取更真實的句子意義（請參閱 TLA 上下一節中的範例）。

例如，意見調查回應指出每個人員的最愛水果（例如 "*Apple and pineapple are the best*"）可能造成擷取 apple 及 pineapple。透過將概念 apple 作都為描述子新增至種類，包含概念 apple（或其所有基礎術語）的所有回應都符合該概念。

然而，如果您對以任何方式簡單瞭解哪些回應提及 *apple* 有興趣，則可以撰寫種類規則（例如 * apple *），並且您還將擷取包含概念（例如 apple、apple sauce 或 french apple tart）的回應。

您還可以透過直接將類型用作描述子（例如 <Fruit>），擷取包含以相同方式鍵入之概念的所有文件或記錄。請注意，您不能搭配使用 * 與類型。

如需相關資訊，請參閱主題 第 69 頁的『擷取結果：概念和類型』。

文字鏈結分析 (TLA) 型樣作為描述子

當您想要擷取較細微的且有微妙差別的構想時，將 TLA 型樣結果用作描述子。TLA 擷取期間分析文字時，文字會逐個句子或子句進行處理，而不是查看整個文字（文件或記錄）。透過同時考量單一句子的所有部分，TLA 可以識別意見、兩個元素之間的關係或否定，例如瞭解更真實的意義。您可以將概念型樣或類型型樣用作描述子。請參閱第 127 頁的『類型型樣和概念型樣』主題，以取得更多資訊。

例如，如果我們具有文字 "*the room was not that clean*"，則可以擷取下列概念：room 及 clean。然而，如果在擷取設定中啟用 TLA 擷取，則 TLA 可以偵測到 clean 以方面方式使用，且實際對應於 not clean，這是概念 dirty 的同義字。在這裡您可以看到，將概念 clean 用作其自己的描述子會符合此文字，但是還可能擷取其他提及 cleanliness 的文件或記錄。因此，可能最好使用具有 dirty 的 TLA 概念型樣作為輸出概念，因為它可能符合此文字，並且很可能是更適當的描述子。

種類商業規則作為描述子

種類規則是根據利用所擷取概念、類型及型樣的邏輯表示式，以及布林運算子，自動將文件或記錄分類至種類的陳述式。例如，您可以撰寫表示式，表示包括在此種類中包含所擷取概念 embassy 但不包含 argentina 的所有記錄。

您可以利用 &、| 及 !() 布林值撰寫及使用種類規則作為種類中的描述子，以表示數個不同的構想。如需有關這些規則之語法及如何撰寫與編輯它們的詳細資訊，請參閱第 101 頁的『使用種類規則』。

- 搭配使用種類規則與 & (AND) 布林運算子，可協助您尋找發生 2 個或更多概念的文件或記錄。& 運算子連接的 2 個或更多概念不需要發生在同一句子或片語中，但是可以發生在同一文件或記錄中的任何位置即可視為符合種類。例如，如果您建立種類規則 food & cheap 作為描述子，它將符合包含文字 "*the food was pretty expensive, but the rooms were cheap*" 的記錄，儘管事實上 food 不是稱為 cheap 的名詞，因為該文字同時包含 food 與 cheap。
- 搭配使用種類規則與 !() (NOT) 布林運算子作為描述子，可協助您尋找部分項發生一些事情而其他未發生的文件或記錄。這可以協助避免分組根據單字看起來相關但根據環境定義看起來不相關的資訊。例如，如果您建立種類規則 <Organization> & !(ibm) 作為描述子，它將符合下列文字 *SPSS Inc. was a company founded in 1967*，而不符合下列文字 *the software company was acquired by IBM.*。
- 搭配使用種類規則與 | (OR) 布林運算子作為描述子，可協助您尋找包含數個概念或種類之一的文件或記錄。例如，如果您建立種類規則 (personnel|staff|team|coworkers) & bad 作為描述子，它將在其中找到這些名詞中任何名詞的任何文件或記錄符合概念 bad。

- 使用種類規則中的類型，讓它們更一般且可能更可部署。例如，如果您要使用旅館資料，且可能對瞭解客戶對旅館人員的想法非常有興趣。相關術語可能包括接待人員、服務人員、服務人員（女性）、服務台、前台等單字。在此情況下，您可以建立稱為 <HotelStaff> 的新類型，並將之前的所有術語新增至該類型。由於可以為每種員工建立一個種類規則，例如 [* waitress * & nice]、[* desk * & friendly]、[* receptionist * & accommodating]，您可以利用 <HotelStaff> 類型建立更一般的單一種類規則，以擷取具有正面旅館工作人員意見且採用形式 [<HotelStaff> & <Positive>] 的所有回應。

附註：將 TLA 型樣包括在種類規則中時，您可以在那些規則中同時使用 + 與 &。如需相關資訊，請參閱主題第 103 頁的『在種類規則中使用 TLA 型樣』。

概念、TLA 或種類規則如何作為描述子以不同方式相符的範例

下列範例示範如何使用概念作為描述子，如何使用概念規則作為描述子，或者使用 TLA 型樣作為描述子影響文件或記錄如何分類。讓我們假設您具有以下 5 筆記錄。

- A：「餐廳工作人員極好，食物美味，房間舒適乾淨。」
- B：「餐廳人員糟糕，但是房間乾淨。」
- C：「房間舒適乾淨。」
- D：「我的房間不太乾淨。」
- E：「乾淨。」

由於記錄包括單字乾淨，並且您想要擷取此資訊，則可以建立下表中顯示的其中一個描述子。根據您嘗試擷取的核心要素，您可以看到使用不同種類的描述子如何可以產生不同的結果。

表 *17.* 範例記錄如何符合描述子

描述子	A	B	C	D	E	說明
clean	相符	相符	相符	相符	相符	描述子是擷取的概念。每筆記錄都包含概念 clean，即使是記錄 D，由於沒有 TLA，不會自動瞭解到根據 TLA 規則 "*not clean*" 表示 dirty。
clean + .	-	-	-	-	相符	描述子是自行代表 clean 的 TLA 型樣。僅符合 TLA 擷取期間擷取 clean 且無相關聯概念的記錄。
[clean]	相符	相符	相符	-	相符	描述子是尋找自身包含 clean 或包含其他內容之 TLA 規則的種類規則。符合找到包含 clean 之 TLA 輸出的所有記錄，而無論 clean 是否鏈結至另一個概念（room），以及是否在任何插槽位置。

關於種類

種類是指一組密切相關的概念、意見或態度。種類也可以透過擷取其重要意義的簡短片語或標籤輕鬆說明，這非常有用。

例如，如果您要分析消費者有關新洗衣皂的意見調查回應，則可以建立標籤為 *odor* 的種類，其中包含說明產品味道的所有回應。然而，此類種類不會區分味道令人愉悅與令人厭煩的產品。由於 IBM SPSS Modeler Text Analytics 能夠擷取使用與適當資源時的意見，您可以再建立兩個種類，以識別享受該味道的回覆者與不喜歡該味道的回覆者。

您可以在「種類與概念」視圖視窗左上方窗格的「種類」窗格中，建立及使用您的種類。每一個種類都是由一個或多個描述子進行定義。**描述子**是概念、類型及型樣，以及已用於定義種類的種類規則。

如果您要查看組成給定種類的描述子，則可以按一下「種類」窗格工具列中的鉛筆圖示，然後展開樹狀結構以查看描述子。或者，選取種類並開啟「種類定義」對話框（**檢視 > 種類定義**）。

當您使用種類建置技術（例如概念併入）自動建置種類時，該技術將概念與類型用作描述子，以建立您的種類。如果您擷取 TLA 型樣，還可以新增型樣或部分型樣作為種類描述子。如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。如果您建置叢集，則可以將叢集中的概念新增至新的或現有種類。最後，您可以手動建立種類規則，以用作種類中的描述子。如需相關資訊，請參閱主題 第 101 頁的『使用種類規則』。

種類內容

除了描述子之外，您也可以編輯種類的內容，從而重新命名種類、新增標籤或新增註釋。

存在下列內容：

- **名稱。**依預設，此名稱顯示在樹狀結構中。使用自動技術建立種類時，會自動為它提供名稱。
- **標籤。** 使用標籤可協助建立更有意義的種類說明，以供在其他產品或者其他表格或圖形中使用。如果您選擇顯示標籤的選項，則在介面中使用標籤以識別種類。
- **代碼。**代號對應於此種類的字碼值。.
- **註釋。**您可以在此欄位中為每一個種類新增簡要說明。「建置種類」對話框產生種類時，會自動將附註新增至此註釋。您也可以透過選取文字並從功能表中選擇**種類 > 新增至註釋**，自動將範例文字直接從「資料」窗格新增至註釋。

資料窗格

在建立種類時，有時您可能想要檢閱正在使用的部分文字資料。例如，如果您建立的種類中對 640 個文件進行了分類，您可能想要查看部分或全部文件來了解實際撰寫的文字。您可以在位於右下方的「資料」窗格中檢閱記錄或文件。如果依預設不可見，請從功能表中選擇**檢視 > 窗格 > 資料**。

「資料」窗格呈現每個文件或記錄一列，對應於「種類」窗格、「擷取結果」窗格或「種類定義」對話框中選取的內容，最多可達到某個顯示限制。依預設，會限制在「資料」窗格中顯示的文件或記錄數，以便讓您更快速地查看資料。但您可以在「選項」對話框中對此進行調整。如果您處理的資料集非常大，則可以透過關閉用來顯示種類的選項來提高顯示速度。如需相關資訊，請參閱主題 第 64 頁的『選項：階段作業標籤』。

註：如果可見窗格能夠容納更多的記錄，則您可以使用窗格底部的控制項在記錄中前後移動，或輸入要跳至的頁碼。

顯示及重新整理資料窗格

「資料」窗格不會自動重新整理顯示畫面，因為使用的資料集較大時，自動資料重新整理可能需要花費一些時間才能完成。因此，每當您在此視圖中的另一個窗格或「種類定義」對話框中進行選取時，請按一下**顯示**來重新整理「資料」窗格的內容。

文字文件或記錄

如果文字資料的形式為記錄且文字長度相對較短，則「資料」窗格中的文字欄位會顯示完整的文字資料。但使用記錄及更大的資料集時，文字欄位直欄會顯示一小段文字，並會在右側開啟「文字預覽」窗格以顯示您在表格中已選取記錄的更多或全部文字。如果文字資料的形式為個別文件，則「資料」窗格會顯示文件的檔名。選取文件時，「文字預覽」窗格會開啟並顯示所選取文件的文字。

顏色及強調顯示

每當您顯示資料時，在那些文件或記錄中找到的概念及描述子會以顏色強調顯示，來協助您在文字中輕鬆識別它們。顏色編碼對應於概念所屬的類型。您還可以將滑鼠移至顏色編碼項目上方來顯示在其下面擷取的概念及為其指派的類型。未擷取的任何文字都以黑色顯示。一般這些未擷取的單字通常為連接詞（*and* 或 *with*）、代詞（*me* 或 *they*）及動詞（*is*、*have* 或 *take*）。

資料窗格直欄

文字欄位直欄一律可見，而您也可以顯示其他直欄。若要顯示其他直欄，請從功能表中選擇**檢視** > **資料窗格**，然後選取您想要在「資料」窗格中顯示的直欄。下列直欄可用於顯示：

- **「文字欄位名稱」(#)/文件**。針對從中擷取概念和類型的文字資料新增直欄。如果資料在文件中，則直欄稱為「文件」，並且只有文件檔名或完整路徑可見。若要查看那些文件的文字，您必須查看「文字預覽」窗格。「資料」窗格中的列數會在此直欄名稱之後的括弧中顯示。由於「選項」對話框中存在用來提高載入速度的限制，因此有時可能不會顯示所有文件或記錄。如果達到上限，則數字後面會接有 **- 上限**。如需相關資訊，請參閱 第 64 頁的『選項：階段作業標籤』。
- **種類**。 列出記錄所隸屬的每一個種類。每當顯示此直欄時，為了顯示最新的資訊，重新整理「資料」窗格可能需要花費更長一點的時間。
- **強制移入**。列出您已在其中強制移入文件的種類。可以透過**編輯** > **強制移入**功能表選項將文件強制移入種類。如需相關資訊，請參閱 第 117 頁的『強制將文件移入種類』。
- **強制移出**。列出您已從中移除文件的種類。可以透過**編輯** > **強制移出**功能表選項強制將文件移出種類。例如，當回應者的諷刺導致某個回應被錯誤分類，可能會使用該選項。如需相關資訊，請參閱 第 117 頁的『強制將文件移入種類』。
- **種類計數**。列出記錄所隸屬的種類數目。
- **相關性等級**。為單一種類中的每一筆記錄提供等級。此等級顯示與種類中的其他記錄相比，該記錄在多大程度上適合該種類。在「種類」窗格（左上方窗格）中選取種類可查看等級。如需相關資訊，請參閱 『種類相關性』。
- **回應旗標**。新增直欄以顯示您可能要使用的任何旗標。在此直欄內按一下以變更您指派給文件的旗標類型。您可使用「完成」旗標或「重要」旗標來標示文件，或是移除旗標。這對於檢閱種類模型的完成度而言非常有用。如需相關資訊，請參閱 第 90 頁的『標示回應』。

種類相關性

為了協助您更好地建置種類，您可以檢閱每一個種類中的文件或記錄相關性，以及文件或記錄屬於之所有種類的相關性。

種類至記錄的相關性

只要文件或記錄顯示在「資料」窗格中，則其屬於的所有種類都會列出在「種類」直欄中。當文件或記錄屬於多個種類時，此直欄中的種類會以相關性最高到最低的相符項順序顯示。首先列出的種類被認為最佳對應此文件或記錄。如需相關資訊，請參閱主題 第 88 頁的『資料窗格』。

記錄至種類的相關性

當您選取一個種類時，可以在「資料」窗格的「相關性等級」直欄中檢閱其每一個筆記錄的相關性。這個相關性等級指出與該種類中的其他記錄相比較而言，文件或記錄與所選取種類的適合程度。若要查看單一種類的記錄等級，請在「種類」窗格（左上方窗格）中選取此種類，則文件或記錄的等級會顯示在直欄中。依預設此直欄不可見，但是您可以選擇顯示它。如需相關資訊，請參閱主題 第 88 頁的『資料窗格』。

記錄的等級數越低，此記錄與所選取種類越適合或越相關，1 表示最適合。如果多筆記錄具有相同的相關性，則每一筆記錄都顯示相同的等級，後面是等號 (=)，表示它們的相關性相等。例如，您可能具有下列等級 1=、1=、3、4，依此類推，這表示對於此種類，有兩筆記錄同樣地被視為最相符。

提示：您可以將最相關記錄的文字新增至種類註釋，以協助提供更好的種類說明。透過選取文字，並從功能表中選擇**種類 > 新增至註釋**，直接從「資料」視窗新增文字。

標示回應

若要協助監視您的進度，您可以在「資料」窗格中使用旗標來標示文件。只有在來源文件包含唯一 ID 時，此功能才可用。如果來源文件不含唯一 ID，則您可以在來源文件與「文字採礦」節點之間新增一個「衍生」節點。

您想要標示文件的原因有多種，包括：

- 標示您已手動檢閱的文件，方便您知道稍後要在哪裡選取它們
- 標示您不確定如何處理的文件

使用旗標標示文件之後，您便可以繼續處理文件。它們僅供您自己記錄保留。您可以選擇下列旗標：

表 *18.* 旗標說明

旗標	說明
	完成旗標表示您認為已完成的文件。
	重要旗標表示您認為重要的文件。

若要使用旗標標示文件：

1. 從「資料」窗格中，在您要標示的文件上按一下滑鼠右鍵。
2. 從快速功能表中選擇**檢視** > **資料窗格** > **回應旗標**，然後選取您要使用的旗標類型（重要旗標或完成旗標）。即會指派選取的旗標。如果「資料」窗格中的「旗標」直欄不可見，它會出現。

若要清除旗標：

1. 從「資料」窗格中，在您要移除其旗標的文件上按一下滑鼠右鍵。
2. 從快速功能表中選擇**標示回應方法** > **清除旗標**。即會移除選取的旗標。

建置種類

由於您可能具有來自文字分析套件的種類，您還可以使用語言與頻率技術數目自動建置種類。您可以透過「建置種類設定」對話框，套用自動化的語言及頻率技術，從概念或從概念型樣產生種類。

通常，種類可以由不同種類的描述子組成（類型、概念、TLA 型樣、規則規則）。當您利用自動化種類建置技術建置種類時，在概念或概念型樣（取決於您選取的輸入）之後命名產生的種類，並且每一個都包含一組描述子。這些描述子可能採用種類規則或概念形式，並且包括技術探索到的所有相關概念。

建置種類之後，您可以透過在「種類」窗格中進行檢閱，或者透過圖形及圖表進行探索，詳細瞭解種類。然後，您可以使用手動技術進行次要調整，移除任何錯誤分類，或者新增可能遺漏的記錄或單字。套用技術之後，分組至種類的概念、類型及型樣仍可用於其他技術。此外，由於使用不同的技術可能產生冗餘或不適當的種類，您還可以合併或刪除種類。如需更多資訊，請參閱主題第 115 頁的『編輯及精簡種類』。

重要事項！ 在早期版本中，共生與同義字規則含括在方括弧中。在此版本中，現在方括弧指出文字鏈結分析型樣結果。共生與同義字規則將封裝在括弧中，例如 (speaker systems|speakers)。

若要建置種類

1. 從功能表中，選擇**種類 > 建置種類**。除非您已選擇永不提示，否則將顯示一個訊息框。
2. 選擇您是想要立即建置，還是首先編輯設定。

- 按一下**立即建置**，以開始使用現行設定建置種類。依預設選取的設定通常足以開始分類處理程序。即會開始種類建置處理程序，並會顯示進度對話框。
- 按一下**編輯**，以檢閱並修改建置設定。

註：可以顯示的種類數目上限為 10,000。如果達到或超出此數目，則會顯示一個警告。如果發生此情況，您應該變更「建置種類」或「延伸種類」選項，以減少建置的種類數目。

輸入

從衍生自類型型樣或類型的描述子建置種類。在該表格中，您可以選取要在種類建置處理程序中包括的個別類型或型樣。

類型型樣。如果您選取類型型樣，則種類建置自型樣，而不是其自己的類型及概念。使用該方法時，會分類包含所選取類型型樣之概念型樣的所有記錄或文件。因此，如果您選取表格中的 <Budget> 及 <Positive> 類型型樣，則可能產生諸如 cost & <Positive> 或 rates & excellent 的種類。

將類型型樣用作自動化種類建置的輸入時，有時技術會識別多個方法以形成種類結構。通常，沒有單一正確的方法來產生種類；然而您可能發現一個結構比其他結構更適合您的分析。為了協助自訂此情況下的輸出，您可以指定一個類型作為偏好的焦點。產生的所有最上層種類都將來自您在這裡選取之類型的概念（且無其他類型）。每個種類都將包含此類型的文字鏈結型樣。在**按型樣類型結構化種類：**欄位中選擇此類型，並且表格將更新以僅顯示包含所選取類型的使用型樣。通常，將為您預先選取 <Unknown>。此結果位於包含所選取類型 <Unknown> 的所有型樣中。該表格以遞減順序顯示類型，並且首先顯示具有最多記錄或文件（**Doc.** 計數）的類型。

類型。如果您選取類型，則將從屬於所選取類型的概念建置種類。因此如果您選取表格中的 <Budget> 類型，則由於 cost 及 price 是指派給 <Budget> 類型的概念，因此可能產生諸如 cost 或 price 的種類。

依預設，僅選取擷取最多記錄或文件的類型。這個預先選取可讓您快速聚焦於最有興趣的類型，並避免建置無興趣的種類。該表格以遞減順序顯示類型，並且首先顯示具有最多記錄或文件（**Doc.** 計數）的類型。依預設，在類型表格中取消選取來自 Opinions 檔案庫的類型。

您選擇的輸入會影響您取得的種類。當您選擇使用「類型」作為輸入時，可以更容易地清楚查看相關概念。例如，如果您使用「類型」作為輸入建置種類，則可能取得諸如 apple、pear、citrus fruits、orange 等概念的種類 Fruit。如果您改為選擇「類型型樣」作為輸入，並例如選取型樣 <Unknown> + <Positive>，則您可能取得種類 fruit + <Positive>，並包含一種或兩種水果，例如 fruit + tasty 及 apple + good。這個第二個結果僅顯示 2 個概念型樣，因為不需要正面識別其他出現的水果。並且由於這可能足以用於您的現行文字資料，在您使用不同文件集的縱向研究中，您可能想要在其他描述子（例如 citrus fruit + positive）中手動新增或使用類型。單獨使用類型作為輸入將協助您尋找所有可能的水果。

技術

由於每個資料集都是唯一的，因此一段時間之後，方法數目及您套用它們的順序可能發生變更。由於各個資料集的文字挖掘目標可能不同，因此您可能需要體驗不同的技術，以查看哪個能夠為給定的文字資料產生最佳結果。

您不需要是這些設定的專家即可使用它們。依預設，已選取最常見且常用的設定。因此，您可以略過進階設定對話框，並直接建置種類。同樣地，如果您在這裡進行變更，則無需每次都返回設定對話框，因為將一律保留最新設定。

選取語言或頻率計數，然後按一下「進階設定」按鈕以顯示所選取技術的設定。沒有任何自動技術將完美分類您的資料；因此我們建議尋找並套用適合您資料的一個或多個自動技術。不能同步使用語言及頻率技術進行建置。

- **進階語言技術**。如需相關資訊，請參閱『進階語言設定』。
- **進階頻率技術**。如需相關資訊，請參閱第 97 頁的『進階頻率設定』。

進階語言設定

當您建置種類時，可以從許多進階語言種類建置技術中進行選擇，例如概念併入和語意網路（僅限英文文字）。這些技術可以個別使用或彼此組合使用以建立種類。

請記住，由於每個資料集都是唯一的，因此一段時間之後，方法數目及其套用順序可能發生變更。由於各個資料集的文字挖掘目標可能不同，因此您可能需要體驗不同的技術，以查看哪個能夠為給定的文字資料產生最佳結果。沒有任何自動技術將完美分類您的資料；因此我們建議尋找並套用適合您資料的一個或多個自動技術。

「進階設定：語言」對話框中提供下列區域及欄位：

輸入及輸出

種類輸入 選取將建置的種類：

- **未用的擷取結果**。此選項容許從未在任何現有種類中使用的擷取結果中建置種類。這會將記錄符合多個種類的趨勢降至最低，並限制產生的種類數目。
- **所有擷取結果**。此選項容許使用任何擷取結果建置種類。未存在任何種類或存在較少種類時，這最有用。

種類輸出 選取將建置之種類的一般結構：

- **具有子種類的階層式**。此選項可讓您建立子種類及子子種類。您可以透過選擇可以建立的層次數目上限（**建立的層次上限**欄位），設定種類的深度。如果您選擇 3，則種類可能包含子種類，且那些子種類也可能具有子種類。
- **純文字種類（僅限單一層次）**。此選項只能建置一個層次的種類，意味著不會產生任何子種類。

分組技術

每一個可用的技術都適用於某些類型的資料和狀況，但是在同一個分析中結合技術來擷取完整範圍的文件或記錄往往非常有用。您可能會看到一個概念在多個種類中或是發現冗餘的種類。

概念併入。這項技術會根據多詞彙概念（複合字）包含的單字是其他概念中單字的子集還是超集來將它們分組，藉此來建置種類。例如，概念 `seat` 將會與 `safety seat`、`seat belt` 和 `seat belt buckle` 群組。如需相關資訊，請參閱主題 第 95 頁的『概念併入』。

語意網路。這項技術會先從每一個概念的單字關係延伸索引識別其可能的觀念，然後藉由將相關概念分組來建立種類。這項技術最適用於概念為語意網路所知，且不會過於含糊不清時。當文字包含特殊化術語或是網路不知道專門術語時，它就比較不是那麼有用。在一個範例中，概念 `granny smith apple` 可以與 `gala apple` 和 `winesap apple` 群組，因為它們是 `granny smith` 的同層級。在另一個範例中，概念 `animal` 可能與 `cat` 和 `kangaroo` 群組，因為它們是 `animal` 的下義詞。在本版本中，這項技術僅提供英文文字。如需相關資訊，請參閱主題 第 95 頁的『語意網路』。

註：僅當您選取**語意網路**時，**搜尋距離上限**選項才可用。

搜尋距離上限 選取您希望在產生種類之前，技術搜尋的距離。值越低，您取得的結果越少 - 然而，這些結果將雜訊程度較低，並且很可能明顯彼此鏈結或相關聯。值越高，您取得的結果可能越多 - 然而，這些結果的依賴或相關程度可能較低。由於此選項在廣域上套用至所有技術，因此其效果在共生及語意網路上最佳。

防止特定概念配對。選取此勾選框，以停止在輸出中將兩個概念分組或配對在一起的處理程序。若要建立或管理概念配對，請按一下**管理配對...**。如需更多資訊，請參閱主題『管理鏈結異常狀況配對』。

可能時使用萬用字元一般化 選取此選項，可讓產品使用星號萬用字元產生一般種類規則。例如，使用萬用字元不會產生多個描述子，例如 [apple tart + .] 及 [apple sauce + .]，而是可能產生 [apple * + .]。如果您使用萬用字元一般化，您通常將取得與先前完全一樣的記錄或文件數目。然而，此選項的優點是減少數目並簡化種類描述子。此外，此選項可讓您更好地使用這些種類對新的文字資料（例如，在縱向/波浪研究中）分類更多記錄或文件。

用於建置種類的其他選項

除了選取要套用的分組技術之外，您可以編輯數個其他建置選項，如下所示：

建立的最上層種類數目上限。使用此選項可限制下一次按一下「建置種類」按鈕時可以產生的種類數目。在某些情況下，如果您將此值設為較高值，然後刪除無興趣的種類，可能會取得更好的結果。

每個種類的描述子及/或子種類的數目下限。使用此選項可定義種類要建立而必須包含的描述子及子種類數目下限。此選項會協助限制建立不會擷取大量記錄或文件的種類。

容許描述子出現在多個種類中 選取之後，此選項容許在接下來建置的多個種類中使用描述子。由於項目通常或「本質上」屬於兩個或多個種類，因此讓它們這樣做通常會讓品質種類更高。如果您未選取此選項，則減少多個種類中記錄的重疊，並根據您具有的資料類型，可能需要此選項。然而，使用大部分類型的資料時，將描述子限制為單一種類通常會導致品質或種類涵蓋面遺失。例如，讓我們假設您具有概念 car seat manufacturer。使用此選項，此概念可能根據文字 car seat 出現在一個種類中，並根據 manufacturer 出現在另一個種類中。然而如果未選取此選項，雖然您可能仍具有兩個種類，但是概念 car seat manufacturer 將根據數個因素（包括出現 car seat 及 manufacturer 每一項的記錄數目），在最符合的種類中顯示為描述子。

解析重複種類名稱的依據 選取如何處理名稱與現有種類名稱相同的新種類或子種類。您可以合併具有相同名稱的新種類（及其描述子）與現有種類。或者，如果在現有種類中找到重複名稱，您可以選擇跳過建立任何種類。

管理鏈結異常狀況配對

種類建置、形成叢集及概念對映期間，內部演算法按已知的關聯分組單字。若要防止將兩個概念配對在一起，或者鏈結在一起，您可以在**建置種類進階設定**對話框、**建置叢集**對話框及**概念對映索引設定**對話框中開啟此特性，然後按一下**管理配對**按鈕。

在產生的**管理鏈結異常狀況**對話框中，您可以新增、編輯或刪除概念配對。每行輸入一個配對。在這裡輸入配對將阻止在建置或延伸種類、形成叢集及概念管理時發生配對。輸入您希望的單字，例如，加重音版本的單字不等於不加重音版本的單字。

例如，如果您要確保不分組 hot dog 及 dog，則可以在表格中新增配對作為單獨行。

關於語言技術

當您建置或延伸種類時，可以從許多進階語言種類建置技術中進行選擇，包括概念根衍生、概念併入、語意網路（僅限英文）及共生規則。這些技術可以個別使用或彼此組合使用以建立種類。

您不需要是這些設定的專家即可使用它們。依預設，已選取最常見且常用的設定。如果您想要，您可以略過此進階設定對話框，並直接建置或延伸種類。同樣地，如果您在這裡進行變更，則無需每次都返回設定對話框，因為它將提醒您前次使用的項目。

然而，請記住，由於每個資料集都是唯一的，因此一段時間之後，方法數目及其套用順序可能發生變更。由於各個資料集的文字挖掘目標可能不同，因此您可能需要體驗不同的技術，以查看哪個能夠為給定的文字資料產生最佳結果。沒有任何自動技術將完美分類您的資料；因此我們建議尋找並套用適合您資料的一個或多個自動技術。

為種類建置的主要自動化語言技術為：

- **概念根衍生**。此技術透過採用概念，並分析是否有任何概念元件在形態上相關來尋找與其相關的其他概念，從而建立種類。如需相關資訊，請參閱主題『概念根衍生』。
- **概念併入**。此技術透過採用概念並尋找包括它的其他概念來建立種類。如需相關資訊，請參閱主題 第 95 頁的『概念併入』。
- **語意網路**。此技術透過從擴充的單字關係索引中識別每一個概念的可能感應，然後透過分組相關概念建立種類。如需相關資訊，請參閱主題 第 95 頁的『語意網路』。此選項僅適用於英文文字。
- **共生**。此技術會建立共生規則，可用於建立新的種類、延伸種類或作為另一個種類技術的輸入。如需相關資訊，請參閱主題 第 96 頁的『共生規則』。

概念根衍生

概念根衍生技術透過採用概念，並分析是否有任何概念元件在形態上相關來尋找與其相關的其他概念，從而建立種類。元件是一個單字。技術會透過查看概念中每一個元件的結尾（字尾），並尋找可能從中衍生的其他概念，從而嘗試分組概念。構想是當單字彼此衍生時，它們很可能共用或意義相近。為了識別結尾，系統會使用語言特定的規則。例如，概念 opportunities to advance 與概念 opportunity for advancement 及 advancement opportunity 分組在一起。

您可以在任何排序的文字上使用概念根衍生。它自己會產生相當少的種類，且每一個種類都趨於包含很少的概念。每一個種類中的概念都是同義字或狀況相關。即使您手動建置種類，也可能發現使用此演算法很有用；它發現的同義字可能是您特別有興趣的那些概念的同義字。

註：您可以透過明確指定概念，防止將概念分組在一起。如需相關資訊，請參閱主題 第 93 頁的『管理鏈結異常狀況配對』。

術語元件化及取消字形變化

套用概念根衍生或概念併入技術時，術語會首先細分為元件（單字），然後將元件取消字形變化。套用技術時，系統會載入概念及其相關聯的術語，並根據分隔字元分割為元件，例如空格、連字號及單引號。例如，術語 system administrator 會分割為元件，例如 {administrator, system}。

然而，某些原始術語的部分可能不會予以使用，並作為停止字組使用。使用英文時，其中部分可忽略的元件可能包括 a、and、as、by、for、from、in、of、on、or、the、to 及 with。

例如，術語 examination of the data 具有元件集 {data, examination}，且 of 及 the 被視為可忽略。此外，元件順序不在元件集中。透過此方法，以下三個術語可以相當：cough relief for child、child relief from a cough 及 relief of child cough，因為它們全都具有相同的元件集 {child, cough, relief}。每一次將一對術語識別為相當時，都會合併對應的概念以形成參照所有術語的新概念。

此外，由於術語的概念可能發生字形變化，因此最初會套用語言特定的規則以識別相當的術語，而無論是否發生字形變化，例如複數形式。透過此方法，術語 level of support 及 support levels 可以識別為相當，因為取消字形變化的單數形式為 level。

概念根顏色如何運作

術語元件化及取消字形變化（請參閱上一節）之後，概念根衍生演算法會分析元件結尾或字尾，以尋找元件根，然後將概念與具有相同或類似根的其他概念分組在一起。使用文字語言特定的一組語言衍生規則識別結尾。例如，這是英文語言文字的衍生規則，說明以字尾 ical 結尾的概念元件可能衍生自具有相同根詞幹且以字尾 ic 結尾的概念。使用此規則（及取消字形變化），演算法將能夠分組 epidemiologic study 及 epidemiological studies。

由於術語已元件化，且已識別可忽略的元件（例如 in 及 of），因此概念根衍生演算法也能夠使用 epidemiological studies 分組概念 studies in epidemiology。

已選擇一組元件衍生規則，以便依此演算法分組的大部分概念均為同義字：概念 epidemiologic studies、epidemiological studies、studies in epidemiology 全部為相當的術語。為了提高完整性，有部分衍生規則容許演算法分組狀況上相關的概念。例如，演算法可以分組概念，例如 empire builder 及 empire building。

概念併入

概念併入技術透過採用概念、使用詞彙系列演算法、識別其他概念中包括的概念來建置種類。構想是當概念中的單字為另一個概念的子集時，它反映基礎語意關係。併入是可以與任何類型的文字搭配使用的強大技術。

此技術非常適合與語意網路組合使用，但是可以單獨使用。當文件或記錄包含許多網域特定的術語或專門術語時，概念併入也可能提供更好的結果。如果您已事前調整定義檔，以便單獨擷取及分組特殊術語（使用同義字），則尤其適用。

概念併入如何運作

套用概念併入演算法之前，會對術語進行元件化及取消字形變化。如需相關資訊，請參閱主題 第 94 頁的『概念根衍生』。接下來，概念併入演算法會分析元件集。對於每一個元件集，演算法會尋找作為第一個元件集之子集的另一個元件集。

例如，如果您具有概念 continental breakfast，其元件集為 {breakfast, continental}，並且您具有概念 breakfast，其元件集為 {breakfast}，則演算法將推斷 continental breakfast 是一種 breakfast，並將這些項分組在一起。

在較大的範例中，如果您的「擷取結果」窗格中具有概念 seat，且您套用此演算法，則概念（例如 safety seat、leather seat、seat belt、seat belt buckle、infant seat carrier 及 car seat laws）也將分組在該種類中。

由於術語已元件化，且已識別可忽略的元件（例如 in 及 of），因此概念併入演算法將辨識概念 advanced spanish course 包括概念 course in spanish。

附註：您可以透過明確指定概念，防止將概念分組在一起。如需相關資訊，請參閱主題 第 93 頁的『管理鏈結異常狀況配對』。

語意網路

在此版本中，語意網路技術僅可用於英文語言文字。

此技術使用單字關係的內建網路建置種類。對於此原因，當術語具體且語意不明時，此技術可以產生非常良好的結果。然而，您不應該預期技術在高度技術/具體的概念之間找到許多鏈結。處理此類概念時，您可能發現概念併入及概念根衍生技術更加有用。

語意網路如何運作

語意網路技術背後的構想是利用已知的單字關係以建立同義字或下位詞的種類。**下位詞**是一個概念，是沒有階層式關係的第二個概念的種類，也稱為 ISA 關係。例如，如果 animal 是一個概念，則 cat 及 kangaroo 是 animal 的下位詞，因為它們是 animal 的種類。

除了同義字及下位詞關係之外，語意網路技術還會檢查 <Location> 類型中所有概念之間的部分及完整鏈結。例如，技術將概念 normandy、provence 及 france 分組至一個種類，因為 Normandy 及 Provence 屬於 France。

語意網路透過識別語意網路中每一個概念的可能意義開始。當概念識別為同義字或下位詞時，它們會分組至單一種類。例如，技術將建立包含以下三個概念的單一種類：eating apple、dessert apple 及 granny smith，因為語意網路包含以下資訊：1) dessert apple 是 eating apple 的同義字，並且 2) granny smith 是 eating apple 的一個種類（表示它是 eating apple 的下位詞）。

許多概念（特別是單一術語）在單獨使用時都語意不明。例如，概念 buffet 可能表示一種餐食或一件傢具。如果概念集包括 meal、furniture 及 buffet，則會強制演算法在使用 meal 或使用 furniture 分組 buffet 之間進行選擇。請瞭解，在部分情況下，演算法進行的選擇可能不適用於一組特定記錄或文件的環境定義。

語意網路技術可以使用某些類型的資料勝過執行概念併入。當語意網路與概念併入同時辨識到 apple pie 是一種 pie 時，只有語意網路辨識到 tart 也是一種 pie。

語意網路將與其他技術一起運作。例如，假設您已同時選取語意網路及併入技術，並且語意網路已利用概念 tutor 分組概念 teacher，（因為 tutor 是一種 teacher）。併入演算法可以利用 tutor 分組概念 graduate tutor，因此，兩個演算法會合作以產生包含全部三個概念的輸出種類：tutor、graduate tutor 及 teacher。

語意網路的選項

有許多其他設定可能對此技術有興趣。

- 變更**搜尋距離上限**。選取您希望在產生種類之前，技術搜尋的距離。值越低，產生的結果越少 - 然而，這些結果將雜訊程度較低，並且很可能明顯彼此鏈結或相關聯。值越高，您取得的結果越多 - 而，這些結果的依賴或相關程度可能較低。

例如，根據距離，演算法會從 Danish pastry 到 coffee roll（其母項）進行搜尋，然後搜尋 bun（祖母項），以及向上到 bread。

如果您認為所產生的種類太大或將太多的項目分組在一起，則透過縮短搜尋距離，此技術會產生較小的種類，這可能更易於使用。

重要事項！ 此外，我們建議您在使用此技術時不套用選項**容納根字元限制下限的拼字錯誤**（在節點的「專家」標籤上或「擷取」對話框中定義）以進行模糊分組，因為有些錯誤分組可能對結果產生較大的負面影響。

共生規則

共生規則可讓您探索及分組在一組文件或記錄內緊密相關的概念。構想是當通常在文件及記錄中一起找到概念時，共生反映可能是種類定義中值的基礎關係。此技術會建立共生規則，可用於建立新的種類、延伸種類或作為另一個種類技術的輸入。如果兩個概念頻繁地一起出現在一組記錄中，且很少單獨地出現在任何其他記錄中，則它們很可能共生。此技術可能產生良好的結果，具有至少包含數百文件或記錄的更大資料集。

例如，如果許多記錄包含單字 price 及 availability，則這些概念可能分組至共生規則 (price & available)。在另一個範例中，如果概念 peanut butter、jelly 及 sandwich 通常一起出現，則它們將分組至一個概念共生規則 (peanut butter & jelly & sandwich)。

重要事項！ 在早期版本中，共生與同義字規則含括在方括弧中。在此版本中，現在方括弧指出文字鏈結分析型樣結果。共生與同義字規則將封裝在括弧中，例如 (speaker systems|speakers)。

共生規則如何運作

此技術會掃描文件或記錄，尋找很可能一起出現的兩個或更多概念。如果兩個或更多概念頻繁地出現在一組文件或記錄中，並且如果它們很少單獨出現在任何其他文件或記錄中，則它們很可能共生。

找到共生概念時，會形成種類規則。這些規則包含使用 & 布林運算子連接的兩個或更多概念。這些規則是邏輯陳述式，如果規則中的一組概念全部共生在該文件或記錄中，這些規則會自動將文件或記錄分類至一個種類。

共生規則的選項

如果您要使用共生規則技術，則可以細部調整影響所產生規則的數個設定：

- 變更**搜尋距離上限**。選取您希望技術搜尋共生的程度。隨著您增加搜尋距離，每一個共生需要的相似性值下限都會降低；因此，可能產生許多共生規則，但是具有較低相似性值的那些項通常重要性極低。隨著您縮短搜尋距離，需要的相似性值下限會提會高；因此，會產生更少共生規則，但是它們很可能更為重要（更加強大）。
- **文件數目下限**。 必須包含給定的一對概念才能視為共生的記錄或文件數目下限；您將此選項設定得越低，越容易找到共生。提高該值會產生更少但更重要的共生。例如，假設在 2 筆記錄中同時找到概念 "apple" 及 "pear"（並且這兩筆記錄未出現在任何其他記錄中）。將**文件**數目下限設為 2（預設值），共生技術將建立種類規則（蘋果與梨）。如果值提高至 3，則不再建立規則。

附註：使用小型資料集（< 1000 個值得預設回應），您可能不會找到任何具有預設值的共生。如果這樣的話，請嘗試提高搜距離值。

附註：您可以透過明確指定概念，防止將概念分組在一起。如需相關資訊，請參閱主題 第 93 頁的『管理鏈結異常狀況配對』。

進階頻率設定

您可以根據直接明確的機械頻率技術建置種類。使用此技術，您可以針對在給定記錄或文件計數上方找到的每一個項目（類型、概念或型樣）建置一個種類。此外，您可以針對所有不太頻繁發生的項目建置單一種類。計數時，我們將包含問題中所擷取概念（以及其任何同義字）、類型或型樣的記錄或文件數目對照整個文字中出現的總數。

分組頻繁發生的項目可能產生有興趣的結果，因為它可能指示一般或重要回應。套用其他技術之後，該技術對於未用擷取結果非常有用。當不存在任何其他種類時，另一個應用程式會在擷取後立即執行此技術，編輯結果以刪除無興趣的種類，然後展開那些種類以便它們符合更多記錄或文件。如需相關資訊，請參閱主題 第 98 頁的『延伸種類』。

您可以不使用此技術，而是透過遞減「擷取結果」窗格中的記錄或文件數目來排序概念或概念型樣，然後將排在前面的項目拖放至「種類」窗格以建立對應的種類。

「進階設定：頻率」對話框內提供下列欄位：

產生種類描述子。選取描述子的輸入類型。如需相關資訊，請參閱主題 第 90 頁的『建置種類』。

- **概念層次**。選取此選項表示將使用概念或概念型樣頻率。如果選取類型作為種類建置的輸入，則將使用概念，如果選取類型型樣，則使用概念型樣。通常，將此技術套用至概念層次將產生更多特定結果，因為概念及概念型樣代表較低層次的測量。

- **類型層次**。選取此選項表示將使用類型或類型型樣頻率。如果選取類型作為種類建置的輸入，則將使用類型，如果選取類型型樣，則使用類型型樣。將此技術套用至類型層次可讓您取得有關所給定呈現的資訊種類的快速視圖。

項目可擁有其自己的種類的**文件計數下限**。此選項可讓您從頻繁發生的項目建置種類。此選項將輸出僅限制為包含在至少 X 個記錄或文件中發生之描述子的那些種類，其中 X 是為此選項輸入的值。

將所有剩餘項目分組至具名種類。此選項可讓您將所有頻繁發生的概念或類型分組至具有您選擇之名稱的單一 'catch-all' 種類。依預設，此種類稱為其他。

種類輸入。選取套用技術的群組：

- **未用的擷取結果**。此選項容許從未在任何現有種類中使用的擷取結果中建置種類。這會將記錄符合多個種類的趨勢降至最低，並限制產生的種類數目。
- **所有擷取結果**。此選項容許使用任何擷取結果建置種類。未存在任何種類或存在較少種類時，這最有用。

解析重複種類名稱的依據。選取如何處理名稱與現有種類名稱相同的新種類或子種類。您可以合併具有相同名稱的新種類（及其描述子）與現有種類。或者，如果在現有種類中找到重複名稱，您可以選擇跳過建立任何種類。

