IBM SPSS Modeler Text Analytics 18.0 用户指南

声明

使用本信息以及其支持的产品之前，请阅读第 201 页的『声明』中的信息。

产品信息

此版本适用于 IBM SPSS Modeler Text Analytics V18.0.0 以及所有后续发行版和修订版，除非在新版本中另有声明。

目录

前言

IBM® SPSS® Modeler Text Analytics 提供强大的文本分析功能，该功能使用高级语言技术和自然语言处理 (NLP) 来快速处理各种各样的非结构化文本数据，并根据此文本抽取和阻止关键概念。此外，IBM SPSS Modeler Text Analytics 还可以将这些概念分组到类别中。

组织内保存的大约 80% 的数据采用文本文档形式，例如报告、Web 页面、电子邮件和呼叫中心注释。文本是使组织能够更好了解其客户行为的关键因素。融合 NLP 的系统可以智能抽取概念，包括复合短语。此外，利用底层语言的知识，可以使用含义和上下文将术语分类到相关组中，如产品、组织或人员。因此，可以快速确定信息与需求的相关性。这些已抽取的概念和类别可与现有结构化数据（如人口统计信息）合并，并应用于 IBM SPSS Modeler 的全套数据挖掘工具中的建模以产生更好且更专注的决策。

语言系统视知识而定 - 其字典中包含的信息越多，结果的质量越高。IBM SPSS Modeler Text Analytics 随附语言资源集，如术语和同义词字典、库以及模板。通过本产品，可以根据上下文开发并优化这些语言资源。语言资源的微调通常是一个迭代式过程，并且对于准确的概念检索和分类而言有必要。此外，还包含特定领域（如 CRM 和基因组学）的定制模板、库和字典。

有关 IBM Business Analytics

IBM Business Analytics 软件提供决策执行者信任的完整、一致和准确的信息，以改善业务性能。商业智能、预测性分析、财务性能和策略管理以及分析应用程序的综合产品服务组合提供了有关当前性能的清楚、及时和可操作洞察力，以及预测将来成果的能力。通过丰富的行业解决方案、成熟的实践和专业的服务，使不同规模的组织可实现最高效率，放心地执行自动化决策，从而实现最佳效果。

IBM SPSS Predictive Analytics 软件作为此产品服务组合的一部分，帮助组织预测将来事件，并主动针对此洞察作出响应以实现更好的业务效果。IBM SPSS 技术具有强大的竞争优势，吸引、保留和增长客户，使全球商业、政府和学术客户依赖此技术来减少欺诈和降低风险。通过在其日常操作中使用 IBM SPSS 软件，组织变为预测企业，可以自动化决策并引导这些决策满足业务目标，获得强大竞争优势。有关更多信息或要联系代表，请访问 http://www.ibm.com/spss。

技术支持

维护客户可联系技术支持。客户可联系技术支持以获取有关使用 IBM Corp. 产品的帮助，或获取有关其中一个受支持硬件环境的安装帮助。要联系技术支持，请参阅 IBM Corp. Web 站点，网址为 http://www.ibm.com/support。请求帮助时，准备提供您自己的姓名、公司名称以及支持协议。

第 1 章 关于 IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics 提供强大的文本分析功能，该功能使用高级语言技术和自然语言处理 (NLP) 来快速处理各种各样的非结构化文本数据，并根据此文本抽取和阻止关键概念。此外，IBM SPSS Modeler Text Analytics 还可以将这些概念分组到类别中。

组织内保存的大约 80% 的数据采用文本文档形式，例如报告、Web 页面、电子邮件和呼叫中心注释。文本是使组织能够更好了解其客户行为的关键因素。融合 NLP 的系统可以智能抽取概念，包括复合短语。此外，利用底层语言的知识，可以使用含义和上下文将术语分类到相关组中，如产品、组织或人员。因此，可以快速确定信息与需求的相关性。这些已抽取的概念和类别可与现有结构化数据（如人口统计信息）合并，并应用于 IBM SPSS Modeler 的全套数据挖掘工具中的建模以产生更好且更专注的决策。

语言系统视知识而定 - 其字典中包含的信息越多，结果的质量越高。IBM SPSS Modeler Text Analytics 随附语言资源集，如术语和同义词字典、库以及模板。通过本产品，可以根据上下文开发并优化这些语言资源。语言资源的微调通常是一个迭代式过程，并且对于准确的概念检索和分类而言有必要。此外，还包含特定领域（如 CRM 和基因组学）的定制模板、库和字典。

部署。您可以使用 IBM SPSS Modeler Solution Publisher 来部署文本挖掘流，以对非结构化数据进行实时评分。能够部署这些流可确保成功对文本挖掘实现进行闭合循环。例如，您的组织现在可通过应用预测模型分析入站或出站调用者的便签式注释，来实时增强市场营销消息的准确性。

注: 要将 IBM SPSS Modeler Text Analytics 与 IBM SPSS Modeler Solution Publisher 一起运行，请将目录 `<install_directory>/ext/bin/spss.TMWBServer` 添加到 `$LD_LIBRARY_PATH` 环境变量。

受支持语言的自动翻译。IBM SPSS Modeler Text Analytics 配合 SDL 的软件即服务 (SaaS) 支持您将受支持的语言列表（包括阿拉伯语、中文和波斯语）内的文本翻译为英语。然后，您可以对已翻译的文本执行文本分析并将这些结果部署至无法理解源语言内容的人员。由于文本挖掘结果自动链接回对应的外语文本，因此贵组织能够将最需要的本机表述者资源集中用于最重要的分析结果。SDL 使用统计翻译算法提供自动语言翻译，此类算法源自 20 人年的高级翻译研究。

升级到 IBM SPSS Modeler Text Analytics V18

从 PASW Text Analytics 或 Text Mining for Clementine 的先前版本升级。

在安装 IBM SPSS Modeler Text Analytics V18 之前，应保存并从当前版本导出要在新版本中使用的任何 TAP、模板和库。建议将这些文件保存到在安装最新版本时将不会删除或覆盖的目录。

安装最新版本的 IBM SPSS Modeler Text Analytics 后，可以上载已保存的 TAP 文件，添加任何已保存的库，或者导入并上载任何已保存的模板以在最新版本中进行使用。

要点: 如果卸载当前版本而未首先保存并导出所需的文件，那么在先前版本中执行的任何 TAP、模板和公共库工作都将丢失，并且无法在 IBM SPSS Modeler Text Analytics V18 中使用。

有关文本挖掘

目前，越来越多的信息以非结构化和半结构化格式进行保存，例如，客户电子邮件、呼叫中心通知、开放式调查响应、新闻订阅源和 Web 表单等。产生的此类大量信息为很多组织带来问题: 可如何收集、研究和利用此信息？

文本*挖掘*是一个分析文本材料集合的过程，用于捕获关键概念和主题，以及发现隐藏的关系和趋势，而无需知道作者用于表示这些概念的精确单词或术语。尽管存在很大差别，但有时也会将文本挖掘与信息检索混淆。虽然精确检索和存储信息具有很大挑战，但抽取和管理信息中包含的质量内容、术语和关系至关重要。

文本挖掘和数据挖掘

针对每篇包含文本的文章，基于语言的文本挖掘会返回概念索引以及有关这些概念的信息。可以将此提取的结构化信息与其他数据源组合使用来解决以下问题:

- 哪些概念会一起出现？
- 这些概念还链接到哪些对象？
- 可以从抽取的信息建立哪些更高级别的类别？
- 概念或类别预测哪些事项？
- 概念或类别如何预测行为？

将文本挖掘与数据挖掘组合比单独使用结构化或非结构化数据提供更深入的洞察力。通常，此过程包含以下步骤:

1. **识别要挖掘的文本。** 准备要挖掘的文本。如果文本存在于多个文件中，请将文件保存到单个位置。针对数据库，确定包含文本的字段。
2. **挖掘文本并抽取结构化数据。** 将文本挖掘算法应用于源文本。
3. **构建概念和类别模型。** 识别关键概念和/或创建类别。通常，将从非结构化数据返回大量概念。识别要评分的基本概念和类别。
4. **分析结构化数据。** 利用传统数据挖掘方法（例如，集群、分类和预测建模）发现概念之间的关系。将抽取的概念与其他结构化数据进行合并，以根据概念预测将来行为。

文本分析和分类

文本分析是一种量性分析，用于从文本抽取有用信息，以将此文本中包含的关键构想或概念分组到相应数目的类别。可以针对所有类型和长度的文本执行文本分析，但分析的方法在某种程度上会有所不同。

较短的记录或文档最容易进行分类，因为它们不那么复杂，通常包含的意思模糊的单词或响应较少。例如，针对较短的开放式调查问题，如果我们要相关人员列出其三个偏好的假期活动，可能会希望看到很多较短的答案，例如，*海滩度假*、*国家公园度假*或*不进行什么活动*。另一方面，较长的开放式响应可能很复杂和冗长，尤其是在响应者受过良好教育，主动且具有足够时间来完成调查表时。如果我们要相关人员在调查中告知其政治信仰或具有有关政治的博客订阅源，那么可能希望获得一些较长的有关所有种类问题和立场的注释。