延伸種類

延伸是一種處理程序，透過該處理程序自動新增或加強描述子以「增加」現有種類。目標是產生更好的種類，擷取最初未指派給該種類的相關記錄或文件。

您選取的自動分組技術將嘗試識別現有種類描述子相關的概念、TLA 型樣及種類規則。然後這些新概念、型樣及種類規則會作為新的描述子新增，或者新增至現有描述子。用於延伸的分組技術包括概念根衍生、概念併入、語意網路（僅限英文）及共生規則。**使用從種類名稱產生的描述子延伸空的種類**方法使用種類名稱中的單字產生描述子，因此種類名稱的敘述性越好，結果越佳。

註：延伸種類時，頻率技術無法使用。

延伸是以互動方式改良種類的良好方法。以下是您可以延伸種類的部分範例：

- 拖曳/釋放概念型樣以在「種類」窗格中建立種類
- 用手建立種類並新增簡式種類規則與描述子
- 改良其中種類具有極好地描述性名稱的預先定義種類檔案
- 精簡您在選擇之 TAP 中的種類之後

您可以多次延伸種類。例如，如果您使用極具描述性的名稱改良預先定義的種類檔案，則可以使用**使用從種類名稱產生的描述子延伸空的種類**選項進行延伸，以取得第一組描述子，然後再次延伸那些種類。然而，在其他情況下，如果描述子延伸得更寬，則多次延伸可能導致種類太過普通。由於建置及延伸分組技術使用類似的基礎演算法，則在建置種類之後直接延伸可能不會產生更有興趣的結果。

提示：

- 如果您嘗試延伸，且不想要使用結果，則一律可以在延伸之後立即復原作業（**編輯 > 復原**）。
- 由於在處理程序期間單獨建置規則，因此延伸可以在一個種類中產生兩個或更多種類規則，完全符合一組文件。如果想要的話，您可以透過手動編輯種類說明，檢閱種類並移除冗餘。如需相關資訊，請參閱主題 第 116 頁的『編輯種類描述子』。

若要延伸種類

1. 在「種類」窗格中，選取您想要延伸的種類。
2. 從功能表中，選擇**種類 > 延伸種類**。除非您已選擇永不提示的選項，否則將顯示一個訊息框。
3. 選擇您是想要立即建置，還是首先編輯設定。

- 按一下**立即延伸**，以開始使用現行設定延伸種類。即會開始處理程序，並會顯示進度對話框。
- 按一下**編輯**，以檢閱並修改設定。

嘗試延伸之後，發現新描述子的種類會在「種類」窗格中加上單字**已延伸**旗標，以便您可以快速識別。「已延伸」文字會保留，直到您再次延伸，以另一種方式編輯種類，或者透過快速功能表清除這些文字為止。

註：可以顯示的種類數目上限為 10,000。如果達到或超出此數目，則會顯示一個警告。如果發生此情況，您應該變更「建置種類」或「延伸種類」選項，以減少建置的種類數目。

建置或延伸種類時可用的每一個技術都完美適合某些類型的資料及狀況，但是它通常有助於將技術結合在同一分析中，以擷取完整範圍的文件或記錄。在互動式工作台中，在您下一次建置種類時，分組至種類的概念及類型仍可用。這意味著您可能會在多個種類中看到一個概念，或者找到冗餘的種類。

「延伸種類：設定」對話框中提供下列區域及欄位：

延伸內容。 選取將用於延伸種類的輸入：

- **未用的擷取結果**。此選項容許從未在任何現有種類中使用的擷取結果中建置種類。這會將記錄符合多個種類的趨勢降至最低，並限制產生的種類數目。
- **所有擷取結果**。此選項容許使用任何擷取結果建置種類。未存在任何種類或存在較少種類時，這最有用。

分組技術

如需這些技術中每一個技術的簡要說明，請參閱第 92 頁的『進階語言設定』。這些技術包括：

- **概念根延伸**
- **語意網路**（僅限英文文字，並且當選取「僅一般化」選項時不使用。）
- **概念併入**
- **共生**及**文件數目下限**子選項。

語意網路技術中永久地排除許多類型，原因是那些類型將不會產生相關結果。它們包括 <Positive>、<Negative>、<IP>、其他非語言類型等。

搜尋距離上限 選取您希望在產生種類之前，技術搜尋的距離。值越低，您取得的結果越少 - 然而，這些結果將雜訊程度較低，並且很可能明顯彼此鏈結或相關聯。值越高，您取得的結果可能越多 - 然而，這些結果的依賴或相關程度可能較低。由於此選項在廣域上套用至所有技術，因此其效果在共生及語意網路上最佳。

防止特定概念配對。選取此勾選框，以停止在輸出中將兩個概念分組或配對在一起的處理程序。若要建立或管理概念配對，請按一下**管理配對...**。如需更多資訊，請參閱主題第 93 頁的『管理鏈結異常狀況配對』。

可能的情況下：選擇是否使用萬用字元簡化延伸、一般化描述子或兩者。

- **延伸及一般化**。此選項將延伸所選取的種類，然後一般化描述子。當您選擇一般化時，產品將使用星號萬用字元在種類中建立種類規則。例如，使用萬用字元不會產生多個描述子，例如 [apple tart + .] 及 [apple sauce + .]，而是可能產生 [apple * + .]。如果您使用萬用字元一般化，您通常將取得與先前完全一樣的記錄或文件數目。然而，此選項的優點是減少數目並簡化種類描述子。此外，此選項可讓您更好地使用這些種類對新的文字資料（例如，在縱向/波浪研究中）分類更多記錄或文件。

- **僅延伸**。此選項將延伸您的種類，而不一般化。首先為手動建立的種類選擇**僅延伸**選項，然後使用**延伸及一般化**選項再次延伸相同的種類，這可能很有用。
- **僅一般化**。此選項將一般化描述子，而不會以任何其他方式延伸您的種類。

註：選取此選項會停用**語意網路**選項；這是因為僅當要延伸描述時，**語意網路**選項才可用。

用於延伸種類的其他選項

除了選取要套用的技術之外，您可以編輯以下任何選項：

據以延伸描述子的項目數目上限。使用項目（概念、類型及其他表示式）延伸描述子時，定義可以新增至單一描述子的項目數目上限。如果您將此限制設為 0，則不能將超過 10 個其他項目新增至現有描述子。如果要新增超出 10 個項目，則在新增技術後，技術會停止新增項目。這樣做可以讓描述子清單更短，但是不保證首先使用最有興趣的項目。您可能偏好使用**可能時使用萬用字元一般化**選項切斷延伸的大小，而不會降低品質。此選項將套用至包含 & (AND) 或 ! (NOT) 的描述子。

同時延伸子種類。此延伸還將延伸所選取種類下方的任何子種類。

使用從種類名稱產生的描述子延伸空的種類。此方法僅套用至具有 0 個描述子的空種類。如果種類已包含描述子，它將不會以此方式延伸。此選項會嘗試根據組成種類名稱的單字，自動建立每一個種類的描述子。系統會掃描種類名稱，以查看名稱中是否有任何單字符合任何擷取的概念。如果識別概念，則它會用於尋找相符的概念型樣，且這兩項都用來形成種類的描述子。種類名稱詳細且是敘述性時，此選項會產生最佳結果。這是用來產生種類描述子的快速方法，這會啟用種類以擷取包含那些描述子的記錄。當您從其他位置匯入種類時，或者當您使用詳細描述性名稱手動建立種類時，此選項最有用。

產生描述子作為。僅當選取之前的選項時，此選項才適用。

- **概念**。選擇此選項，以概念形式產生所產生的描述子，而無論是否已從來源文字中擷取它們。
- **型樣**。選擇此選項，以型樣形式產生所產生的描述子，而無論是否已擷取所產生的形樣或任何型樣。

手動建立種類

除了使用自動化種類建置技術及規則編輯器建立種類之外，您還可以手動建立種類。存在下列手動方法：

- 建立您將逐個新增元素的空種類。如需相關資訊，請參閱主題 『建立新的或重新命名種類』。
- 將術語、類型及型樣拖曳至種類窗格。如需相關資訊，請參閱主題 第 101 頁的『透過拖放建立種類』。

建立新的或重新命名種類

您可以建立空的種類，從而在其中新增概念及類型。您還可以重新命名您的種類。

若要建立新的空種類

1. 前往「種類」窗格。
2. 從功能表中，選擇**種類 > 建立空的種類**。即會開啟「種類內容」對話框。
3. 在「名稱」欄位中輸入此種類的名稱。
4. 按一下**確定**以接受名稱，然後關閉對話框。即會關閉對話框，並在窗格中顯示新的種類名稱。

現在您可以開始新增至此種類。如需相關資訊，請參閱主題 第 116 頁的『將描述子新增至種類』。

要重新命名種類

1. 選取一個種類，然後選擇**種類 > 重新命名種類**。即會開啟「種類內容」對話框。
2. 在「名稱」欄位中輸入此種類的新名稱。

3. 按一下**確定**以接受名稱，然後關閉對話框。即會關閉對話框，並在窗格中顯示新的種類名稱。

透過拖放建立種類

拖放技術是手動的，而不是基於演算法。您可以透過拖曳，在「種類」窗格中建立種類：

- 將概念、類型或型樣從「擷取結果」窗格擷取至「種類」窗格。
- 將概念從「資料」窗格擷取至「種類」窗格。
- 整個列從「資料」窗格至「種類」窗格。這將建立種類，由該列中包含的所有所擷取概念及型樣組成。

附註：「擷取結果」窗格支援多重選項，可協助拖放多個元素。

重要事項！ 您無法從「資料」窗格拖放未擷取自文字的概念。如果您要強制擷取在資料中找到的概念，則必須將此概念新增至類型。然後再次執行擷取。新的擷取結果將包含您剛剛新增的概念。然後，您可以將它用在種類中。如需相關資訊，請參閱主題 第 77 頁的『將概念新增至類型』。

若要使用拖放建立種類：

1. 從「擷取結果」窗格或「資料」窗格中，選取一個或多個概念、型樣、類型、記錄或部分記錄。
2. 按下滑鼠按鈕的同時將元素拖曳至現有種類或窗格區域，以建立新的種類。
3. 當您達到要釋放元素的區域時，鬆開滑鼠按鈕。即會將元素新增至「種類」窗格。修改的種類以特殊背景顏色顯示。此顏色稱為**種類意見背景**。如需相關資訊，請參閱主題 第 64 頁的『設定選項』。

附註：系統會自動命名所產生的種類。如果您要變更名稱，則可以進行重新命名。如需相關資訊，請參閱主題 第 100 頁的『建立新的或重新命名種類』。

如果您要查看指派種類的記錄，請在「種類」窗格中選取該種類。資料窗格會自動重新整理，並顯示該種類的所有記錄。

使用種類規則

您可以透過許多方法建立種類。其中一個方法是定義種類規則以表達構想。種類規則是根據利用所擷取概念、類型及型樣的邏輯表示式，以及布林運算子，自動將文件或記錄分類至種類的陳述式。例如，您可以撰寫表示式，表示包括在此種類中包含所擷取概念 `embassy` 但不包含 `argentina` 的所有記錄。

由於部分種類規則是在利用分組技術建置種類時自動產生，例如共生及概念根衍生（**種類 > 建置設定 > 進階設定：語言**），您也可以利用對文字及環境定義的種類理解，在編輯器中手動建立種類規則。每一個規則都連接至單一種類，以便每一個符合規則的文件或記錄都會在該種類中進行評分。

種類規則會協助加強文字挖掘結果的品質與生產力，並容許您利用更準確的特異性對回應進行分類，從而進一步定量分析。您的經驗與商業知識可讓您更好地瞭解您的資料與環境定義。您可以利用此瞭解將該知識轉換為種類規則，透過結合所擷取元素與布林邏輯，從而更有效且準確地分類您的文件或記錄。

能夠建立這些規則可讓您將商業知識分層至產品的擷取技術，從而提高編碼精確度、效率及生產力。

附註：如需規則如何符合文字的範例，請參閱第 106 頁的『種類規則範例』

種類規則語法

由於部分種類規則是在利用分組技術建置種類時自動產生，例如共生及概念根衍生（**種類 > 建置設定 > 進階設定：語言**），您也可以在編輯器中手動建立種類規則。每一個規則都是單一種類的描述子；因此，每一個符合規則的文件或記錄都會自動在該種類中進行評分。

附註：如需規則如何符合文字的範例，請參閱第 106 頁的『種類規則範例』

當您建立或編輯規則時，您必須在規則編輯器中開啟它。您可以新增概念、類型或型樣，以及使用萬用字元以延伸相符項。當您使用擷取概念、類型及型樣時，能夠從尋找所有相關概念獲益。

重要事項！ 若要避免一般錯誤，我們建議直接將概念從「擷取結果」窗格、「文字鏈結分析」窗格或「資料」窗格拖放到規則編輯器，或者在可能時透過環境定義功能表新增它們。

辨識概念、類型及型樣時，圖示會顯示在文字旁邊。

表 *19.* 擷取圖示

圖示	說明
	擷取的概念
	擷取的類型
	擷取的型樣

規則語法及運算子

下表包含您用來定義規則語法的字元。搭配使用這些字元與概念、類型及型樣以建立規則。

表 *20.* 受支援的語法

字元	說明
&	"and" 布林值。例如，a & b 包含 a 及 b，例如： - invasion & united states - 2016 & olympics - good & apple
|	"or" 布林值是內含的，表示如果找到任何或所有元素，則會相符。例如，a | b 包含 a 或 b，例如： - attack | france - condominium | apartment
!()	"not" 布林值。例如，!(a) 未包含 a。例如， !(good & hotel) 、assassination & !(austria) 或 !(gold) & !(copper)
*	萬用字元根據使用方式，代表從單一字元到完整單字的任何項。如需相關資訊，請參閱主題 第 104 頁的『在種類規則中使用萬用字元』。
()	表示式定界字元。首先評估括弧內的所有表示式。
+	用來形成順序特定型樣的型樣連接器。呈現時，必須使用方括弧。如需相關資訊，請參閱主題 第 103 頁的『在種類規則中使用 TLA 型樣』。
[]	如果您要根據種類規則內的所擷取 TLA 型樣相符，則需要型樣定界字元。方括弧內的內容是指 TLA 型樣，且根據簡式共生永不符合概念或類型。如果您未擷取此 TLA 型樣，則不會具有任何相符項。如需相關資訊，請參閱主題 第 103 頁的『在種類規則中使用 TLA 型樣』。如果您希望符合概念及類型而不是型樣，請使用方括弧。 附註：在較舊的版本中，使用方括弧圍繞種類建置技術產生的共生及同義字規則。在所有新版本中，現在方括弧指出顯示 TLA 型樣。共生技術及同義字產生的規則將封裝在括弧中，例如 (speaker systems|speakers)。

& 與 | 運算子是可交換的，例如 a & b = b & a 及 a | b = b | a。

使用反斜線跳出字元

如果您的概念包含也是語法字元的任何字元，則必須在該字元之前放置一個反斜線，以便適當地解譯規則。反斜線 (\) 字元用來跳出字元，否則具有特殊意義。拖放至編輯器時，會自動為您使用反斜線。

如果您想要下列規則語法字元不僅僅作為規則語法，則之前必須加上反斜線：
& ! | + < > ( ) [ ] *

例如，由於概念 r&d 包含 "and" 運算子 (&)，則在鍵入規則編輯器時需要反斜線，例如：r\&d。

在種類規則中使用 TLA 型樣

可以在種類規則中明確地定義文字鏈結分析型樣，以容許您取得甚至更多特定及環境定義結果。當您在種類規則中定義型樣時，會略過更簡單的概念擷取結果，以及只有根據所擷取文字鏈結分析型樣結果相符的文件及記錄。

重要事項！ 為了在種類規則中使用 TLA 型樣符合文件，您必須在已啟用文字鏈結分析的情況下執行擷取。種類規則將尋找在該處理程序期間找到的相符項。如果您未選擇在「文字挖掘」節點的「模型」標籤中探索 TLA 結果，則可以選擇在互動式階段作業內的擷取設定中啟用 TLA 擷取，然後重新擷取。如需相關資訊，請參閱 主題 第 70 頁的『擷取資料』。

使用方括弧定界。如果您要在種類規則內部使用 TLA 型樣，則必須含括在 [ ] 方括弧中。如果您希望根據所擷取 TLA 型樣進行比對，則需要型樣定界字元。由於種類規則可能包含類型、概念或型樣，因此方括弧向規則澄清方括弧內的內容是指擷取的 TLA 型樣。如果您未擷取此 TLA 型樣，則不會具有任何相符項。如果您在「種類」窗格中看到沒有方括弧的型樣，例如 apple + good，這可能表示型樣已直接新增至種類規則編輯器外部的種類。例如，如果您直接將概念型樣從文字鏈結分析視圖新增至種類，則它將不會顯示方括弧。然而，在種類規則內使用型樣時，您必須在種類規則內將型樣封裝在方括弧中，例如 [banana + !(good)]。

在型樣內使用 + 號。在 IBM SPSS Modeler Text Analytics 中，您最多可以具有 6-part 或 -slot 型樣。若要指出順序很重要，請使用 + 號連接每一個元素，例如 [company1 + acquired + company2]。以下順序很重要，因為它會變更公司所獲得的意義。順序並非由句子結構判定，而是由 TLA 型樣輸出的結構判定。例如，如果您具有文字 "*I love Paris*" 且想要擷取此構想，則 TLA 型樣很可能是 [paris + like] 或 [<Location> + <Positive>]，而不是 [<Positive> + <Location>]，因為預設意見資源通常將意見放置在 2 部分型樣中的第二個位置。因此它可能對於直接將型樣用作種類中的描述子以避免問題很有用。然而，如果您需要將型樣用作更複雜的陳述式的一部分，請特別注意「文字鏈結分析」視圖中呈現的型樣內元素的順序，因為順序對於是否能夠找到相符項非常重要。

例如，讓我們假設您具有以下兩個樣本文字表示式："I *like pineapple*" 及 "I hate *pineapple*. However, I *like* strawberries"。表示式 like & pineapple 將符合這兩個文字，因為它是概念表示式，而不是文字鏈結規則（未含括在方括弧中）。表示式 pineapple + like 只符合 "I *like pineapple*"，因為在第二段文字中，單字 *like* 與 *strawberries* 相關聯。

使用型樣分組。您可以使用自己的型樣簡化規則。讓我們假設您想要擷取以下三個表示式：cayenne peppers + like、chili peppers + like 及 peppers + like。您可以將它們分組至單一種類規則，例如 [* peppers & like]。如果您具有另一個表示式 hot peppers + good，則可以利用規則分組那四項，例如 [* peppers + <Positive>]。

型樣中的順序。為了更好地組織輸出，您利用產品安裝的範本中提供的文字鏈結分析規則嘗試使用同一順序輸出基本型樣，而無論句子中的單字順序為何。例如，如果您具有包含文字 "*Good presentations.*" 的記錄，且另

一筆記錄包含 "*the presentations were good*"，則兩段文字都透過同一規則相符，且在概念型樣結果中的輸出順序與 presentation + good 相同，而不是與 presentation + good 以及 good + presentation 相同。並且在如範例中的兩個插槽型樣中那樣，依預設，「意見」庫中指派給類型的概念將呈現在輸出的最後，例如 apple + bad。

表 *21.* 型樣語法與布林值使用

表示式	符合文件或記錄
[ ]	包含任何 TLA 型樣。如果您希望根據所擷取 TLA 型樣進行比對，則種類規則中需要型樣定界字元。方括弧內的內容是指 TLA 型樣，而不是簡式概念及類型。如果您未擷取此 TLA 型樣，則不會具有任何相符項。 如果您要建立未包括任何型樣的規則，則可以使用 !([ ])。
[a]	包含至少一個元素為 a（而無論其在型樣中的位置）的項目型樣。例如，[deal] 可以符合 [deal + good]，或者僅符合 [deal + .]
[a + b]	包含概念型樣。例如，[deal + good]。 附註：如果您只想要擷取此型樣，而不新增任何其他元素，我們建議您直接將型樣新增至種類，而不是制定規則。
[a + b + c]	包含概念型樣。+ 號表示相符元素的順序很重要。例如，[company1 + acquired + company2]。
[<A> + <B>]	包含在第一個插槽中具有類型 <A>，且在第二個插槽中具有類型 <B> 的所有型樣，並且正好有兩個插槽。+ 號表示相符元素的順序很重要。例如，[<Budget> + <Negative>]。 附註：如果您只想要擷取此型樣，而不新增任何其他元素，我們建議您直接將型樣新增至種類，而不是制定規則。
[<A> & <B>]	包含具有類型 <A> 及類型 <B> 的所有類型型樣。例如，[<Budget> & <Negative>]。永不擷取此 TLA 型樣；然而，如此撰寫時，它實際上等於 [<Budget> + <Negative>]|[<Negative> + <Budget>]。相符元素的順序並不重要。此外，其他元素可能使用型樣，但是它必須至少具有 <Budget> 及 <Negative>。
[a + .]	包含一個型樣，其中 a 是唯一的概念，並且該型樣中的任何其他插槽內都沒有任何項。例如，[deal + .] 符合唯一輸出是概念 deal 的概念型樣。如果您新增概念 deal 作為種類描述子，您將取得使用交易作為概念的所有記錄，且包括有關交易的正面陳述式。然而，使用 [deal + .] 將僅符合代表 deal，沒有任何其他關係或意見，且不符合 deal + fantastic 的那些記錄型樣結果。 附註：如果您只想要擷取此型樣，而不新增任何其他元素，我們建議您直接將型樣新增至種類，而不是制定規則。
[<A> + <>]	包含一個型樣，其中 <A> 是唯一的類型。例如，[<Budget> + <>] 符合唯一的輸出是類型為 <Budget> 的概念的型樣。 附註：僅當您將 <> 放置在類型型樣中的型樣 + 符號之後時，它才表示空的類型，例如 [<Budget> + <>]，而不是 [price + <>]。 附註：如果您只想要擷取此型樣，而不新增任何其他元素，我們建議您直接將型樣新增至種類，而不是制定規則。
[a + !(b)]	至少包含一個型樣，其中包括概念 a，但不包括概念 b。必須至少包括一個型樣。 例如，[price + !(high)] 或者對於類型，[!(<Fruit>|<Vegetable>) + <Positive>]
!([<A> & <B>])	不包含特定型樣。例如，!([<Budget> & <Negative>])。

附註：如需規則如何符合文字的範例，請參閱第 106 頁的『種類規則範例』

在種類規則中使用萬用字元

萬用字元可以新增至規則中的概念，以便延伸相符的功能。星號 * 萬用字元可以放置在單一之前及/或之後，以指出如何可以相符概念。有兩種類型的萬用字元使用：

- **附加字萬用字元**。這些萬用字元作為直接的字首或字尾，而不會使用任何空格分隔字串與星號。例如，operat* 可以符合 *operat*、*operate*、*operates*、*operations*、*operational* 等。
- **單字萬用字元**。這些萬用字元作為概念的字首或字尾，並在概念與星號之間加上空格。例如，* operation 可以符合 *operation*、*surgical operation*、*post operation* 等。此外，單字萬用字元可以與附加字萬用字元一起使用，例如，* operat* *，這可能符合 *operation*、*surgical operation*、*telephone operator*、*operatic aria* 等。如您在這個最後一個範例中看到的一樣，我們建議小心使用萬用字元，以便不會過於廣泛地強制轉型網路，並擷取不想要的相符項。

異常狀況！

- 萬用字元永不代表自己。例如，不會接受 (apple | *)。
- 萬用字元永不用來符合類型名稱。<Negative*> 根本不會符合任何類型名稱。
- 您可以透過萬用字元從相符的概念中過濾掉某些類型。系統會自動使用指派給概念的類型。
- 萬用字元永遠不能位於單字序列的中間，而無論它是否以單字 (open* account) 或獨立式元件 (open * account) 結尾或開頭。您不能在類型名稱中使用萬用字元。例如，word* word，例如 apple* recipe 不符合 applesauce recipe 或根本不符合任何其他項。然而，apple* * 將符合 *applesauce recipe*、*apple pie*、*apple* 等。在另一個範例中，word * word（例如 apple * toast）不符合 *apple cinnamon toast*，或者根本不符合任何項，因此星號出現在兩個其他單字之間。然而，apple * 將符合 *apple cinnamon toast*、*apple*、*apple pie* 等。

表 *22.* 萬用字元用法

表示式	符合文件或記錄
*apple	包含以所撰寫字母結尾，但可能具有任何數目的字母作為字首的概念。例如：*apple 以字母 *apple* 結尾，但可以採用字首，例如： - apple - pineapple - crabapple
apple*	包含以所撰寫字母開頭，但可能具有任何數目的字母作為字尾的概念。例如：apple* 以字母 *apple* 開頭，但是可以採用一個字尾或不採用任何字尾，例如： - apple - applesauce - applejack 例如，apple* & !(pear* | quince)（其包含以字母 apple 開頭的概念，但不是以字母 *pear* 開頭的概念，或者概念 quince）不符合：apple & quince 但可能符合： - applesauce - apple & orange
product	包含具有所撰寫字母 product，但可能具有任何數目的字母作為字首及/或字尾的概念。 例如：*product* 可能符合： - product - byproduct - unproductive

表 22. 萬用字元用法 (繼續)

表示式	符合文件或記錄
* loan	包含具有單字 loan，但可能是在之前放置另一個單字的複合字的概念。例如，* loan 可能符合： - loan - car loan - home equity loan 例如，[* delivery + <Negative>] 包含一個概念，在第一個位置以單字 delivery 結尾，且在第二個位置包含類型 <Negative>，可能符合下列概念型樣： - package delivery + slow - overnight delivery + late
event *	包含具有單字 event 的概念，但可能是在之後具有另一個單字的複合字。例如，event * 可能符合： - event - event location - event planning committee
* apple *	包含以任何單字開頭，之後是單字 apple，且之後可能具有另一個單字的概念。* 表示 0 或 n，因此它也符合 apple。例如，* apple * 可以符合： - gala applesauce - granny smith apple crumble - famous apple pie - apple 例如，[* reservation* * + <Positive>]，其在第一個位置包含具有單字 reservation 的概念（而無論其在概念中的位置為何），且在第二個位置中包含類型 <Positive>，可能符合概念型樣： - reservation system + good - online reservation + good

附註：如需規則如何符合文字的範例，請參閱『種類規則範例』

種類規則範例

為了協助示範規則如何根據用來進行表達的語法以不同的方式與記錄比對，請考量下列範例。

範例記錄

假設您具有兩筆記錄：

- **記錄 A**：「當我檢查我的錢包時，發現丟失 *5* 美元。」
- **記錄 B**：「在野餐區域找到 *5* 美元，但是丟了毯子。」

以下兩個表格顯示可能為概念及類型擷取的項目，以及概念型樣及類型型樣。

從範例擷取的概念及類型

表 23. 擷取概念及類型的範例

擷取的概念	概念鍵入為
wallet	<Unknown>
missing	<Negative>
USD5	<Currency>
blanket	<Unknown>
picnic area	<Unknown>

從範例擷取的 TLA 型樣

表 *24.* 擷取 *TLA* 型樣輸出的範例

擷取的概念型樣	擷取的類型型樣	從記錄
picnic area + .	<Unknown> + <>	記錄 B
wallet + .	<Unknown> + <>	記錄 A
blanket + missing	<Unknown> + <Negative>	記錄 B
USD5 + .	<Currency> + <>	記錄 B
USD5 + missing	<Currency> + <Negative>	記錄 A

種類規則相符的可能性

下表包含可以在種類規則編輯器中輸入的部分語法。並非這裡的所有規則都適用，且並非全部都符合相同的記錄。請參閱不同的語法如何影響相符的記錄。

表 *25.* 樣本規則

規則語法	結果
USD5 & missing	符合記錄 A 與 B，因為它們都包含所擷取的概念 missing 及所擷取的概念 USD5。這相當於： (USD5 & missing)
missing & USD5	符合記錄 A 與 B，因為它們都包含所擷取的概念 missing 及所擷取的概念 USD5。這相當於： (missing & USD5)
missing & <Currency>	符合記錄 A 與 B，因為它們都包含所擷取的概念 missing，且一個概念符合類型 <Currency>。這相當於： (missing & <Currency>)
<Currency> & missing	符合記錄 A 與 B，因為它們都包含所擷取的概念 missing，且一個概念符合類型 <Currency>。這相當於： (<Currency> & missing)
[USD5 + missing]	符合 A，但是**不**符合 B，因為記錄 B 未產生任何包含 USD5 + missing 的 TLA 型樣輸出（請參閱上表）。這相當於 TLA 型樣輸出： USD5 + missing
[missing + USD5]	既不符合記錄 A 也不符合記錄 B，因為沒有任何所擷取 TLA 型樣（請參閱上表）符合這裡表達的順序，且第一個位置中具有 missing。這相當於 TLA 型樣輸出： USD5 + missing
[missing & USD5]	符合 A，但**不**符合 B，因為未從記錄 B 擷取任何此類 TLA 型樣。使用字元 & 指出相符時該順序不重要；因此，此規則會尋找符合 [missing + USD5] 或 [USD5 + missing] 的型樣。只有記錄 A 的 [USD5 + missing] 具有相符項。
[missing + <Currency>]	既不符合記錄 A 也不符合記錄 B，因為沒有任何所擷取 TLA 型樣符合此順序。這沒有相當項，因為 TLA 僅基於術語 (USD5 + missing) 或類型 (<Currency> + <Negative>)，但是未混合概念及類型。
[<Currency> + <Negative>]	符合記錄 A 但不符合記錄 B，因為未從記錄 B 擷取任何 TLA 型樣。這相當於 TLA 輸出： <Currency> + <Negative>

表 25. 樣本規則 (繼續)

規則語法	結果
[<Negative> + <Currency>]	既不符合記錄 A 也不符合記錄 B，因為沒有任何所擷取 TLA 型樣符合此順序。在 Opinions 範本中，依預設，找到具有 *opinion* 的 *topic* 時，*topic* (<Currency>) 佔用第一個插槽位置，*opinion* (<Negative>) 佔用第二個插槽位置。

建立種類規則

當您建立或編輯規則時，必須在規則編輯器中開啟規則。您可以新增概念、類型或型樣，以及使用萬用字元以延伸相符項。當您使用已辨識的概念、類型及型樣時，您會因它找到所有相關概念而獲益。例如，當您使用概念時，其所有相關聯術語、複數形式及同義字也符合規則。同樣地，當您使用類型時，規則還會擷取其所有概念。

您可以透過編輯現有規則或用滑鼠右鍵按一下種類名稱並選擇**建立規則**，從而開啟規則編輯器。

您可以使用快速功能表、拖放，或者手動將概念、類型及型樣輸入編輯器。然後，結合這些項與布林運算子（&、!()、|）及方括弧，以形成您的規則表示式。為了避免一般錯誤，我們建議直接將概念從「擷取結果」窗格或「資料」窗格拖放至規則編輯器。請特別注意規則的語法以避免錯誤。如需相關資訊，請參閱主題 第 101 頁的『種類規則語法』。

附註：如需規則如何符合文字的範例，請參閱第 106 頁的『種類規則範例』。

若要建立規則

1. 如果您尚未擷取任何資料，或者您的擷取過期，請立即執行。如需相關資訊，請參閱主題 第 70 頁的『擷取資料』。

   附註：如果您過濾擷取後，不再有任何概念可見，則當您嘗試建立或編輯種類規則時，會顯示一則錯誤訊息。為了防止發生此情況，請修改您的擷取過濾器，以便概念可用。
2. 在「種類」窗格中，選取您要在其中新增規則的種類。
3. 從功能表中，選擇**種類 > 建立規則**。即會在視窗中開啟種類規則編輯器窗格。
4. 在「規則名稱」欄位中，輸入規則的名稱。如果您未提供名稱，則表示式將自動用作名稱。您可以稍後重新命名此規則。
5. 在較大的表示式文字欄位中，您可以：
   - 直接在欄位中輸入文字，或者從另一個窗格進行拖放。僅使用所擷取的概念、類型及型樣。例如，如果您輸入單字 cats，但是「擷取結果」窗格中僅顯示單數形式 cat，則編輯器將無法辨識 cats。在這個最後一種情況下，單數形式可能自動包括複數，否則您可以使用萬用字元。如需相關資訊，請參閱主題 第 101 頁的『種類規則語法』。
   - 選取您要新增至規則的概念、類型或型樣，並使用功能表。
   - 新增布林運算子以將規則中的元素鏈結在一起。使用工具列按鈕，以將型樣的 "and" 布林值 **&**、"or" 布林值 **|**、"not" 布林值 **!()**、括弧 **()** 及方括弧 **[ ]** 新增至您的規則。
6. 按一下**測試規則**按鈕，以驗證您的規則是否形式完整。如需相關資訊，請參閱主題 第 101 頁的『種類規則語法』。找到的文件或記錄數顯示在文字**測試結果**旁邊的括弧中。在此文字的右側，您可以看到規則中已辨識的元素，或者任何錯誤訊息。如果類型、型樣或概念旁邊的圖形顯示紅色問號，這指出元素不符合任何已知的擷取。如果它不符合，則規則將不會找到任何結果。
7. 若要測試規則的某個部分，請選取該部分，然後按一下**測試選擇**。

8. 請進行必要的變更，並在您發現問題時重新測試規則。
9. 完成後，按一下**儲存並關閉**以再次儲存您的規則，然後關閉編輯器。新的規則名稱顯示在種類中。

編輯及刪除規則

已建立並儲存規則之後，您隨時可以編輯該規則。如需相關資訊，請參閱主題 第 101 頁的『種類規則語法』。

如果您不再想要規則，則可以進行刪除。

要編輯規則

1. 在「種類定義」對話框的「描述子」表格中，選取規則。
2. 從功能表中，選擇**種類 > 編輯規則**，或者按兩下規則名稱。即會利用選取的規則開啟編輯器。
3. 利用擷取結果及工具列按鈕對規則進行任何變更。
4. 重新測試您的規則，以確保它會傳回預期的結果。
5. 按一下**儲存並關閉**，再次儲存您的規則並關閉編輯器。

要刪除規則

1. 在「種類定義」對話框的「描述子」表格中，選取規則。
2. 從功能表中，選擇**編輯 > 刪除**。即會從種類刪除規則。

匯入及匯出預先定義的種類

如果您將自己的種類儲存在 Microsoft Excel (*.*xls*, *.*xlsx*) 檔中，則可以將它們匯入 IBM SPSS Modeler Text Analytics。

您還可以將開啟的互動式工作台階段作業中的種類匯出至 Microsoft Excel（*.*xls*、*.*xlsx*）檔案。當您匯出種類時，可以選擇包括或排除部分其他資訊，例如描述子及分數。如需相關資訊，請參閱主題 第 112 頁的『匯出種類』。

如果預先定義的種類沒有代碼或者您想要新的代碼，則可以透過從功能表中選擇**種類 > 管理種類 > 自動產生代碼**，為種類窗格中的一組種類產生一組新的代碼。這將移除所有現有代碼，並全部自動重新編號。

匯入預先定義的種類

您可以將預先定義的種類匯入至 IBM SPSS Modeler Text Analytics。匯入之前，確保預先定義的種類檔案位於 Microsoft Excel（*.*xls*、*.*xlsx*）檔案中，並使用其中一種受支援的格式進行結構化。您還可以選擇讓產品自動為您偵測格式。支援的格式如下：

- **純文字清單格式**：如需相關資訊，請參閱主題第 110 頁的『純文字清單格式』。
- **壓縮格式**：如需更多資訊，請參閱主題第 111 頁的『壓縮格式』。
- **縮排的格式**：如需更多資訊，請參閱主題第 111 頁的『縮排的格式』。

若要匯入預先定義的種類

1. 從互動式工作台功能表中，選擇**種類 > 管理種類 > 匯入預先定義的種類**。即會顯示「匯入預先定義的種類」精靈。
2. 從「查看範圍」下拉清單中，選取檔案所位於的磁碟機及資料夾。
3. 從清單中選取檔案。檔案的名稱即會顯示在「檔案名稱」文字框中。
4. 從清單中選取包含預先定義的種類的工作表。工作表名稱即會顯示在「工作表」欄位中。
5. 若要開始選擇資料格式，請按**下一步**。

6. 選擇檔案的格式，或者選擇選項以容許產品嘗試自動偵測格式。自動偵測最適合最常見的格式。
   - **純文字清單格式**：如需相關資訊，請參閱主題『純文字清單格式』。
   - **壓縮格式**：如需更多資訊，請參閱主題第 111 頁的『壓縮格式』。
   - **縮排的格式**：如需更多資訊，請參閱主題第 111 頁的『縮排的格式』。
7. 若要定義其他匯入選項，請按**下一步**。如果您選擇自動偵測格式，則系統會將您引導至最終步驟。
8. 如果一個或多個列包含直欄標頭或者其他額外的資訊，請在**開始匯入所在列**選項中選取您要開始匯入的列號。例如，如果您的種類名稱從第 7 列開始，則必須為此選項輸入數字 7，從而正確地匯入檔案。
9. 如果您的檔案包含種類代碼，請選擇選項**包含種類代碼**。這樣做會協助精靈適當地辨識您的資料。
10. 檢閱以顏色編碼的資料格及圖註，從而確保已正確地識別資料。在檔案中偵測到的錯誤會以紅色顯示，並在格式預覽表格下方進行參照。如果選取錯誤的格式，請返回並選擇另一個。如果您需要對檔案進行更正，請進行那些變更，並透過再次選取該檔案以重新啟動精靈。您必須先更正所有錯誤，然後才能完成精靈。
11. 若要檢閱匯入的種類集及子種類集，以及定義如何為這些種類建立描述子，請按**下一步**。
12. 檢閱將在表格中匯入的種類集。如果您未看到預期作為描述子看到的關鍵字，則可能是匯入期間未辨識它們。請確保它們已適當地加上字首，並顯示在正確的資料格中。
13. 選擇您希望如何處理階段作業中預先存在的種類。
    - **取代所有現有種類**。此選項會清除所有現有種類，然後在原位置僅使用新匯入的種類。
    - **附加至現有種類**。此選項將匯入種類，並將任何共用種類與現有種類進行合併。新增至現有種類時，您需要判定希望如何處理任何重複項。一個選擇（選項：**合併**）是將所匯入種類與現有種類進行合併（如果它們共用種類名稱）。另一個選擇（選項：**從匯入中排除**）是禁止匯入種類（如果已存在具有相同名稱的種類）。
14. **匯入關鍵字作為描述子**可讓您匯入在資料中識別的關鍵字作為相關聯種類的描述子。
15. **透過衍生描述子延伸種類**可讓您從代表種類、子種類名稱的單字，以及/或組成註釋的單字中產生描述子。如果單字符合確切的結果，會將這些單字作為描述子新增至種類。種類名稱或註釋詳細且是敘述性時，此選項會產生最佳結果。這是用來產生種類描述子的快速方法，這會啟用種類以擷取包含那些描述子的記錄。
    - **從**欄位可讓您從衍生描述子的文字、名稱或種類及子種類、註釋中的單字中進行選取。
    - **作為**欄位可讓您選擇以概念或 TLA 樣式的形式建立這些描述子。如果未進行 TLA 擷取，則會在此精靈中停用**樣式**的選項。
16. 若要將預先定義的種類匯入至「種類」窗格，請按一下**完成**。

純文字清單格式

在純文字清單格式中，只有一個最上層的種類，而沒有任何階層，表示沒有任何子種類或子網路。種類名稱位於單一直欄中。

此格式的檔案中可以包含下列資訊：

- 選用**代碼**直欄包含唯一識別每一個種類的數值。如果您指定資料檔包含代碼（**內容設定**步驟中的**包含種類代碼**選項），則包含每一個種類之唯一代碼的直欄必須直接存在於種類名稱左側的資料格中。如果您的資料未包含代碼，但是您想要稍後建立部分代碼，則一律可以稍後產生代碼（**種類 > 管理種類 > 自動產生代碼**）。
- 必要**種類名稱**直欄包含種類的所有名稱。利用此格式進行匯入需要此直欄。
- 選用**註釋**位於種類名稱右側的資料格中。這個註釋包含說明您的種類/子種類的文字。
- 選用**關鍵字**可以作為種類的描述子匯入。為了進行辨識，這些關鍵字必須存在於直接位於相關聯種類/子種類名稱下方的資料格中，並且關鍵字必須使用底線 (_) 字元作為字首，例如 `_firearms, weapons / guns`。

關鍵字資料格可以包含一個或多個用來說明每一個種類的單字。這些單字將作為描述子匯入，或者根據您在精靈的最後一步指定的項目予以忽略。稍後，會將描述子與從文字擷取的結果進行比較。如果找到相符項目，則會將該記錄或文件評分至包含此描述子的種類。

表 *26.* 具有代碼、關鍵字及註釋的純文字清單格式

直欄 A	**直欄 B**	**直欄 C**
種類代碼（選用）	種類名稱	註釋
	_描述子/關鍵子清單（選用）	

壓縮格式

壓縮格式的結構化類似於純文字清單格式，除了壓縮格式與階層式種類搭配使用。因此，需要代碼層次直欄，以定義每一個種類及子種類的階層式層次。

此格式的檔案中可以包含下列資訊：

- 必要的**代碼層次**直欄包含數字，指出該列中後續資訊的階層式位置。例如，如果指定值 1、2 或 3，並且您同時具有種類與子種類，則 1 用於種類，2 用於子種類，且 3 用於子子種類。如果您只有種類與子種類，則 1 用於種類，2 用於子種類。依此類推，直到想要的種類深度。
- 選用**代碼**直欄包含唯一識別每一個種類的值。如果您指定資料檔包含代碼（**內容設定**步驟中的**包含種類代碼**選項），則包含每一個種類之唯一代碼的直欄必須直接存在於種類名稱左側的資料格中。如果您的資料未包含代碼，但是您想要稍後建立部分代碼，則一律可以稍後產生代碼（**種類 > 管理種類 > 自動產生代碼**）。
- 必要**種類名稱**直欄包含種類及子種類的所有名稱。利用此格式進行匯入需要此直欄。
- 選用**註釋**位於種類名稱右側的資料格中。這個註釋包含說明您的種類/子種類的文字。
- 選用**關鍵字**可以作為種類的描述子匯入。為了進行辨識，這些關鍵字必須存在於直接位於相關聯種類/子種類名稱下方的資料格中，並且關鍵字必須使用底線 (_) 字元作為字首，例如 `_firearms, weapons / guns`。關鍵字資料格可以包含一個或多個用來說明每一個種類的單字。這些單字將作為描述子匯入，或者根據您在精靈的最後一步指定的項目予以忽略。稍後，會將描述子與從文字擷取的結果進行比較。如果找到相符項目，則會將該記錄或文件評分至包含此描述子的種類。

表 *27.* 具有代碼的壓縮格式範例

直欄 A	**直欄 B**	**直欄 C**
階層式代碼層次	種類代碼（選用）	種類名稱
階層式代碼層次	子種類代碼（選用）	子種類名稱

表 *28.* 無代碼的壓縮格式範例

直欄 A	**直欄 B**
階層式代碼層次	種類名稱
階層式代碼層次	子種類名稱

縮排的格式

在縮排的檔案格式中，內容為階層式，這表示它包含種類以及一個或多個層次的子種類。此外，其結構已縮排以表示此階層。檔案中的每一列都包含一個種類或子種類，但是子種類從種類縮排，而所有子子種類從子種類縮排，依此類推。您可以在 Microsoft Excel 中手動建立此結構，或者使用從另一個產品匯出的結構，並儲存至 Microsoft Excel 格式。