使用 IBM SPSS Modeler Text Analytics 的一个关键优势是，可以在较短的时间段内从这些较长的文本源抽取关键概念和创建有洞察力的类别。通过使用自动化语言和统计方法的组合，以针对每个文本分析过程阶段获取最可靠的结果，从而利用此优势。

语言处理和 NLP

管理所有此类非结构化文本数据的主要问题在于，没有针对编写文本提供计算机可进行理解的标准规则。针对每个文档和每个文本部分，语言以及随之而来的含义都有所不同。唯一正确检索和识别此类非结构化数据的方法是，分析语言从而了解其含义。提供了多种不同自动化方法来从非结构化信息抽取概念。这些方法可划分为两种类型：语言和非语言。

一些组织已尝试根据统计和神经网络利用自动化非语言解决方案。相比人工阅读，使用计算机技术时，这些解决方案可更快速地扫描和分类关键概念。但是，此类解决方案将没那么精确。多数基于统计信息的系统仅需计算单词的出现次数以及计算其与相关概念的统计近似值。这将生成很多不相关的结果或过于杂乱，且会丢失应该找到的结果（这称为静默）。

为了补偿其在精确度方面的缺陷，一些解决方案利用可帮助区分相关结果和不相关结果的复杂非语言规则。这称为*基于规则的文本挖掘*。

另一方面，*基于语言的文本挖掘*将自然语言处理 (NLP)（对人类语言的计算机辅助分析）的原则应用于单词、短语、语法或文本结构的分析。利用 NLP 的系统可智能抽取概念（包括复合短语）。而且，对基础语言的了解可实现通过使用含义和上下文，将概念分类为相关组（例如，产品、组织或人员）。

基于语言的文本挖掘可模拟相关人员行为，通过将不同单词形式识别为具有相似含义以及分析句子结构来提供理解文本的框架，从而了解文本中具有的含义。针对基于统计信息的系统，此方法加快了速度，提高了成本效益，同时提供了更高的准确度，需要的人工干预也少很多。

要使用除了日语之外的所有语言说明抽取过程期间基于统计信息和基于语言的方法之间的差异，请考虑每种方法响应有关 `reproduction of documents` 的查询的方式。基于统计信息的解决方案和基于语言的解决方案将必须扩展单词 `reproduction` 以包含同义词，例如，`copy` 和 `duplication`。否则，将忽略相关信息。但是，如果基于统计信息的解决方案尝试执行此类同义词操作（即，搜索具有相同含义的其他术语），那么还可能会包含术语 `birth`，同时生成很多不相关的结果。对于语言的理解可降低文本模糊性，通过定义使基于语言的文本挖掘方法更加可靠。

通过感知分析器使用基于语言的方法可抽取更有意义的表达。对于表情的分析和捕获可降低文本模糊性，通过定义使基于语言的文本挖掘方法更加可靠。

理解抽取过程如何工作可帮助您在微调语言资源（库、类型和同义词等）时作出关键决策。抽取过程中会执行以下步骤:

- 将源数据转换为标准格式
- 识别候选术语
- 识别等效类和同义词整合
- 分配类型
- 建立索引，在请求时，使用辅助分析器执行模式匹配

步骤 1. 将源数据转换为标准格式

在此第一个步骤中，导入的数据将转换为可用于将来分析的统一格式。此转换在内部执行，不会更改原始数据。

步骤 2. 识别候选术语

理解语言资源的角色对于在语言抽取期间识别候选术语很重要。每次执行抽取时，都会使用语言资源。这些资源以模板、库和编译资源的形式存在。库包含单词、关系和其他用于指定或调整抽取的信息的列表。无法查看或编辑编译资源。但是，可以在模板编辑器或者如果您处于交互式工作台会话中，可以在资源编辑器中编辑剩余资源。

编译资源为 IBM SPSS Modeler Text Analytics 中抽取引擎的核心内部组件。这些资源包括一个常规字典，其中包含具有词类代码（名词、动词和形容词等）的基本形式的列表。

除了这些编译资源，本产品还提供了多个库，可用于补充编译资源中的类型和概念定义以及提供同义词。这些库以及任何您创建的定制库由多个字典组成。这些字典包括类型字典、同义词字典和排除字典。