- **最上層種類代碼及種類名稱**分別佔用直欄 A 與 B。或者，如果未呈現任何代碼，則種類名稱位於直欄 A 中。
- **子種類代碼及子種類名稱**分別佔用直欄 B 與 C。或者，如果未呈現任何代碼，則子種類名稱位於直欄 B 中。子種類是種類的成員。如果您沒有最上層種類，則不能具有子種類。

表 *29.* 具有代碼的縮排結構

直欄 A	**直欄 B**	**直欄 C**	**直欄 D**
種類代碼（選用）	種類名稱		
	子種類代碼（選用）	子種類名稱	
		子子種類代碼（選用）	子子種類名稱

表 *30.* 無代碼的縮排結構

直欄 A	**直欄 B**	**直欄 C**
種類名稱		
	子種類名稱	
		子子種類名稱

此格式的檔案中可以包含下列資訊：

- 選用**代碼**必須是唯一識別每一個種類或子種類的值。如果您指定資料檔包含代碼（**內容設定**步驟中的**包含種類代碼**選項），則包含每一個種類或子種類之唯一代碼的直欄必須直接存在於種類/子種類名稱左側的資料格中。如果您的資料未包含代碼，但是您想要稍後建立部分代碼，則一律可以稍後產生代碼（**種類 > 管理種類 > 自動產生代碼**）。
- 每一個種類及子種類的必要**名稱**。子種類必須從種類縮排一個資料格，位於右側的一個單獨列中。
- 選用**註釋**位於種類名稱右側的資料格中。這個註釋包含說明您的種類/子種類的文字。
- 選用**關鍵字**可以作為種類的描述子匯入。為了進行辨識，這些關鍵字必須存在於直接位於相關聯種類/子種類名稱下方的資料格中，並且關鍵字必須使用底線 (_) 字元作為字首，例如 `_firearms, weapons / guns`。關鍵字資料格可以包含一個或多個用來說明每一個種類的單字。這些單字將作為描述子匯入，或者根據您在精靈的最後一步指定的項目予以忽略。稍後，會將描述子與從文字擷取的結果進行比較。如果找到相符項目，則會將該記錄或文件評分至包含此描述子的種類。

重要事項！ 如果您在一個層次使用代碼，則必須包括每一個種類及子種類的代碼。否則，匯入處理程序將失敗。

匯出種類

您還可以將開啟的互動式工作台階段作業中的種類匯出至 Microsoft Excel（**.xls*、**.xlsx*）檔案格式。該資料將主要匯出自「種類」窗格的現行內容或種類內容。因此，如果您計劃還匯出 **Docs.** 分數值，我們建議您再次評分。

表 *31.* 種類匯出選項

一律匯出...	**選擇性地匯出...**
• 種類節點（如果存在的話） • 種類（及子種類）名稱 • 代碼層次（如果存在的話）（純文字/壓縮格式） • 直欄標題（純文字/壓縮格式）	• **文件**。分數 • 種類註釋 • 描述子名稱 • 描述子計數

重要事項！ 當您匯出描述子時，它們會轉換為字串，並在字首加上底線。如果您重新匯入至此產品，則無法識別型樣描述子、種類規則描述子與一般概念描述子。如果您要在此產品中重複使用這些種類，我們強烈建議您改為建立文字分析套件 (TAP) 檔案，因為 TAP 格式會將所有描述子保留為目前定義的狀況，以及保留所有種類、代碼及使用的語言資源。可以同時在 IBM SPSS Modeler Text Analytics 與 IBM SPSS Text Analytics for Surveys 中使用 TAP 檔。如需相關資訊，請參閱主題 『使用文字分析套件』。

要匯出預先定義的種類

1. 從互動式工作台功能表中，選擇**種類 > 管理種類 > 匯出種類**。即會顯示「匯出種類」精靈。
2. 選擇位置，並輸入將匯出的檔案名稱。
3. 在「檔名」文字框中輸入輸出檔的名稱。
4. 若要選擇您要用來匯出種類資料的格式，請按**下一步**。
5. 從下列格式中進行選擇：
   - **純文字或壓縮清單格式**：如需相關資訊，請參閱主題第 110 頁的『純文字清單格式』。純文字清單不包含任何子種類。如需相關資訊，請參閱主題 第 111 頁的『壓縮格式』。壓縮清單格式包含階層式種類。
   - **縮排的格式**：如需更多資訊，請參閱主題第 111 頁的『縮排的格式』。
6. 若要開始選擇要匯出的內容並檢閱提出的資料，請按**下一步**。
7. 檢閱所匯出檔案的內容。
8. 選取或取消選取要匯出的其他內容設定，例如**註釋**或**描述子名稱**。
9. 若要匯出種類，請按一下**完成**。

使用文字分析套件

文字分析套件（也稱為 TAP）可用作文字回應種類的範本。使用 TAP 是透過最小人為介入分類文字資料的簡單方法，原因是它包含快速及自動為大量記錄編製代碼所需要的預先建置的種類集及文字資源。使用語言資源，系統會分析及發掘文字資料以擷取主要概念。根據在文字中找到的主要概念及型樣，可以將記錄分類到您在 TAP 中選取的種類集。您可以建立自己的 TAP 或更新一個。

TAP 由下列元素組成：

- **種類集**。種類集主要由預先定義的種類、種類代碼、每一個種類的描述子以及整個種類集的名稱組成。描述子是語言元素（概念、類型、型樣及規則），例如術語 *cheap* 或型樣 *good price*。描述子用來定義種類，以便在文字符合任何種類描述子時，文件或記錄會放置到種類中。
- **語言資源**。語言資源是一組檔案庫及進階資源，已調整以擷取主要概念及型樣。而這些擷取概念及型樣用作描述子，容許將記錄放置到種類集的種類中。

您可以建立自己的 TAP、更新一個或載入文字分析套件。

選取 TAP 並選擇的種類集之後，SPSS Modeler Text Analytics 可以擷取並分類您的記錄。

註：可以透過可交換的方式在 SPSS Text Analytics for Surveys 與 SPSS Modeler Text Analytics 之間建立及使用 TAP。然而，請注意，根據您是否直接從 SPSS Modeler Text Analytics 載入文字分析套件 (TAP)，或者您是否從 IBM SPSS Text Analytics for Surveys 載入 TAP，在 SPSS Modeler Text Analytics 中評分規則可能有所不同。我們建議您使用在 SPSS Modeler Text Analytics 內建立的 TAP；這是因為在 IBM SPSS Text Analytics for Surveys 中建立的 TAP 可能係利用不同版本的語言資源進行建立。

建立文字分析套件

只要您具有至少含一個種類及部分資源的階段作業，就可以從開啟的互動式工作台階段作業的內容建立文字分析套件 (TAP)。可以在 TAP 中建立的種類及描述子（概念、類型、規則或 TLA 型樣輸出）集，並在資源編輯器中開啟所有語言資源。

您可以看到為其建立資源的語言。語言係在範本編輯器或資源編輯器的「進階資源」標籤中進行設定。

若要建立文字分析套件

1. 從功能表中，選擇**檔案 > 文字分析套件 > 建立套件**。即會開啟「建立套件」對話框。
2. 瀏覽至您將儲存 TAP 的目錄。依預設，TAP 會儲存到產品安裝目錄的 \TAP 子目錄。
3. 在**檔名**欄位中輸入 TAP 的名稱。
4. 在**套件標籤**欄位中輸入標籤。當您輸入檔名時，此名稱會自動顯示為標籤，但是您可以變更此標籤。
5. 若要從 TAP 中排除一個種類集，請取消選取**包括**勾選框。這樣做將確保它未新增至套件。依預設，TAP 中針對每個問題包括一個種類集。TAP 中必須一律至少具有一個種類集。
6. 重新命名任何種類集。依預設，**新建種類集**直欄包含一般名稱，這透過將 Cat_ 字首新增至文字變數名稱予以產生。在資料格中按一下即可讓名稱可編輯。輸入或按一下其他位置即會套用重新命名。如果您重新命名種類集，則名稱只會在 TAP 中變更，而不會在開啟的階段作業中變更變數名稱。
7. 如果想要的話，使用種類集表格右側的方向鍵對種類集重新排序。
8. 按一下**儲存**，以建立文字分析套件。即會關閉對話框。

載入文字分析套件

配置文字採礦建模節點時，您必須指定將在擷取期間使用的資源。您可以不選擇資源範本，而是選擇文字分析套件 (TAP) 或 SPSS Text Analytics for Surveys 專案 (.tas)，從而不僅將其資源，還將種類集複製到節點。如果您選取 .tas 檔，它將轉換為 TAP。

當以互動方式建立種類模型時，由於您可以將種類集用作分類的起始點，因此 TAP 最有興趣。當您執行串流時，會啟動互動式工作台階段作業，並且這個種類集會顯示在「種類」窗格中。透過此方法，您可以立即使用這些種類對文件及記錄進行評分，然後繼續精簡、建置及展開這些種類，直到它們滿足您的需要為止。如需相關資訊，請參閱 第 84 頁的『用來建立種類的方法及策略』。

從第 14 版開始，您還可以按一下**載入**，然後選擇 TAP，從而查看為其定義此 TAP 中資源的語言。

載入 TAP 或 TAS

1. 編輯文字採礦建模節點。
2. 在「模型」標籤上，選擇**複製資源來源**區段中的文字分析套件。
3. 按一下**載入**。即會開啟「載入文字分析套件」對話框。
4. 瀏覽至包含您要複製到節點之資源及種類集的 TAP 或 SPSS Text Analytics for Surveys 專案 (.tas) 的位置。依預設，它們會儲存至您產品安裝目錄的 \TAP 子目錄中。
5. 在**檔名**欄位中輸入 TAP 的名稱。系統會自動顯示標籤。
6. 選取您要使用的種類集。這是將顯示在互動式工作台階段作業中的種類集。然後，您可以手動調整並改進這些種類，或者使用「建置」或「延伸」種類選項。
7. 按一下**載入**以將文字分析套件或 SPSS Text Analytics for Surveys 專案的內容複製到節點。即會關閉對話框。當載入內容時，內容會複製到節點；因此，除非您明確更新及重新載入內容，否則將不會反映您對外部資源及種類所做的任何變更。

更新文字分析套件

如果您改良種類集、語言資源或建立全新的種類集，則可以更新文字分析套件 (TAP)，讓稍後重複使用這些改良更為容易。若要這樣做，您必須在包含您要放置於 TAP 中之資訊的開啟的階段作業中。當您更新時，您可以選擇附加種類集、取代資源、變更套件層次或重新命名/重新排序種類集。

更新文字分析套件

1. 從功能表中，選擇**檔案 > 文字分析套件 > 更新套件**。即會顯示「更新套件」對話框。
2. 瀏覽至包含您要更新之文字分析套件的目錄。
3. 在**檔名**欄位中輸入 TAP 的名稱。
4. 若要將 TAP 內的語言資源取代為現行階段作業的語言資源，請選取**將此套件中的資源取代為開啟的階段作業中的資源**選項。這通常對更新語言資源有意義，因為它們用來擷取用於建立種類定義的主要概念及型樣。具有最近的語言資源可讓您在分類記錄時取得最佳結果。如果您未選取此選項，則已在套件中的語言資源會保持不變更。
5. 若要僅更新語言資源，請確保您選取**將此套件中的資源取代為開啟的階段作業中的資源**選項，並且僅選取已在 TAP 中的現行種類集。
6. 若要將開啟的階段作業中的新種類集納入 TAP，請選取要新增之每一個種類集的勾選框。您可以新增一個、多個種類集，或者不新增任何種類集。
7. 若要從 TAP 中移除種類集，請取消選取對應的**併入**勾選框。由於您正在新增改良的種類集，您可能選擇移除已在 TAP 中的種類集。若要這樣做，請在「現行種類集」直欄中取消選取對應種類集的**併入**勾選框。TAP 中必須一律至少具有一個種類集。
8. 必要的話，重新命名種類集。在資料格中按一下即可讓名稱可編輯。輸入或按一下其他位置即會套用重新命名。如果您重新命名種類集，則名稱只會在 TAP 中變更，而不會在開啟的階段作業中變更變數名稱。如果兩個種類集具有相同的名稱，則該名稱將以紅色顯示，直到您更正重複項為止。
9. 若要建立新的套件，且階段作業內容合併所選取 TAP 的內容，請按一下**另存為新檔**。即會顯示「另存為文字分析套件」。請參閱下列指示。
10. 按一下**更新**，以儲存您對所選取 TAP 進行的變更。

儲存文字分析套件

1. 瀏覽至您將儲存 TAP 檔的目錄。依預設，TAP 檔會儲存到安裝目錄的 \TAP 子目錄。
2. 在**檔名**欄位中輸入 TAP 檔的名稱。
3. 在**套件標籤**欄位中輸入標籤。當您輸入檔名時，此名稱會自動儲存為標籤。然而，您可以重新命名此標籤。您必須具有一個標籤。
4. 按一下**儲存**以建立新的套件。

編輯及精簡種類

建立部分種類之後，您將總是想要進行檢查，並進行部分調整。除了精簡語言資源之外，您還應該透過尋找方法結合或清除其定義，以及檢查部分已分類的文件或記錄，從而檢閱您的種類。您還可以檢閱種類中的文件或記錄並進行調整，以便利用擷取細微差別及區別的方法定義種類。

您可以使用內建、自動化的種類建置技術來建立種類；然而，您很可能想要對這些種類執行數個 tweak。使用一個或多個技術之後，許多新的種類會顯示在視窗中。然後，您可以在種類中檢閱資料並進行變更，直到您適應種類定義為止。如需相關資訊，請參閱主題 第 87 頁的『關於種類』。

以下是用來精簡種類的部分選項，下列頁面中說明了其中大部分選項：

將描述子新增至種類

使用自動化技術之後，您將很可能仍具有未在任何種類定義中使用的延伸結果。您應該在「延伸結果」窗格中檢閱此清單。如果您找到要移至某個種類的元素，則可以將它們新增至現有種類或新的種類。

若要將概念或類型新增至種類

1. 從「擷取結果」及「資料」窗格內，選取您要新增至新種類或現有種類的元素。
2. 從功能表中，選擇**種類 > 新增至種類**。「所有種類」對話框會顯示一組種類。選取您要新增所選取元素的種類。如果您要將元素新增至新種類，請選取**新種類**。新的種類會使用選取的第一個元素的名稱出現在「種類」窗格中。

編輯種類描述子

已建立部分種類之後，您可以開啟每一個種類，以查看組成其定義的所有描述子。在「種類定義」對話框內部，您可以對種類描述子進行許多編輯。此外，如果種類顯示在種類樹狀結構中，您也可以在這裡使用它們。

要編輯種類

1. 選取您要在「種類」窗格中編輯的種類。
2. 從功能表中，選擇**檢視 > 種類定義**。即會開啟「種類定義」對話框。
3. 選取您要編輯的描述子，然後按一下對應的工具列按鈕。

下表說明您可以用來編輯種類定義的每一個工具列按鈕。

表 *32.* 工具列按鈕及說明

圖示	說明
	從種類中刪除所選取的描述子。
	將所選取的描述子移至新的或現有種類。
	將 & 種類規則形式的所選取描述子移至種類。如需相關資訊，請參閱主題 第 101 頁的『使用種類規則』。
	將每一個所選取描述子作為其自己的新種類移動
 **顯示**	根據所選取的描述子，更新「資料」窗格及「視覺化」窗格中顯示的內容

移動種類

如果您要將種類放置到另一個現有種類中，或者將描述子移至另一個種類中，您可以移動它。

要移動種類

1. 在「種類」窗格中，選取您要移至另一個種類的種類。
2. 從功能表中，選擇**種類 > 移至種類**。功能表會在清單頂端，呈現含最近建立之種類的一組種類。選取您要移動所選取概念的目標種類的名稱。

- 如果您看到所尋找的名稱，請進行選取，所選取的元素即會新增至該種類。
- 如果您未看到，請選取**更多**以顯示「所有種類」對話框，並從清單中選取該種類。

壓縮種類

當您具有含種類及子種類的階層式種類時，可以壓縮結構。當您壓縮種類時，該種類之子種類中的所有描述子都會移至所選取的種類，並且現在刪除空的子種類。透過此方法，現在，用來符合子種類的所有文件都分類至所選取種類。

要壓縮種類

1. 在「種類」窗格中，選取您要壓縮的種類（最上層或子種類）。
2. 從功能表中，選擇**種類 > 壓縮種類**。系統會刪除子種類，並將描述子合併至所選取種類。

合併或結合種類

如果您要將兩個或更多現有種類結合至新的種類，您可以合併它們。當您合併種類時，會使用一般名稱建立新的種類。種類描述子中使用的所有概念、類型及型樣都移至這個新的種類。稍後，您可以透過編輯種類內容重新命名此種類。

要合併種類或部分種類

1. 在「種類」窗格中，選取您要合併在一起的元素。
2. 從功能表中，選擇**種類 > 合併種類**。即會顯示「種類內容」對話框，您可以在其中輸入新建立的種類的名稱。所選取的種類會作為子種類合併至新的種類。

強制將文件移入種類

強制將文件移入種類或從種類中移出可讓您置換由自動種類建置技術建立的種類定義，而無需變更實際的種類定義。您可能會發現，雖然文件包含用來定義特定種類的術語，但文件本身不應在該種類中。在此情況下，您可以強制從該種類中移出文件，而不必從種類定義中移除術語。

在以下特殊案例中會使用強制：文件適合（或不適合）某個種類，但基於某種原因（例如，它包含特殊術語）被指派（或未指派）給該種類。例如，當回應者在回應時使用諷刺（例如，「披薩太棒了，我確信所有人都愛燒焦了的冷披薩。」）時可能會發生此情況。假設您有一個種類稱為 `Pos: [<Food> + <Positive>]`，用來擷取關於飯店所供應食物的正面意見，這樣上述回應便會指派給該種類。在此情況下，您可能想要強制將此回應移出該種類。

強制移入或移出種類

1. 從「資料」窗格中，選取您要強制移入或移出特定種類的文件。
2. 從功能表中選擇**種類** > **強制移入**或**種類** > **強制移出**。子功能表會顯示您可以從中選取的種類清單。
3. 選取您要強制將此文件移入或從中移出的種類。如果您建立了多個種類，部分種類可能在子功能表中不可見。

- 在此情況下，選取子功能表底端的**尚有**。即會開啟「所有種類」對話框，您可以從中選取種類並按一下**確定**套用變更。
- 如果您要強制將文件移入新的種類，請選取**建立空白種類**。即會在種類樹狀結構中出現使用通用名稱的新種類。

當某個種類包含一或多個強制的文件時，會在樹狀結構的種類名稱之下顯示名為**強制移入**或**強制移出**的虛擬種類。

清除強制狀態

1. 從「資料」窗格中，選取您不再想要強制移入或移出種類的文件。
2. 從功能表中選擇**種類 > 強制移入**以強制移入，或選擇**種類 > 強制移出**以強制移出。強制移入文件或從中移出文件的種類前面會出現勾號。
3. 在子功能表中選取您要移除強制的已勾選種類。即會移除勾號，不再強制該文件。

清除所有強制狀態

1. 從「資料」窗格中選取包含**強制移入**或**強制移出**的記錄。
2. 從功能表中選擇**種類** > **全部清除** > **強制移入**或**種類** > **全部清除** > **強制移出**。即會清除文件上的強制狀態，且文件不再強制移入或移出種類。

註：只有在來源文字包含唯一 ID 時，此功能才可用。如果來源文字不含唯一 ID，則您可以在來源文件與「文字採礦」節點之間新增一個「衍生」節點。此功能只會影響互動式階段作業的執行。當您部署類別模型以進行非互動式評分時，此部分資訊不會保留或使用，因為它基於文件 ID。

刪除種類

如果您不再想要保留某個種類，則可以刪除它。

要刪除種類

1. 在「種類」窗格中，選取您要刪除的種類。
2. 從功能表中，選擇**編輯 > 刪除**。

第 10 章 分析叢集

您可以在「叢集」視圖中建置及探索概念叢集（**視圖** > **叢集**）。叢集是叢集演算法根據相關概念在文件/記錄集中出現的頻率，以及它們在同一份文件中一起出現的頻率（又稱為出現次數），所產生的相關概念分組。叢集中的每一個概念都與叢集中的至少一個其他概念一起出現。叢集的目標是要群組一起出現的概念，而種類的目標則是要根據文件或記錄包含的文字如何符合每一個種類的描述子（概念、規則、型樣）來群組它們。

好的叢集是指其概念穩固鏈結且經常出現，並具有幾個與其他叢集中的概念之鏈結的叢集。處理較大型的資料集時，此技術可能會導致遠遠較長的處理時間。

叢集作業是一開始先分析一組概念，並尋找在文件中經常出現的概念的一項程序。在文件中出現的兩個概念會被視為概念對組。接下來，叢集作業程序會評量每一個概念對組的相似性值，其作法為將對組在其中一起出現的文件數與每一個概念在其中出現的文件數相比較。如需相關資訊，請參閱主題 第 121 頁的『計算相似性鏈結值』。

最後，叢集作業程序會藉由聚集將類似的概念群組成叢集，並將它們的鏈結值和「建置叢集」對話框中定義的設定納入考量。對於聚集，意指新增概念或是將較小的叢集合併成較大的叢集，直到叢集飽和為止。當要再合併概念或較小的叢集將會導致叢集超出「建置叢集」對話框中的設定（概念、內部鏈結或外部鏈結數）時，就表示叢集已飽和。叢集會取用叢集內具有與叢集內的其他概念的最高整體鏈結數之概念的名稱。

在結尾時，並不是所有的概念對組最終都會在同一個叢集中在一起，因為另一個叢集中可能有更穩固的鏈結，或是飽和度可能會防止合併它們在其中出現的叢集。基於此原因，所以有內部和外部鏈結兩者。

- 內部鏈結是指叢集內的概念對組之間的鏈結。並非所有的概念都會在叢集中彼此鏈結。不過，每一個概念至少都會鏈結到叢集內的一個其他概念。
- 外部鏈結是指各別叢集（一個概念在一個叢集內，而一個概念在另一個叢集外）中的概念對組之間的鏈結。

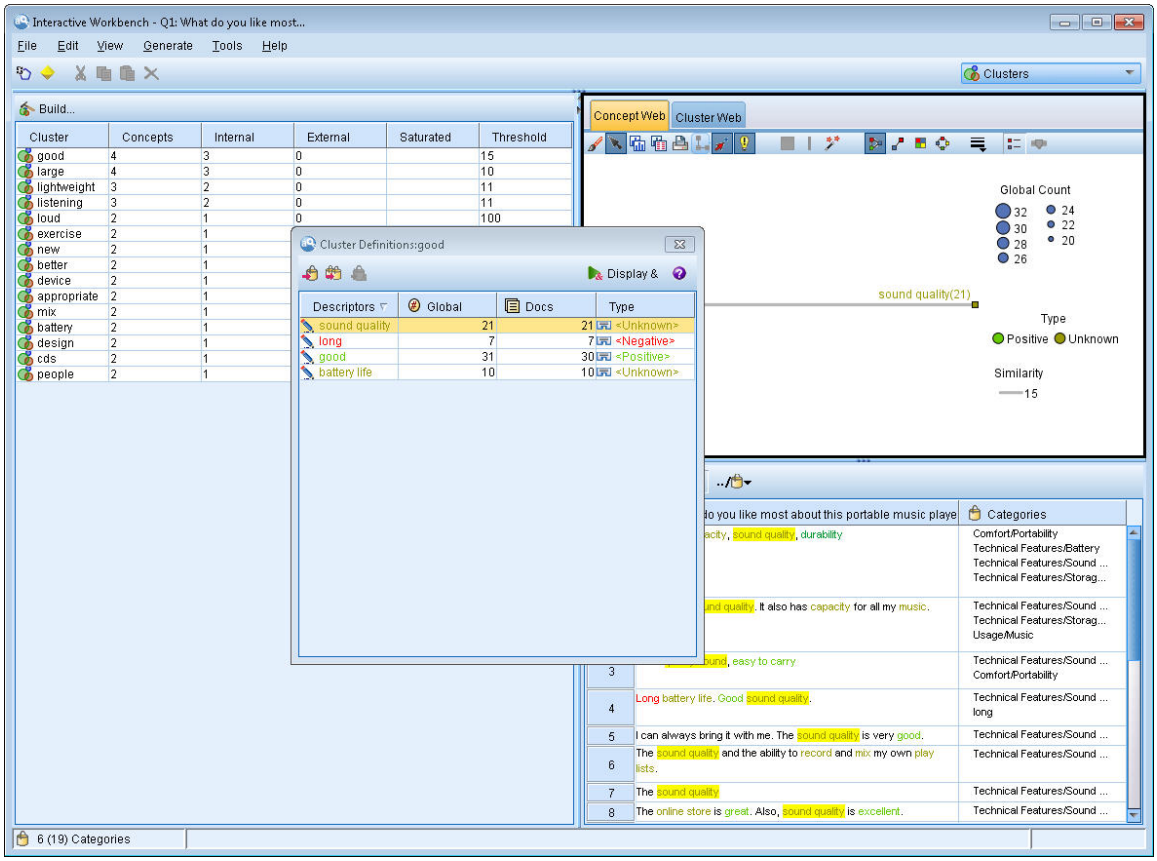


圖 *30.* 「叢集」視圖

「叢集」視圖編排成三個窗格，從「視圖」功能表中選取其名稱就可以隱藏或顯示其中的每一個窗格：

- **「叢集」窗格** 您可以在此窗格中建置及管理叢集。如需相關資訊，請參閱主題 第 122 頁的『探索叢集』。
- **「視覺化」窗格** 您可以在此窗格中以視覺化方式探索叢集以及它們的互動方式。如需相關資訊，請參閱主題 第 136 頁的『叢集圖形』。
- **「資料」窗格** 您可以探索及檢閱對應於「叢集定義」對話框中的選取項目的文件或記錄內包含的文字。如需相關資訊，請參閱主題 第 122 頁的『叢集定義』。

建置叢集

當您最先存取「叢集」視圖時，會看不到任何叢集。您可以透過功能表（**工具 > 建置叢集**）或按一下工具列上的**建置...** 按鈕來建置叢集。這個動作會開啟「建置叢集」對話框，您可以在其中定義用於建置叢集的設定和限制。

附註：每當擷取結果不再符合資源時，這個窗格會如同「擷取結果」窗格一樣變成黃色。您可以重新擷取以取得最新的擷取結果，且黃色著色將會消失。不過，每次執行擷取時，都會清除「叢集」窗格，您就得重建叢集。同樣地，叢集並不會從一個階段作業儲存到另一個階段作業。

「建置叢集」對話框內有下列區域和欄位可用：

輸入

「輸入」表格 叢集是從某些類型所衍生的描述子所建置。在此表格中，您可以選取在建置程序中要包含的類型。依預設會預先選取佔有最多記錄或文件的類型。

選取叢集的概念： 選取選取要用於叢集作業之概念的方法。減少概念數就可以加速叢集作業程序。您可以使用前幾名概念的數目、前幾名概念的百分比或是使用所有概念來進行叢集作業：

- **根據文件計數的數目** 當您選取**前幾名概念數目**時，請輸入要考量進行叢集作業的概念數目。系統會根據那些具有最高文件計數值的概念來選擇概念。文件計數是指其中出現概念的文件或記錄數。上限值是 150,000。
- **根據文件計數的百分比** 當您選取**前幾名概念百分比**時，請輸入要考量進行叢集作業的概念百分比。系統會根據具有最高文件計數值的概念的這個百分比來選擇概念。

輸出限制

可建立的叢集數目上限 這個值是可在「叢集」窗格中產生及顯示的叢集數目上限。在叢集作業程序期間，已飽和的叢集會在未飽和的叢集之前存在，也因此產生的許多叢集將會已飽和。為了查看更多未飽和的叢集，您可以將此設定變更為大於已飽和叢集數目的值。

叢集中的概念數目上限 這個值是叢集可以包含的概念數目上限。

叢集中的概念數目下限 這個值是必須鏈結才能建立叢集的概念數目下限。

內部鏈結數目上限 這個值是叢集可以包含的內部鏈結數目上限。內部鏈結是指叢集內的概念對組之間的鏈結。

外部鏈結數目上限 這個值是與叢集外的概念的鏈結數目上限。外部鏈結是指各別叢集內的概念對組之間的鏈結。

鏈結值下限 這個值是要考量概念對組進行叢集作業可接受的最小鏈結值。鏈結值是使用相似性公式所計算。請參閱『計算相似性鏈結值』主題，以取得更多資訊。

防止特定概念配對。選取此勾選框可停止處理程序在輸出中將兩個概念一併分組或配對。若要建立或管理概念配對，請按一下**管理配對**。如需相關資訊，請參閱主題 第 93 頁的『管理鏈結異常狀況配對』。

計算相似性鏈結值

只知道概念對組在其中出現的文件數本身並無法瞭解兩個概念類似的程度。在這些情況下，相似性值可能非常有用。相似性鏈結值是使用共現的文件計數與關係中的每一個概念的個別文件計數相比較所測量得來。計算相似性時，度量單位是在其中發現概念或概念對組的文件數（文件計數）。在文件中「發現」概念或概念對組是指它在文件中至少出現一次。您可以選擇讓「概念」圖形中的線厚度代表圖形中的相似性鏈結值。

此演算法顯示那些最穩固的關係，這表示概念在文字資料中一起出現的傾向遠高於它們獨立出現的傾向。 此演算法會在內部產生範圍從 0 到 1 的相似性係數，其中值 1 表示兩個概念永遠會一起出現，絕不會分開出現。然後會將相似性係數結果乘以 100 並捨入到最近的整數。使用了下圖中顯示的公式計算相似性係數。

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

圖 *31.* 相似性係數公式

其中：

- C_I 是指其中出現概念 I 的文件或記錄數。
- C_J 是其中出現概念 J 的文件或記錄數。
- C_{IJ} 是指在其中，概念對組 I 和 J 在文件集中出現的文件或記錄數。

例如，假設您有 5,000 份文件。讓 I 和 J 為擷取的概念，並讓 IJ 為出現 I 和 J 的概念對組。下表提出了兩個實務來示範係數和鏈結值的計算方式。

表 33. 概念頻率範例

概念/對組	實務 A	實務 B
概念：I	在 20 份文件中出現	在 30 份文件中出現
概念：J	在 20 份文件中出現	在 60 份文件中出現
概念對組：IJ	在 20 份文件中出現	在 20 份文件中出現
相似性係數	1	0.22222
相似性鏈結值	100	22

在實務 A 中，概念 I 和 J 以及對組 IJ 在 20 份文件中出現，產生相似性係數 1，這表示這些概念總是一起出現。此對組的相似性鏈結值將會是 100。

在實務 B中，概念 I 在 30 份文件中出現，而概念 J 在 60 份文件中出現，但是對組 IJ 僅在 20 份文件中出現。因此，相似性係數為 0.22222。此對組的相似性鏈結值將會向下捨入至 22。

探索叢集

建置叢集之後，您可以在「叢集」窗格中看到一組結果。針對每一個叢集，表格中會提供下列資訊：

- **叢集。** 這是叢集的名稱。叢集是以內部鏈結數目最高的概念命名。
- **概念數。** 這是叢集中的概念數目。如需相關資訊，請參閱主題 『叢集定義』。
- **內部。** 這是叢集中的內部鏈結數目。內部鏈結是指叢集內的概念對組之間的鏈結。
- **外部。** 這是叢集中的外部鏈結數目。外部鏈結是指當一個概念在一個叢集中，而另一個概念在另一個叢集中時，概念對組之間的的鏈結。
- **已飽和。** 如果存在一個符號，這表示這個叢集很可能較大，但是應已超出一或多個限制，因此該叢集的叢集作業程序已結束，並被視為已飽和。在叢集作業程序結束時，已飽和的叢集會在未飽和的叢集之前存在，也因此產生的許多叢集將會已飽和。為了查看更多未飽和的叢集，您可以將**要建立的叢集數目上限**設定變更為大於已飽和叢集數目的值，或是減少**鏈結值下限**。如需相關資訊，請參閱主題 第 120 頁的『建置叢集』。
- **臨界值。** 對於叢集中所有出現的概念對組，這是叢集中所有對組的最低相似性鏈結值。如需相關資訊，請參閱主題 第 121 頁的『計算相似性鏈結值』。具有高臨界值的叢集表示，在該叢集中的概念具有較高的整體相似性，且比起叢集中其臨界值較低的那些概念更緊密相關。

如果要進一步瞭解給定的叢集，您可以選取它，位於右側的「視覺化」窗格就會顯示兩個圖形，幫助您探索叢集。如需相關資訊，請參閱主題 第 136 頁的『叢集圖形』。您也可以將表格的內容剪下並貼入另一個應用程式中。

每當擷取結果不再符合資源時，這個窗格會如同「擷取結果」窗格一樣變成黃色。您可以重新擷取以取得最新的擷取結果，且黃色著色將會消失。不過，每次執行擷取時，都會清除「叢集」窗格，您就得重建叢集。同樣地，叢集並不會從一個階段作業儲存到另一個階段作業。

叢集定義

您可以查看叢集內的所有概念，方法為在「叢集」窗格中選取它，然後開啟「叢集定義」對話框（**視圖 > 叢集定義**）。

所選叢集中的所有概念會出現在「叢集定義」對話框中。如果您在「叢集定義」對話框中選取一或多個概念，然後按一下**顯示 &**，「資料」窗格將會顯示在其中所有的所選概念一起出現的所有記錄或文件。不過，當您在「叢集」窗格中選取叢集時，「資料」窗格不會顯示任何文字記錄或文件。如需有關「資料」窗格的一般資訊，請參閱中的。

在此對話框中選取概念也會變更概念 Web 圖形。如需相關資訊，請參閱主題 第 136 頁的『叢集圖形』。同樣地，當您在「叢集定義」對話框中選取一或多個概念時，「視覺化」窗格也會顯示來自那些概念的所有外部和內部鏈結。

直欄說明

會顯示圖示，以便您可以輕易識別每一個描述子。

表 *34.* 直欄和描述子圖示

直欄	**說明**
描述子	概念的名稱。
 **廣域**	顯示此描述子出現在整個資料集中的次數，又稱為廣域頻率。
 **文件**	顯示此描述子在其中出現的文件或記錄數，又稱為文件頻率。
類型	顯示描述子所屬的一或多個類型。如果描述子是種類規則，則此直欄中不會顯示任何類型名稱。

工具列動作

從這個對話框中，您還可以選取要在種類中使用的一或多個概念。有數種方式可以這麼做，但是最有趣的是選取在叢集中出現的概念，然後將它們新增為種類規則。如需相關資訊，請參閱主題 第 96 頁的『共生規則』。您可以使用工具列按鈕來新增概念至種類。

表 *35.* 用來新增概念至種類的工具列按鈕

圖示	**說明**
	將選取的概念新增至新的或現有的種類
	將 & 種類規則表單中的所選概念新增至新的或現有的種類。如需相關資訊，請參閱主題 第 101 頁的『使用種類規則』。
	新增每一個選取的概念作為其自己的新種類
	根據選取的描述子更新「資料」窗格和「視覺化」窗格中顯示的內容

附註：您也可以使用快速功能表，將概念新增至類型作為同義字或作為排除項目。

第 11 章 探索文字鏈結分析

在「文字鏈結分析 (TLA)」視圖中，您可以探索文字鏈結分析型樣結果。文字鏈結分析是型樣比對技術，可讓您定義型樣規則並將這些規則與實際擷取的概念及在文字中找到的關係進行比較。

例如，擷取組織相關的構想對您來說可能不夠有趣。使用 TLA，您也可以瞭解此組織與其他組織之間的鏈結或組織內的人員。您還可以使用 TLA 來擷取產品的相關意見，或針對部分語言擷取基因之間的關係。

擷取部分 TLA 型樣結果之後，您可以在「文字鏈結分析」視圖的「類型型樣」和「概念型樣」窗格中檢閱它們。如需相關資訊，請參閱主題 第 127 頁的『類型型樣和概念型樣』。您可以進一步在此視圖中的「資料」或「視覺化」窗格中探索它們。最重要的可能是您可以將它們新增至種類。

如果您尚未選擇這麼做，則可以按一下**擷取**並在「擷取設定」對話框中選擇**啟用文字鏈結分析型樣擷取**。如需相關資訊，請參閱主題 第 126 頁的『擷取 TLA 型樣結果』。

必須在您使用的資源範本或程式庫中定義部分 TLA 型樣規則，以便擷取 TLA 型樣結果。您可以在 IBM SPSS Modeler Text Analytics 隨附的某些資源範本中使用 TLA 型樣。您可以擷取的關係及型樣類型完全取決於在資源中定義的 TLA 規則。您可以定義自己的 TLA 規則. 型樣由巨集、單字清單及字隙組成，從而形成與您輸入的文字進行比較的布林查詢或規則。如需相關資訊，請參閱主題 第 187 頁的第 18 章, 『關於文字鏈結規則』。

每當 TLA 型樣規則與文字相符時，此文字可以作為型樣擷取，並作為輸出資料重組。然後結果會在「文字鏈結分析」視圖窗格中可見。每一個窗格都可以透過從「檢視」功能表中選取其名稱來隱藏或顯示：

- **類型型樣和概念型樣窗格**。您可以在這兩個窗格中建置和探索型樣。如需相關資訊，請參閱主題 第 127 頁的『類型型樣和概念型樣』。
- **視覺化窗格**。您可以透過視覺化的方式探索型樣中的概念和類型如何在此窗格中互動。如需相關資訊，請參閱主題 第 138 頁的『文字鏈結分析圖形』。
- **資料窗格**。您可以探索和檢閱對應於另一個窗格中選取內容的文件和記錄內包含的文字。如需相關資訊，請參閱主題 第 128 頁的『資料窗格』。

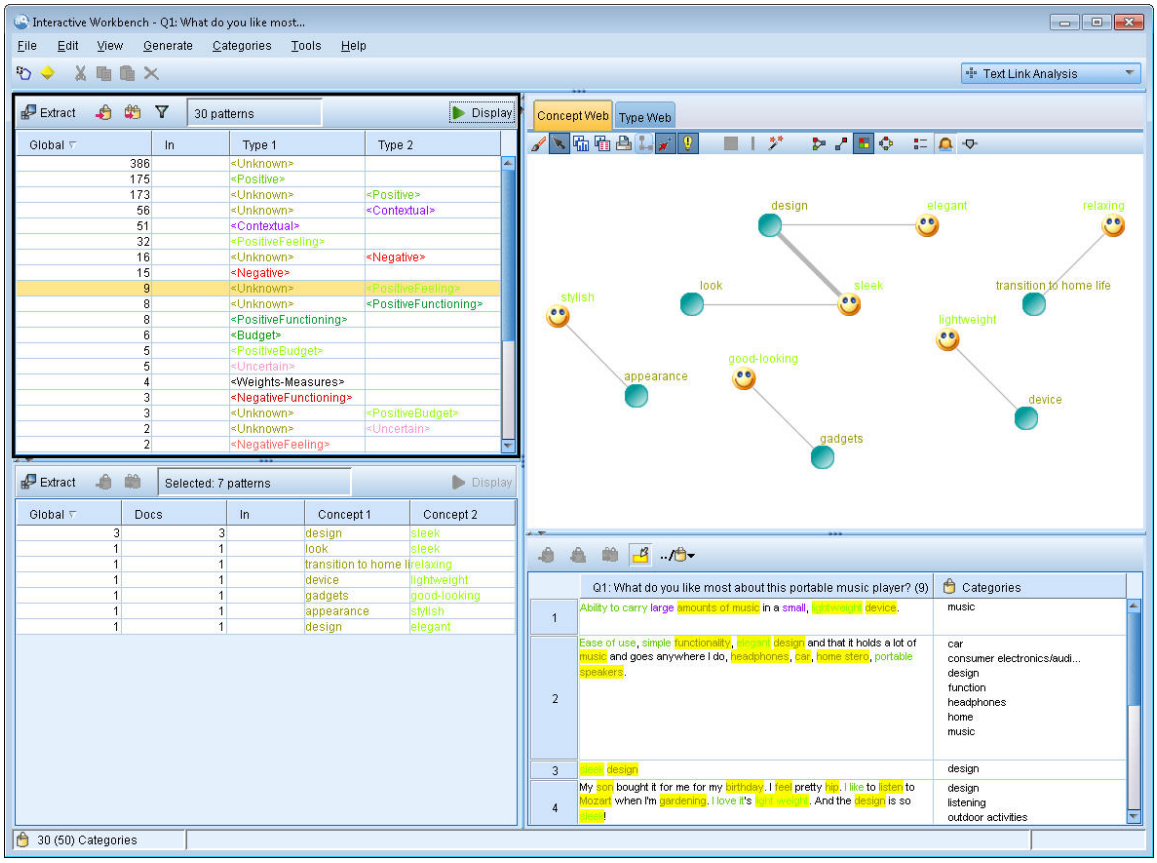


圖 32. 文字鏈結分析視圖

擷取 TLA 型樣結果

擷取程序會產生一組概念和類型，以及「文字鏈結分析 (TLA)」型樣（如果啟用）。如果您擷取 TLA 型樣，則可以在「文字鏈結分析」視圖中看到那些型樣。每當擷取結果與資源不同步時，「型樣」窗格都會變成黃色，指出重新擷取會產生不同的結果。

您只能選擇使用選項**啟用文字鏈結分析型樣擷取**，在節點設定或「擷取」對話框中擷取這些型樣。如需相關資訊，請參閱主題 第 70 頁的『擷取資料』。

附註：在資料集大小與完成擷取程序所花費時間之間有一種關係。如需效能統計資料及建議，請參閱安裝指示。您可以一律考量插入「樣本」節點上游或最佳化機器的配置。

擷取資料

1. 從功能表中，選擇**工具 > 擷取**。或者，按一下**擷取**工具列按鈕。
2. 變更您想要使用的任何選項。請記住，必須選取此標籤上的選項**啟用文字鏈結分析型樣擷取**以及在範本中有 TLA 規則，才能擷取 TLA 型樣結果。如需相關資訊，請參閱主題 第 70 頁的『擷取資料』。
3. 按一下**擷取**以開始擷取程序。

擷取開始後，進度對話框即會開啟。如果您想要中斷擷取，請按一下**取消**。完成擷取時，對話框即會關閉，而結果會在窗格中出現。如需相關資訊，請參閱主題 第 127 頁的『類型型樣和概念型樣』。

類型型樣和概念型樣

型樣由兩部分組成，概念和類型的組合。嘗試探索特定主題的相關意見或概念之間的關係時，型樣極其有用。例如，擷取競爭者的產品名稱對您來說可能不夠有趣。在此情況下，您可以查看擷取的型樣來了解是否可以找到一些範例，其中文件或記錄包含表示產品良好、惡劣或昂貴的文字。

型樣可由最多六種類型或六個概念組成。基於此原因，兩個型樣窗格中的列都會包含最多六個位置。每一個位置對應於元素在 TLA 型樣規則中的特定位置，就像在語言資源中定義的一樣。在互動式工作台中，如果位置不包含值，則不會在表格中顯示。例如，如果最長的型樣結果包含的位置不超過四個，則不會顯示最後兩個。如需相關資訊，請參閱主題 第 187 頁的第 18 章, 『關於文字鏈結規則』。

擷取型樣結果時，它們先在類型層次分組，然後分成概念型樣。基於此原因，有兩個不同的結果窗格：**類型型樣**（左上方）和**概念型樣**（左下方）。若要查看傳回的所有概念型樣，請選取全部類型型樣。然後底端的概念型樣窗格會顯示所有概念型樣，最多可達到等級值上限（在「過濾器」對話框中定義）。

類型型樣 此窗格會呈現型樣結果，由一個或多個符合 TLA 型樣規則的相關類型組成。類型型樣顯示為 <Organization> + <Location> + <Positive>，可提供特定位置中組織的相關正面意見。語法如下所示：

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

概念型樣 此窗格會針對在其上的「類型型樣」窗格中目前選取的所有類型型樣，在概念層次呈現型樣結果。概念型樣接在建築物之後，例如，hotel + paris + wonderful。語法如下所示：

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

當型樣結果使用的位置小於上限六個時，僅會顯示必要的位置（或直欄）數。在兩個已填充位置之間找到的任何位置都會捨棄，以便型樣 <Type1>+<>+<Type2>+<>+<>+<> 可由 <Type1>+<Type3> 代表。針對概念型樣，這會是 concept1+.+concept2（其中 . 代表空值）。

就像「種類和概念」視圖中的擷取結果一樣，您可以在這裡檢閱結果。如果您查看希望對組成這些型樣的類型及概念進行的精簡，您可在「種類和概念」視圖中的「擷取結果」窗格中進行，或直接在「資源」編輯器中進行，然後重新擷取型樣。每當在種類定義中依現狀或作為部分規則使用概念、類型或型樣時，種類或規則圖示都會在「型樣或擷取結果」表格中的**在**直欄中出現。

註：如果可見窗格能夠容納更多的結果，則您可以使用窗格底部的控制項在結果中前後移動，或輸入要跳至的頁碼。

過濾 TLA 結果

當您使用的資料集非常大時，擷取程序可能會產生上百萬的結果。對於許多使用者來說，這個數量會使有效檢閱結果的難度增加。但您可以過濾這些結果以便聚焦於那些最有趣的結果。您可以變更「過濾器」對話框中的設定來限制顯示哪些型樣。所有這些設定會一起使用。

在 TLA 視圖中，「過濾器」對話框包含下列區域及欄位。

依頻率過濾 您可以過濾以僅顯示具有某個廣域或文件頻率值的那些結果。

- **廣域頻率**是型樣在整個文件或記錄集中出現的總次數，會在**廣域**直欄中顯示。
- **文件頻率**是型樣在其中出現的文件或記錄總數，會在**文件數**直欄中顯示。

例如，如果型樣在 500 個記錄中出現了 300 次，我們會說，此型樣的廣域頻率為 300，文件頻率為 500。

依據比對文字 您還可以過濾來僅顯示符合您在這裡所定義規則的那些結果。在**比對文字**欄位中輸入要比對的字元集，並透過識別位置號碼或全部來選取是在概念還是類型名稱中尋找此文字。然後選取在其中套用比對的條

件（您無需使用角括弧來表示類型名稱的開頭或結尾）。從下拉清單中選取**和**或**或**，以便規則符合兩個陳述式或其中一個，並以與第一個文字相符陳述式相同的方式定義第二個。

表 *36.* 比對文字條件

條件	**說明**
包含	如果字串在某個位置出現則符合文字。（預設選項）
開始於	僅在概念或類型以指定文字開頭時符合文字。
結尾是	僅在概念或類型以指定文字結尾時符合文字。
完全相符	整個字串必須符合概念或類型名稱。

結果在型樣窗格中顯示

假設您使用英文版本的軟體；這裡是如何根據過濾器在「型樣」窗格工具列上顯示結果的部分範例。

圖 *33.* 過濾結果範例 *1*

在此範例中，工具列顯示傳回的型樣數因為過濾器中指定的等級上限而受到限制。如果紫色圖示存在，這表示符合型樣數上限。如需相關資訊，請移至圖示上方。請參閱之前**依據等級**過濾器的說明。

圖 *34.* 過濾結果範例 *2*

在此範例中，工具列顯示使用比對文字過濾器（請參閱放大鏡圖示）限制了結果。您可以移至該圖示上方來查看比對文字是什麼。

過濾結果

1. 從功能表中，選擇**工具** > **過濾器**。「過濾器」對話框即會開啟。
2. 選取並精簡您要使用的過濾器。
3. 按一下**確定**以套用過濾器並查看新結果。

資料窗格

當您擷取並探索文字鏈結分析型樣時，您可能想要檢閱正在使用的部分資料。例如，您可能想要查看在其中探索到一組型樣的實際記錄。您可以在位於右下方的「資料」窗格中檢閱記錄或文件。如果依預設不可見，請從功能表中選擇**檢視** > **窗格** > **資料**。

「資料」窗格會呈現每個文件或記錄一列，對應於視圖中選取的內容，最多可達到某個顯示限制。依預設，會限制在「資料」窗格中顯示的文件或記錄數，以便讓您更快地查看資料。但您可以在「選項」對話框中對此進行調整。如需相關資訊，請參閱 第 64 頁的『選項：階段作業標籤』。

註：如果可見窗格能夠容納更多的結果，則您可以使用窗格底部的控制項在結果中前後移動，或輸入要跳至的頁碼。

顯示及重新整理資料窗格

「資料」窗格不會自動重新整理顯示畫面，因為使用的資料集較大時，自動資料重新整理可能需要花費一些時間才能完成。因此，每當您在此視圖中選取類型或概念型樣時，都可以按一下**顯示**來重新整理「資料」窗格的內容。

文字文件或記錄

如果文字資料的形式為記錄且文字長度相對較短，則「資料」窗格中的文字欄位會顯示完整的文字資料。但使用記錄及更大的資料集時，文字欄位直欄會顯示一小段文字，並會在右側開啟「文字預覽」窗格以顯示您在表格中已選取記錄的更多或全部文字。如果文字資料的形式為個別文件，則「資料」窗格會顯示文件的檔名。選取文件時，「文字預覽」窗格會開啟並顯示所選取文件的文字。

顏色及強調顯示

每當您顯示資料時，在那些文件或記錄中找到的概念及描述子會以顏色強調顯示，來協助您在文字中輕鬆識別它們。顏色編碼對應於概念所屬的類型。您還可以將滑鼠移至顏色編碼項目上方來顯示在其下面擷取的概念及為其指派的類型。未擷取的任何文字都以黑色顯示。一般這些未擷取的單字通常為連接詞（*and* 或 *with*）、代詞（*me* 或 *they*）及動詞（*is*、*have* 或 *take*）。

資料窗格直欄

文字欄位直欄一律可見，而您也可以顯示其他直欄。若要顯示其他直欄，請從功能表中選擇**檢視** > **資料窗格**，然後選取您想要在「資料」窗格中顯示的直欄。下列直欄可用於顯示：

- **「文字欄位名稱」(#)/文件**。針對從中擷取概念和類型的文字資料新增直欄。如果資料在文件中，則直欄稱為「文件」，並且只有文件檔名或完整路徑可見。若要查看那些文件的文字，您必須查看「文字預覽」窗格。「資料」窗格中的列數會在此直欄名稱之後的括弧中顯示。由於「選項」對話框中存在用來提高載入速度的限制，因此有時可能不會顯示所有文件或記錄。如果達到上限，則數字後面會接有 **- 上限**。如需相關資訊，請參閱 第 64 頁的『選項：階段作業標籤』。
- **種類**。 列出記錄所隸屬的每一個種類。每當顯示此直欄時，為了顯示最新的資訊，重新整理「資料」窗格可能需要花費更長一點的時間。
- **強制移入**。列出您已在其中強制移入文件的種類。可以透過**編輯** > **強制移入**功能表選項將文件強制移入種類。如需相關資訊，請參閱 第 117 頁的『強制將文件移入種類』。
- **強制移出**。列出您已從中移除文件的種類。可以透過**編輯** > **強制移出**功能表選項強制將文件移出種類。例如，當回應者的諷刺導致某個回應被錯誤分類，可能會使用該選項。如需相關資訊，請參閱 第 117 頁的『強制將文件移入種類』。
- **種類計數**。列出記錄所隸屬的種類數目。
- **相關性等級**。為單一種類中的每一筆記錄提供等級。此等級顯示與種類中的其他記錄相比，該記錄在多大程度上適合該種類。在「種類」窗格（左上方窗格）中選取種類可查看等級。如需相關資訊，請參閱 第 89 頁的『種類相關性』。
- **回應旗標**。新增直欄以顯示您可能要使用的任何旗標。在此直欄內按一下以變更您指派給文件的旗標類型。您可使用「完成」旗標或「重要」旗標來標示文件，或是移除旗標。這對於檢閱種類模型的完成度而言非常有用。如需相關資訊，請參閱 第 90 頁的『標示回應』。

標示回應

若要協助監視您的進度，您可以在「資料」窗格中使用旗標來標示文件。只有在來源文件包含唯一 ID 時，此功能才可用。如果來源文件不含唯一 ID，則您可以在來源文件與「文字採礦」節點之間新增一個「衍生」節點。

您想要標示文件的原因有多種，包括：

- 標示您已手動檢閱的文件，方便您知道稍後要在哪裡選取它們
- 標示您不確定如何處理的文件

使用旗標標示文件之後，您便可以繼續處理文件。它們僅供您自己記錄保留。您可以選擇下列旗標：

表 *37.* 旗標說明

旗標	說明
	完成旗標表示您認為已完成的文件。
	重要旗標表示您認為重要的文件。

若要使用旗標標示文件：

1. 從「資料」窗格中，在您要標示的文件上按一下滑鼠右鍵。
2. 從快速功能表中選擇**檢視** > **資料窗格** > **回應旗標**，然後選取您要使用的旗標類型（重要旗標或完成旗標）。即會指派選取的旗標。如果「資料」窗格中的「旗標」直欄不可見，它會出現。

若要清除旗標：

1. 從「資料」窗格中，在您要移除其旗標的文件上按一下滑鼠右鍵。
2. 從快速功能表中選擇**標示回應方法** > **清除旗標**。即會移除選取的旗標。

Type Reassignment Rule

Type Reassignment Rule (TRR) 旨在將一系列類型、巨集及/或記號轉換為具有特定類型的新概念。尤其是可在「意見」範本中使用它們以捕捉極性變更的意見。例如，在片語 "not that bad" 中，單字 "bad" 是一個 *negative* 意見。但在此上下文中，實際意義是 "not bad" – 它是一個 *positive* 意見。

直到 18.2 版，此極性變更由特定「文字鏈結分析 (TLA)」規則進行管理：

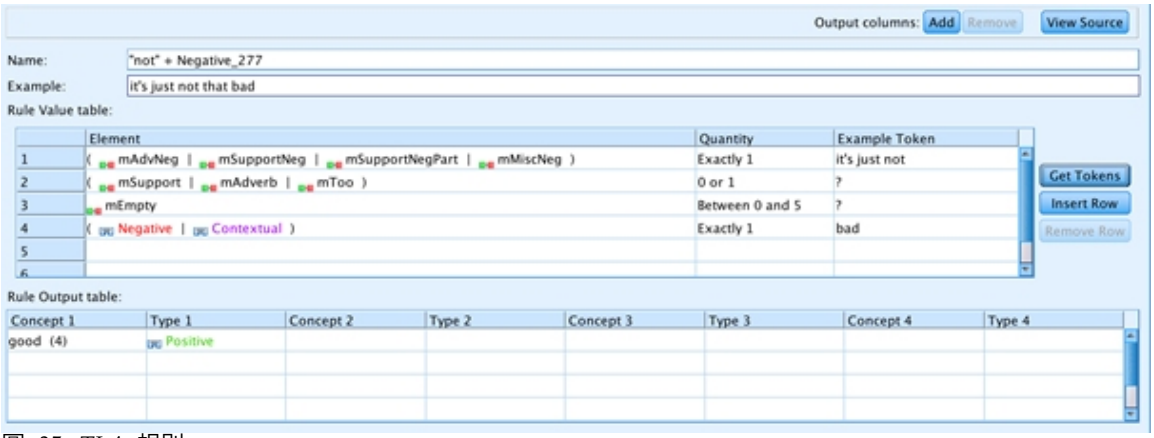


圖 *35. TLA* 規則

由於不同的意見類型（`Positive`、`PositiveAttitude`、`PositiveBudget`、`PositiveCompetence`、`PositiveFeeling`、`PositiveFunctioning`、`PositiveRecommendation`、`Negative`、`NegativeAttitude`、`NegativeBudget`、`NegativeCompetence`、`NegativeFeeling`、`NegativeFunctioning`、`NegativeRecommendation` 和 `Contextual`），因此它涉及撰寫特定的 TLA 規則：

- 針對每一種類型。例如：

```
"not + xxx + <NegativeBudget>" => "<PositiveBudget>"
```

或

```
"not + xxx + <PositiveAttitude>" => "<NegativeAttitude>"
```

- 針對許多語法上下文。例如：

```
* topic + negation + opinion ("hotel wasn't good")
* negation + opinion + topic ("it was not a good hotel")
* negation + opinion ("not very good")
* topic + opinion + negation + opinion ("hotel was well-located but  not that good")
* 2 topics + negation + opinion ("room and swimming pool weren't always clean")
* ...
```

從 18.2 版開始，新方法旨在「捕捉」此類序列（任何否定 + 任何空白單字 + 特定意見），選取要出現在新概念中的單字（任何標準化否定– 例如，"not" – 以及意見），並為這個新概念定義類型（又稱為「虛擬術語」）。然後，這個新概念可用在 TLA 規則中。

因此，下列規則將符合包含一個主題後接一個意見的任何序列，無論意見是術語 (comfortable) 還是虛擬術語 (not economical)，也無論特定意見的子類型為何（屬性、預算等）。

```
#@# Bed was extremely comfortable
[pattern(190)]
name=topic + opinion_190
value=$mTopic ($mEmpty|$mToo){0,3} ($mOpinionPos|$mOpinionNeg|$Contextual)
output(1)=$1\t#1\t$3\t#3
```

除了變更意見的極性之外，您還可以使用 TRR 來協助細部調整字典。例如，假設您有一個類型 Anatomy（擁有的身體部分為 heart、chest、breast 和 adrenal gland 等），以及另一個類型 MedicalProcedures（擁有的程序為 biopsy、needle biopsy、MRI 和 CT scan 等）。幾乎不太可能正確地列出與某個器官相關聯的所有醫療程序。因此，您可以建立兩個 TRR 來識別下列圖例中所示的可能醫療程序。執行擷取動作之後，您可以針對類型 PotentialMedicalProcedures 新增一個過濾器，檢閱候選術語，然後將其新增至 MedicalProcedures 類型。

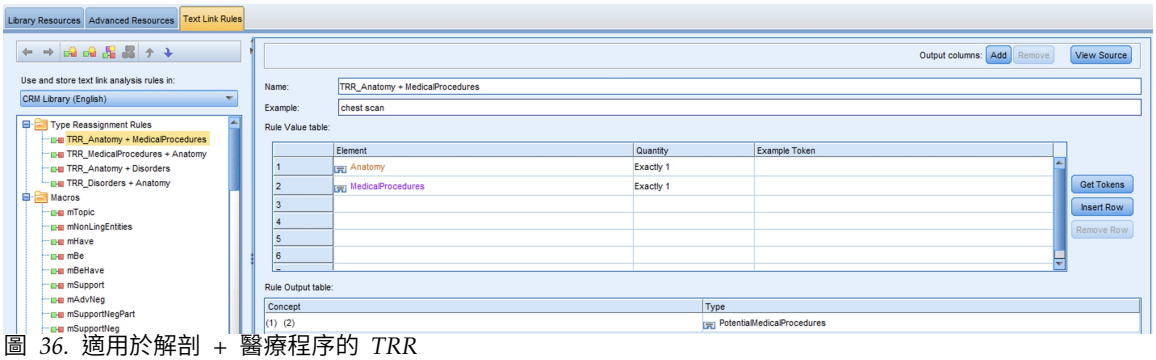


圖 *36.* 適用於解剖 + 醫療程序的 *TRR*

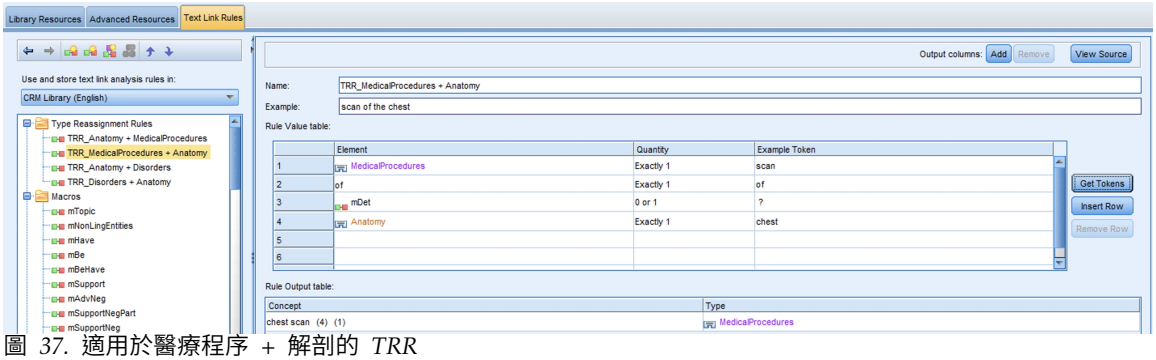


圖 37. 適用於醫療程序 + 解剖的 *TRR*

語法

```
#@# not that expensive
[typeReassignmentRule]
name=TRR_"not" NegativeBudget
value=$mAllNeg ($mAdverb|$mBe|$mHave|$mSupport|$mDet|that|more|$mQuant){0,3} $NegativeBudget
output=not $3\tPositiveBudget
```

- "name" 必須唯一 (TRR_"not" NegativeBudget)。它無法用於巨集或 TLA 規則中。只能使用輸出中定義的類型。
- "value" 是一系列相符元素。元素可為類型 ($NegativeBudget)、巨集 ($mAllNeg) 或記號 (more)。部分元素可能是必要元素、選用元素或者具有特定的數量。
- "output" 是概念 + 類型的單一配對 (not $3\tPositiveBudget)。請注意，您可以在輸出中使用可用的類型（定義在範本中的類型），您也可以建立一個新類型。

  輸出類型也可以參照相符元素（例如，#2）。如果值與輸出之間的類型沒有變更，此功能尤其有用。例如：

  ```
  #@# could not have been any more pleased
  [typeReassignmentRule]
  name=TRR_"couldn't be more" opinion
  value=$mNotNeg ($mOpinionPos|$mOpinionNeg|$Contextual)
  output=$2\t#2
  ```

如同在 TLA 規則中一樣，必須在更通用的 TRR 之前定義一個更具體的 TRR。若要確保所有 TRR 都依正確的順序定義，您可以使用「取得記號」功能來循序測試每個 TRR。如果 TRR 不符，但符合其他定義，則您可以上下移動它。

特殊案例

在部分案例中，必須仍然能夠存取序列的個別元素而不是 TRR。這樣做一般會關注協調而不是否定。在片語 "not that fashionable or eyecatching" 中，協調 "or" 不容許用於探索它，在此上下文中，"eyecatching" 實際表示 "not eyecatching"。

在此情況下，建議使用特定的規則，例如：

```
#@# not that fashionable or eyecatching
[pattern(263)]
name="not" + 2 Positive_263
value=($mAdvNeg|$mSupportNeg|$mMiscNeg) @{0,1} $PositiveFeeling or $PositiveFeeling
output(1)=not $3\tNegativeFeeling
output(2)=not $5\tNegativeFeeling
```

雖然規則 (($mAdvNeg|$mSupportNeg|$mMiscNeg) @{0,1} $PositiveFeeling) 的第一部分可能符合 TRR，但 TLA 規則將具有優先順序。

如果您撰寫更通用的規則（例如下列範例），則仍然適用存在於 18.1.1 版及更早版本中的相同限制。新建的概念（虛擬概念）可能有不正確的類型（<Negative> 而不是 <NegativeFeeling>），您可能想要以包含兩個不同類型的 TLA 概念來結尾。暫行解決方法是使用正確的類型來建立對應術語（不是 xxx）。

```
#@# not that fashionable or eyecatching
[pattern(263)]
name="not" + 2 Positive_263
value=($mAdvNeg|$mSupportNeg|$mMiscNeg) @{0,1} $mPos or $mPos
output(1)=not $3\tNegative
output(2)=not $5\tNegative
```

優點

- 使用 TRR 的主要優點是具有更少 TLA 規則。
- TRR 的一個不太明顯的優點是確保虛擬術語將主要導致出現正確的類型（但請記住之前提到的限制）。在過去，由於遺漏部分特定的 TLA 規則，因此 "not + positiveXXX" 被分類為 Negative 而不是 NegativeXXX。
- 如果使用者新增特定的意見類型（例如，NegativeNoise），則不必抄寫特定的 TLA 規則來反轉極性。使用者只需要建立相關的 TRR。

第 12 章 視覺化圖形

「種類和概念」視圖、「叢集」視圖及「文字鏈結分析」視圖都在視窗的右上角有一個視覺化窗格。您可以使用此窗格來以視覺化的方式探索資料。下列圖形及圖表是可用的。

- **種類和概念視圖。**此視圖有三個圖形及圖表：種類長條、種類 *Web* 及種類 *Web* 表格。在此視圖中，僅會您按一下**顯示**時更新圖形。如需相關資訊，請參閱主題 『種類圖形與圖表』。
- **叢集視圖。**此視圖有兩個 Web 圖形：概念 *Web* 圖形和叢集 *Web* 圖形。如需相關資訊，請參閱主題 第 136 頁的『叢集圖形』。
- **文字鏈結分析視圖。**此視圖有兩個 Web 圖形：概念 *Web* 圖形和類型 *Web* 圖形。如需相關資訊，請參閱主題 第 138 頁的『文字鏈結分析圖形』。

如需用於編輯圖形的所有一般工具列及選用區的相關資訊，請參閱線上說明或檔案 *ModelerSPOnodes.pdf* 中關於「編輯圖形」的章節，可作為產品下載的一部分提供。

種類圖形與圖表

建置種類時，請務必抽出時間檢閱種類定義、它們包含的文件或記錄以及種類如何重疊。視覺化窗格提供對種類的數個視景。「視覺化」窗格位於「種類與概念」視圖的右上角。如果尚不可見，則可以從「視圖」功能表（**視圖 > 窗格 > 視覺化**）存取此窗格。

在此視圖中，視覺化窗格在文件或記錄種類中提供有關共同性的三個視景。此窗格中的圖表及圖形可用來分析種類結果，以及協助細部調整種類或報告。精簡種類時，您可以使用此窗格檢閱種類定義，從而不涵蓋太過類似（例如，它們共用超過 75% 的文件或記錄）或太過特殊的種類。如果兩個種類太過類似，則可能協助您決定結合兩個種類。或者，您可能決定透過從一個種類或另一個種類移除某些描述子，從而精簡種類定義。

根據在「擷取結果」窗格、「種類窗格」或「種類定義」對話框中選取的項目，您可以檢視此窗格中每一個標籤上文件/記錄與種類之間的對應互動。每一個都代表類似的資訊，但是透過不同的方式或利用不同的詳細程度。然而，為了重新整理目前選取的圖形，在您進行選擇的窗格或對話框的工具列上按一下**顯示**。

「種類與概念」視圖中的「視覺化」窗格提供下列圖形及圖表：

- **種類長條圖。**表格與長條圖呈現對應於您的選擇及相關聯種類之文件/記錄之間的重疊。長條圖還呈現種類中的文件/記錄與文件/記錄 如需相關資訊，請參閱主題 『種類長條圖』。
- **種類 Web 圖形。**此圖形代表種類的文件/記錄重疊，而文件/記錄根據其他窗格中的選擇屬於這些種類。如需相關資訊，請參閱主題 第 136 頁的『種類 Web 圖形』。
- **種類 Web 表格。**此表格代表的資訊與「種類 Web」標籤相同，但使用表格格式。該表格包含三個直欄，可以透過按一下直欄標頭進行排序。如需相關資訊，請參閱主題 第 136 頁的『種類 Web 表格』。

如需相關資訊，請參閱主題 第 81 頁的第 9 章, 『分類文字資料』。

種類長條圖

此表格顯示一個表格與長條圖，顯示對應於您選擇的文件/記錄與相關聯種類之間的重疊。長條圖還呈現種類中的文件/記錄與文件/記錄總數的比例。您不能編輯此圖表的佈置。然而，您可以透過按一下直欄標頭排序直欄。

該圖表包含下列直欄：

- **種類。**此直欄呈現您選擇中種類的名稱。依預設，首先列出您選擇中最常見的種類。
- **列。**此直欄以視覺化方式呈現給定種類中文件或記錄的數目與文件或記錄總數之間的比例。
- **選擇 %。**此直欄根據某個種類的文件或記錄總數與選擇中呈現的文件或記錄總數的比例，呈現一個百分比。
- **文件。** 此直欄代表給定種類之選擇中文件或記錄數目。

種類 Web 圖形

此標籤顯示種類 Web 圖形。Web 代表種類的文件/記錄重疊，而文件/記錄根據其他窗格中的選擇屬於這些種類。如果存在種類標籤，這些標籤會顯示在圖形中。您可以使用此窗格中的工具列按鈕選擇一個圖形佈置（網路、圓形、導向或網格）。

在 Web 中，每一個節點代表一個種類。使用您的滑鼠，您可以在窗格內選取及移動節點。根據您選擇中該種類的文件或記錄數目，節點的大小代表相對大小。兩個種類之間線條的粗細與顏色表示它們具有之一般文件或記錄的數目。如果您將滑鼠移至「探索」模式下的節點上方，則「工具提示」會顯示種類的名稱（或標籤），以及種類中文件或記錄的總數。

註：依預設，為您可以移動節點的圖形啟用「探索」模式。然而，您可以切換至「編輯」模式，以編輯圖形佈置，包括顏色、字型、圖註等。如需相關資訊，請參閱第 138 頁的『使用圖形工具列及選用區』。

如果您使用**複製視覺化資料**按鈕複製圖形資料，並將它貼上至試算表或文字編輯器，則將看到資料是給定的直欄標頭，例如 V1、V2 至 V7。這些直欄包含下列資訊：

- **V1、V2** 這些值對應於畫面座標（分別為 X 與 Y）。
- **V3、V5** 列出種類概念。
- **大小、V6** 顯示在其中找到概念的文件數目。
- **V7** 目前未用。

種類 Web 表格

此表格顯示與「種類 Web」標籤相同的資訊，但使用表格格式。該表格包含三個直欄，可以透過按一下直欄標頭進行排序：

- **計數。** 此直欄呈現兩個種類之間共用或一般的文件或記錄數目。
- **種類 1。**此直欄呈現第一個種類的名稱，後面是它包含的文件或記錄總數，顯示在括弧中。
- **種類 2。**此直欄呈現第二個種類的名稱，後面是它包含的文件或記錄總數，顯示在括弧中。

叢集圖形

在建置叢集之後，您可以在「視覺化」窗格中的 Web 圖形中，以視覺方式探索它們。「視覺化」窗格提供兩個有關叢集作業的視景：「概念 Web」圖形和「叢集 Web」圖形。此窗格中的 Web 圖形可用來分析叢集作業結果，以及輔助揭露您可能想要新增至種類的一些概念和規則。「視覺化」窗格位於「叢集」視圖的右上角。如果還看不到它，您可以從「視圖」功能表存取此窗格（**視圖 > 窗格 > 視覺化**）。在「叢集」窗格中選取叢集後，就可以自動在「視覺化」窗格中顯示對應的圖形。

附註：依預設，圖形是處於互動式/選擇模式，在該模式中可以移動節點。不過，您可以在「編輯」模式中編輯圖形佈置，包括顏色和字型、圖註等等。如需相關資訊，請參閱主題 第 138 頁的『使用圖形工具列及選用區』。

「叢集」視圖有兩個 Web 圖形。

- **概念 Web 圖形**。這個圖形會呈現所選叢集內的所有概念以及在該叢集外的鏈結概念。這個圖形可以幫助您瞭解叢集內的概念如何鏈結以及所有外部鏈結。如需相關資訊，請參閱主題 『概念 Web 圖形』。
- **叢集 Web 圖形**。這個圖形會呈現選取的叢集，同時所選叢集之間的所有外部鏈結會顯示為點虛線。如需相關資訊，請參閱主題 『叢集 Web 圖形』。

如需相關資訊，請參閱主題 第 119 頁的第 10 章, 『分析叢集』。

概念 Web 圖形

此標籤會顯示 Web 圖形，其中顯示所選取叢集內的所有概念以及該叢集外部的鏈結概念。此圖形可以協助您查看叢集內的概念如何鏈結和任何外部鏈結。叢集中的每一個概念都表示為一個根據類型顏色進行顏色編碼的節點。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。

會繪製叢集內概念之間的內部鏈結，且每一個鏈結的線條粗細度與每一個概念配對共現的文件數或相似性鏈接值直接相關，視圖形工具列上的選項而定。還會顯示某個叢集的概念與該叢集外部的那些概念之間的外部鏈結。

如果在「叢集定義」對話框中選取了概念，則概念 Web 圖形會顯示那些概念以及與那些概念相關聯的任何內部及外部鏈結。其他概念之間不包括其中一個已選取概念的任何鏈結都不會在該圖形上出現。

註：依預設，圖形處於互動/選取模式，您可以在其中移動節點。但您可以在「編輯」模式中編輯圖形佈置，包括顏色及字型、圖註等。如需相關資訊，請參閱 第 138 頁的『使用圖形工具列及選用區』。

如果您使用**複製視覺化資料**按鈕複製圖形資料，並將其貼上到試算表或文字編輯器，您會看到為資料提供了直欄標頭，例如，V1、V2 一直到 V7。這些直欄包含下列資訊：

- **V1、V2** 這些值對應於畫面座標（分別為 X 及 Y）。
- **V3、V6** 列出概念類型。
- **V4、V5** 顯示概念標籤。
- **V7** 目前未用。

叢集 Web 圖形

此標籤會顯示 web 圖形，其中顯示所選取的叢集。所選取叢集之間的外部鏈結以及其他叢集之間的任何鏈結都顯示為點虛線。在叢集 Web 圖形中，每一個節點代表整個叢集，在它們之間繪製的線條粗細度代表兩個叢集之間的外部鏈結數。

重要事項！ 為了顯示叢集 Web 圖形，您必須已使用外部鏈結建置了叢集。外部鏈結是不同叢集中概念配對之間的鏈結（一個叢集內的概念和另一個叢集中的概念）。

例如，比如說有兩個叢集。Cluster A 有三個概念：A1、A2 和 A3。Cluster B 有兩個概念：B1 和 B2。下列概念是鏈結的：A1-A2、A1-A3、A2-B1（外部）、A2-B2（外部）、A1-B2（外部）及 B1-B2。這表示在叢集 Web 圖形中，線條粗細度會代表三個外部鏈結。

附註：依預設，圖形處於互動/選取模式，您可以在其中移動節點。但您可以在「編輯」模式中編輯圖形佈置，包括顏色及字型、圖註等。如需相關資訊，請參閱主題 第 138 頁的『使用圖形工具列及選用區』。

文字鏈結分析圖形

在擷取「文字鏈結分析 (TLA)」型樣之後，您可以在「視覺化」窗格的 Web 圖形中以視覺化的方式探索它們。視覺化窗格會提供兩個 TLA 型樣視景：概念（型樣）Web 圖形及類型（型樣）Web 圖形。此窗格中的 Web 圖形可用來以視覺化的方式表示型樣。「視覺化」窗格位於「文字鏈結分析」的右上角。如果已經不可見，則可以從「檢視」功能表（**檢視 > 窗格 > 視覺化**）中存取此窗格。如果未選取任何內容，則圖形區域是空的。

附註：依預設，圖形處於互動/選取模式，您可以在其中移動節點。但您可以在「編輯」模式中編輯圖形佈置，包括顏色及字型、圖註等。如需相關資訊，請參閱主題 『使用圖形工具列及選用區』。

「文字鏈結分析」視圖有兩個 Web 圖形。

- **概念 Web 圖形**。此圖形會呈現所選取型樣中的所有概念。概念圖形中的線寬及節點大小（如果類型圖示不顯示）會顯示所選取表格中的廣域出現次數。如需相關資訊，請參閱主題 『概念 Web 圖形』。
- **類型 Web 圖形**。此圖形會呈現所選取型樣中的所有類型。該圖形中的線寬及節點大小（如果類型圖示不顯示）會顯示所選取表格中的廣域出現次數。節點由類型顏色或圖示代表。如需相關資訊，請參閱主題 『類型 Web 圖形』。

如需相關資訊，請參閱主題 第 125 頁的第 11 章, 『探索文字鏈結分析』。

概念 Web 圖形

這個 Web 圖形會呈現目前選取內容中表示的所有概念。例如，如果您選取的類型型樣有三個相符的概念型樣，則此圖形會顯示三組鏈結的概念。概念圖形中的線寬及節點大小代表廣域頻率計數。該圖形會以視覺化的方式表示與型樣窗格中所選取內容相同的資訊。每一個概念的類型由顏色或圖示呈現，視您在圖形工具列上選取的內容而定。如需相關資訊，請參閱主題 『使用圖形工具列及選用區』。

類型 Web 圖形

這個 Web 圖形會呈現目前選取內容的每一個類型型樣。例如，如果您選取了兩個概念型樣，則此圖形會顯示所選取型樣中每個類型一個節點，以及在相同型樣中找到的那些節點之間的鏈結。線寬及節點大小代表集的廣域頻率計數。該圖形會以視覺化的方式表示與型樣窗格中所選取內容相同的資訊。除了圖形中出現的類型名稱以外，還會透過其顏色或類型圖示識別其類型，視您在圖形工具列上選取的內容而定。如需相關資訊，請參閱主題 『使用圖形工具列及選用區』。

使用圖形工具列及選用區

對於每一個圖形，有一個工具列，為您提供部分一般選用區的快速存取，您可以從中利用圖形執行許多動作。每一個視圖（種類與概念、叢集及「文字鏈結分析」）都具有略微不同的工具列。您可以在探索視圖模式與編輯視圖模式之間進行選擇。

使用「探索」模式可讓您以分析方式探索視覺化所表示的資料和值，「編輯」模式則可讓您變更視覺化的配置及外觀。例如，您可以變更字型和顏色來搭配您組織樣式使用手冊的外觀。若要選取此模式，請從功能表中選擇**檢視** > **視覺化窗格** > **編輯模式**（或者按一下工具列圖示）。

在「編輯」模式中，有數個影響視覺化不同配置方式的工具列。如果其中有您用不到的工具，您可以隱藏這些工具列以增加對話框中的空間，如此一來才能顯示圖表。若要選取或取消選取工具列，請在「視圖」功能表上按一下相關工具列或選用區名稱。

如需用於編輯圖形之所有一般工具列及選用區的相關資訊，請參閱線上說明或檔案 *ModelerSPOnodes.pdf* 中有關「編輯視覺化」的小節，您的產品下載會隨副該檔案。

表 *38.* 文字分析工具列按鈕

按鈕/清單	說明
	啟用「編輯」模式。切換至「編輯」模式，以變更圖形的外觀，例如放大字型、變更顏色以符合您公司的樣式手冊或移除標籤與圖註。
	啟用「探索」模式。依預設，會開啟「探索」模式，這表示您可以在圖形周圍移動及拖曳節點，以及將游標移至圖形物件上方以顯示其他「工具提示」資訊。
	為「種類與概念」視圖以及「文字鏈結分析」視圖中的圖形選取一種 Web 顯示畫面。 • **圓形佈置** 可以套用至任何圖形的一般佈置。它會佈置一個圖形，假設鏈結未導向，並將所有節點視為相同。節點僅放置在圓圈周界周圍。 • **網路佈置** 可以套用至任何圖形的一般佈置。它會佈置一個圖形，假設鏈結未導向，並將所有節點視為相同。節點會自由放置在佈置內。 • **引導佈置** 應該僅用於引導圖形的佈置。此佈置會產生類似於樹狀結構，從根節點向下到葉節點，並按顏色組織。利用此佈置可能很好地顯示階層式資料。 • **網格佈置** 可以套用至任何圖形的一般佈置。它會佈置一個圖形，假設鏈結未導向，並將所有節點視為相同。節點僅放置在空間內的網格點上。
	鏈結大小呈現。選擇圖形中線條粗細代表什麼。這僅適用於「叢集」視圖。「叢集」Web 圖形僅顯示叢集之間的外部鏈結數目。您可以在下列各項之間選擇： • **相似性** 粗細指出兩個叢集之間的外部鏈結數目 • **共生** 粗細指出發生描述子共生的文件數目。
	一個切換按鈕，按下時顯示圖註。未推上該按鈕時，不會顯示圖註。
	一個切換按鈕，按下時顯示圖形中的類型圖示，而不是類型顏色。這僅適用於「文字鏈結分析」視圖。
	一個切換按鈕，按下時在圖形下方顯示「鏈結滑塊」。您可以透過滑動箭頭過濾結果。
	將顯示所選取最上層種類而不是其子種類的圖形。
	將顯示所選取最下層種類的圖形。
	此選項會控制如何在輸出中顯示子種類的名稱。 • **完整種類路徑** 此選項將輸出種類的名稱及母項種類的完整路徑，如果適用，使用斜線來分隔種類名稱與子種類名稱。 • **簡要種類路徑** 此選項將僅輸出種類的名稱，但是使用省略符號顯示問題中種類的母項種類數目。 • **底端層次種類** 此選項將僅輸出種類的名稱，但不顯示完整路徑或母項種類。

第 13 章 階段作業資源編輯器

IBM SPSS Modeler Text Analytics 會快速且精確地從文字資料中擷取主要概念。此擷取程序主要依賴語言資源來指定如何從文字資料中擷取資訊。依預設，這些資源來自資源範本。

IBM SPSS Modeler Text Analytics 隨附了一組特殊化的**資源範本**，其中包含一組形式為程式庫和進階資源的語言及非語言資源，來協助定義如何處理和擷取資料。如需相關資訊，請參閱主題 第 145 頁的第 14 章, 『範本及資源』。

在節點對話框中，您可以將範本資源的副本載入到節點中。進入互動式工作台階段作業內部之後，如果您想的話，則可以針對此節點的資料具體地自訂這些資源。在互動式工作台階段作業期間，您可以在資源編輯器視圖中使用資源。每當啟動互動式階段作業時，會使用在節點對話框中載入的資源執行擷取，除非您已快取節點中的資料及擷取結果。

在資源編輯器中編輯資源

透過資源編輯器可存取用來產生互動式工作台階段作業擷取結果（概念、類型及型樣）的資源集。此編輯器與範本編輯器非常相似，不同的是在資源編輯器中，您可編輯此階段作業的資源。當您完成在資源上的工作和任何其他工作時，您可以更新建模節點來儲存此工作，以便能夠在後續互動式工作台階段作業中還原。如需相關資訊，請參閱主題 第 66 頁的『更新建模節點及儲存』。

如果您想要直接在用來將資源載入節點的範本上工作，我們建議您使用範本編輯器。您可以在資源編輯器內部執行的許多作業就像在範本編輯器中一樣執行，例如：

- **使用程式庫**。如需相關資訊，請參閱主題 第 155 頁的第 15 章, 『使用檔案庫』。
- **建立類型字典**。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。
- **將術語新增至字典**。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。
- **建立同義字**。如需相關資訊，請參閱主題 第 170 頁的『定義同義字』。
- **匯入及匯出範本**。如需相關資訊，請參閱主題第 151 頁的『匯入及匯出範本』。
- **發佈程式庫**。如需相關資訊，請參閱主題第 160 頁的『發佈程式庫』。

使用於英文、法文、德文、義大利文、葡萄牙文及西班牙文文字

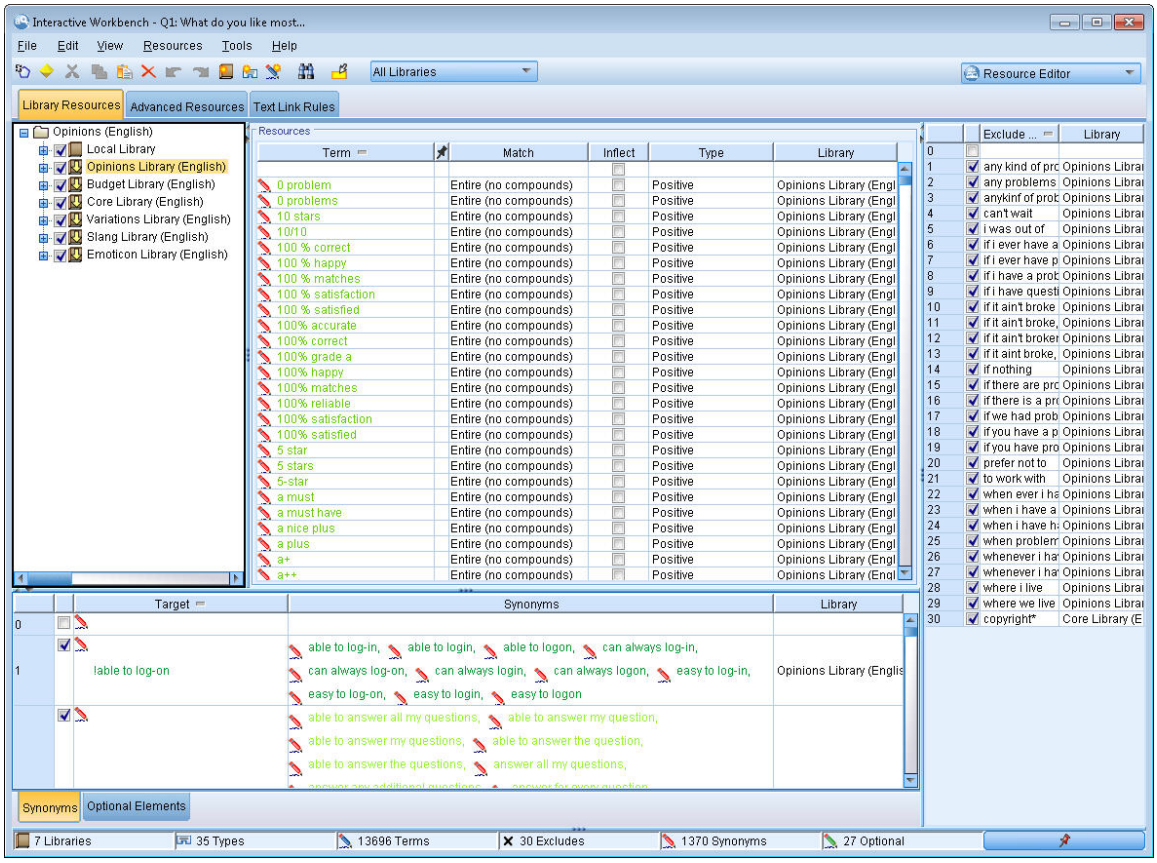


圖 *38.* 資源編輯器視圖

建立及更新範本

只要您對資源進行變更，並且想要在未來重複使用它們，則可以將資源儲存為範本。這樣做時，您可以選擇使用現有範本名稱或透過提供新的名稱進行儲存。然後，只要您在未來載入此範本，則將能夠取得相同的資源。如需相關資訊，請參閱主題 第 23 頁的『從範本及 TAP 複製資源』。

附註：您還可以發佈及共用檔案庫。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。

若要建立（或更新）範本

1. 從資源編輯器視圖的功能表中，選擇**資源 > 建立資源範本**。即會開啟「建立資源範本」對話框。
2. 如果您要建立新的範本，則在「範本名稱」欄位中輸入新的名稱。如果您要使用目前載入的資源改寫現有範本，請選取表格中的範本。
3. 按一下**儲存**，以建立範本。