导入和转换数据后，抽取引擎将开始识别要抽取的候选术语。候选术语是用于识别文本中概念的单词或一组单词。处理文本期间，将使用词类模式抽取器识别单个单词（**单术语**）和复合单词（**多术语**）。之后，会使用感知文本链接分析识别候选感知关键字。

注: 先前提及的常规编译字典中的术语表示可能无用地或在语言上模糊为单术语的所有单词。识别单术语时会从抽取排除这些单词。但是，确定词类或查看较长的候选复合词（多术语）时，会对其进行重新评估。

步骤 3. 识别等效类和同义词整合

识别候选单术语和多术语后，软件会使用标准化字典来识别等效类。等效类是短语的基本形式或相同短语的两个变体的单一形式。 要确定哪个概念用于等效类，抽取引擎会按列出的顺序应用以下规则:

- 库中用户指定的形式。
- 预编译资源所定义的最常见形式。

步骤 4. 分配类型

接下来，会将类型分配给抽取的概念。类型为概念的语义分组。将在此步骤中同时使用编译资源和库。类型为更高级别的概念、肯定词和否定词、名字、位置和组织等。请参阅主题第 161 页的『类型字典』，以获取更多信息。

请注意，日语资源具有一组不同类型。

语言系统具有知识敏感性，其字典中包含的信息越多，结果质量也就越高。修改字典内容（例如，同义词定义）可简化生成的信息。这通常是一个执行精确概念检索时所需的迭代式过程。 NLP 是 IBM SPSS Modeler Text Analytics 的核心元素。

抽取的工作方式

从响应抽取关键概念和构想时， IBM SPSS Modeler Text Analytics 依赖于基于语言的文本分析。此方法加快了基于统计信息的系统的处理速度，节省了成本。还提供了更高的准确度，同时需要的人工干预也少很多。基于语言的文本分析基于称为自然语言处理（也称为计算语言）的学习字段。

要点: 针对日语文本，抽取过程将遵循一组不同步骤。

理解抽取过程如何工作可帮助您在微调语言资源（库、类型和同义词等）时作出关键决策。抽取过程中会执行以下步骤:

- 将源数据转换为标准格式
- 识别候选术语
- 识别等效类和同义词整合
- 分配类型
- 建立索引
- 匹配模式和事件抽取

步骤 1. 将源数据转换为标准格式

在此第一个步骤中，导入的数据将转换为可用于将来分析的统一格式。此转换在内部执行，不会更改原始数据。

步骤 2. 识别候选术语

理解语言资源的角色对于在语言抽取期间识别候选术语很重要。每次执行抽取时，都会使用语言资源。这些资源以模板、库和编译资源的形式存在。库包含单词、关系和其他用于指定或调整抽取的信息的列表。无法查看或编辑编译资源。但是，可以在模板编辑器或者如果您处于交互式工作台会话中，可以在资源编辑器中编辑编辑剩余资源（模板）。

编译资源为 IBM SPSS Modeler Text Analytics 中抽取引擎的核心内部组件。这些资源包括一个常规字典，其中包含具有词类代码（名词、动词、形容词、副词、分词、连词、限定词或介词）的基本形式的列表。这些资源包括保留的内置类型，用于将很多抽取的术语分配给以下类型: `<Location>`、`<Organization>` 或 `<Person>`。请参阅主题第 162 页的『内置类型』，以获取更多信息。

除了这些编译资源，本产品还提供了多个库，可用于补充编译资源中的类型和概念定义以及提供其他类型和同义词。这些库以及任何您创建的定制库由多个字典组成。这些字典包括类型字典、替换字典（同义词和可选元素）和排除字典。请参阅主题第 153 页的第 16 章，『处理库』，以获取更多信息。

导入和转换数据后，抽取引擎将开始识别要抽取的候选术语。候选术语是用于识别文本中概念的单词或一组单词。处理文本期间，会将编译资源中不存在的单个单词（*单术语*）视为候选术语抽取。将使用词类模式抽取器识别候选复合单词（*多术语*)。例如，多术语 `sports car` 遵循“形容词和名称”词类模式，具有两个组成部分。多术语 `fast sports car` 遵循“形容词、形容词和名称”词类模式，具有三个组成部分。

注: 先前提及的常规编译字典中的术语表示可能无用地或在语言上模糊为单术语的所有单词。识别单术语时会从抽取排除这些单词。但是，确定词类或查看较长的候选复合词（多术语）时，会对其进行重新评估。