重要事項！ 由於當您在節點中選取範本時會載入範本，而不是在執行串流時載入，因此如果您要取得最新的變更，請確保在使用該資源範本的任何其他節點中重新載入該範本。如需相關資訊，請參閱主題 第 150 頁的『載入之後更新節點資源』。

切換資源範本

如果您要將階段作業中目前載入的資源取代為另一個範本中那些項的副本，您可以切換至那些資源。這樣做將改寫階段作業中目前載入的所有資源。如果您要切換資源以具有部分預先定義的「文字鏈結分析 (TLA)」型樣規則，請確保選取已在 TLA直欄中標記它們的範本。

當您想要還原階段作業工作（種類、型樣及資源），但是想要從範本中載入更新的資源副本而不失去其他階段作業工作時，切換資源特別有用。您可以選取要將其內容複製到 資源編輯器 的範本，然後按一下**確定**。這代表您取代此 階段作業中的資源。如果您要在下一次啟動互動式工作台階段作業時保持這些變更，請確保您在階段作業結尾更新建模節點。

註：如果您在互動式階段作業期間切換至另一個範本的內容，則節點中列出的範本名稱仍將是所載入及所複製的最後一個範本的名稱。為了從這些資源或其他階段作業工作中獲益，請在現有階段作業之前更新您的建模節點，並在節點中選取**使用階段作業工作**選項。如需相關資訊，請參閱主題 第 66 頁的『更新建模節點及儲存』。

若要切換資源

1. 從資源編輯器視圖的功能表中，選擇**資源 > 切換資源範本**。即會開啟「切換資源」對話框。
2. 從表格中顯示的項目中選取您要使用的範本。
3. 按一下**確定**以放棄目前載入的那些資源，並在所選取範本中的原位置載入那些資源的副本。如果您已對資源進行變更，並且想要儲存檔案庫以供未來使用，則可以在切換之前發佈、更新及共用它們。請參閱第 159 頁的『共用程式庫』主題，以取得更多資訊。

第 14 章 範本及資源

IBM SPSS Modeler Text Analytics 會快速且精確地從文字資料中擷取主要概念。此擷取程序主要依賴語言資源來指定如何從文字資料中擷取資訊。如需相關資訊，請參閱主題 第 4 頁的『擷取如何運作』。您可以在資源編輯器視圖中對這些資源進行細部調整。

當您安裝軟體時，您還會取得一組特殊化的資源。透過這些隨附資源，您可從針對特定語言和特定應用程式的多年研究及細部調整中得到好處。因為隨附資源不能一律完全地適合資料的環境定義，您可以編輯這些資源範本，甚至建立並使用對組織的資料進行唯一細部調整的自訂程式庫。這些資源提供了各種形式，每一種形式都可以在階段作業中使用。資源可以在下列位置中找到：

- **資源範本**。範本由一組程式庫、類型及部分進階資源組成，它們共同形成了一組特殊化的資源，可適合特定的網域或環境定義，例如，產品意見。
- **文字分析套件 (TAP)**。除了儲存在範本中的資源以外，TAP 還將一個或多個使用那些資源產生的特殊化種類集組合在一起，因此種類和資源都儲存在一起，並可重複使用。如需相關資訊，請參閱主題 第 113 頁的『使用文字分析套件』。
- **程式庫**。程式庫會用作 TAP 和範本的建置區塊。它們還可以個別新增至階段作業中的資源。每一個程式庫由用來定義及管理類型、同義字和排除清單的數個字典組成。雖然程式庫也會個別隨附，但它們是在範本和 TAP 中預先包裝在一起的。如需相關資訊，請參閱主題 第 155 頁的第 15 章, 『使用檔案庫』。

附註：在擷取期間，還會使用一些編譯的內部資源。這些編譯的資源包含大量補足「核心」程式庫中類型的定義。無法編輯這些編譯的資源。

透過資源編輯器可存取用來產生擷取結果（概念、類型及型樣）的資源集。您可以在資源編輯器中執行的作業有很多，它們包括：

- **使用程式庫**。如需相關資訊，請參閱主題 第 155 頁的第 15 章, 『使用檔案庫』。
- **建立類型字典**。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。
- **將術語新增至字典**。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。
- **建立同義字**。如需相關資訊，請參閱主題 第 170 頁的『定義同義字』。
- **更新 TAP 中的資源**。如需相關資訊，請參閱主題 第 115 頁的『更新文字分析套件』。
- **製作範本**。如需相關資訊，請參閱主題 第 142 頁的『建立及更新範本』。
- **匯入及匯出範本**。如需相關資訊，請參閱主題 第 151 頁的『匯入及匯出範本』。
- **發佈程式庫**。如需相關資訊，請參閱主題 第 160 頁的『發佈程式庫』。

範本編輯器與資源編輯器

使用和編輯範本、程式庫及其資源有兩種主要的方法。您可以在範本編輯器或資源編輯器中使用語言資源。

範本編輯器

透過範本編輯器，您可建立及編輯資源範本，而無需互動式工作台階段作業，也不必相依於特定的節點或串流。您可以使用此編輯器來建立或編輯資源範本，然後再將它們載入「文字鏈結分析」節點和「文字採礦」建模節點。

範本編輯器可透過**工具 > 文字分析範本編輯器**功能表中的 IBM SPSS Modeler 主要工具列存取。

資源編輯器

透過可在互動式工作台階段作業內存取的資源編輯器，您可在特定節點和資料集的環境定義中使用資源。當您將「文字採礦」建模節點新增至串流時，您可以載入資源範本內容的副本或文字分析套件（種類集和資源）的副本來控制如何擷取文字進行文字採礦。啟動互動式工作台階段作業時，除了建立種類、擷取文字鏈結分析型樣和建立種類模型，您還可以在整合的資源編輯器視圖中針對該階段作業的資料對資源進行細部調整。如需相關資訊，請參閱主題 第 141 頁的『在資源編輯器中編輯資源』。

每當您在互動式工作台階段作業中使用資源時，那些變更僅會套用至該階段作業。如果您想要儲存工作（資源、種類、型樣等）以便能夠在後續階段作業中繼續，則必須更新建模節點。如需相關資訊，請參閱主題 第 66 頁的『更新建模節點及儲存』。

如果您想要將變更儲存回其內容已複製到建模節點的原始範本，以便這個更新的範本可以載入到其他節點，則您可以從資源中製作範本。如需相關資訊，請參閱主題 第 142 頁的『建立及更新範本』。

註： 如果您變更範本或程式庫，並將它們儲存至備份目錄，然後升級 IBM SPSS Modeler Text Analytics 的版本，則將會提供您匯入自訂範本和程式庫的選項。當您第一次執行 SPSS Modeler Text Analytics 串流或是在升級之後開啟「資源編輯器」時，預設範本和程式庫會複製到您的機器。畫面上會顯示**已存範本**警告或**已存程式庫**警告（或兩者），以及在產品升級過程中所更新的範本和/或程式庫清單，並提供您從儲存自訂範本和程式庫所在的目錄匯入它們的選項。在警告訊息中按一下**確定**之後，您可以隨時開啟**管理資源範本**對話框或**管理程式庫**對話框來選擇您要匯入哪些自訂範本或程式庫。

編輯器介面

您在範本編輯器或資源編輯器中執行的作業與管理及細部調整語言資源有關。這些資源以範本及程式庫的形式儲存。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。

程式庫資源標籤

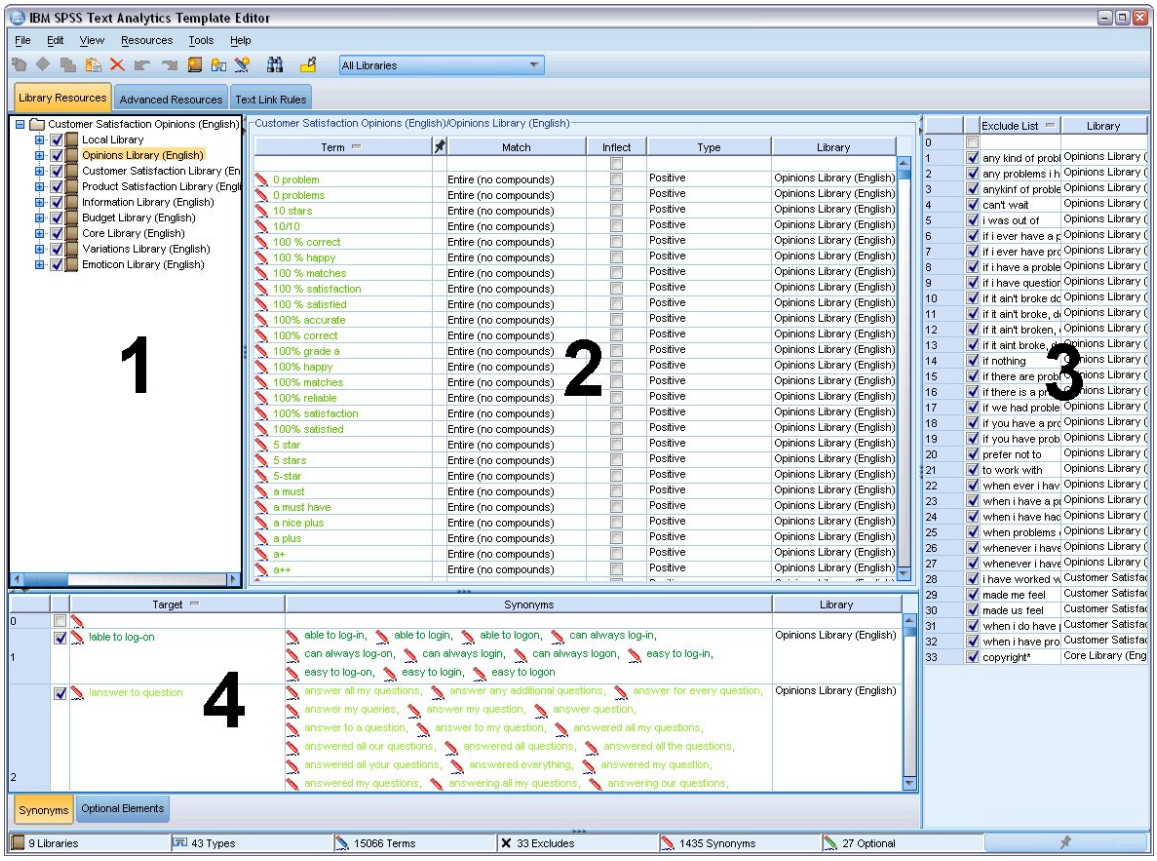


圖 *39.* 文字採礦範本編輯器

介面組織為四個部分，如下所示：

1.「檔案庫樹狀結構」窗格。此窗格位於左上角，顯示檔案庫的樹狀結構。您可以在此樹狀結構中啟用及停用檔案庫，以及透過在樹狀結構中選取一個檔案庫以在其他窗格中過濾視圖。您可以使用快速功能表，在此樹狀結構中執行許多作業。如果您在樹狀結構中展開檔案庫，則可以查看其包含的類型集。如果您要僅聚焦於特定檔案庫，還可以透過**視圖**功能表過濾此清單。

2.「類型定義檔」窗格中的術語清單。此窗格位於檔案庫樹狀結構的右側，顯示樹狀結構中選取之檔案庫的類型定義檔術語清單。**類型定義檔**是根據一個標籤或類型、名稱分組的術語集合。當擷取引擎讀取您的文字資料時，它會對在文字中找到的單字與類型定義檔中的術語進行比較。如果所擷取概念顯示為類型定義檔中的一個術語，則會指派該類型名稱。您可以將類型定義檔視為具有某些共同點的特定術語定義檔。例如，「核心」檔案庫中的 <Location> 類型包含諸如 new orleans、great britain 及 new york 等概念。這些術語全部代表地理位置。檔案庫可以包含一個或多個類型定義檔。如需相關資訊，請參閱主題 第 163 頁的『類型字典』。

3.「排除定義檔」窗格。此窗格位於右側，顯示將從最終擷取結果中排除的術語集合。此排除定義檔中顯示的術語不會顯示在「擷取結果」窗格中。排除的術語可以儲存在您選擇的檔案庫中。然而，「排除定義檔」窗格顯示檔案庫樹狀結構中可見的所有檔案庫的所有已排除術語。如需更多資訊，請參閱主題第 172 頁的『排除字典』。

4.「替代定義檔」窗格。此窗格位於左下方，顯示同義字及選用元素，每一個都位於自己的標籤中。同義字及選用元素協助在最終擷取結果中根據一個前導或目標、概念分組類似的術語。此定義檔可能包含已知同義字及使用定義的同義字與元素，以及與正確拼字配對的常見錯誤拼字。同義字定義及選用元素可以儲存在您選擇的檔案庫中。然而，替代定義檔窗格會顯示在檔案庫樹狀結構中可見的所有檔案庫的所有內容。由於此窗格顯示

所有檔案庫中的所有同義字或選用元素，因此樹狀結構中所有檔案庫的替代項都在此窗格中顯示在一起。檔案庫只能包含一個替代定義檔。如需相關資訊，請參閱主題 第 169 頁的『替代/同義字字典』。

附註：

- 如果您要過濾以便只看到單一檔案庫的相關資訊，則可以使用工具列上的下拉清單變更檔案庫視圖。它包含稱為**所有檔案庫**的最上層項目，以及每一個個別檔案庫的其他項目。如需相關資訊，請參閱主題 第 157 頁的『檢視程式庫』。

進階資源標籤

從編輯器視圖的第二個標籤可取得進階資源。您可以在此標籤中檢閱和編輯進階資源。如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。

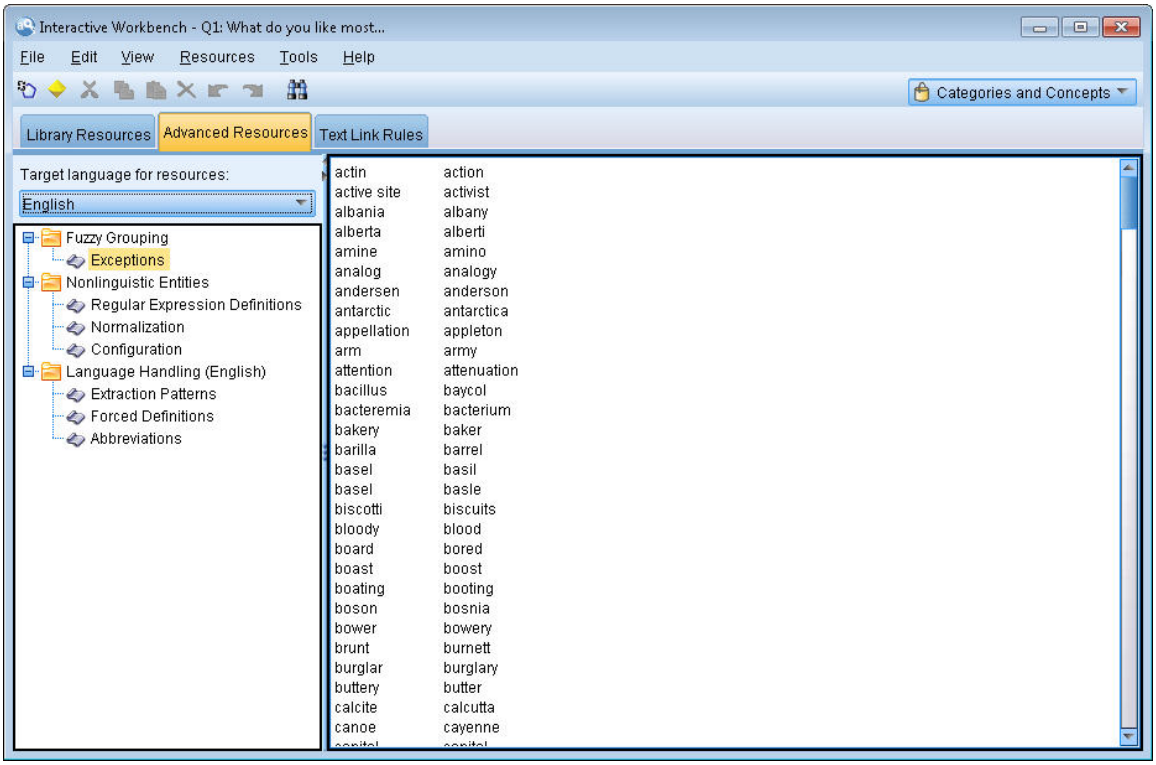


圖 *40.* 文字挖掘範本編輯器 - 進階資源標籤

文字鏈結規則標籤

自第 14 版以來，文字鏈結分析規則可在編輯器視圖的自己的標籤中編輯。您可以在規則編輯器中工作、建立自己的規則，甚至執行模擬來查看您的規則如何影響 TLA 結果。如需相關資訊，請參閱主題 第 187 頁的第 18 章, 『關於文字鏈結規則』。

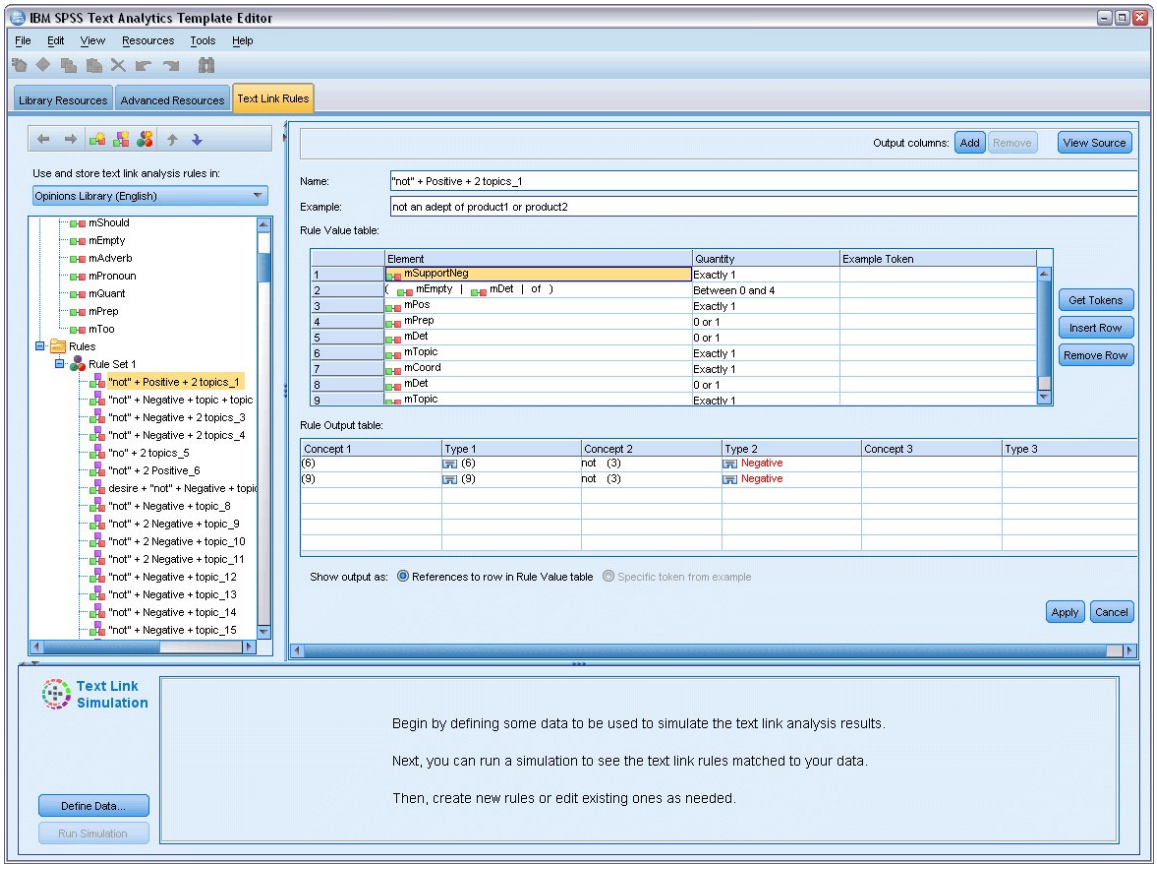


圖 *41.* 文字採礦範本編輯器 - 文字鏈結規則標籤

開啟範本

啟動範本編輯器時，系統會提示您開啟範本。同樣，您可以從「檔案」功能表中開啟範本。如果您想要包含部分文字鏈結分析 (TLA) 規則的範本，請確定選取的範本在 TLA 直欄中有圖示。為其建立範本的語言會在「語言」直欄中顯示。

如果您想要匯入表格中未顯示的範本，或如果您想要匯出範本，則可以使用「開啟範本」對話框中的按鈕。如需相關資訊，請參閱主題 第 151 頁的『匯入及匯出範本』。

開啟範本

1. 從範本編輯器中的功能表中，選擇**檔案 > 開啟資源範本**。「開啟資源範本」對話框即會開啟。
2. 從表格中顯示的那些範本中選取您想要使用的範本。
3. 按一下**確定**以開啟此範本。如果您目前在編輯器中開啟了另一個範本，按一下「確定」會放棄該範本並顯示您在這裡選取的範本。如果您已對資源進行了變更並想要儲存程式庫以供將來使用，則可以在開啟另一個程式庫之前發佈、更新及共用它們。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。

儲存範本

在範本編輯器中，您可以儲存對範本所做的變更。您可以選擇使用現有範本名稱或透過提供新名稱進行儲存。

如果您對之前已載入節點的範本進行變更，則必須將範本內容重新載入到節點才能取得最新變更。如需相關資訊，請參閱主題 第 23 頁的『從範本及 TAP 複製資源』。

或者，如果您使用「文字採礦」節點的「模型」標籤中的選項**使用儲存的互動式工作**，則表示您使用的是之前互動式工作台階段作業中的資源，您必須從互動式工作台階段作業內切換到此範本的資源。如需相關資訊，請參閱主題 第 143 頁的『切換資源範本』。

附註：您也可以發佈並共用程式庫。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。

儲存範本

1. 從範本編輯器中的功能表中，選擇**檔案 > 儲存資源範本**。「儲存資源範本」對話框即會開啟。
2. 如果您想要將此範本儲存為新範本，請在「範本」名稱欄位中輸入新名稱。如果您想要使用目前載入的資源改寫現有範本，請在表格中選取範本。
3. 如果需要，請輸入說明以在表格中顯示註解或註釋。
4. 按一下**儲存**以儲存範本。

重要事項！ 因為範本或 TAP 中的資源會載入/複製到節點中，如果您對範本進行變更並想要在現有串流中從這些變更中得到好處，則必須透過重新載入資源來進行更新。如需相關資訊，請參閱主題 『載入之後更新節點資源』。

載入之後更新節點資源

依預設，當您將節點新增至串流時，會將預設範本中的一組資源載入並內嵌到節點中。如果您變更範本或使用 TAP，載入它們時，那些資源的副本則會改寫資源。因為範本及 TAP 未直接鏈結至節點，所以您對範本或 TAP 進行的任何變更都不會在現有的點選中自動提供。為了從那些變更中得到好處，您只能更新該節點中的資源。可以透過兩種方式中的一種來更新資源。

方法 1：在模型標籤中重新載入資源

如果您想要使用新的或更新的範本或 TAP 更新節點中的資源，則可以在節點的「模型」標籤中重新載入它。透過重新載入，您會使用更新的副本來取代節點中的資源副本。為了方便起見，更新時間和日期連同原始範本的名稱都會在「模型」標籤上出現。如需相關資訊，請參閱主題 第 23 頁的『從範本及 TAP 複製資源』。

但如果您使用「文字採礦」建模節點中的互動式階段作業資料，並且在「模型」標籤上選取了**使用階段作業工作**選項，則會使用儲存的階段作業工作及資源，並會停用**載入**按鈕。停用它是因為在互動式工作台階段作業期間，您選擇了**更新建模節點**選項並保留了種類、資源及其他階段作業工作。在這種情況下，如果您想要變更或更新那些資源，則可以嘗試下一個在資源編輯器中切換資源的方法。

方法 2：在資源編輯器中切換資源

在互動式階段作業期間無論您何時想要使用不同的資源，都可以使用「切換資源」對話框交換那些資源。當您想要重複使用現有種類工作但取代資源時，此對話框特別有用。在此情況下，您可以在「文字採礦」建模節點的「模型」標籤上選取**使用階段作業工作**選項。這樣做會停用透過節點對話框重新載入範本的功能，而是保留您在階段作業期間進行的設定及變更。然後您可以透過執行串流啟動互動式工作台階段作業，並在資源編輯器中切換資源。如需相關資訊，請參閱主題 第 143 頁的『切換資源範本』。

為了保留階段作業工作（包括資源）以供後續階段作業使用，您必須從互動式工作台階段作業內更新建模節點，以便將資源（及其他資料）保存回節點。如需相關資訊，請參閱主題 第 66 頁的『更新建模節點及儲存』。

附註：如果您在互動式階段作業期間切換至另一個範本的內容，則節點中列出的範本名稱仍為載入並複製的最後一個範本的名稱。為了從這些資源或其他階段作業工作中得到好處，請在結束階段作業之前更新建模節點。

管理範本

您可能想要經常對範本執行的還有一些基本的管理作業，例如，重新命名範本、匯入及匯出範本，或刪除已作廢的範本。這些作業會在「管理範本」對話框中執行。匯入及匯出範本可讓您與其他使用者共用範本。如需相關資訊，請參閱主題 『匯入及匯出範本』。

附註：您無法重新命名或刪除此產品安裝（或隨附）的範本。因此，如果您想要重新命名，則可以開啟已安裝的範本，並使用您選擇的名稱製作新範本。您可以刪除自訂範本；但如果您嘗試刪除隨附的範本，則會將其重設為原始安裝的版本。

重新命名範本

1. 從功能表中，選擇**資源 > 管理資源範本**。「管理範本」對話框即會開啟。
2. 選取您要重新命名的範本，並按一下**重新命名**。名稱方框會在表格中變成可編輯的欄位。
3. 鍵入新名稱並按 Enter 鍵。確認對話框即會開啟。
4. 如果您對名稱變更感到滿意，請按一下**是**。如果不滿意，請按一下**否**。

刪除範本

1. 從功能表中，選擇**資源 > 管理資源範本**。「管理範本」對話框即會開啟。
2. 在「管理範本」對話框中，選取您要刪除的範本。
3. 按一下「**刪除**」。確認對話框即會開啟。
4. 按一下**是**以刪除，或按一下**否**以取消要求。如果您按一下**是**，則會刪除範本。

匯入及匯出範本

您可以透過匯入及匯出範本來與其他使用者或機器共用範本。範本儲存在內部資料庫中，但可以作為 **.lrt* 檔案匯出至硬碟。

因為您在某些情況下可能想要匯入或匯出範本，有數個對話框可提供那些功能。

- 範本編輯器中的「開啟範本」對話框
- 「文字採礦」建模節點及「文字鏈結分析」節點中的「載入資源」對話框。
- 範本編輯器及資源編輯器中的「管理範本」對話框。

匯入範本

1. 在對話框中，按一下**匯入**。「匯入範本」對話框即會開啟。
2. 選取要匯入的資源範本檔案 (**.lrt*)，然後按一下**匯入**。您可以用另一個名稱儲存正在匯入的範本，或改寫現有範本。對話框即會關閉，而範本現在會在表格中出現。

匯出範本

1. 在對話框中，選取您要匯出的範本，然後按一下**匯出**。「選取目錄」對話框即會開啟。
2. 選取您要匯出的目錄，然後按一下**匯出**。此對話框即會關閉，而範本會匯出並攜帶副檔名 (**.lrt*)

結束範本編輯器

在範本編輯器中完成工作時，您可以儲存工作並結束該編輯器。

結束範本編輯器

1. 從功能表中，選擇**檔案 > 關閉**。「儲存後關閉」對話框即會開啟。
2. 選取**將變更儲存到範本**以便在關閉編輯器之前儲存開啟的範本。
3. 如果您想要在關閉編輯器之前，在開啟的範本中發佈任何程式庫，請選取**發佈程式庫**。如果您選取此選項，系統會提示您選取要發佈的程式庫。如需相關資訊，請參閱主題 第 160 頁的『發佈程式庫』。

備份資源

作為一種安全措施，您可能想要經常備份資源。

重要事項！ 當您還原時，產品中資源的整個內容都會抹除乾淨，只有備份檔的內容可以存取。這包括任何開啟的工作。

附註：您只能備份及還原至軟體的相同主要版本。例如，如果您從第 15 版備份，則無法將該備份還原到第 16 版。

備份資源

1. 從功能表中，選擇**資源 > 備份工具 > 備份資源**。「備份」對話框。
2. 輸入備份檔的名稱，然後按一下**儲存**。對話框即會關閉，而備份檔會建立。

還原資源

1. 從功能表中，選擇**資源 > 備份工具 > 還原資源**。警示會警告您還原將改寫資料庫的現行內容。
2. 請按一下**是**繼續進行。這時會開啟對話框。
3. 選取您要還原的備份檔，然後按一下**開啟**。對話框即會關閉，而資源會在應用程式中還原。

匯入資源檔

如果您已直接在此產品外的資源檔中進行變更，則可以將它們匯入到選取的檔案庫中，方法為選取該檔案庫，然後進行匯入。當您匯入目錄時，也可以將所有受支援的檔案匯入到特定的開啟檔案庫中。您只能匯入 `*.txt` 檔案。

每一個匯入的檔案中，每行只能包含一個項目，且若是內容的結構為：

- 單字或詞組清單（每行一個）。檔案會匯入作為類型字典的詞彙清單，其中的類型字典取用檔案的名稱去掉副檔名。
- 一份項目清單（如 `term1 <TAB> term2`），然後會匯入它作為同義字清單，其中 `term1` 是基礎詞彙集，而 `term2` 是目標詞彙。

匯入單一資源檔

1. 從功能表中，選擇**資源 > 匯入檔案 > 匯入單一檔案**。即會開啟「匯入檔案」對話框。
2. 選取要匯入的檔案，然後按一下**匯入**。檔案內容會轉換成內部格式並新增至檔案庫。

匯入目錄中的所有檔案

1. 從功能表中，選擇**資源 > 匯入檔案 > 匯入整個目錄**。即會開啟「匯入目錄」對話框。

2. 選取您要所有的資源檔在其中從**匯入**清單匯入。如果您選取**預設值**選項，則會使用目錄名稱作為其名稱來建立新的檔案庫。
3. 選取要從中匯入檔案的目錄。將不會讀取子目錄。
4. 按一下**匯入**。這時對話框會關閉，並且來自那些匯入的資源檔的內容會出現在編輯器中的字典和進階資源檔表單中。

第 15 章 使用檔案庫

擷取引擎用來從文字資料中擷取和群組詞彙的資源一律包含一或多個檔案庫。您可以在位於 範本編輯器 和 資源編輯器 左上側的檔案庫樹狀結構中看到檔案庫集。檔案庫由三種字典組成： 「類型」、「替代」和「排除」。請參閱第 163 頁的第 16 章, 『關於程式庫字典』主題，以取得更多資訊。

您選擇的 TAP 中的資源範本或資源包含數個檔案庫，可讓您立即開始從文字資料中擷取概念。不過，您也可以建立您自己的檔案庫及發佈它們，讓您可以重複使用它們。如需相關資訊，請參閱主題 第 160 頁的『發佈程式庫』。

例如，假設您經常使用與汽車產業相關的文字資料。在分析資料之後，您決定您想要建立一些自訂的資源來處理產業特有的詞彙或專門術語。利用 範本編輯器，您可以建立新的範本，以及在其中建立檔案庫來擷取及群組汽車術語。由於您將會再度需要此檔案庫中的資訊，因此您將檔案庫發佈到可在**管理檔案庫**對話框中存取的中央儲存庫，以便可在不同的串流階段作業中獨立重複使用它。

假設您也對分組不同子產業特有的術語感興趣，例如電子裝置、引擎、冷卻系統，或甚至特定的製造商或市場。您可以針對每一個群組各建立一個檔案庫，然後發佈檔案庫，以便它們可以與多組文字資料搭配使用。如此一來，您就可以新增與文字資料的環境定義最為對應的檔案庫。

附註：您可以在「進階資源」標籤中配置及管理其他資源。某些資源適用於所有檔案庫及管理非語言實體、模糊分組異常狀況等等。此外，您還可以在「文字鏈結規則」標籤中編輯文字鏈結分析型樣規則，這些規則為檔案庫所特有。 如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。

隨附的程式庫

依預設，IBM SPSS Modeler Text Analytics 安裝了數個程式庫。 您可以使用這些預先格式化的程式庫來存取數以千計預先定義的術語及同義字以及多個不同的類型。這些隨附的程式庫經過細部調整，可適合數個不同的地區並會以數種不同的語言提供。

程式庫有很多，但最常用的如下所示：

- **本端程式庫。**用來儲存使用者定義的字典。它是依預設新增至所有資源的空程式庫。它也包含空的類型字典。直接從「種類和概念」視圖、「叢集」視圖及「文字鏈結分析」視圖中對資源進行變更或精簡（例如，將單字新增至類型）時，它極其有用。在此情況下，那些變更和精簡會自動儲存在資源編輯器的程式庫樹狀結構中列出的第一個程式庫中；依預設，這是本端程式庫。您不能發佈此程式庫，因為它是階段作業資料的特定程式庫。如果您想要發佈其內容，則必須先重新命名該程式庫。
- **核心程式庫。**可在大部分情況下使用，因為它包含基本的五種內建類型，代表人員、位置、組織、產品和不明。雖然您可能看到在它的其中一個類型字典中只列出了幾個術語，但「核心」程式庫中代表的類型實際會補足在文字採礦產品隨附的內部編譯資源中找到的強韌類型。這些內部編譯資源針對每一個類型包含數以千計的術語。基於此原因，雖然您可能在類型字典術語清單中看不到術語，但仍可以使用「核心」類型進行擷取和歸類。這說明了只有 *John* 在「核心」程式庫中的 `<Person>` 類型字典中出現時，如何擷取 *George* 等名字並作為 `<Person>` 鍵入。同樣，如果您不包括「核心」程式庫，則仍可以在擷取結果中看到這些類型，因為仍由擷取引擎使用包含這些類型的編譯資源。
- **意見程式庫。**最經常用來從文字資料中擷取意見及觀感。此程式庫包括數以千計的單字，當與其他術語一起使用時代表指出主題相關意見的屬性、限定元及喜好設定。此程式庫包括大量內建類型、同義字及排

除。它還包括大量用於文字鏈結分析的型樣規則。若要從此程式庫中的文字鏈結分析規則及其產生的型樣結果中得到好處，必須在「文字鏈結規則」標籤中指定此程式庫。請參閱第 187 頁的第 18 章, 『關於文字鏈結規則』主題，以取得更多資訊。

- **預算程式庫。**用來擷取指代某物成本的術語。此程式庫包括多個單字及詞組，代表關於某物價格或品質的形容詞、限定元及判斷。
- **變異程式庫。**用來包括某些語言變異需要同義字定義來將其適當分組的情況。此程式庫僅包括同義字定義。

雖然在範本外部隨附的部分程式庫與部分範本中的內容類似，但範本已特別轉向特定應用程式並包含其他進階資源。我們建議您嘗試使用針對您使用的文字資料類型而設計的範本，並對那些資源進行變更而不只是將個別程式庫新增至更通用的範本。

編譯資源也會隨附於 IBM SPSS Modeler Text Analytics。 它們一律在擷取程序期間使用，並包含針對預設程式庫中內建類型字典的大量補足定義。因為這些資源已經編譯，所以無法進行檢視或編輯。然而，您可以將這些編譯資源歸類的術語強制到任何其他字典中。如需相關資訊，請參閱主題 第 168 頁的『強制術語』。

建立程式庫

您可以建立任何數目的程式庫。建立新的程式庫之後，您可以開始在此程式庫中建立類型字典，並輸入術語、同義字及排除。

建立程式庫

1. 從功能表中，選擇**資源 > 新建程式庫**。即會開啟程式庫內容對話框。
2. 在「名稱」文字框中輸入程式庫的名稱。
3. 如果需要，請在「註釋」文字框中輸入註解。
4. 如果您要現在發佈此程式庫，然後再在程式庫中輸入內容，請按一下**發佈**。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。您也可以稍後隨時發佈。
5. 按一下**確定**以建立程式庫。對話框即會關閉，而程式庫會在樹狀結構視圖中出現。如果您在樹狀結構中展開程式庫，您會看到程式庫中已自動包括空的類型字典。您可以立即開始在其中新增術語。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。

新增公用程式庫

如果您想要從另一個階段作業資料重複使用程式庫，只要它是公用程式庫，就可將其新增至現行資源。公用程式庫是已發佈的程式庫。如需相關資訊，請參閱主題 第 160 頁的『發佈程式庫』。

當您新增公用程式庫時，本端副本會內嵌至階段作業資料中。您可以對此程式庫進行變更；但如果您想要共用變更，則必須重新發佈程式庫的公用版本。

新增公用程式庫時，如果在一個程式庫和另一個本端程式庫中的術語與類型之間探索到任何衝突，則「解決衝突」對話框可能會出現。您必須解決這些衝突或接受提出的解決方案才能完成此作業。如需相關資訊，請參閱主題 第 161 頁的『解決衝突』。

註：如果您一律在啟動互動式工作台階段作業時更新程式庫或在關閉時發佈，則不太可能有不同步的程式庫。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。

新增程式庫

1. 從功能表中，選擇**資源 > 新增程式庫**。「新增程式庫」對話框即會開啟。
2. 在清單中選取程式庫。

3. 按一下「**新增**」。如果在新增程式庫和已存程式庫之間發生任何衝突，系統會要求您在完成作業之前驗證衝突解決方案或進行變更。如需相關資訊，請參閱主題 第 161 頁的『解決衝突』。

尋找術語及類型

您可以使用「尋找」功能在編輯器的各窗格中進行搜尋。在編輯器中，您可以從功能表中選擇**編輯 > 尋找**，「尋找」工具列即會出現。您可以使用此工具列來一次尋找一個出現項目。透過再按一下**尋找**，您可以尋找後續出現的搜尋術語。

搜尋時，編輯器僅搜尋「尋找」工具列的下拉清單中列出的程式庫。如果選取了**所有程式庫**，則程式會搜尋編輯器中的一切內容。

開始搜尋時，會從有焦點的區域開始。然後會繼續將每一個區段搜尋一遍，重新循環直到回到作用中的資料格。您可以使用方向箭頭反轉搜尋的順序。您也可以選擇搜尋是否區分大小寫。

在視圖中尋找字串

1. 從功能表中，選擇**編輯 > 尋找**。「尋找」工具列即會顯示。
2. 輸入您要搜尋的字串。
3. 按一下**尋找**按鈕以開始搜尋。然後會強調顯示下一個出現的術語或類型。
4. 再按一下該按鈕以逐個移動出現項目。

在術語中使用星號

如果您處理的是粘著型語言，透過將其他單字複合在一起而不使用岔斷空格來建立新單字，則在術語中使用星號 (*) 尤其有用。例如，德文單字 *Übernachtungspreis* 由以下單字組成：*Übernachtung + s + Preis*。

比如，如果您在類型 Budget 的術語中搜尋 preis*，則會符合擷取的概念，例如 preiserhöhung。同樣，*preis 將符合 Übernachtung，*preis* 將符合 Übernachtungspreiserhöhung。

檢視程式庫

您可以顯示一個特定程式庫或所有程式庫的內容。處理多個程式庫時或想要在發佈特定程式庫之前檢閱其內容時，這可能很有幫助。變更視圖僅會影響您在此「程式庫資源」標籤中看到的內容，但不會使任何程式庫無法在擷取期間使用。如需相關資訊，請參閱主題 第 158 頁的『停用本端程式庫』。

預設視圖是**所有程式庫**，這會顯示樹狀結構中的所有程式庫及其在其他窗格中的內容。您可以使用工具列的下拉清單或透過功能表選項（**檢視 > 程式庫**）變更此選項 檢視單一程式庫時，其他程式庫中的所有項目都會從視圖中消失，但在擷取期間仍會讀取。

變更程式庫視圖

1. 從「程式庫資源」標籤中的功能表中，選擇**檢視 > 程式庫**。這時會開啟包含所有本端程式庫的功能表。
2. 選取您想要查看的程式庫或選取**所有程式庫**選項以查看所有程式庫的內容。會根據您選取的內容過濾視圖的內容。

管理本端程式庫

本端程式庫是互動式工作台階段作業內部或範本內部的程式庫，與公用程式庫相對。如需相關資訊，請參閱主題 第 158 頁的『管理公用程式庫』。您可能想要執行的還有一些基本的本端程式庫管理作業，包括：重新命名、停用或刪除本端程式庫。

重新命名本端程式庫

您可以重新命名本端程式庫。如果您重新命名本端程式庫，則會取消其與公用版本（如果存在）的關聯。這表示無法再與公用版本共用後續變更。您可以用新名稱重新發佈此本端程式庫。這也表示，您將無法使用對此本端版本所做的任何變更來更新原始公用版本。

附註：您無法重新命名公用程式庫。

1. 從功能表中，選擇**編輯 > 程式庫內容**。「程式庫內容」對話框即會開啟。

重新命名本端程式庫

1. 在樹狀結構視圖中，選取您要重新命名的程式庫。
2. 在「名稱」文字框中輸入程式庫的新名稱。
3. 按一下**確定**以接受程式庫的新名稱。對話框即會關閉，而程式庫名稱會在樹狀結構視圖中更新。

停用本端程式庫

如果您想要從擷取程序中暫時排除程式庫，則可以取消選取樹狀結構視圖中程式庫名稱左側的勾選框。這表示您想要保留程式庫，但想要在檢查衝突時及擷取期間忽略其內容。

停用程式庫

1. 在程式庫樹狀結構窗格中，選取您要停用的程式庫。
2. 按一下空格鍵。名稱左側的勾選框即會清除。

刪除本端程式庫

您可以移除程式庫而不刪除程式庫的公用版本，反之亦可。刪除本端程式庫會僅從階段作業中刪除程式庫及其所有內容。刪除程式庫的本端版本不會從其他階段作業中刪除該程式庫或公用版本。如需相關資訊，請參閱主題 『管理公用程式庫』。

刪除本端程式庫

1. 在樹狀結構視圖中，選取您要刪除的程式庫。
2. 從功能表中，選擇**編輯 > 刪除**以刪除程式庫。即會移除程式庫。
3. 如果您之前從未發佈此程式庫，則會開啟一條訊息，詢問您是希望刪除還是保留此程式庫。按一下**刪除**以繼續，如果希望保留此程式庫，則按一下**保留**。

附註：必須一律保留一個程式庫。

管理公用程式庫

為了重複使用本端程式庫，您可以透過「管理程式庫」對話框（**資源 > 管理程式庫**）依次發佈、使用和查看它們。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。您可能想要執行的一些基本的公用程式庫管理作業包括：匯入、匯出或刪除公用程式庫。您無法重新命名公用程式庫。

匯入公用程式庫

1. 在「管理程式庫」對話框中，按一下**匯入...**。「匯入程式庫」對話框即會開啟。
2. 請選取您想要匯入的程式庫檔案 (**.lib*)，如果您還想要在本端新增此程式庫，請選取**將程式庫新增至現行專案**。
3. 按一下**匯入**。對話框即會關閉。如果相同名稱的公用程式庫已存在，系統會要求您重新命名正在匯入的程式庫或改寫現行公用程式庫。

匯出公用程式庫

您可以將公用程式庫匯出至 *.lib* 格式以便共用。

1. 在「管理程式庫」對話框中，選取清單中您要匯出的程式庫。
2. 按一下**匯出**。「選取目錄」對話框即會開啟。
3. 選取您要匯出的目錄，然後按一下**匯出**。對話框即會關閉，而程式庫檔案 (**.lib*) 會匯出。

刪除公用程式庫

您可以移除本端程式庫而不刪除程式庫的公用版本，反之亦可。但如果從此對話框中刪除了程式庫，則無法再將其新增至任何階段作業資源，直到再次發佈本端版本。

如果您刪除該產品已安裝的程式庫，則會還原原始安裝的版本。

1. 在「管理程式庫」對話框中，選取您要刪除的程式庫。您可以透過按一下適當的標頭，對清單進行排序。
2. 按一下**刪除**以刪除程式庫。IBM SPSS Modeler Text Analytics 會驗證程式庫的本端版本是否與公用程式庫相同。如果相同，則會移除程式庫而無任何警示。如果程式庫版本不同，則會開啟警示來詢問您是想要保留還是移除公用版本。

共用程式庫

程式庫可讓您以一種在多個互動式工作台階段作業之間輕鬆共用的方式使用資源。程式庫可能存在兩種狀態，或版本。可在編輯器及部分互動式工作台階段作業中編輯的程式庫稱為**本端程式庫**。例如，在互動式工作台階段作業中使用時，您可能會在蔬菜程式庫中進行大量變更。如果您的變更對其他資料有用，則可以透過建立蔬菜程式庫的**公用程式庫**版本來提供這些資源。公用程式庫，顧名思義，可用於任何互動式工作台階段作業中的任何其他資源。

您可以在「管理程式庫」對話框中查看公用程式庫。此公用程式庫版本存在後，您可以將其新增至其他環境定義中的資源，以便共用這些自訂語言資源。

隨附的程式庫一開始是公用程式庫。可以在這些程式庫中編輯資源，然後建立新的公用版本。然後可以在其他互動式工作台階段作業中存取那些新版本。

隨著您繼續使用程式庫並進行變更，您的程式庫版本將失去同步。在某些情況下，本端版本可能比公用版本更新，而在其他情況下，公用版本可能比本端版本更新。如果公用版本從另一個互動式工作台階段作業內更新，公用版本和本端版本也都可以包含另一個所不包含的變更。如果程式庫版本失去同步，則您可以再次使其同步化。同步化程式庫版本由重新發佈和/或更新本端程式庫組成。

每當您啟動或關閉互動式工作台階段作業時，系統都會提示您同步化需要更新或重新發佈的任何程式庫。此外，您還可以透過在樹狀結構視圖中程式庫名稱旁邊出現的圖示或透過檢視「程式庫內容」對話框，來輕鬆識別本端程式庫的同步化狀態。您也可以選擇透過功能表選項隨時這麼做。下表說明了五種可能的狀態及其關聯的圖示。

表 *39.* 本端程式庫同步化狀態

圖示	本端程式庫狀態說明
	未發佈 — 從未發佈過本端程式庫。

表 *39.* 本端程式庫同步化狀態 *(*繼續*)*

圖示	本端程式庫狀態說明
	已同步 — 本端和公用程式庫版本是相同的。這也適用於本端程式庫，該程式庫無法發佈，因為它用來僅包含階段作業特定的資源。
	已過期 — 公用程式庫版本比本端版本更新。您可以使用變更來更新本端版本。
	更新 — 本端程式庫版本比公用版本更新。您可以將本端版本重新發佈到公用版本。
	不同步 — 本端程式庫和公用程式庫都包含另一個所不包含的變更。您必須決定是更新還是發佈本端程式庫。如果您更新，您會遺失前次更新或發佈以來所做的變更。如果您選擇發佈，則會改寫公用版本中的變更。

附註：如果您一律在啟動互動式工作台階段作業時更新程式庫或在關閉時發佈，則不太可能有不同步的程式庫。

無論何時您認為程式庫中的變更會對可能也包含此程式庫的其他串流 有好處，都可以重新發佈該程式庫。因此，如果變更會對其他串流有好處，您就可以在那些串流中更新本端版本。這樣，您可以透過建立新的程式庫及/或將任何數目的公用程式庫新增至資源，針對套用至資料的每一個環境定義或網域建立串流。

如果程式庫的公用版本已共用，則本端版本和公用版本之間極有可能會產生差異。每當您從互動式工作台階段作業啟動或關閉並發佈或者從範本編輯器時，都會顯示一條訊息，讓您發佈和/或更新其版本與「管理程式庫」對話框中的版本不同步的任何程式庫。如果公用程式庫版本比本端版本更新，則會開啟一個對話框，詢問您是否希望進行更新。您可以選擇是依現狀保留本端版本而不使用公用版本進行更新，還是將更新合併到本端程式庫中。

發佈程式庫

如果您從未發佈過特定的程式庫，則必須在資料庫中建立本端程式庫的公用副本才能發佈。如果您正在重新發佈程式庫，則本端程式庫的內容會取代現有公用版本的內容。在重新發佈之後，您可以在任何其他串流階段作業中更新此程式庫，以便其本端版本與公用版本同步。即使您可以發佈程式庫，本端版本也一律會儲存在階段作業中。

重要事項！ 如果您對本端程式庫進行變更，而在此期間程式庫的公用版本也有所變更，則會將您的程式庫視為不同步。我們建議您先使用公用變更來更新本端版本，進行您想要的任何變更，然後再次發佈本端版本以使兩個版本相同。如果您先進行變更並發佈，則會改寫公用版本中的任何變更。

將本端程式庫發佈至資料庫

1. 從功能表中，選擇**資源 > 發佈程式庫**。「發佈程式庫」對話框即會開啟，依預設會選取需要發佈的所有程式庫。
2. 選取您想要發佈或重新發佈的每一個程式庫左側的勾選框。
3. 按一下**發佈**以將程式庫發佈到「管理程式庫」資料庫。

更新程式庫

每當您啟動或關閉互動式工作台階段作業時，都可以更新或發佈與公用版本不再同步的任何程式庫。如果公用程式庫版本比本端版本更新，則會開啟一個對話框，詢問您是否希望更新程式庫。您可以選擇是保留本端版本

而不使用公用版本進行更新，還是使用公用版本取代本端版本。如果程式庫的公用版本比本端版本更新，則可以更新本端版本以使其內容與公用版本的內容同步。更新表示將在公用版本中找到的變更納入本端版本。

附註：如果您一律在啟動互動式工作台階段作業時更新程式庫或在關閉時發佈，則不太可能有不同步的程式庫。如需相關資訊，請參閱主題 第 159 頁的『共用程式庫』。

更新本端程式庫

1. 從功能表中，選擇**資源 > 更新程式庫**。「更新程式庫」對話框即會開啟，依預設會選取需要更新的所有程式庫。
2. 選取您想要發佈或重新發佈的每一個程式庫左側的勾選框。
3. 按一下**更新**以更新本端程式庫。

解決衝突

本端與公用程式庫衝突

每當您啟動串流階段作業時，IBM SPSS Modeler Text Analytics 都會對本端程式庫與「管理程式庫」對話框中所列的程式庫執行比較。如果階段作業中的本端程式庫與已發佈版本不同步，則「程式庫同步化警告」對話框會開啟。您可以從下列選項中選擇，來選取您要在這裡使用的程式庫版本：

- **檔案的所有本端程式庫。**此選項會依現狀保留所有本端程式庫。您可以一律稍後重新發佈或更新它們。
- **此機器上的所有已發佈程式庫。**此選項會使用在資料庫中找到的版本來取代顯示的本端程式庫。
- **所有更近的程式庫。**此選項會使用資料庫中更新的公用版本來取代任何較舊的本端程式庫。
- **其他。**此選項會讓您透過在表格中選擇來手動選取想要的版本。

強制術語衝突

每當您新增公用程式庫或更新本端程式庫時，都可能會在此程式庫中的術語和類型與資源中其他程式庫中的術語和類型之間發現衝突及重複項目。如果發生這種情況，系統會要求您在「編輯強制術語」對話框中完成作業之前，驗證提出的衝突解決方案或進行變更。如需相關資訊，請參閱主題 第 168 頁的『強制術語』。

「編輯強制術語」對話框包含每一對衝突的術語或類型。替代的背景顏色可用來以視覺化的方式識別每一對衝突。這些顏色可以在「選項」對話框中變更。如需相關資訊，請參閱主題 第 65 頁的『選項：顯示標籤』。「編輯強制術語」對話框包含兩個標籤：

- **重複項。**此標籤包含在程式庫中找到的重複術語。如果大頭圖釘圖示在術語之後出現，則表示術語此次出現是強制的。如果黑色 X 圖示出現，則表示在擷取期間將忽略術語此次出現，因為在其他位置強制。
- **使用者定義。**此標籤包含在類型字典術語窗格中手動強制的全部術語清單以及未解決的衝突。

附註：「編輯強制術語」對話框會在您新增或更新程式庫之後開啟。如果您取消離開此對話框，則不會取消程式庫的更新或新增。

解決衝突

1. 在「編輯強制術語」對話框中，針對您想要強制的術語選取「使用」直欄中的圓鈕。
2. 完成時，按一下**確定**以套用強制術語並關閉對話框。如果您按一下**取消**，您將取消在此對話框中進行的變更。

第 16 章 關於程式庫字典

用來擷取文字資料的資源以範本及程式庫的形式儲存。程式庫可以由三個字典組成。

- **類型字典**包含在一個標籤或類型名稱下面分組的術語集合。當擷取引擎讀取文字資料時，它會將文字中找到的單字與類型字典中定義的術語進行比較。在擷取期間，類型的術語及同義字的字形變化形式在稱為概念的目標術語下面分組。擷取的概念會指派給它們在其中作為術語出現的類型字典。您可以在編輯器的左上方和中心窗格 - 程式庫樹狀結構窗格和術語窗格中管理類型字典。如需相關資訊，請參閱主題 『類型字典』。
- **替代字典**包含定義為同義字或選用元素的單字集合，用來在一個目標術語（在最終擷取結果中稱為概念）下面分組相似術語。您可以使用「同義字」標籤及「選用」標籤在編輯器的左下方窗格中管理替代字典。如需相關資訊，請參閱主題 第 169 頁的『替代/同義字字典』。
- **排除字典**包含會從最終擷取結果中移除的術語及類型集合。您可以在編輯器最右側的窗格中管理排除字典。如需相關資訊，請參閱主題 第 172 頁的『排除字典』。

如需相關資訊，請參閱主題 第 155 頁的第 15 章, 『使用檔案庫』。

類型字典

類型字典 *(type dictionary)* 是由類型名稱（或標籤）以及詞彙清單所組成。類型字典是在編輯器中的「檔案庫資源」標籤的左上角和中間窗格中管理。您可以透過功能表中的**檢視 > 資源編輯器**存取此視圖，如果您在互動式工作台階段作業中。否則，您可以在範本編輯器中編輯特定範本的字典。

當擷取引擎讀取文字資料時，它會將在文字中發現的單字與類型字典中定義的詞彙相互比較。詞彙是指語言資源中的類型字典中的單字或詞組。

當單字符合某個詞彙時，會將它指派給該詞彙的類型名稱。中在擷取期間讀取了資源時，在文字中發現的詞彙就會歷經數個處理步驟，之後再變成「擷取結果」窗格中的概念。如果擷取引擎判定多個屬於相同類型字典的詞彙是同義的，則會將它們群組在最常出現的詞彙下，並在「擷取結果」窗格中稱為概念 *(concept)*。比方說，如果詞彙 `question` 和 `query` 可能在結尾時出現在概念名稱 `question` 下。

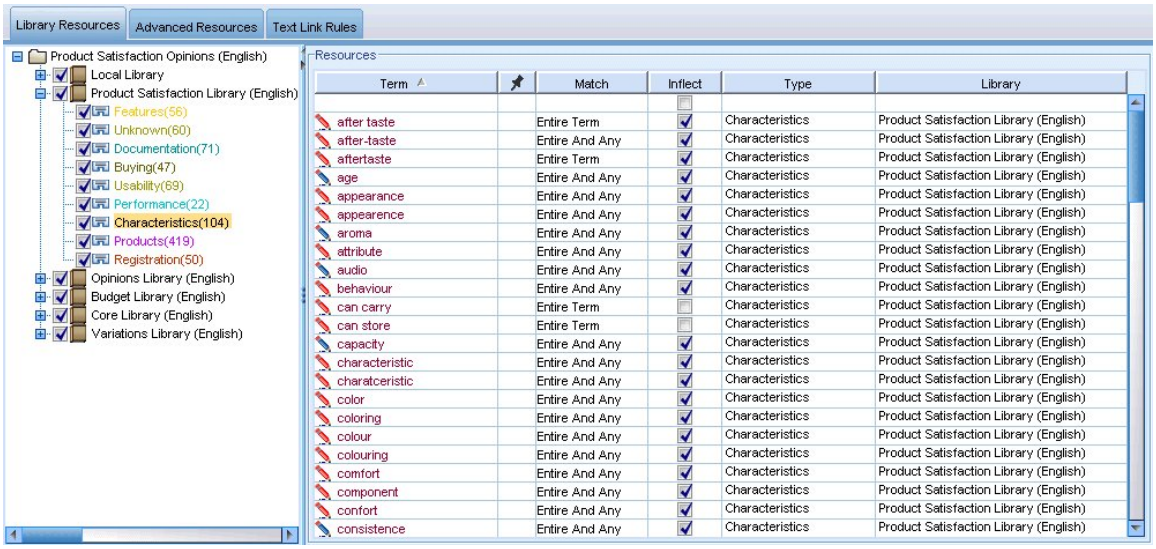


圖 *42.* 檔案庫樹狀結構與詞彙窗格

類型字典清單會顯示於左側的檔案庫樹狀結構窗格。每一個類型字典的內容會出現在中間窗格中。類型字典由不止一個詞彙清單所組成。文字資料中的單字和單字詞組符合類型字典中定義之詞彙的方式，取決於所定義的符合選項。**符合選項**在文字資料中的候選單字或詞組方面指定詞彙錨定的方式。請參閱第 166 頁的『新增術語』主題，以取得更多資訊。

此外，您還可以擴充類型字典中的詞彙，方法為指定您是否要自動產生及新增受影響的詞彙表單到字典。藉由產生受影響的表單，您可以自動將單數詞彙的複數表單、複數詞彙的單數表單和形容詞新增到類型字典。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。

註：對於大部分語音，未在任何類型定義檔中找到但擷取自文字的概念會自動鍵入為 `<Unknown>`

在詞彙中使用星號

如果您正在處理透過將其他單字複合在一起（沒有岔斷空格）來創造新單字的膠著語，則在詞彙中使用星號 (*) 特別有用。例如，德文單字 *Übernachtungspreis* 是由下列幾項組成：*Übernachtung* + *s* + *Preis*：

舉例而言，如果您在詞彙中搜尋類型 `Budget` 中的 `preis*`，它將符合擷取的概念，如 `preiserhöhung`。同樣地，`*preis` 將會符合 `Übernachtung`，而 `*preis*` 將會符合 `Übernachtungspreiserhöhung`。

內建類型

IBM SPSS Modeler Text Analytics 隨附了一組形式為隨附程式庫及編譯資源的語言資源。隨附程式庫包含一組內建類型字典，例如，`<Location>`、`<Organization>`、`<Person>` 及 `<Product>`。

擷取引擎使用這些類型字典將類型指派給它擷取的概念，例如，將類型 `<Location>` 指派給概念 `paris`。雖然在內建類型字典中定義了大量的術語，但它們不會涵蓋每一種可能性。因此，您可以新增至這些字典或建立自己的字典。如需特定隨附類型字典內容的說明，請閱讀「類型內容」對話框中的註釋。選取樹狀結構中的類型，並從快速功能表中選擇**編輯 > 內容**。

註：

除了隨附程式庫之外，編譯資源（擷取引擎也使用）包含補足內建類型字典的大量定義，但它們的內容在產品中不可見。然而，您可以將編譯字典歸類的術語強制到任何其他字典中。如需相關資訊，請參閱 第 168 頁的『強制術語』。

建立類型

您可以建立類型字典來協助對相似術語進行分組。在擷取程序期間探索到此字典中出現的術語時，會將它們指派給此類型名稱並在概念名稱下面擷取。每當建立程式庫時，一律會包括空的類型程式庫，以便您可以立即開始輸入術語。

如果您分析的文字與食品相關並想要對與蔬菜相關的術語進行分組，則可以建立自己的 <Vegetables> 類型字典。然後如果您認為 carrot、broccoli 和 spinach 等術語是會在文字中出現的重要術語，則可以新增它們。然後，在擷取期間，如果找到上述任何術語，則會作為概念進行擷取，並指派給 <Vegetables> 類型。

您不必定義單字或表示式的每一種形式，因為您可以選擇產生術語的字形變化形式。透過選擇此選項，擷取引擎會在術語的其他形式中自動將單數或複數形式辨識為屬於此類型。因為您不可能想要動詞或形容詞的字形變化形式，所以當類型主要包含名詞時，此選項特別有用。

「類型內容」對話框包含下列欄位。

名稱。為您正在建立的類型字典指定的名稱。我們建議您不要在類型名稱中使用空格，特別是在兩個或多個類型名稱以相同的單字開頭的情況。

註：類型名稱及符號使用有一些限制。例如，名稱內不要使用 "@" 或 "!" 等符號。

預設比對。預設比對屬性指示擷取引擎如何將此術語與文字資料進行比對。每當您將術語新增至此類型字典時，這是為其自動指派的比對屬性。您一律可以在術語清單中手動變更比對選項。選項包括：**整個術語**、**開頭**、**結尾**、**任何**、**開頭或結尾**、**整個和開頭**、**整個和結尾**、**整個和（開頭或結尾）**以及**整個（無複合字）**。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。

新增至。此欄位指出您會在其中建立新類型字典的程式庫。

依預設產生字形變化的形式。此選項會指示擷取引擎使用文法變形來擷取新增至此字典的術語的相似形式並進行分組，例如，術語的單數或複數形式。當類型主要包含名詞時，此選項特別有用。選取此選項時，新增至此類型的所有新術語都會自動具有此選項，但您可以在清單中對其進行手動變更。

字型顏色。此欄位可讓您從介面中的其他結果中識別出此類型中的結果。如果您選取**使用母項顏色**，則預設類型顏色也會用於此類型字典。此預設顏色在選項對話框中設定。如需相關資訊，請參閱主題 第 65 頁的『選項：顯示標籤』。如果您選取**自訂**，請從下拉清單中選取顏色。

註釋。此欄位是選用性欄位，可以用於任何註解或說明。

建立類型字典

1. 選取您希望在其中建立新類型字典的程式庫。
2. 從功能表中，選擇**工具 > 新建類型**。「類型內容」對話框即會開啟。
3. 在**名稱**文字框中輸入類型字典的名稱，並選擇您想要的選項。
4. 按一下**確定**以建立類型字典。新類型會在程式庫樹狀結構窗格中可見，並在中心窗格中出現。您可以立即開始新增術語。如需相關資訊，請參閱 第 166 頁的『新增術語』。

註：這些指示向您顯示如何在資源編輯器視圖或範本編輯器內進行變更。請記住，您也可以從其他視圖中的「擷取結果」窗格、「資料」窗格、「種類」窗格或「叢集定義」對話框直接執行此類型的細部調整。如需相關資訊，請參閱主題 第 76 頁的『精簡擷取結果』。

新增術語

程式庫樹狀結構窗格會顯示程式庫，並可展開以顯示它們所包含的類型字典。在中間窗格中，術語清單會顯示所選取程式庫或類型字典中的術語，視樹狀結構中的選取內容而定。

在資源編輯器中，您可以直接在術語窗格中或透過「新增術語」對話框將術語新增至類型字典。您新增的術語可以是單字，也可以是複合字。您會一律在清單頂端找到空白列以容許您新增術語。

註：這些指示向您顯示如何在資源編輯器視圖或範本編輯器內進行變更。請記住，您也可以從其他視圖中的「擷取結果」窗格、「資料」窗格、「種類」窗格或「叢集定義」對話框直接執行此類型的細部調整。如需相關資訊，請參閱主題 第 76 頁的『精簡擷取結果』。

術語直欄

在此直欄中，將單字或複合字輸入到資料格中。術語的顯示顏色視在其中儲存或強制術語的類型的顏色而定。您可以在「類型內容」對話框中變更類型顏色。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。

強制直欄

在此直欄中，透過將大頭圖釘圖示放入此資料格，擷取引擎就會知道忽略相同術語在其他程式庫中出現的任何其他項目。如需相關資訊，請參閱主題 第 168 頁的『強制術語』。

比對直欄

在此直欄中，選取比對選項來指示擷取引擎如何將此術語與文字資料進行比對。如需範例，請參閱表格。您可以透過編輯類型內容來變更預設值。如需相關資訊，請參閱主題 第 165 頁的『建立類型』。從功能表中，選擇**編輯 > 變更比對**。下列是基本比對選項，因為這些選項也可以進行組合：：

- **開頭**。如果字典中的術語與從文字中所擷取概念中的第一個單字相符，則會指派此類型。例如，如果您輸入 `apple`，則 `apple tart` 會符合。
- **結尾**。如果字典中的術語與從文字中所擷取概念中的最後一個單字相符，則會指派此類型。例如，如果您輸入 `apple`，則 `cider apple` 會符合。
- **任何**。如果字典中的術語與從文字中所擷取概念的任何一個單字相符，則會指派此類型。例如，如果您輸入 `apple`，則**任何**選項會將 `apple tart`、`cider apple` 及 `cider apple tart` 以相同的方式歸類。
- **整個術語**。如果從文字中擷取的整個概念與字典中的確切術語相符，則會指派此類型。新增術語作為**整個術語**、**整個和開頭**、**整個和結尾**、**整個和任何**或**整個（無複合字）**會強制擷取術語。

  此外，因為 `<Person>` 類型僅擷取兩部分的名稱，例如，*edith piaf* 或 *mohandas gandhi*，如果您在未提及姓氏時嘗試擷取名字，則您可能想要明確地將名字新增至此類型字典。例如，如果您想要擷取 *edith* 的所有實例作為名字，則您應使用**整個術語**或**整個和開頭**將 `edith` 新增至 `<Person>` 類型。
- **整個（無複合字）**。如果從文字中擷取的整個概念與字典中的確切術語相符，則會指派此類型，且會停止擷取來禁止擷取將該術語與較長的複合詞進行比對。例如，如果您輸入 `apple`，則**整個（無複合字）**選項會將 `apple` 歸類，且不會擷取複合字 `apple sauce`，除非在其他某個位置強制。

在下表中，假設術語 `apple` 在類型字典中。視比對選項而定，如果在文字中找到會擷取和歸類的概念，則此表格會顯示它們。

表 *40.* 比對範例

比對選項 **術語：**  **apple**	**擷取的概念**			
	apple	**apple** *tart*	*ripe* **apple**	*homemade* **apple** *tart*
整個術語	✔			
開頭		✔		
結尾			✔	
開頭或結尾		✔	✔	
整個和開頭	✔	✔		
整個和結尾	✔		✔	
整個和（開頭或結尾）	✔	✔	✔	
任何		✔	✔	✔
整個和任何	✔	✔	✔	✔
整個（無複合字）	✔	從未擷取	從未擷取	從未擷取

字形變化直欄

在此直欄中，選取擷取引擎是否應在擷取期間產生此術語的字形變化形式，以便它們全都一併分組。此直欄的預設值在「類型內容」中定義，但您可以直接在直欄中逐個變更此選項。從功能表中，選擇**編輯 > 變更字形變化**。

類型直欄

在此直欄中，從下拉清單中選取類型字典。會根據程式庫樹狀結構窗格中的選取內容過濾類型清單。清單中的第一個類型一律為在程式庫樹狀結構窗格中選取的預設類型。從功能表中，選擇**編輯 > 變更類型**。

程式庫直欄

在此直欄中，在其中儲存了術語的程式庫會出現。您可以將術語拖放至程式庫樹狀結構窗格中的另一個類型來變更其程式庫。

將單一術語新增至類型字典

1. 在程式庫樹狀結構窗格中，選取您要新增術語的類型字典。
2. 在中心窗格中的術語清單中，在第一個可用的空資料格中鍵入術語並為此術語設定您想要的任何選項。

將多個術語新增至類型字典

1. 在程式庫樹狀結構窗格中，選取您要新增術語的類型字典。
2. 從功能表中，選擇**工具 > 新建術語**。「新增術語」對話框即會開啟。
3. 透過鍵入術語或複製並貼上一組術語，輸入您想要新增至所選取類型字典的術語。如果您輸入多個術語，則必須使用「選項」對話框中定義的定界字元區隔它們，或在新的一行上新增每一個術語。如需相關資訊，請參閱主題 第 64 頁的『設定選項』。
4. 按一下**確定**以將術語新增至字典。比對選項會自動設為此類型程式庫的預設選項。對話框即會關閉，而新術語會在字典中出現。

強制術語

如果您想將術語指派給特定類型，您可以將其新增至對應的類型字典。但如果有多個名稱相同的術語，則擷取引擎必須知道應該使用的類型。因此，系統會提示您選取應該使用的類型。這稱為將術語強制到類型中。從（內部、不可編輯的）編譯字典中置換類型指派時，此選項極其有用。一般而言，我們建議完全避免重複術語。

強制不會移除此術語的其他出現項目；更確切地說，擷取引擎會忽略它們。您可以稍後透過強制或取消強制術語來變更應使用的出現項目。您可能還必須在新增公用程式庫或更新公用程式庫時將術語強制到類型字典中。

您可以在術語窗格的第二欄「強制」直欄中查看強制或忽略的術語。如果大頭圖釘圖示出現，這表示已經強制術語的這個出現項目。如果黑色 X 圖示出現，這表示在擷取期間將忽略術語的這個出現項目，因為已在其他位置強制。此外，當您強制術語時，它會以在其中強制它的類型的顏色顯示。這表示如果您將 Type 1 和 Type 2 中的術語強制到了 Type 1，則無論何時您在視窗中看到此術語，它都會以針對 Type 1 定義的字型顏色顯示。

您可以按兩下圖示以變更狀態。如果該術語在其他位置出現，則「解決衝突」對話框會開啟以容許您選取應該使用的出現項目。

重新命名類型

您可以透過編輯類型內容重新命名類型字典或變更其他字典設定。

重要：我們建議您不要在類型名稱中使用空格，特別是在兩個或多個類型名稱以相同的單字開頭的情況。我們還建議您不要重新命名「核心」或「意見」程式庫中的類型，也不要變更預設的比對屬性。

重新命名類型

1. 在程式庫樹狀結構窗格中，選取您要重新命名的類型字典。
2. 按一下滑鼠右鍵，然後從快速功能表中選擇**類型內容**。「類型內容」對話框即會開啟。

3. 在「名稱」文字框中輸入類型字典的新名稱。
4. 按一下**確定**以接受新名稱。新的類型名稱會在程式庫樹狀結構窗格中可見。

移動類型

您可以將類型字典拖至程式庫中的另一個位置，或拖至樹狀結構中的另一個程式庫。

在程式庫中將類型重新排序

1. 在程式庫樹狀結構窗格中，選取您要移動的類型字典。
2. 從功能表中，選擇**編輯 > 上移**以在程式庫樹狀結構窗格中將類型字典向上移動一個位置，或**編輯 > 下移**以將其向下移動一個位置。

將類型移至另一個程式庫

1. 在程式庫樹狀結構窗格中，選取您要移動的類型字典。
2. 按一下滑鼠右鍵，然後從快速功能表中選擇**類型內容**。「類型內容」對話框即會開啟。（您也可以將類型拖放至另一個程式庫）。
3. 在「新增至」清單框中，選取您想要將類型字典移動到的程式庫。
4. 按一下**確定**。對話框即會關閉，而類型現在位於您選取的程式庫中。

停用及刪除類型

如果您想要暫時移除類型字典，則可以透過取消選取程式庫樹狀結構窗格中字典名稱左側的勾選框來停用它。這表示您想要在程式庫中保留字典但想要在衝突檢查期間及擷取程序期間忽略其內容。

您也可以從程式庫中永久地刪除類型字典。

停用類型字典

1. 在程式庫樹狀結構窗格中，選取您要停用的類型字典。
2. 按一下空格鍵。類型名稱左側的勾選框即會清除。

刪除類型字典

1. 在程式庫樹狀結構窗格中，選取您要刪除的類型字典。
2. 從功能表中，選擇**編輯 > 刪除**以刪除類型字典。

替代/同義字字典

替代字典是由詞彙組成的集合，可幫助將類似的詞彙群組在一個目標詞彙下。替代字典是在「檔案庫資源」標籤的底端窗格中管理。您可以透過功能表中的**檢視 > 資源編輯器**存取此視圖，如果您在互動式工作台階段作業中。否則，您可以在範本編輯器中編輯特定範本的字典。

您可以在這個字典中定義兩種形式的替代：同義字和選用元素。按一下此窗格中的標籤即可在它們之間切換。

在文字資料上執行擷取之後，您可能會發現數個本身為同義字的概念或是其他概念的受影響表單。透過識別選用元素和同義字，您可以強制擷取引擎將這些對映到一個單一目標詞彙。

使用同義字和選用元素進行替換會減少「擷取結果」窗格中的概念數，因為將它們結合在一起成為具有較高頻率 Doc. 計數的更有意義、代表性的概念。

同義字

同義字會將兩個或多個具有相同意義的單字產生關聯。您也可以使用同義字來將詞彙與它們的縮寫群組，或是將一般誤拼的單字與正確的拼字群組。您可以在「同義字」標籤上定義這些同義字。

同義字定義由兩個部分組成。第一個部分是**目標**詞彙，這是您要擷取引擎將所有同義字群組在其下的詞彙。除非使用此目標詞彙作為另一個目標詞彙的同義字，或是除非它已被排除，否則它很可能變成出現在「擷取結果」窗格中的概念。第二個部分是同義字清單，這些同義字將會群組在目標詞彙下。

比方說，如果您要 automobile 被 vehicle 所取代，則 automobile 是同義字，而 vehicle 是目標詞彙。

您可以在**同義字**直欄中輸入任何同義字，但是如果在擷取期間找不到該單字，且詞彙具有含 Entire 的符合選項，則無法進行任何替代。不過，要將同義字群組在此詞彙下並不需要擷取目標詞彙。

選用元素

選用元素識別複合詞彙中在擷取期間可以忽略的選用單字，使得即使類似的詞彙在文字中顯得有點不同，也能讓它們放在一起。選用元素是單一單字，如果從複合詞彙中加以移除，則可以建立與另一個詞彙的相符項。這些單一單字可能會出現在複合詞彙內的任何地方：在開頭、中間或結尾。您可以在「選用」標籤上定義選用元素。

比方說，如果要將詞彙 ibm 和 ibm corp 群組在一起，您應宣告要將 corp 視為此案例中的選用元素處理。在另一個範例中，如果您要將詞彙 access 指定為選用元素，並在擷取期間同時發現 internet access speed 和 internet speed，則會在最常出現的詞彙下將它們群組在一起。

定義同義字

在「同義字」標籤上，您可以在表格頂端的空行中輸入同義字定義。首先定義目標術語及其同義字。您也可以選取希望在其中儲存此定義的程式庫。在擷取期間，所有出現的同義字項目在最終擷取中都會在目標術語下面進行分組。如需相關資訊，請參閱主題 第 166 頁的『新增術語』。

例如，如果文字資料包括大量電信資訊，您可能會有下列術語：cellular phone、wireless phone 及 mobile phone。在此範例中，您可能想要將 cellular 及 mobile 定義為 wireless 的同義字。如果您定義這些同義字，則會將每次擷取的 cellular phone 及 mobile phone 出現項目視為與 wireless phone 相同的術語，並會一起出現在術語清單中。

建置類型字典時，您可以輸入某個術語，然後想出三個或四個該術語的同義字。在這種情況下，您可以先將所有術語和目標術語輸入到替代字典中，然後拖曳同義字。

同義字替代也會套用至同義字的字形變化形式（例如，複數形式）。視環境定義而定，您可能想要對如何替換術語施加限制。某些字元可以用來對同義字處理的範圍進行限制：

- **驚嘆號 (!)**。直接在同義字之前加上驚嘆號 !synonym 時，這指出目標術語不會替換同義字的字形變化形式。但直接在目標術語之前加上驚嘆號 !target-term 則表示，您不希望複合目標術語或變式的任何部分收到進一步的替代。
- **星號 (*)**。直接在同義字之後放置星號，例如 synonym*，表示您想讓目標術語取代此單字。例如，如果您將 manage* 定義為同義字，將 management 定義為目標，則目標術語 associate management 會取代 associate managers。您也可以在單字之後添加空格及星號 (synonym *)，例如 internet *。如果您將目標定義為 internet，將同義字定義為 internet * * 和 web *，則 internet 會取代 internet access card 和 web portal。您在此字典中不能使用星號萬用字元作為單字或字串的開頭。
- **脫字符號 (^)**。在同義字之前放置脫字符號及空格，例如，^ synonym，表示同義字分組僅適用於術語以同義字開頭的情況。例如，如果您將 ^ wage 定義為同義字，將 income 定義為目標且兩個術語都已擷取，則會

在術語 income 下面將它們一併分組。但是如果擷取了 minimum wage 及 income，則不會一併分組，因為 minimum wage 不以 wage 開頭。空格必須放在此符號與同義字之間。

- **錢幣符號 ($)**。同義字後面放置空格及錢幣符號，例如 synonym $，表示同義字分組僅適用於術語以同義字結尾的情況。例如，如果您將 cash $ 定義為同義字，將 money 定義為目標且兩個術語都已擷取，則會在術語 money 下面將它們一併分組。但是如果擷取了 cash cow 及 money，則不會一併分組，因為 cash cow 不以 cash 結尾。空格必須放在此符號與同義字之間。
- **脫字符號 (^) 和錢幣符號 ($)**。如果脫字符號和錢幣符號一起使用，例如，^ synonym $，則術語僅在完全相符時符合同義字。這表示在擷取的術語中同義字前後都沒有單字出現，同義字分組才能發生。例如，您可能想要將 ^ van $ 定義為同義字，將 truck 定義為目標，以便只有 van 與 truck 分在一組，而 marie van guerin 將保留不變。此外，每當您使用脫字符號 (^) 和錢幣符號定義同義字且此單字在來源文字中的某個位置出現時，就會自動擷取同義字。

新增同義字項目

1. 顯示替代窗格時，按一下左下角的**同義字**標籤。
2. 在表格頂端的空行中，在「目標」直欄中輸入目標術語。您輸入的目標術語會以顏色顯示。此顏色代表術語出現在其中的類型或在其中強制它的類型，如果是這樣的話。如果術語以黑色顯示，這表示它在任何類型字典中都不出現。
3. 按一下目標右側的第二個資料格，並輸入同義字集。使用「選項」對話框中定義的廣域定界字元區隔每一個項目。如需相關資訊，請參閱主題 第 64 頁的『設定選項』。您輸入的術語會以顏色顯示。此顏色代表術語出現在其中的類型。如果術語以黑色顯示，這表示它在任何類型字典中都不出現。
4. 按一下最後一個資料格以選取您要在其中儲存此同義字定義的程式庫。

註：這些指示向您顯示如何在資源編輯器視圖或範本編輯器內進行變更。請記住，您也可以從其他視圖中的「擷取結果」窗格、「資料」窗格、「種類」窗格或「叢集定義」對話框直接執行此類型的細部調整。如需相關資訊，請參閱主題 第 76 頁的『精簡擷取結果』。

定義選用元素

在「選用」標籤上，您可以為想要的任何程式庫定義選用元素。這些項目針對每一個程式庫一併分組。一將程式庫新增至程式庫樹狀結構窗格，空的選用元素行就會立即新增至「選用」標籤。

所有項目都會自動轉換為小寫單字。擷取引擎會將項目與文字中的小寫及大寫單字進行比對。

註：會使用「選項」對話框中定義的定界字元定界術語。如需相關資訊，請參閱主題 第 64 頁的『設定選項』。如果您正在輸入的選用元素包括的定界字元與術語的一部分相同，則必須在它前面加上反斜線。

新增項目

1. 顯示替代窗格時，按一下編輯器左下角的「選用」標籤。
2. 針對您要為其新增此項目的程式庫，按一下「選用元素」直欄中的資料格。
3. 輸入選用元素。使用「選項」對話框中定義的廣域定界字元區隔每一個項目。如需相關資訊，請參閱主題 第 64 頁的『設定選項』。

停用及刪除替代項目

您可以透過在字典中停用項目來暫時移除它。透過停用項目，會在擷取期間忽略該項目。

您也可以在替代字典中刪除任何已作廢項目。

停用項目

1. 在字典中，選取您要停用的項目。
2. 按一下空格鍵。項目左側的勾選框即會清除。

附註：您也可以取消選取項目左側的勾選框來停用它。

刪除同義字項目

1. 在字典中，選取您要刪除的項目。
2. 從功能表中，選擇**編輯 > 刪除**或按鍵盤上的 **Delete** 鍵。字典中不會再有該項目。

刪除選用元素項目

1. 在字典中，按兩下您要刪除的項目。
2. 手動刪除術語。
3. 按 Enter 鍵以套用變更。

排除字典

排除字典是單字、詞組或局部字串的清單。擷取會忽略或排除符合或包含排除字典中項目的任何術語。排除字典會在編輯器的右窗格中管理。您新增至此清單的術語一般為補充性單字或詞組，為了連續性在文字中使用但不會實際上為文字增添任何重要內容，並可能會弄亂擷取結果。透過將這些術語新增至排除字典，您可以確定永遠不會擷取它們。

排除字典會在編輯器中「程式庫資源」標籤的右上方窗格中管理。您可以透過功能表中的**檢視 > 資源編輯器**存取此視圖，如果您在互動式工作台階段作業中。否則，您可以在範本編輯器中編輯特定範本的字典。

在排除字典中，您可以在表格頂端的空行中輸入單字、詞組或局部字串。您可以使用星號作為萬用字元，將字串作為一個或多個單字甚至局部單字新增至排除字典。在排除字典中宣告的項目會用於禁止擷取概念。如果在介面中的其他位置也宣告了某個項目，例如在類型字典中，則在其他字典中顯示時會加上刪除線，指出目前已排除。此字串沒有必要在文字資料中出現，也沒必要作為要套用的任何類型字典的一部分宣告。

註：如果您將概念新增至還在同義字項目中充當目標的排除字典，則還會排除目標及其所有同義字。如需相關資訊，請參閱主題 第 170 頁的『定義同義字』。

使用萬用字元 (*)

可以使用星號萬用字元來表示您希望將排除項目視為局部字串。擷取引擎找到的任何術語如果包含以在排除字典中所輸入字串開頭或結尾的單字，則會從最終擷取中排除。但有兩種情況不允許使用萬用字元：

- 橫線字元 (-) 前面加星號萬用字元，例如 *-
- 單引號 (') 前面加星號萬用字元，例如 *'s

表 41. 排除項目的範例

項目	範例	結果
單字	*next*	如果概念（或其術語）包含單字 *next*，則不會擷取它們。
詞組	*for example*	如果概念（或其術語）包含詞組 *for example*，則不會擷取它們。
局部	*copyright**	會排除符合或包含單字 *copyright* 變異的任何概念（或其術語），例如，copyrighted、copyrighting、copyrights 或 copyright 2010。
局部	**ware*	會排除符合或包含單字 *ware* 變異的任何概念（或其術語），例如，freeware、shareware、software、hardware、beware 或 silverware。

新增項目

- 在表格頂端的空行中，輸入術語。您輸入的術語會以顏色顯示。此顏色代表術語出現在其中的類型。如果術語以黑色顯示，這表示它在任何類型字典中都不出現。

停用項目

您可以透過在排除字典中停用項目來暫時移除它。透過停用項目，會在擷取期間忽略該項目。

1. 在排除字典中，選取您要停用的項目。
2. 按一下空格鍵。項目左側的勾選框即會清除。

註：您也可以取消選取項目左側的勾選框來停用它。

刪除項目

您可以在排除字典中刪除任何不需要的項目。

1. 在排除字典中，選取您要刪除的項目。
2. 從功能表中，選擇**編輯 > 刪除**。字典中不會再有該項目。

第 17 章 關於進階資源

除了類型以外（不包括替代字典），您還可以使用多種進階資源設定，例如，模糊分組設定或非語言類型定義。您可以在 範本編輯器 或 資源編輯器 視圖中的「進階資源」標籤中處理這些資源。

當您跳至「進階資源」標籤時，可以編輯下列資訊：

- **資源的目標語言。**用於選取建立和調整資源使用的語言。如需相關資訊，請參閱主題 第 176 頁的『資源的目標語言』。
- **模糊分組（異常狀況）。**用於從模糊分組（拼寫錯誤更正）演算法中排除單字配對。如需相關資訊，請參閱主題 第 177 頁的『模糊分組』。
- **非語言類實體。**用於啟用和停用可以擷取的非語言類實體，以及在其擷取期間套用的正規表示式和正規化規則。如需相關資訊，請參閱主題 第 177 頁的『非語言實體』。
- **語言處理。**用於宣告結構化語句（擷取型樣和強制定義），以及對所選語言使用縮寫的特殊方式。如需相關資訊，請參閱主題第 182 頁的『語言處理』。

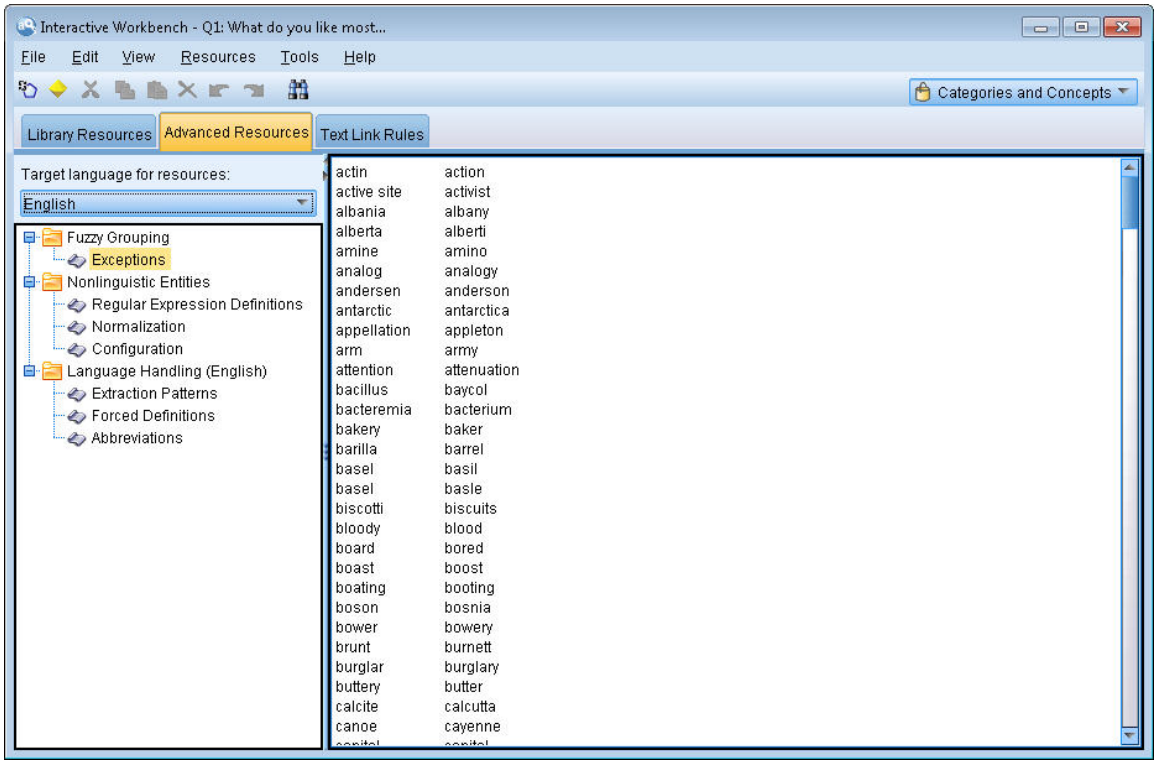


圖 *43.* 文字挖掘範本編輯器 - 進階資源標籤

註：您可以使用「尋找/取代」工具列來快速尋找資訊或是對區段進行一致變更。如需相關資訊，請參閱第 176 頁的『取代』。

編輯進階資源

1. 尋找並選取您要編輯的資源區段。內容出現在右窗格中。
2. 必要的話，使用功能表或工具列按鈕來剪下、複製或貼上內容。

3. 使用本節中的格式化規則來編輯您要變更的檔案。您的變更完成之後就會立即儲存。使用工具列上的復原或重做箭頭來回復至前一項變更。

尋找

在某些情況下，您可能需要在特定區段中快速尋找資訊。比方說，如果您執行文字鏈結分析，則可能會有數百個巨集和型樣定義。您可以使用「尋找」功能快速尋找特定的規則。如果要搜尋某個區段中的資訊，您可以使用「尋找」工具列。

使用尋找功能

1. 尋找並選取您要搜尋的資源區段。內容出現在編輯器的右窗格中。
2. 從功能表中，選擇**編輯 > 尋找**。「尋找」工具列即會出現在「編輯進階資源」對話框的右上角。
3. 在文字框中輸入您要搜尋的單字字串。您可以使用工具列按鈕來控制區分大小寫、部分比對和搜尋方向。
4. 按一下**尋找**以開始搜尋。如果發現相符項目，會在視窗中強調顯示該文字。
5. 再按一下**尋找**以尋找下一個相符項目。

註：在「文字鏈結規則」標籤中工作時，只有在檢視原始碼時才有「尋找」選項可用。

取代

在某些情況下，您可能需要對進階資源進行更廣泛的更新。「取代」功能可以幫助您對內容進行一致的更新。

使用取代功能

1. 尋找並選取您要在其中搜尋及取代的資源區段。內容出現在編輯器的右窗格中。
2. 從功能表中，選擇**編輯 > 取代**。即會開啟「取代」對話框。
3. 在**尋找目標**文字框中，輸入您要搜尋的單字字串。
4. 在**取代為**文字框中，輸入要用來取代所發現文字的字串。
5. 如果您只要尋找或取代完整的單字，請選取**僅符合完整單字**。
6. 如果您只要尋找或取代完全符合大小寫的單字，請選取**大小寫相符**。
7. 按一下**尋找下一個**以尋找相符項目。如果發現相符項目，會在視窗中強調顯示該文字。如果您不要取代此相符項目，請再按一下**尋找下一個**，直到找到您要取代的相符項目。
8. 按一下**取代**以取代選取的相符項目。
9. 按一下**取代**以取代區段中的所有相符項目。即會開啟一則訊息，含有已完成的取代數。
10. 進行取代完成時，按一下**關閉**。即會關閉對話框。

註：如果取代錯誤，可以復原取代，方法為關閉對話框，然後從功能表中選擇**編輯 > 復原**。您必須對您要復原的每個變更都執行此動作一次。

資源的目標語言

資源是針對特定的文字語言所建立。已針對其調整這些資源的語言定義於「進階資源」標籤。必要的話，您可以切換至另一種語言，方法為在**資源的目標語言**組合框中選取該語言。此外，這裡列出的語言將會顯示為您使用這些資源所建立的任何文字分析套件的語言。

重要：您將很少需要在資源中變更語言。這麼做可能會在資源不再符合擷取語言時導致問題。縱使很少採用，不過如果您因為預期會有採用多種語言的文字而計劃在擷取期間使用所有語言選項，則可能會變更語言。透過

變更語言，您可以（舉例而言）存取您感興趣的第二語言之擷取型樣、縮寫和強制執行定義的語言處理資源。不過，請記住，在發佈或儲存您所做的資源變更或執行另一個擷取之前，請將語言設定回您有意擷取的主要語言。

模糊分組

在「文字採礦」節點和「擷取設定」中，如果您選取**針對根字元限制下限容納拼字**，則您已啟用模糊分組演算法。

模糊分組可幫助群組一般誤拼的單字或緊密拼寫的單字，其作法為暫時除去所有母音（第一個母音除外）以及從擷取的單字加倍或三倍子音，然後比較它們以查明它們是否相同。在擷取程序期間，模糊分組功能會套用到擷取的詞彙，並比較結果以判定是否找到任何相符項目。如果是如此，則原始詞彙會在最終擷取清單中群組在一起。它們會群組在資料中最常出現的詞彙之下。

註：如果正在比較的兩個詞彙已指派給不同的類型（<Unknown> 類型除外），則模糊分組技術不會套用到此對組。換句話說，詞彙必須屬於相同類型或 <Unknown> 類型，才能套用此技術。

如果您啟用了此功能，並發現兩個拼字類似的單字不正確地群組在一起，則您可能需要將它們排除在模糊分組之外。您可以藉由將不正確符合的對組輸入到「進階資源」標籤中的「異常狀況」區段中來執行此動作。如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。

下列範例示範模糊分組如何執行。如果已啟用模糊分組，則這些單字似乎是相同的，並以下列方式相符：

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

在前述範例中，您將最有可能想要將 mountain 和 montana 從群組在一起排除。因此，您可以在「異常狀況」區段中以下列方式輸入它們：

```
mountain      montana
```

重要：在某些情況下，模糊分組異常狀況並不會阻止 2 個單字配對，因為正在套用某些同義字規則。在該情況下，您可能想要嘗試使用驚嘆號萬用字元 (!) 來輸入同義字，以禁止單字在輸出中變成同義。如需相關資訊，請參閱第 170 頁的『定義同義字』。

模糊分組異常狀況的格式化規則

- 每行僅定義一個異常狀況對組。
- 使用簡式或複合字。
- 僅使用小寫字元的單字。將忽略大寫字母的單字。
- 使用 TAB 字元來區隔對組中的每一個單字。

非語言實體

使用某種資料時，您可能會對擷取日期、社會安全碼、百分比或其他非語言實體非常有興趣。這些實體會在配置檔中明確宣告，您可以在其中啟用或停用實體。如需相關資訊，請參閱主題 第 181 頁的『配置』。為了最佳化來自擷取引擎的輸出，會正規化來自非語言實體的輸入，以根據預先定義格式群組相似的實體。如需相關資訊，請參閱主題 第 180 頁的『正規化』。

註：您可以在擷取設定中開啟及關閉非語言實體擷取。

可用的非語言實體

您可以擷取下列表格中的非語言實體。類型名稱在括弧中。

表 *42.* 可以擷取的非語言實體

位址	(<Address>)
氨基酸	(<Aminoacid>)
貨幣	(<Currency>)
日期	(<Date>)
延遲	(<Delay>)
數字	(<Digit>)
電子郵件位址	(<email>)
HTTP/URL 位址	(<url>)
IP 位址	(<IP>
組織	(<Organization>)
百分比	(<Percent>)
乘積	(<Product>)
蛋白質	(<Gene>)
電話號碼	(<PhoneNumber>)
時間	(<Time>)
美國社會安全碼	(<SocialSecurityNumber>)
重量與測量	(<Weights-Measures>)

清除文字以進行處理

在非語言實體擷取發生之前，會清除輸入文字。在此步驟期間，會完成下列暫時變更，以便可以識別及擷取非語言實體，如下：

- 任何兩個以上空格的序列會被單一空格取代。
- 列表會被空格取代。
- 單一的行尾字元或序列字元會被空格取代，而多個行尾序列會被標示為段落結尾。行尾可以由換行 (CR) 和換行 (LF) 表示，甚或兩者一起表示。
- 會暫時除去及忽略 HTML 和 XML 標籤。

正規表示式定義

擷取非語言實體時，您可能想要編輯或新增至用來識別正規表示式的正規表示式定義。這是在「進階資源」標籤中的**正規表示式定義**區段中完成。如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。

此檔案會拆開成不同的區段。第一個區段稱為 [macros]。除了該區段之外，每一個非語言實體都可以存在另一個區段。您可以將區段新增到此檔案。在每一個區段內，會將規則予以編號（*regexp1*、*regexp2*，依此類推）。這些規則必須從 1–*n* 循序編號。編號上的任何岔斷都會導致完全暫停此檔案的處理。

在某些情況下，實體為語言相依。如果實體取 0 以外的值作為配置檔中的語言參數，則它會被視為語言相依。如需相關資訊，請參閱主題 第 181 頁的『配置』。當實體是語言相依時，則必須使用語言作為區段名稱的字首，如 [english/PhoneNumber]。當以值 2 代表語言提供給 PhoneNumber 實體時，該區段將會包含僅適用於英文電話號碼的規則。

重要事項！ 如果您在編輯器中變更此檔案或任何其他檔案，且擷取引擎不再如所需般運作，請使用工具列上的**重設為原始設定**選項，將檔案重設為原始運送內容。此檔案需要對正規表示式有特定程度的熟悉度。如果您需要此範圍內的另外協助，請聯絡 IBM Corp. 以取得協助。

特殊字元 . [] {} () \ * + ? | ^ $

所有字元都符合它們自己本身，但是下列特殊字元除外，這些字元用於表示式中的特定用途：.[{()*+?|^$ 如果要照此使用這些字元，則必須在定義中的它們前面放反斜線 (\)。

比方說，如果您嘗試擷取網址，則完全停止字元對於實體非常重要，因此，您必須在它前面放反斜線，如：

```
www\.[a-z]+\.[a-z]+
```

重複運算子與限量元 ? + * {}

若要讓定義更有彈性，您可以使用正規表示式的數個標準萬用字元。它們是 * ? +

- 星號 * 表示有零或多個前置字串。例如，ab*c 符合 "*ac*"、"*abc*"、"*abbbc*"，依此類推。
- 加號 + 表示有一或多個前置字串。例如，ab+c 符合 "*abc*"、"*abbc*"、"*abbbc*"，但是不符合 "*ac*"。
- 問號 *?* 表示有零或一個前置字串。例如，modell?ing 符合 "*modeling*" 和 "*modeling*" 兩者。
- 以方括弧 *{}* 限制重複指出重複的範圍。例如，

  [0-9]{n} 符合恰恰重複 *n* 次的數字。例如，[0-9]{4} 將符合"*1998*"，但是既不符合"*33*"也不符合"*19983*"。

  [0-9]{n,} 符合重複 *n* 或多次的數字。例如，[0-9]{3,} 將符合"*199*"或"*1998*"，但是不符合"*19*"。

  [0-9]{n,m} 符合重複 *n* 到 *m* 次（內含）的數字。例如，[0-9]{3,5} 將符合"*199*"、"*1998*"或"*19983*"，但是既不符合"*19*"，也不符合"*199835*"。

選用空格和連字號

在某些情況下，您會想要在定義中包含選用空格。比方說，如果您要擷取 "*uruguayan pesos*"、"*uruguayan peso*"、"*uruguay pesos*"、"*uruguay peso*"、"*pesos*" 或 "*peso*" 之類的貨幣，則您將需要處理可能有兩個以空格區隔的單字的事實。在此情況下，應將此定義撰寫為 (uruguayan |uruguay)?pesos?。由於 *uruguayan* 或 *uruguay* 在與 *pesos/peso* 搭配使用時後面有一個空格，因此必須在選用序列內定義選用空格 (uruguayan |uruguay)。如果空格不在選用序列中（如 (uruguayan|uruguay)? pesos?），則在"*pesos*"或"*peso*"上將不會符合它，因為將會需要空格。

如果您正在清單中尋找包括連字號字元 (-) 的一系列事物，則必須最後定義連字號。比方說，如果您正在尋找逗點 (,) 或連字號 (-)，請使用 [,-]，切勿使用 [-,]。

清單和巨集中的字串順序

您應一律在較短序列之前定義最長的序列，否則將永遠不會讀取最長的序列，因為比對將在較短的序列上發生。比方說，如果您在尋找字串"*billion*"或"*bill*"，則必須在"*bill*"之前定義"*billion*"。因此舉例而言為 (billion|bill) 而不是 (bill|billion)。這也適用於巨集，因為巨集是字串清單。

定義區段中的規則順序

每一行定義一個規則。在每一個區段內，會將規則予以編號（*regexp1*、*regexp2*，依此類推）。這些規則必須從 1–*n* 循序編號。編號上的任何岔斷都會導致完全暫停此檔案的處理。如果要停用某個項目，請在用來定義正規表示式的每一行的開始處放置 # 符號。如果要啟用某個項目，請移除該行前面的 # 字元。

在每一個區段中，最特定的規則必須在最一般的規則之前定義，以確保適當的處理。比方說，如果您在表單"*month year*"和表單"*month*"中尋找某個日期，則必須在"*month*"規則之前定義"*month year*"規則。以下是它該有的定義方式：

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

and not

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

在規則中使用巨集

每當在數個規則中使用特定的序列時，您可以使用巨集。如此一來，如果您需要變更此序列的定義，則只需要變更它一次，不需要變更參照它的所有規則。例如，假設您具有下列巨集：

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

每當您參照巨集的名稱時，都必須用 `$()` 括住它，例如：`regexp1=$(MONTH)`

所有的巨集都必須定義於 `[macros]` 區段。

正規化

擷取非語言實體時，會正規化所遇到的實體，以根據預先定義的格式來群組相似的實體。例如，貨幣符號及其相等單字會被視為相同。正規化項目會儲存庫在「進階資源」標籤中的**正規化**區段中。如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。此檔案會拆開成不同的區段。

重要事項！ 此檔案僅適用於進階使用者。會需要變更此檔案是極不可能的。如果您需要此範圍內的另外協助，請聯絡 IBM Corp. 以取得協助。

正規化的格式化規則

- 每一行僅新增一個正規化項目。
- 嚴格遵循此檔案中的各個區段。無法新增任何新區段。
- 如果要停用某個項目，請在該行的開頭處放置一個 # 符號。如果要啟用某個項目，請移除該行前面的 # 字元。

正規化中的英文日期

依預設，在美式日期格式中可以辨識英文範本中的日期；亦即：月, 日, 年。如果您需要將該格式變更為日, 月, 年格式，請停用 "format:US" 這一行（方法為在該行的開頭處新增 #）並啟用 "format:UK"（方法為從該行移除 #）。

配置

您可以啟用及停用要在非語言實體配置檔中擷取的非語言實體類型。藉由停用不需要的實體，可以減少所需的處理時間。這是在「進階資源」標籤中的**配置**區段中完成。如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。如果已啟用非語言擷取，則擷取引擎會在擷取程序期間讀取此配置檔，以判定應擷取那些非語言實體類型。

此檔案的語法如下：

#name<TAB>*Language*<TAB>*Code*

表 *43.* 配置檔的語法

直欄標籤	**說明**
#name	將在用於擷取非語言實體的兩個其他必要檔案中藉以參照非語言實體的用語。這裡所用的名稱區分大小寫。
語言	文件 的語言。最好是選取特定的語言；不過，有**任何**選項存在。可能的選項有：0 = 任何，這在每當正規表示式不是某種語言所特有時使用，可用於採用不同語言的數個範本，例如 IP/URL/電子郵件位址；1 = 法文；2 = 英文；4 = 德文；5 = 西班牙文；6 = 荷蘭文；8 = 葡萄牙文；10 = 義大利文。
代碼	詞類代碼。除了少數情況下，大部分實體的值都是"s"。可能的值有：s = 停用字詞；a = 形容詞；n = 名詞。如果已啟用，則會先擷取非語言實體，並會套用擷取型樣以識別它在較大的環境定義中的角色。例如，賦予了一個值"a"給百分比。假設擷取了 30% 作為非語言實體。它會被識別為形容詞。如果您的文字包含 "30% 加薪，則"30%"非語言實體符合詞類型樣"ann"（形容詞 名詞 名詞）。

定義實體的順序

在此檔案中宣告實體的順序非常重要，會影響它們的擷取方式。它們會依照列出的順序套用。變更順序將會改變結果。最特定的非語言實體必須在較一般的非語言實體之前定義。

例如，非語言實體"Aminoacid"的定義方式為：

```
regexp1=($(AA)-?$(NUM))
```

其中 $(AA) 對應於"(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)"，這些是對應於特定胺基酸的特定 3 字母序列。

另一方面，非語言實體 "Gene" 較為一般，並由下式定義：

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

如果 "Gene" 在「配置」區段中定義於 "Aminoacid" 前面，則 "Aminoacid" 永遠不會符合，因為 "Gene" 中的 regexp3 將永遠先符合。

配置的格式化規則

- 使用 TAB 字元來區隔直欄中的每一個項目。
- 不要刪除任何行。
- 遵循之前表格中顯示的語法。
- 如果要停用某個項目，請在該行的開始處放置 # 符號。如果要啟用某個實體，請移除該行前面的 # 字元。

語言處理

當今所用的每一種語言都有特殊的表達概念、建構句子及使用縮寫的方式。在「語言處理」區段中，您可以編輯擷取型樣、強制執行那些型樣的定義，以及宣告您已在「語言」下拉清單中選取之語言的縮寫。

- 擷取型樣
- 強制執行的定義
- 縮寫

擷取型樣

從文件中擷取資訊時，擷取引擎會將一組詞性擷取型樣套用至文字中的一「堆」單字，以識別候選術語（單字及詞組）進行擷取。您可以新增或修改擷取型樣。

詞性包括文法元素，例如，名詞、形容詞、過去分詞、限定詞、介系詞、對等連接詞、名字、首字母及小品詞。上述一系列元素組成了詞性擷取型樣。在 IBM Corp. 文字採礦產品中，每個詞性由單一字元代表，以便更輕鬆地定義型樣。例如，形容詞由小寫字母 *a* 代表。受支援代碼集依預設在每一個預設擷取型樣區段的頂端出現，連同一組型樣及每一個型樣的範例來協助您了解使用的每一個代碼。

擷取型樣的格式化規則

- 每行一個型樣。
- 在行的開頭使用 # 來停用型樣。

您列出擷取型樣的順序非常重要，因為給定的單字序列僅由擷取引擎讀取一次，並指派給引擎為其找到相符項的前幾個擷取型樣。

支援的詞性代碼

下表是在英文編譯字典中定義的所有受支援的詞性代碼。

在特定範本中使用的所有詞性都會在**進階資源** > **擷取型樣**的頂端列出。

基本資源範本與意見範本之間的主要差異是在基本範本中使用最小限定詞 ("d") 和介系詞 ("c") 時，會在意見範本中使用其延伸的對等項目（"e" 和 "r"）。而且在意見範本中，同時具有 "a" 和 "Q" 詞性的所有單字只會處理為 "Q"。"0"、"1" 及 "2" 在所有意見範本中使用都會受限制。請參閱**進階資源** > **語言處理（英文）** > **強制定義及擷取型樣**。

其他英文範本可能使用未在字典中列出的部分詞性（例如，「市場情報」範本中的 "w" 和 "W"）。但在這種情況下，那些詞性會指派給**進階資源** > **強制定義**下面的特定單字。

表 *44.* 支援的詞性代碼

代碼	意義	範例
a	形容詞	腹部的，藍色的...
A	未用	未用
b	副詞	經常、通常、非常...
B	未用	未用
c	介系詞	「的」
C	拼錯單字的內碼	
d	限定詞	「該」
D	未用	未用

表 *44.* 支援的詞性代碼 *(*繼續*)*

代碼	意義	範例
e	延伸	限定詞：該、一個、我的、您的...
E	未用	未用
f	名字	John、Mary...
F	未用	未用
g	未用	未用
G	國籍形容詞	法國的、美國的...
h	未用	未用
H	未用	未用
i	未用	未用
I	未用	未用
j	未用	未用
J	未用	未用
k	未用	未用
K	未用	未用
l	未用	未用
L	未用	未用
m	名詞或不明	狗、ibm
M	未用	未用
n	名詞	狗
N	未用	未用
o	對等連接詞	「和」、「&」
O	未用	未用
p	過去分詞	已放棄、已附加...
P	未用	未用
q	未用	未用
Q	限定元	昂貴、小、良好...
r	延伸介系詞	的、之中、針對、從...
R	未用	未用
s	停止字組 (stop word)	我們不想擷取的任何單字
S	未用	未用
t	標題	夫人、太太、上尉、旅長...
T	技術形容詞	限制腫瘤的... （所有的 "T" 也是 "a"）
u	定義不明、不在字典中	
U	未用	未用
v	動詞	eat、eats、ate、eating, ...
V	不定式動詞	eat...
w	未用	未用
W	未用	未用
x	助動詞	be
X	未用	未用

表 *44.* 支援的詞性代碼 *(*繼續*)*

代碼	意義	範例
y	小品詞	von、di、de... （用來擷取人名：John von Doe）
Y	未用	未用
z	未用	未用
Z	未用	未用
0	意見副詞	僅在「意見」中。請參閱**進階資源 > 語言處理（英文） > 強制定義**。
1	「意見」中的「to」	請參閱**進階資源 > 語言處理（英文） > 強制定義**
2	特定的限定元	僅在「意見」中。請參閱**進階資源 > 語言處理（英文） > 強制定義**。
3	未用	未用
4	未用	未用
5	未用	未用
6	未用	未用
7	未用	未用
8	未用	未用
9	未用	未用

強制定義

從文件中擷取資訊時，擷取引擎會掃描文字並識別它遇到的每一個單字的詞性。在某些情況下，根據環境定義一個單字可能適合數個不同的角色。如果您想要強制某個單字充當特定的詞性角色或從處理中完全排除該單字，則可以在「進階資源」標籤的**強制定義**區段中這麼做。如需相關資訊，請參閱主題 第 175 頁的第 17 章, 『關於進階資源』。

若要為給定單字強制詞性角色，您必須使用下列語法向此區段新增一行：

term:*code*

表 *45.* 語法說明

項目	說明
term	術語名稱。
code	代表詞性角色的單一字元代碼。每個單元術語您可以列出最多六個不同的詞性代碼。此外，您還可以使用小寫代碼 s 停止將單字擷取到複合字/詞組，例如，additional:s。

強制定義的格式化規則

- 每個單字一行。
- 術語不能包含冒號。
- 使用小寫 s 作為詞性代碼來停止一起擷取單字。
- 每行使用最多六個詞性代碼。支援的詞性代碼會在「擷取型樣」區段中顯示。如需相關資訊，請參閱主題 第 182 頁的『擷取型樣』。
- 針對部分相符，在字串結尾使用星號字元 (*) 作為萬用字元。例如，如果您輸入 add*:s，則永遠不會擷取 add、additional、additionally、addendum 及 additive 等單字作為術語或複合字術語的一部分。但如果明確宣告單字相符作為編譯字典或強制定義中的術語，則仍會進行擷取。例如，如果您同時輸入 add*:s 及 addendum:n，則如果在文字中找到 addendum 時仍會擷取。

縮寫

當擷取引擎處理文字時，它一般會將找到的任何句點都視為句子結束的指示。這通常是正確的；但當文字中包含縮寫時，句點字元的這種處理則不適用。

如果您從文字中擷取術語並發現某些縮寫處理錯誤，則應在本節中明確宣告該縮寫。

附註：如果縮寫已在同義字定義中出現或定義為類型字典中的術語，則無需在這裡新增縮寫項目。

縮寫的格式化規則

- 每行定義一個縮寫。

第 18 章 關於文字鏈結規則

文字鏈結分析 (TLA) 是一項型樣比對技術，用來利用一組規則來擷取在文字中發現的關係。當已啟用文字鏈結分析進行擷取時，會將文字資料與這些規則相比較。當發現相符項目時，會擷取並呈現文字鏈結分析型樣。這些規則定義於「文字鏈結規則」標籤。

例如，擷取代表有關組織的簡單想法的概念可能不足以讓您感到興趣，但是藉由使用 TLA，您還可以瞭解不同組織之間的鏈結或是與組織相關聯的人員。TLA 還可用來擷取關於一些主題的意見，例如人們對於某給定的產品或經驗有什麼感覺。

若要獲取 TLA 的優點，您必須擁有包含文字鏈結 (TLA) 規則的資源。選取範本時，您可以透過範本在 TLA 直欄中是否有圖示，來瞭解那些範本具有 TLA 規則。

在擷取程序的型樣比對階段期間，會在文字資料中找到文字鏈結分析型樣。在此階段期間，會將規則與文字資料比較，當發現相符項目時，會擷取此資訊作為型樣。有時候，您可能想要從文字鏈結分析中取得更多內容或是變更內容相符的方式。在這些情況下，您可以修正規則來調整它們合乎您的特定需求。這項作業是在「文字鏈結規則」標籤中執行。

註：從 18.2 版開始，Type Reassignment Rule (TRR) 可用。TRR 會將一系列類型、巨集及/或記號轉換為具有特定類型的新概念。可在「意見」範本中使用它們以捕捉極性變更的意見。如需相關資訊，請參閱 第 130 頁的『Type Reassignment Rule』。

可使用文字鏈結規則的地方

您可以直接在範本編輯器或資源編輯器視圖中的「文字鏈結規則」標籤中編輯及建立規則。為幫助您瞭解規則可能會如何符合文字，您可以在此標籤中執行模擬。在模擬期間，只會在樣本模擬資料上執行擷取，並會套用文字鏈結規則以查明是否有任何型樣相符。然後模擬窗格中就會顯示任何符合文字的規則。根據相符項目，您可以選擇編輯規則和巨集來變更文字符合的方式。

有別於其他進階資源，TLA 規則為檔案庫所特有；因此，您一次只能使用來自一個檔案庫的 TLA 規則。從範本編輯器或資源編輯器內，跳至**文字鏈結規則**標籤。在此標籤中，您可以指定您的範本中包含要使用或編輯之 TLA 規則的檔案庫。基於此原因，我們強烈建議您將所有的規則儲存在一個檔案庫中，除非有不希望這麼做的極特定原因。

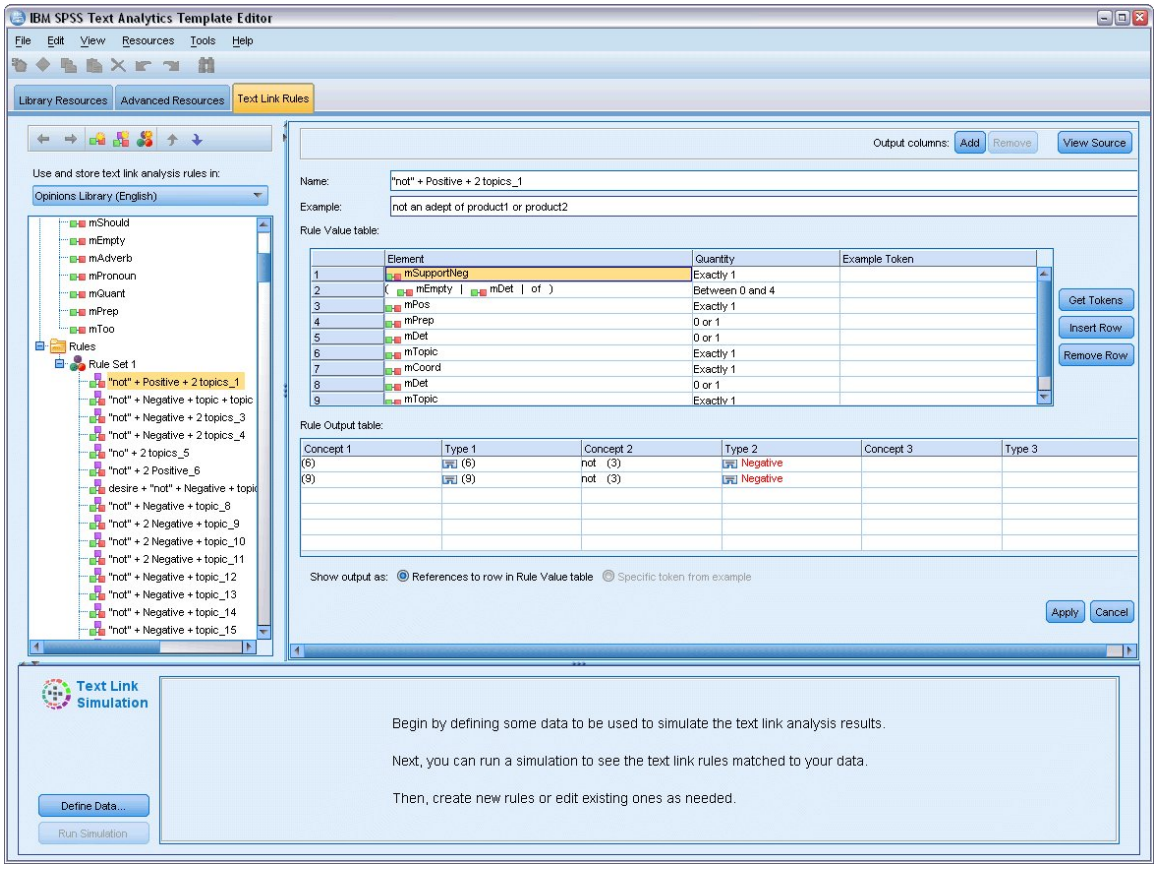


圖 *44.* 「文字鏈結規則」標籤

從何處開始

有數種方式可開始在「文字鏈結規則」標籤編輯器中工作：

- 從使用某範例文字來模擬結果開始，然後根據現行規則集從模擬資料擷取型樣的方式，來編輯或建立比對規則。
- 從頭開始建立新規則或編輯現有的規則。
- 直接在程式碼視圖中工作。

何時編輯或建立規則

縱使每一個範本隨附的文字鏈結分析規則對於從文字擷取許多簡單或複雜關係往往已經足夠了，但是有時您可能想要對這些規則進行一些變更，或是建立您自己的一些規則。例如：

- 透過建立新的規則或巨集，攫取到現有規則並未擷取的構想或關係。
- 變更您新增至資源之類型的預設行為。這通常需要您編輯像是 mTopic 或 mNonLingEntities 之類的巨集。如需相關資訊，請參閱主題 第 193 頁的『特殊巨集：mTopic、mNonLingEntities、SEP』。
- 將新類型新增至現有的文字鏈結分析規則和巨集。比方說，如果您認為類型 <Organization> 過於廣泛，則您可以針對位於數個不同企業部門（如 <Pharmaceuticals>、<Car Manufacturing>、<Finance> 等等）內的組織建立新的類型。在此情況下，您必須編輯文字鏈結分析規則和/或建立巨集，以將這些新類型納入考量並據此處理它們。

- 將類型新增至現有的文字鏈結分析規則。 例如，假設您有一個規則會擷取下列文字 john doe called jane doe，但是您要這個擷取電話通訊的規則也擷取電子郵件交換。您可以將電子郵件的非語言實體類型新增至規則，使它也會擷取像是以下的文字：johndoe@ibm.com emailed janedoe@ibm.com。
- 稍微修改現有的規則，而不是建立新的規則。例如，假設您有一個規則會比對下列文字 xyz is very good，但是您要這個規則也擷取 xyz is very, very good。

模擬文字鏈結分析結果

為了幫助定義新的文字鏈結規則，或是幫助瞭解某些句子在文字鏈結分析期間是如何相符，取一段樣本文字並執行模擬往往非常有用。在模擬期間，只會在樣本模擬資料上使用現行語言資源集和現行擷取設定來執行擷取。目標是取得模擬的結果及使用這些結果來改善規則、建立新的規則，或是更充分的瞭解相符如何發生。對於每一段文字（句子、單字或子句，視上下文而定），模擬輸出會顯示記號集合以及任何揭露該文字中的型樣的 TLA 規則。**記號**定義為在擷取程序期間所識別的任何單字或單字詞組。

有別於其他進階資源，TLA 規則為檔案庫所特有；因此，您一次只能使用來自一個檔案庫的 TLA 規則。從範本編輯器或資源編輯器內，跳至**文字鏈結規則**標籤。在此標籤中，您可以指定您的範本中包含要使用或編輯之 TLA 規則的檔案庫。基於此原因，我們強烈建議您將所有的規則儲存在一個檔案庫中，除非有不希望這麼做的極特定原因。

重要事項！ 我們強烈建議，如果您使用資料檔，請確保它包含的文字較為簡短，以便將處理時間減至最少。模擬的目標是要瞭解一段文字是如何解譯，以及瞭解規則符合這段文字的情況。這項資訊將會幫助您撰寫及編輯規則。請使用文字鏈結分析節點，或是使用已啟用 TLA 擷取的互動式階段作業來執行串流，以取得更完整資料集的結果。這項模擬僅供測試和規則編寫之用。

定義用於模擬的資料

為幫助您瞭解規則可能會如何符合文字，您可以使用樣本資料來執行模擬。第一個步驟是定義資料。

定義資料

1. 在**文字鏈結規則**標籤底端的模擬窗格中，按一下**定義資料**。或者，如果先前未定義任何資料，請從功能表中選擇**工具集** > **執行模擬**。即會開啟「模擬資料」精靈。
2. 選取下列一項來指定資料類型：
   - **直接貼上或輸入文字** 系統會提供一個文字框，供您從剪貼簿貼上某文字，或是手動輸入要處理的所需文字。您可以每一行輸入一個句子，或是使用標點符號來分開句子，如句點或逗點。輸入文字後，您就可以按一下**執行模擬**來開始模擬。
   - **指定檔案資料來源** 這個選項指出您要處理包含文字的檔案。按**下一步**以進行精靈步驟，您可以在其中定義要處理的檔案。選取檔案後，您就可以按一下**執行模擬**來開始模擬。下列是受支援的檔案類型：.txt 和 .text。在模擬期間，會「依現狀」讀取您選擇的資料檔。整個檔案的處理方式，與如同您已將「檔案清單」節點連接到「文字採礦」節點的方式相同。

   重要：我們強烈建議，如果您使用資料檔，請確保它包含的文字較為簡短，以便將處理時間減至最少。模擬的目標是要瞭解一段文字是如何解譯，以及瞭解規則符合這段文字的情況。這項資訊將會幫助您撰寫及編輯規則。請使用文字鏈結分析節點，或是使用已啟用 TLA 擷取的互動式階段作業來執行串流，以取得更完整資料集的結果。這項模擬僅供測試和規則編寫之用。
3. 若要開始模擬程序，請按一下**執行模擬**。即會出現進度對話框。如果您是在互動式階段作業中，則在模擬期間使用的擷取設定，是那些目前在互動式階段作業中選定的設定（請參閱「概念與種類」視圖中的**工具集** > **擷取設定**）。如果您是在範本編輯器中，則在模擬期間使用的擷取設定是預設擷取設定，這與「文字鏈結分析」節點的「匯出」標籤中顯示的那些設定相同。如需相關資訊，請參閱第 190 頁的『瞭解模擬結果』。

瞭解模擬結果

為幫助您瞭解規則可能會如何符合文字，您可以使用樣本資料來執行模擬及檢閱結果。您可以從那裡將規則集變更為更適合您的資料。當擷取和模擬程序完成時，將會向您呈現模擬的結果。

針對在擷取期間所確認的每一個「句子」，系統會呈現您數份資訊，其中包括確切的「句子」、在此輸入文字句子中發現的記號細分，最後是任何符合該句子中的文字的規則。「**句子**」的意思是指單字、句子或子句，視擷取程式如何將文字細分成可讀取的片段而定。

記號定義為在擷取程序期間識別的任何單字或詞組。例如，在句子 *My uncle lives in New York* 中，可能會在擷取期間找到下列記號：*my*、*uncle*、*lives*、*in* 及 *new york*。此外，可以將 *uncle* 作為概念擷取並歸類為 <Unknown>，還可以將 *new york* 作為概念擷取並歸類為 <Location>。所有概念都是記號，但並非所有記號都是概念。記號也可以是其他巨集、文字字串及字隙。只有那些歸類的單字或詞組可以是概念。

當您在互動式階段作業或資源編輯器中工作時，您是在概念層次上工作。TLA 規則較為精細，且即使句子中的個別記號從未被擷取及轉型，仍然可以在規則的定義中使用它們。能夠使用不是概念的記號可提供規則在文字中擷取複式關係上更大的彈性。

如果模擬資料中有多個句子，則按一下**下一個**和**上一個**，就可以在結果之間往前和往回移動。

在句子不符合所選檔案庫（請查看此標籤中的樹狀結構上方的檔案庫名稱）中的任何 TLA 規則的那些情況下，結果會被視為未符合，並且會啟用**下一個未符合結果**和**前一個未符合結果**以讓您知道沒有任何規則發現相符項目的文字，以及讓您能快速導覽這些實例。

在建立新規則、編輯規則或是變更資源或擷取設定之後，您可能需要重新執行模擬。如果要重新執行模擬，請在模擬窗格中按一下**執行模擬**，將會再度使用相同的輸入資料。

模擬結果中會顯示下列欄位和表格：

輸入文字。擷取程序從您在精靈中定義的模擬資料所識別的實際「句子」。句子的意思是指單字、句子或子句，視擷取程式如何將文字細分成可讀取的片段而定。

系統視圖。。擷取程序已識別的記號集合。

- **輸入文字記號**。 在輸入文字中發現的每一個記號。本主題稍早已定義記號。
- **轉型為**。 如果記號已被識別為概念並轉型，則此直欄中會顯示相關聯的類型名稱（如 <Unknown>、<Person>、<Location>）。
- **符合的巨集**。 如果記號符合現有的巨集，則此直欄中會顯示相關聯的巨集名稱。

規則符合輸入文字。 此表格顯示比對輸入文字符合的任何 TLA 規則。針對每一個符合的規則，您將會在**規則輸出**直欄中看到該規則的名稱，以及該規則的相關聯輸出值（概念 + 類型對組）。 您可以按兩下符合的規則名稱，在模擬窗格上方的編輯器窗格中開啟規則。

產生規則按鈕。如果您在模擬窗格中按一下這個按鈕，模擬窗格上方的規則編輯器窗格中將會開啟一個新規則。它將會以輸入文字作為其範例。同樣地，也會自動在**「規則值」表格**中的「元素」直欄中插入在模擬期間任何轉型為或符合巨集的記號。如果記號已轉型到並且符合巨集，則巨集值是將在規則中使用的值，以簡化規則。比方說，如果您是使用「基本英文」資源，則在模擬期間可以將句子"*I like pizza*"轉型為 <Unknown> 並符合巨集 mTopic。在此情況下，mTopic 將用來作為所產生規則中的元素。如需相關資訊，請參閱主題 第 194 頁的『使用文字鏈結規則』。

在樹狀結構中導覽規則和巨集

在擷取期間執行文字鏈結分析時，將會使用**文字鏈結規則**標籤中所選定儲存在檔案庫中的文字鏈結規則。

有別於其他進階資源，TLA 規則為檔案庫所特有；因此，您一次只能使用一個檔案庫中的 TLA 規則。從範本編輯器或資源編輯器內，跳至**文字鏈結規則**標籤。在此標籤中，您可以指定您的範本中包含要使用或編輯之 TLA 規則的檔案庫。基於此原因，我們強烈建議您將所有的規則儲存在一個檔案庫中，除非有不希望這麼做的強烈或特定原因。

您可以在「文字鏈結規則」標籤中指定您要在哪一個檔案庫中工作，方法為在此標籤中的**使用及儲存下列檔案庫中的文字鏈結分析規則：**下拉清單中選取該檔案庫。在擷取期間執行文字鏈結分析時，將會使用**文字鏈結規則**標籤中所選定儲存在檔案庫中的文字鏈結規則。因此，如果您在多個檔案庫中定義文字鏈結規則（TLA 規則），則只會使用在其中發現 TLA 規則的第一個檔案庫進行文字鏈結分析。基於此原因，我們強烈建議您將所有的規則儲存在一個檔案庫中，除非有不希望這麼做的極特定原因。

當您在樹狀結構中選取某個巨集或規則時，右側的編輯器窗格中會顯示其內容。如果您用滑鼠右鍵按一下樹狀結構中的任何項目，將會開啟快速功能表，顯示有哪些其他作業可以執行，例如：。

- 在樹狀結構中建立新巨集，並在右側的編輯器中開啟它。
- 在樹狀結構中建立新規則，並在右側的編輯器中開啟它。
- 在樹狀結構中建立新的規則集。
- 剪下、複製及貼上項目以簡化編輯。
- 刪除巨集、規則和規則集，以從資源中移除它們。
- 停用巨集、規則和規則集，以指出在處理期間應忽略它們。
- 將規則上移或下移以影響處理順序。

樹狀結構中的警告數

警告會隨黃色三角形顯示於樹狀結構中，在那裡通知您可能有問題。請將滑鼠指標移至錯誤的巨集或規則上來顯示蹦現說明。在大部分的情況下，您將會看到像是如下的訊息：**Warning: No example provided; Enter an example**，因此您需要輸入範例。

如果您遺漏範例，或是範例不符合規則，您將無法使用「取得記號」功能，因此我們建議您每個規則僅輸入一個範例。

當規則以黃色強調顯示時，表示 TLA 編輯器不知道某個類型或巨集。此訊息將會類似於：**Warning: Unknown type or macro**。這是要通知您將會由程式碼視圖中的 `$something` 所定義的項目（例如 `$myType`）既不是檔案庫中的舊式類型，也不是巨集。

如果要更新語法檢查程式，您需要切換至另一個規則或巨集；不需要重新編譯任何項目。因此，比方說，如果規則 A 因為遺漏範例而顯示警告，則您需要新增範例、按一下上方或下方的規則，然後返回規則 A 檢查它現在是否正確。

使用巨集

巨集可以簡化文字鏈結分析規則的外觀，因為它可讓您使用 OR 運算子 (|) 將類型、其他巨集和文字（單字）字串群組在一起。使用巨集的優點是，不但可以在多個文字鏈結分析規則中重複使用巨集來簡化它們，還可以讓您在一個巨集中進行更新，而無需在所有的文字鏈結分析規則中都進行更新。大部分隨附的 TLA 規則皆包含預設巨集。巨集會出現在「文字鏈結規則」標籤的最左邊窗格中的樹狀結構頂端。

模擬結果中會顯示下列欄位和表格：

名稱。識別此巨集的唯一名稱。我們建議您將小寫 m 放在巨集名稱前面，以幫助您在規則中快速識別巨集。當您在規則中手動參照巨集時（透過行內編輯或是在程式碼視圖中），您必須使用 $ 字元字首，以便擷取程序知道要尋找此特殊名稱。不過，如果您拖放巨集名稱或是透過快速功能表新增它，則此產品會自動將它辨識為巨集且不會新增任何 $。

巨集值表格。

- 若干列，代表此巨集可以代表的所有可能的值。這些值有區分大小寫。
- 這些值可以包含一個類型、文字字串、單字間隙或巨集或其組合。如需相關資訊，請參閱主題 第 200 頁的『受支援的規則和巨集元素』。
- 如果要在巨集中輸入元素的值，請按兩下您要在其中工作的列。即會出現可編輯的文字框，您可以在其中輸入類型參照、巨集參照、文字字串或單字間隙。或者，在資料格中按一下滑鼠右鍵以顯示上下文功能表，提供一般巨集、類型名稱和非語言類型名稱。如果要參照類型或巨集，您必須在巨集或類型名稱前面放一個'$'字元，例如 $mTopic 用於巨集 mTopic。結合引數時，您必須使用括弧 ( ) 來群組引數和字元 | 以指出布林 OR。
- 您可以在「巨集值」表格中使用在其右邊的按鈕來新增或移除列。
- 請在其自己的列中輸入每一個元素。比方說，如果您要建立代表 3 個文字字串之一的巨集（如 am OR was OR is），可以在視圖中的各別列上輸入每一個文字字串，這樣「巨集」表格將會包含 3 列。

建立及編輯巨集

您可以建立新巨集或編輯現有的巨集。請遵循巨集編輯器的準則和說明。如需相關資訊，請參閱主題 第 191 頁的『使用巨集』。

建立新的巨集

1. 從功能表中，選擇**工具 > 新增巨集**。或者，按一下樹狀結構工具列中的「新增巨集」圖示，在編輯器中開啟新巨集。
2. 輸入唯一名稱及定義巨集值元素。
3. 完成時按一下**套用**，以檢查是否有錯誤。

編輯巨集

1. 按一下樹狀結構中的巨集名稱。該巨集即會在右側的編輯器窗格中開啟。
2. 進行變更。
3. 完成時按一下**套用**，以檢查是否有錯誤。

停用及刪除巨集

停用巨集

如果您要在處理期間忽略某個巨集，可以停用它。這麼做可能會在任何仍然參照此已停用巨集的規則中導致警告或錯誤。請小心刪除及停用巨集。

1. 按一下樹狀結構中的巨集名稱。該巨集即會在右側的編輯器窗格中開啟。
2. 用滑鼠右鍵按一下名稱。
3. 從快速功能表中，選擇**停用**。巨集圖示即會變成灰色，且巨集本身會變成無法編輯。

刪除巨集

如果您要除去巨集，可以刪除它。這麼做可能會在任何仍然參照此巨集的規則中導致錯誤。請小心刪除及停用巨集。

1. 按一下樹狀結構中的巨集名稱。該巨集即會在右側的編輯器窗格中開啟。
2. 用滑鼠右鍵按一下名稱。
3. 從快速功能表中，選擇**刪除**。該巨集即會從清單中消失。

檢查是否有錯誤、儲存和取消

套用巨集變更

如果您在巨集編輯器外按一下，或是按一下**套用**，就會自動掃描巨集是否有錯誤。如果發現錯誤，則您將需要先修正它，然後再繼續進行到應用程式的另一個部分。

不過，如果是偵測到的較不嚴重的錯誤，則只會提出警告。比方說，如果您的巨集包含對於類型或其他巨集的不完整或未參照的定義，則會顯示警告訊息。在按一下**套用**後，任何未更正的警告都會導致警告圖示出現在左窗格中的「規則與巨集」樹狀結構內的巨集名稱左側。

套用巨集並不表示已永久儲存巨集。套用將會導致驗證處理程序檢查是否有錯誤和警告。

在互動式工作台階段作業內儲存資源

1. 如果要儲存您在互動式工作台階段作業期間對資源所做的變更，讓您可以在下次執行串流取得它們，您必須：
   - 更新建模節點，以確定在下次執行串流時可以取得這些相同的資源。如需相關資訊，請參閱主題 第 66 頁的『更新建模節點及儲存』。然後儲存串流。如果要儲存串流，請在更新建模節點之後，在主要IBM SPSS Modeler視窗中執行此動作。
2. 如果要儲存您在互動式工作台階段作業期間對資源所做的變更，以便可以在其他串流中使用它們，您可以：

- 更新您所用的範本或製作新的範本。如需相關資訊，請參閱主題 第 142 頁的『建立及更新範本』。這不會儲存對現行節點的變更（請參閱前一個步驟）
- 或者，更新您所用的 TAP。如需相關資訊，請參閱主題 第 115 頁的『更新文字分析套件』。

在範本編輯器內儲存資源

1. 首先，發佈檔案庫。如需相關資訊，請參閱主題 第 160 頁的『發佈程式庫』。
2. 然後，透過功能表中的**檔案 > 儲存資源範本**來儲存範本。

取消巨集變更

1. 如果您想要捨棄變更，請按一下**取消**。

特殊巨集：mTopic、mNonLingEntities、SEP

「意見」範本（和讚範本）以及「基本資源」範本隨附兩個特殊巨集，稱為 `mTopic` 和 `mNonLingEntities`。

mTopic

依預設，巨集 `mTopic` 會群組範本中隨附的所有很可能會以某個意見連接的類型，例如下列 *Core* 檔案庫類型：`<Person>`、`<Organization>`、`<Location>` 等等，前提是該類型不是意見類型（例如，`<Negative>` 或 `<Positive>`）或是定義為「進階資源」中的非語言實體的類型。

每當您在「意見」範本（或類似的範本）中建立新的類型時，此產品會假設除非在另一個巨集或是「進階資源」標籤的非語言實體區段中指定了這個類型，否則會以處理巨集 `mTopic` 中定義的其他類型的相同方式處理它。

假設您在資源中從「意見」範本建立新的類型：<Vegetables> 和 <Fruit>。在無需進行任何變更下，您的新類型會被視為 mTopic 類型處理，因此您可以自動揭露關於新類型的正面、負面、中立和環境定義選項。例如，在擷取期間，"*I enjoy broccoli, but I hate grapefruit*" 這個句子會產生下列 2 個輸出型樣：

```
broccoli <Vegetables> + like <Positive>

grapefruit <Fruit> + dislike <Negative>
```

不過，如果您要以處理 mTopic 中的其他類型的不同方式來處理那些類型，可以將類型名稱新增至現有的類型（如 mPos），這會群組所有的正面意見類型，或是建立新的巨集，可以稍後在一或多個規則中加以參照。

重要事項！ 如果您建立 <Vegetables> 之類的新類型，則這個新類型將會被包含作為 mTopic 中的類型，不過，在巨集定義將不會明確看到這個類型名稱。

mNonLingEntities

同樣地，如果您在「進階資源」標籤的**非語言實體**區段中新增非語言實體，則除非另行指定，否則會自動將它們作為 mNonLingEntities 處理。如需相關資訊，請參閱主題 第 177 頁的『非語言實體』。

SEP

您也可以使用預設巨集 SEP，此巨集對應於本端機器上定義的廣域分隔字元，一般是逗點 (,)。

使用文字鏈結規則

文字鏈結分析規則是一個布林查詢，用來對句子執行比對。文字鏈結分析規則包含下列一或多個引數：類型、巨集、文字字串或單字間隙。您必須至少擁有一個文字鏈結分析規則，才能擷取 TLA 結果。

「規則編輯器」的「文字鏈結規則」標籤中會顯示下列區域和欄位：

名稱欄位。文字鏈結規則的唯一名稱。

範例欄位。您可以選擇包含將會被此規則所擷取的範例句子或單字序列。我們建議使用範例。在此編輯器中，您將能夠從此範例文字中產生記號，以瞭解它如何符合規則以及它將如何被輸出。**記號**定義為在擷取程序期間識別的任何單字或詞組。例如，在句子 *My uncle lives in New York* 中，可能會在擷取期間找到下列記號：*my*、*uncle*、*lives*、*in* 及 *new york*。此外，可以將 *uncle* 作為概念擷取並歸類為 <Unknown>，還可以將 *new york* 作為概念擷取並歸類為 <Location>。所有概念都是記號，但並非所有記號都是概念。記號也可以是其他巨集、文字字串及字隙。只有那些歸類的單字或詞組可以是概念。

「規則值」表格。此表格包含規則的元素，用於將規則與句子比對。您可以在此表格中使用在其右邊的按鈕來新增或移除列。此表格由 3 個直欄組成：

- **元素**直欄。以一個類型、文字字串、單字間隙（<任何記號>）或巨集或其組合輸入值。如需相關資訊，請參閱主題 第 200 頁的『受支援的規則和巨集元素』。按兩下元素資料格可直接輸入資訊。或者，在資料格中按一下滑鼠右鍵以顯示上下文功能表，提供一般巨集、類型名稱和非語言類型名稱。請記住，如果您是藉由在資料格中鍵入資訊來將它輸入其中，請在巨集或類型名稱前面放一個'$'字元，例如 $mTopic 用於巨集 mTopic。您建立元素列的順序對於規則將會如何符合文字至關重大。結合引數時，您必須使用括弧 ( ) 來群組引數和字元 | 以指出布林 OR。請記住，值區分大小寫。
- **數量**直欄。此欄指出要發生相符所必須發現元素的次數數目下限和上限。比方說，如果您要在從 0 到 3 個單字的兩個其他元素之間定義間隙或一系列的單字，您可以從清單中選擇**介於 0 與 3 之間**，或是直接在對話框中輸入數字。預設值是「**整整 1**」。在某些情況下，您會想要使某個元素成為選用元素。如果是這種情

況，則它將會有數量下限 0 及大於 0 的數量上限（亦及 0 或 1、介於 0 與 2 之間）。請注意，規則中的第一個元素不可為選用元素，表示它的數量不可為 0。

- **範例記號**直欄。如果您按一下**取得記號**，則程式會將**範例**文字拆開成記號，並使用那些記號，以那些符合您所定義的元素的記號填入這個直欄。您也可以選擇在輸出表格中查看這些記號。

「規則輸出」表格 此表格中的每一列皆定義 TLA 型樣輸出將會如何出現在結果中。規則輸出最多可能產生六個「概念/類型」直欄對組的型樣，每一個都代表一個插槽。例如，類型型樣 <Location> + <Positive> 是兩插槽型樣，表示它是由 2 個「概念/類型」直欄對組所組成。

註：「規則值」表格的**元素**直欄中的詞彙，或是**「規則輸出」表格**的**概念**直欄中的詞彙的開頭不可為下列任何字元：`, #, %, ^, *, _, -, :, <, >, /, \, or "。

就像語言賦予我們以許多不同的方式自由表達相同的基本想法，因此您可以定義若干個規則來擷取相同的基本想法。例如，文字 *"Paris is a place I love"* 與文字 *"I really, really like Paris and Florence"* 代表相同的基本想法，即喜歡巴黎，但是表達的方式不同，將需要兩個不同的規則才會擷取兩者。然而，如果類似的想法群組在一起，則處理型樣結果會較為容易。基於此原因，縱使您可能有 2 個不同的規則以擷取這 2 個詞組，您可以對兩個規則定義相同的輸出（如類型型樣 <Location> + <Positive>），讓它代表全部兩個文字。如此一來，您可以看到輸出就不會總是模擬在原始文字中發現的單字的結構或順序。再者，這樣的類型型樣可能會符合其他詞組，並且可能產生 paris + like 和 tokyo + like 之類的概念型樣。

為幫助您快速定義輸出而錯誤較少，您可以使用快速功能表來選擇要在輸出中查看的元素。或者，您也可以從「規則值」表格拖放元素到輸出中。比方說，如果您有規則包含對於「規則值」表格的第 2 列中的 mTopic 巨集的參照，並且您希望該值能在輸出中，您可以只要將 mTopic 的元素拖/放到「規則輸出」表格中的第一個直欄對組即可。這麼做將會自動移入您所選對組的「概念」和「類型」兩者。或者，如果您要輸出從規則值表格的第三個元素（第 3 列）所定義的類型開始，則將該類型從「規則值」表格拖曳到輸出表格中的**類型 1** 資料格。此表格將會更新，顯示在括弧中的列參照 (3)。

或者，您可以在表格中手動輸入這些參照，方法為在每一個**概念**直欄中按兩下您要輸出的資料格，然後輸入 $ 符號後接列號碼（如 $2）來參照「規則值」表格的第 2 列中所定義的元素。當您手動輸入資訊時，必須也定義**類型**直欄、輸入 # 符號後接列號碼（如 #2）來參照「規則值」表格的第 2 列中所定義的元素。

再者，您甚至還可以結合方法。假設您在「規則值」表格的第 4 列中具有類型 <Positive>。您可以將它拖曳到 Type 2 直欄，然後按兩下 Concept 2 直欄中的資料格，然後在它前面手動輸入單字 *'not'*。然後輸出直欄將會在表格中顯示 not (4)，而若是您是在編輯模式或來源模式中，則會顯示 not $4。然後您可以在「類型 1」直欄中按一下滑鼠右鍵，並選取（舉例而言）稱為 mTopic 的巨集。這樣此輸出就可能導致如下的概念型樣：car + bad。

大部分的規則都只有一個輸出列，但是有時候可能會有及想要有多個輸出。在此情況下，請在「規則輸出」表格中，每列各定義一個輸出。

重要：請記住，在擷取 TLA 型樣期間會執行其他語言處理作業。因此，當輸出讀取 t$3\t#3 時，這表示在套用所有的語言處理程序（同義字及其他分組）之後，該型樣最終將顯示第三個元素的最終概念以及第三個元素的最終類型。

- **將輸出顯示為**。依預設，會選取選項**參照「規則值」表格中的列**，並且會藉由使用「規則值」表格中所定義的列數值參照來顯示輸出。如果您先前按一下「取得記號」，並在「規則值」表格中的「範例記號」直欄中具有記號，則您可以選擇透過選擇此選項，來查看這些特定記號的輸出。

註：如果輸出表格中未顯示足夠的概念/類型輸出對組，您可以按一下編輯器工具列中的「新增」按鈕來新增另一個對組。如果目前顯示 3 個對組，然後您按一下新增，則會多 2 個直欄（概念 4 和類型 4）新增至表格。這表示現在您將會在所有規則的輸出表格中看到 4 個對組。您也可以移除未用的對組，前提是此檔案庫中的規則集中沒有其他規則使用該對組。

範例規則

假設您的資源包含下列文字鏈結分析規則，並且您已啟用 TLA 結果擷取：

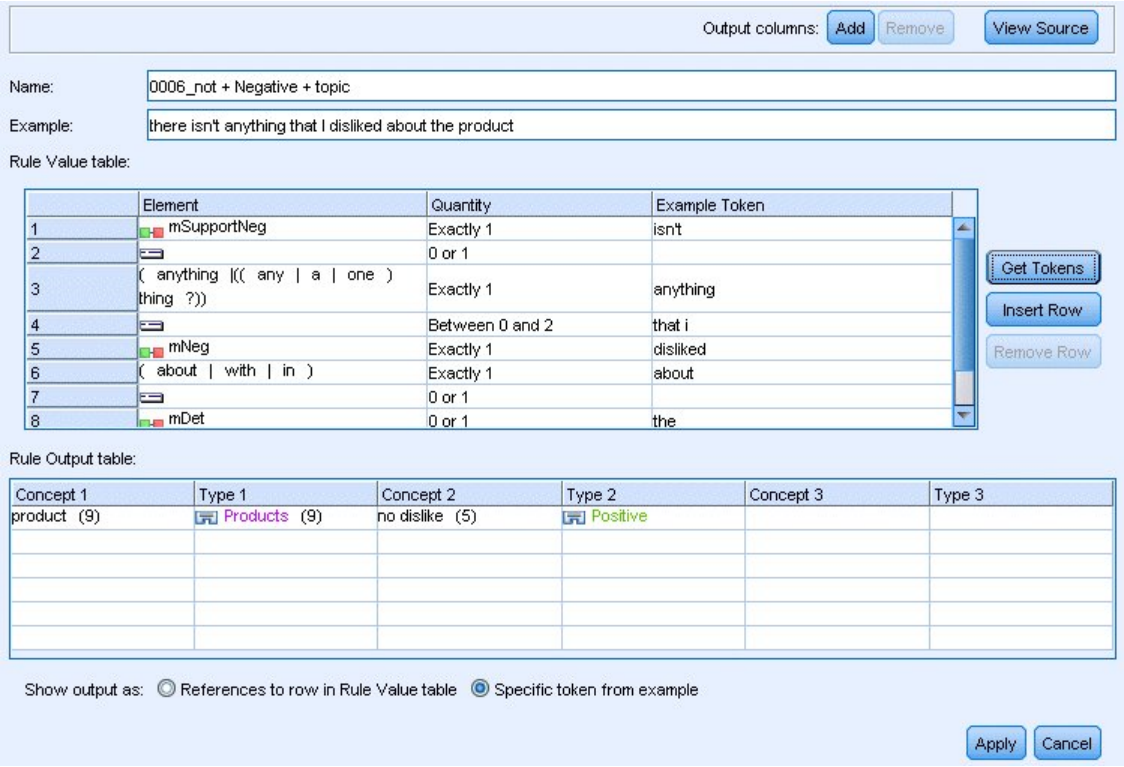


圖 *45.* 「文字鏈結規則」標籤：規則編輯器

每當您擷取時，擷取引擎都會讀取每一個句子，並且會嘗試比對下列序列：。

表 *46.* 擷取序列範例

元素（列）	引數的說明
1	巨集 mPos 或 mNeg 所代表的其中一個類型或類型 <Uncertain> 中的概念。
2	轉型為巨集 mTopic 所代表的其中一個類型的概念。
3	巨集 mBe 所代表的其中一個單字。
4	選用元素、0 或 1 個單字又稱為單字間隙或 <任何記號>
5	轉型為巨集 mTopic 所代表的其中一個類型的概念。

此輸出表格全部顯示從此規則所要的是型樣，其中任何概念或類型都對應於**「規則值」表格**中的第 5 列中所定義的 mTopic 巨集 + 任何概念或類型都對應於**「規則值」表格**中的第 1 列中所定義的 mPos、mNeg 或 <Uncertain>。 . 這可能是 sausage + like 或 <Unknown> + <Positive>。

建立及編輯規則

您可以建立新規則或編輯現有的規則。請遵循規則編輯器的準則和說明。如需相關資訊，請參閱主題 第 194 頁的『使用文字鏈結規則』。

建立新的規則

1. 從功能表中，選擇**工具 > 新增規則**。或者，按一下樹狀結構工具列中的「新增規則」圖示，在編輯器中開啟新規則。
2. 輸入唯一名稱及定義規則值元素。
3. 完成時按一下**套用**，以檢查是否有錯誤。

編輯規則

1. 按一下樹狀結構中的規則名稱。該規則即會在右側的編輯器窗格中開啟。
2. 進行變更。
3. 完成時按一下**套用**，以檢查是否有錯誤。

停用及刪除規則

停用規則

如果您要在處理期間忽略某個規則，可以停用它。請小心刪除及停用規則。

1. 按一下樹狀結構中的規則名稱。該規則即會在右側的編輯器窗格中開啟。
2. 用滑鼠右鍵按一下名稱。
3. 從快速功能表中，選擇**停用**。規則圖示即會變成灰色，且規則本身會變成無法編輯。

刪除規則

如果您要除去規則，可以刪除它。請小心刪除及停用規則。

1. 按一下樹狀結構中的規則名稱。該規則即會在右側的編輯器窗格中開啟。
2. 用滑鼠右鍵按一下名稱。
3. 從快速功能表中，選擇**刪除**。該規則即會從清單中消失。

檢查是否有錯誤、儲存和取消

套用規則變更

如果您在規則編輯器外按一下，或是按一下**套用**，就會自動掃描規則是否有錯誤。如果發現錯誤，則您將需要先修正它，然後再繼續進行到應用程式的另一個部分。

不過，如果是偵測到的較不嚴重的錯誤，則只會提出警告。比方說，如果您的規則包含對於類型或巨集的不完整或未參照的定義，則會顯示警告訊息。在按一下**套用**後，任何未更正的警告都會導致警告圖示出現在左窗格中的樹狀結構內的規則名稱左側。

套用規則並不表示已永久儲存規則。套用將會導致驗證處理程序檢查是否有錯誤和警告。

在互動式工作台階段作業內儲存資源

1. 如果要儲存您在互動式工作台階段作業期間對資源所做的變更，讓您可以在下次執行串流取得它們，您必須：

- 更新建模節點，以確定在下次執行串流時可以取得這些相同的資源。如需相關資訊，請參閱主題 第 66 頁的『更新建模節點及儲存』。然後儲存串流。如果要儲存串流，請在更新建模節點之後，在主要IBM SPSS Modeler視窗中執行此動作。

2. 如果要儲存您在互動式工作台階段作業期間對資源所做的變更，以便可以在其他串流中使用它們，您可以：
   - 更新您所用的範本或製作新的範本。如需相關資訊，請參閱主題 第 142 頁的『建立及更新範本』。這不會儲存對現行節點的變更（請參閱前一個步驟）
   - 或者，更新您所用的 TAP。如需相關資訊，請參閱主題 第 115 頁的『更新文字分析套件』。

在範本編輯器內儲存資源

1. 首先，發佈檔案庫。如需相關資訊，請參閱主題 第 160 頁的『發佈程式庫』。
2. 然後，透過功能表中的**檔案 > 儲存資源範本**來儲存範本。

取消規則變更

1. 如果您想要捨棄變更，請按一下編輯器窗格中的**取消**。

規則的處理順序

在擷取期間執行文字鏈結分析時，將會依序對著每一個規則比對「句子」（子句、單字、詞組），直到找到相符項目或是已耗盡所有規則為止。樹狀結構中的位置指定嘗試規則的順序。最佳作法指出您應將規則從最特定的排序到最一般的。 最特定的規則應位於樹狀結構的頂端。如果要變更特定規則或規則集的順序，請從「規則與巨集樹狀結構」快速功能表中選取**上移**或**下移**，或是選取工具列中的上移鍵或下移鍵。

如果您是在程式碼視圖中，則無法藉由在編輯器中來回移動規則來變更它們的順序。規則出現在程式碼視圖中越高，就會越快處理它。我們強烈建議只在樹狀結構中重新排序規則以避免複製/貼上問題。

重要事項！ 在舊版 IBM SPSS Modeler Text Analytics 中，要求您擁有唯一的數值規則 ID。在 18.2.1 版中，只能藉由在樹狀結構中上移或下移規則，或是以它們在程式碼視圖中的位置來指示處理順序。

例如，假設您的文字包含兩個句子：

I love anchovies

I love anchovies and green peppers

此外，假設兩個文字鏈結分析規則存在，具有下列值：

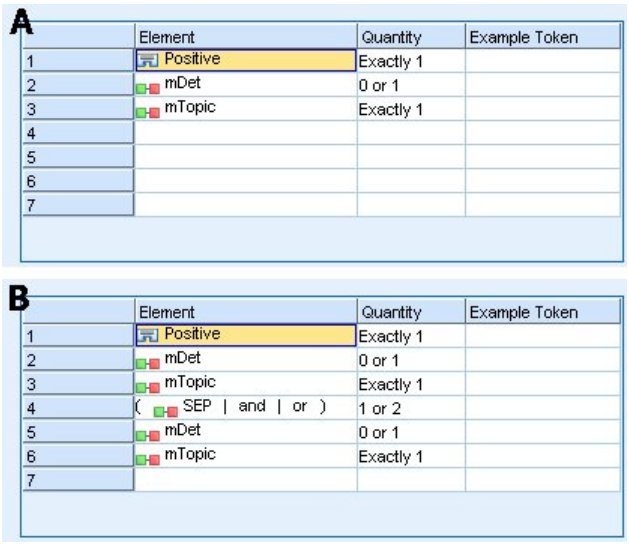


圖 *46. 2* 範例規則

在程式碼視圖中，規則值看來可能如下所示：

A: `value = $Positive $mDet? $mTopic`

B: `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

如果規則 **A** 在樹狀結構中比規則 **B** 還高（較接近頂端），則會先處理規則 **A**，並且會先以 `$Positive $mDet? $mTopic` 比對句子 *I love anchovies and green peppers*，它將會產生不完整的型樣輸出 `(anchovies + like)`，因為是以未尋找 2 個 `$mTopic` 相符項目的規則比對它。

因此，如果要擷取文字的真實核心要素，則較為特定的規則（在此案例中為 **B**）必須放在樹狀結構中較一般的規則（在此案例中為 **A**）高的地方。

使用規則集（多個傳遞）

規則集是一種在「規則與巨集樹狀結構」中將一組相關的規則分組在一起的有用方式，以執行多個傳遞處理。規則集除名稱之外沒有定義本身，並用來將規則組織成有意義的群組。在某些環境定義中，文字會過於豐富且各異，而無法在單一傳遞中處理。例如，當使用安全情報資料時，此文字可能包含透過聯絡方法（*x* 打電話給 *y*）、透過家庭關係（*y* 的姊夫 *x*）、透過金錢交換（*x* 電匯 *100* 姊夫給 *y*）等方式揭露的個人之間的鏈結。在此情況下，建立特殊化的幾組文字鏈結分析規則非常有用，其中的每一組都聚焦於某種類型的關係，例如一個用於揭露聯絡人、另一個用於揭露家庭成員，依此類推。

如果要建立規則集，請從「規則與巨集樹狀結構」快速功能表或工具列中選取「建立規則集」。然後就可以直接在樹狀結構上的「規則集」節點下建立新的規則，或是將現有的規則移至「規則集」。

當您使用資源執行擷取，其中規則群組到規則集中，則會強制擷取引擎通過文字進行多個傳遞，以符合每一個傳遞中的不同類型的型樣。如此一來，「句子」可以符合每一個規則集中的規則，然而如果沒有規則集，則它只能符合單一規則。

附註：每個規則集最多可以新增 512 個規則。

建立新的規則集

1. 從功能表中，選擇**工具 > 新增規則集**。或者，按一下樹狀結構工具列中的「新增規則集」圖示。規則集即會出現在規則樹狀結構中。
2. 將新規則新增到此規則集，或是將現有的規則移入規則集。

停用規則集

1. 用滑鼠右鍵按一下樹狀結構中的規則集名稱。
2. 從快速功能表中，選擇**停用**。規則集圖示即變成灰色，並且在處理期間也會停用及忽略該規則集內包含的所有規則。

刪除規則集

1. 用滑鼠右鍵按一下樹狀結構中的規則集名稱。
2. 從快速功能表中，選擇**刪除**。這時會從資源中刪除規則集以及它包含的所有規則。

受支援的規則和巨集元素

在文字鏈結分析規則和巨集中，接受下列引數作為值參數：

巨集

您可以直接在文字鏈結分析規則中或另一個巨集內使用巨集。如果您要手動或是從程式碼視圖內輸入巨集名稱（與從快速功能表中選取巨集名稱相反），請務必在名稱前面放一個錢幣符號字元 ($)，如 $mTopic。巨集名稱區分大小寫。透過快速功能表選取巨集時，您可以選擇現行「文字鏈結規則」標籤中所定義的任何巨集。

類型

您可以直接在文字鏈結分析規則或巨集中使用類型。如果您要手動或是在程式碼視圖中輸入類型名稱（與從快速功能表中選取類型相反），請務必在類型名稱前面放一個錢幣符號字元 ($)，如 $Person。類型名稱區分大小寫。如果您使用快速功能表，可以從正在使用的現行資源集中選擇任何類型。

如果您參照無法辨識的類型，將會收到警告訊息，且規則在「規則與巨集樹狀結構」中將會有警告圖示，直到您更正它為止。

文字字串

如果要包含從未擷取的資訊，您可以定義擷取引擎將搜尋的文字字串。所有擷取的單字或詞組都已指派給某個類型，因此基於此原因，不能在文字字串中使用它們。如果您使用已擷取的單字，則即使其類型為 <Unknown>，它還是會被忽略。

文字字串可以是一或多個單字。定義文字字串清單時，適用下列規則：

- 用括弧括住字串清單，如 (his)。如果有文字字串選擇，則每一個字串都必須以 OR 運算子區隔，如 (a|an|the) 或 (his|hers|its)。
- 使用單一或複合字。
- 以 | 字元（如同布林 OR）區隔清單中的每一個單字。
- 如果您要同時符合單數和複數表單，則輸入它們兩者。不會自動產生字形變化。
- 僅使用小寫。
- 如果要重複使用文字字串，請將它們定義為巨集，然後在其他巨集和文字鏈結分析規則中使用該巨集。
- 如果字串包含句點（完全停止）或連字號，則必須包含它們。比方說，如果要比對文字中的 a.k.a，請輸入句點，並以 a.k.a 作為文字字串。

排除運算子

使用 ! 作為排除運算子，以停止任何否定表示式占用特定插槽。您只能透過行內資料格編輯（在「規則值」或「巨集值」表格中按兩下資料格）或在程式碼視圖中手動新增排除運算子。比方說，如果您將 $mTopic @{0,2} !($Positive) $Budget 新增至文字鏈結分析規則，您正在尋找包含下列項目的文字：(1) 指派給 mTopic 巨集中的任何類型的詞彙、(2) 長度為零到兩個單字的單字間隙、(3) 沒有指派給 <Positive> 類型的詞彙實例，以及 (4) 指派給 <Budget> 類型的詞彙。這可能會擷取 "*cars have an inflated price tag*"，但是會忽略 "*store offers amazing discounts*"。

如果要使用此運算子，您必須按兩下元素資料格，在資料格中手動輸入驚嘆和括弧。

單字間隙（<任何記號>）

單字間隙又稱為 <Any Token>，定義在兩個元素之間可能存在的記號的數值範圍。比對極類似的詞組（由於存在其他限定詞、介系詞片語、形容詞或其他這類的單字而可能稍微不同）時，單字間隙非常有用。

表 *47.* 「規則值」表格中沒有單字間隙的元素範例

#	元素
1	 不明
2	 mBeHave
3	 Positive

附註：在程式碼視圖中，此值定義為：$Unknown $mBeHave $Positive

此值將符合像是 "*the hotel staff was nice*"的句子，其中 *hotel staff* 屬於類型 <Unknown>，*was* 在巨集 mBeHave 下，而 *nice* 為 <Positive>。但是它將不符合"*the hotel staff was very nice*"。

表 *48.* 「規則值」表格中具有 <任何記號> 單字間隙的元素範例

#	元素
1	 不明
2	 mBeHave
3	
4	 Positive

附註：在程式碼視圖中，此值定義為：$Unknown $mBeHave @{0,1} $Positive

如果您將單字間隙新增至規則值，它將同時符合"*the hotel staff was nice*"和"*the hotel staff was very nice*"。

在程式碼視圖中或使用行內編輯時，單字間隙的語法為 @{#,#}，其中 @ 表示單字間隙，而 {#,#} 定義在之前元素與之後元素之間接受的單字數目下限和上限。例如，@{1,3} 表示如果有至少一個單字存在，但是在兩個已定義元素之間不出現超過三個單字，則在那兩個元素之間可以達成相符。@{0,3} 表示如果有 0、1、2 或 3 個單字存在，但是不超過 3 個單字，則在兩個已定義元素之間可以達成相符。

在來源模式下檢視和工作

針對每一個規則和巨集，TLA 編輯器都會產生基礎原始碼，供擷取程式用來比對及產生 TLA 輸出。如果您偏好使用程式碼本身，可以按一下位於編輯器頂端的「檢視原始碼」按鈕來檢視此原始碼及直接編輯它。「程式碼」視圖將跳至目前選取的規則或巨集並強調顯示。不過，我們建議使用編輯器窗格以減少錯誤的機會。

當檢視或編輯原始碼完成時，請按一下**結束原始碼**。如果您對規則產生無效的語法，系統將會在結束程式碼視圖之前要求您修正它。

重要：如果您在程式碼視圖中編輯，強烈建議您一次編輯一個規則和巨集。在編輯巨集之後，請藉由擷取來驗證結果。如果您滿意結果，建議您先儲存範本再進行另一個變更。如果您不滿意結果或是發生錯誤，請回復為儲存的資源。

程式碼視圖中的巨集

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

表 *49.* 巨集項目

[macro]	每一個巨集都必須以標示 [macro] 的行開頭以表示巨集的開始。
name	巨集定義的名稱。每一個名稱都必須是唯一的。
value	一或多個類型、文字字串、單字間隙或巨集的組合。如需相關資訊，請參閱主題第 200 頁的『受支援的規則和巨集元素』。結合引數時，您必須使用括弧 ( ) 來群組引數和字元 | 以指出布林 OR。

除了有關巨集的區段中所涵蓋的準則和語法之外，程式碼視圖還有幾個在編輯器視圖中工作時並不需要的其他準則。在來源模式中工作時，巨集也必須遵循下列準則：

- 每一個巨集都必須以標示 [macro] 的行開頭以表示巨集的開始。
- 如果要停用元素，請在每一行前面放一個註解指示符 (#)。

範例。此範例定義一個稱為 mTopic 的定義。mTopic 的值表示存在符合下列其中一個類型的詞彙： <Product>、<Person>、<Location>、<Organization>、<Budget> 或 <Unknown>。

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

程式碼視圖中的規則

```
[pattern(ID)]
name = pattern_name
value = [$type_name|macro_name|word_gaps|literal_strings]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

表 *50.* 規則項目

[pattern (<ID>)]	指出該文字鏈結分析規則開始，並提供用來決定處理順序的唯一數值 ID。
name	提供此文字鏈結分析規則的唯一名稱。
value	提供要符合文字的語法和引數。如需相關資訊，請參閱主題 第 200 頁的『受支援的規則和巨集元素』。
輸出	在文字中探索到的結果相符型樣的輸出格式。此輸出並非總是類似來源文字中元素的確切原始位置。此外，將每一個輸出放在分開的行上，一個給定的文字鏈結分析規則就可以有多個輸出行。 輸出的語法： • 以 Tab 代碼 \t 區隔輸出，如 $1\t#1\t$3\t#3 • $ 和數字要求發現符合位於該位置之值參數中所定義的引數的詞彙。因此 $1 表示符合定義給值的第一個引數的詞彙。 • # 和數字要求位於該位置之元素的類型名稱。如果某個項目是文字字串清單，將會指派類型 <Unknown>。 • 值 Null\tNull 將不會建立任何輸出。

除了有關規則的區段中所涵蓋的準則和語法之外，程式碼視圖還有幾個在編輯器視圖中工作時並不需要的其他準則。在來源模式中工作時，規則也必須遵循下列準則：

- 每當定義了兩個以上的元素時，無論它們是否為選用項目，都必須用括弧括住它們（例如，($Negative|$Positive) 或 ($mCoord|$SEP)?)。$SEP 代表逗點。
- 文字鏈結分析規則中的第一個元素不可為選用元素。例如，開頭不可為 value = $mTopic? 或 value = @{0,1}。
- 您可以將數量（或實例計數）關聯到記號。這在撰寫包含所有案例的唯一規則（而不是針對每一個案例撰寫各別的規則）時非常有用。比方說，如果您嘗試符合 ,（逗點）或 and，可以使用文字字串 ($SEP|and)。如果您透過新增數量來延伸此作法，以致文字字串變成 ($SEP|and){1,2}，則您現在將會符合下列任何實例：," "and" ", and"。
- 在文字鏈結分析規則 value 中的巨集名稱與 $ 和 ? 字元之間不支援空格。
- 在文字鏈結分析規則 output 中不支援空格。
- 如果要停用元素，請在每一行前面放一個註解指示符 (#)。

範例。假設您的資源包含下列 TLA 文字鏈結分析規則，並且您已啟用 TLA 結果擷取：

```
## Jean Doe was the former HR director of IBM in France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
 (of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

每當您擷取時，擷取引擎都會讀取每一個句子，並且會嘗試比對下列序列：。

表 *51.* 擷取序列範例

位置	引數的說明
1	人員的名稱 ($Person)，
2	下列項目的一或兩項：逗點 ($SEP)、限定詞 ($mDet)、輔助動詞 ($mSupport)、字串"then"或"as"，
3	0 或 1 個字 (@{0,1})
4	函數 ($Function)

表 *51.* 擷取序列範例 *(*繼續*)*

位置	引數的說明
5	下列其中一個字串："of"、"with"、"for"、"in"、"to"或"at"、
6	0 或 1 個字 (@{0,1})
7	組織的名稱 ($Organization)
8	0、1 或 2 個字 (@{0,2})
9	位置的名稱 ($Location)

此範例文字鏈結分析規則將會比對句子或詞組，如：

IBM 在法國的人力資源主管 *Jean Doe*

Jean Doe 是 *IBM* 在法國的前人力資源主管

IBM 指派 *Jean Doe* 擔任 *IBM* 在法國的人力資源主管

此範例文字鏈結分析規則將會產生下列輸出：

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

其中：

- jean doe 是對應於 $1 的詞彙（文字鏈結分析規則中的第一個元素），而 <Person> 是 jean doe (#1) 的類型，
- hr director 是對應於 $4 的詞彙（文字鏈結分析規則中的第四個元素），而 <Function> 是 hr director (#4) 的類型，
- ibm 是對應於 $7 的詞彙（文字鏈結分析規則中的第七個元素），而 <Organization> 是 ibm (#7) 的類型，
- france 是對應於 $9 的詞彙（文字鏈結分析規則中的第九個元素），而 <Location> 是 france (#9) 的類型

程式碼視圖中的規則集

```
[set(<ID>)]
```

其中 [set (<ID>)] 指出規則集開始，並提供用來決定規則集處理順序的唯一數值 ID。

範例。下列句子包含個人、他們在公司內的職責以及該公司的合併/收購活動的相關資訊。

```
Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO of Org2
Ltd.
```

您可以撰寫一個含有數個輸出的規則來處理所有可能的輸出，如：

```
## Org1 Inc entered into a definitive merger agreement with Org2 Ltd, said
John Doe, CEO of Org2 Ltd.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

將會產生下列 2 個輸出型樣：

- org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd<Organization>
- john doe <Person> + ceo <Function> + org2 ltd<Organization>

重要事項！ 請記住，在擷取 TLA 型樣期間會執行其他語言處理作業。在此情況下，在擷取程序的同義字分組階段期間，merger 會群組在 merges with 下。而且由於 merges with 屬於 <ActiveVerb> 類型，因此此類型名稱是最終 TLA 型樣輸出中出現的名稱。因此，當輸出讀取 t$3\t#3 時，這表示在套用所有的語言處理程序（同義字及其他分組）之後，該型樣最終將顯示第三個元素的最終概念以及第三個元素的最終類型。

取代如前述一樣撰寫複雜的規則，管理及使用兩個規則可能會較為簡單。第一個規則專用於查明公司之間的合併/收購：

```
[set(1)]
## Org1 Inc has entered into a definitive merger agreement with Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

這會產生 org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd <Organization>

第二個規則專用於個人/職責/公司：

```
[set(2)]
## said John Doe, CEO of Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

這會產生 john doe <Person> + ceo <Function> + org2 ltd <Organization>

注意事項

這項資訊是針對全球供應的產品與服務所開發。

在其他國家，IBM 不見得有提供本文件所提及之各項產品、服務或功能。請洽詢當地的 IBM 業務代表，以取得當地目前提供的產品和服務之相關資訊。本文件在提及 IBM 的產品、程式或服務時，不表示或暗示只能使用 IBM 的產品、程式或服務。只要未侵犯 IBM 之智慧財產權，任何功能相當之產品、程式或服務皆可取代 IBM 之產品、程式或服務。不過，任何非 IBM 之產品、程式或服務，使用者必須自行負責作業之評估和驗證責任。

本文件所說明之主題內容，IBM 可能擁有其專利或專利申請案。提供本文件不代表提供這些專利的授權。您可以書面提出授權查詢，來函請寄到：

IBM Director of Licensing
IBM CorporationNorth Castle Drive, MD-NC119
Armonk, NY 10504-1785US

如果是有關雙位元組 (DBCS) 資訊的授權查詢，請洽詢所在國的 IBM 智慧財產部門，或書面提出授權查詢，來函請寄到：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

IBM 僅以「現狀」提供本書，而不提供任何明示或默示之保證（包括但不限於可售性或符合特定效用的保證）。有些轄區不允許放棄在特定交易中的明示或默示保證，因此，這項聲明對您可能不適用。

本資訊中可能會有技術上或排版印刷上的訛誤。因此，IBM 會定期修訂；並將修訂後的內容納入新版中。IBM 隨時會改進及/或變更本出版品所提及的產品及/或程式，不另行通知。

本書對於非 IBM 網站的援引只是為了方便而提供，並不對這些網站作任何認可。該些網站上的內容並非本 IBM 產品內容的一部分，用戶使用該網站時應自行承擔風險。

IBM 得以各種 IBM 認為適當的方式使用或散布 貴客戶提供的任何資訊，而無需對 貴客戶負責。

如果本程式之獲授權人為了 (i) 在個別建立的程式和其他程式（包括本程式）之間交換資訊，以及 (ii) 相互使用所交換的資訊，因而需要相關的資訊，請洽詢：

IBM Director of Licensing
IBM CorporationNorth Castle Drive, MD-NC119
Armonk, NY 10504-1785US

這些資訊可能可以使用，但必須遵循適當的條款，在某些情況中需要付費。

IBM 基於 IBM 客戶合約、IBM 國際程式授權合約或雙方之任何同等合約的條款，提供本文件所提及的授權程式與其所有適用的授權資料。

所引用的客戶範例為說明用途。實際的績效會因不同的配置與作業狀況而異。

本書所提及之非 IBM 產品資訊，係一由產品的供應商，或其出版的聲明或其他公開管道取得。IBM 並未測試過這些產品，也無法確認這些非 IBM 產品的執行效能、相容性或任何對產品的其他主張是否完全無誤。有關非 IBM 產品的性能問題應直接洽詢該產品供應商。

IBM 不須通知即可變更或撤銷與 IBM 未來方向或目的相關的陳述，亦僅代表其目標及方針。

本資訊中含有日常商業活動所用的資料及報告範例。為了盡可能完整地說明，範例中包括了個人、公司行號、品牌以及產品等的名稱。 所有這些名稱都是虛構的，實際個人或商業企業的任何類似項目都純屬巧合。

商標

IBM、IBM 標誌及 ibm.com 是 International Business Machines Corp. 在世界許多管轄區註冊的商標或註冊商標。其他產品及服務名稱可能是 IBM 或其他公司的商標。IBM 商標最新清單可於下列網站之「著作權與商標資訊」("Copyright and trademark information") 網頁上取得，網址如下：www.ibm.com/legal/copytrade.shtml。

Intel、Intel 標誌、Intel Inside、Intel Inside 標誌、Intel Centrino、Intel Centrino 標誌、Celeron、Intel Xeon、Intel SpeedStep、Itanium 及 Pentium 是 Intel Corporation 或其子公司在美國及其他國家或地區的商標或註冊商標。

Linux 是 Linus Torvalds 在美國及（或）其他國家或地區的註冊商標。

Microsoft、Windows、Windows NT 及 Windows 標誌是 Microsoft Corporation 在美國及/或其他國家或地區的商標。

UNIX 是 The Open Group 在美國及其他國家或地區的註冊商標。

Java 和所有以 Java 為基礎的商標及標誌是 Oracle 及（或）其子公司的商標或註冊商標。

其他產品及服務名稱可能是 IBM 或其他公司的商標。

索引

索引順序以中文字，英文字，及特殊符號之次序排列。

〔二劃〕

〔三劃〕

〔四劃〕

〔五劃〕

〔六劃〕

〔七劃〕

〔八劃〕

〔九劃〕

〔十劃〕

〔十一劃〕

〔十二劃〕

〔十三劃〕

〔十四劃〕

〔十五劃〕

〔十六劃〕

〔十七劃〕

〔十八劃〕

〔十九劃〕

〔二十三劃〕

A

F

H

I

L

M

N

O

S

T

U

W

〔特殊字元〕

Printed in Taiwan