最后，将使用特殊算法来处理大写字母字符串（例如，职位），以便可抽取这些特殊模式。

步骤 3. 识别等效类和同义词整合

识别候选单术语和多术语后，软件会使用一组算法来进行比较和识别等效类。等效类是短语的基本形式或相同短语的两个变体的单一形式。将短语分配给等效类是为了确保不会将诸如 `president of the company` 和 `company president` 的短语视为单独概念。要确定哪个概念用于等效类（即，将 `president of the company` 还是 `company president` 用作引导术语），抽取引擎会按列出的顺序应用以下规则:

- 库中用户指定的形式。
- 正文中最常见的形式。
- 正文中最简短的形式（通常与基本形式对应）。

步骤 4. 分配类型

接下来，会将类型分配给抽取的概念。类型为概念的语义分组。将在此步骤中同时使用编译资源和库。类型为更高级别的概念、肯定词和否定词、名字、位置和组织等。用户可定义其他类型。请参阅主题第 161 页的『类型字典』，以获取更多信息。

步骤 5. 建立索引

通过在文本位置和每个等效类的代表术语之间建立指针，对整组记录或文档建立索引。这假设候选概念的所有屈折变化形式实例都已以候选基本形式建立了索引。针对每种基本形式计算全局频率。

步骤 6. 匹配模式和事件抽取

IBM SPSS Modeler Text Analytics 不仅可发现类型和概念，还可发现其之间的关系。本产品具有多种算法和库，且提供了抽取类型和概念之间关系模式的能力。这在尝试发现特定意见（例如，产品反映）或人员或对象（例如，政治团体或基因组之间的链接）之间关系链接时尤其有用。

分类工作方式

在 IBM SPSS Modeler Text Analytics 中创建类别模型时，可选择多种不同方法来创建类别。由于每个数据集唯一，方法数和应用这些方法的顺序会发生变化。由于其他人对于结果的解释可能会有所不同，因此可能需要试验不同方法，以了解哪种方法可针对文本数据产生最佳效果。在 IBM SPSS Modeler Text Analytics 中，可在工作台会话中创建类别模型，在此会话中，可浏览和进一步微调类别。

在本指南中，**类别构建**是指通过使用一个或多个内置方法生成类别定义和分类，**分类**指评分或添加标签过程，通过此过程将唯一标识（名称/标识/值）分配给每个记录或文档的类别定义。

类别构建期间，抽取的概念和类型用作类别的构建块。构建类别时，如果记录或文档包含匹配类别定义的元素，那么会将其自动分配给类别。

IBM SPSS Modeler Text Analytics 为您提供了多种自动化类别构建方法，帮助您快速对文档或记录进行分类。

分组方法

每个可用方法都明确适合于特定类型的数据和情况，但通常有助于组合同一分析中的方法以捕获完整范围的文档或记录。您可能会看到属于多个类别的概念或找到冗余类别。

概念根派生。此方法采用某个概念并通过分析任何概念组件是词法相关还是共享根来查找与其相关的其他概念，从而创建类别。此方法对于识别同义复合词概念非常有用，因为所生成的每个类别中的概念是同义词或含义紧密相关。它处理变长数据并生成更小数量的紧凑类别。例如，概念 `opportunities to advance` 将与概念 `opportunity for advancement` 和 `advancement opportunity` 分组在一起。请参阅主题第 97 页的『概念根派生』以获取更多信息。此选项不适用于日语文本。

语义网络。此方法首先从每个概念的广泛单词关系索引确定其可能含义，然后通过将相关概念分组来创建类别。此方法在概念对于语义网络已知且不太模糊时最适用。在文本包含对于网络未知的专用术语或行话时帮助不大。在一个示例中，概念 `granny smith apple` 可能会与 `gala apple` 和 `winesap apple` 分组在一起，因为它们是 granny smith 的同代。在另一个示例中，概念 `animal` 可能与 `cat` 和 `kangaroo` 分组在一起，因为它们是 `animal` 的下义词。此方法仅在本发行版中适用于英语文本。请参阅主题第 99 页的『语义网络』以获取更多信息。

概念包含。此方法通过基于多术语概念（复合词）包含属于另一个类别中单词的子集还是超集的单词将其分组来构建类别。例如，概念 `seat` 将会与 `safety seat`、`seat belt` 和 `seat belt buckle` 分组在一起。请参阅主题第 98 页的『概念包含』以获取更多信息。

共生。此方法根据在文本中找到的共生来创建类别。构想是当通常在文档和记录中同时找到概念和概念模式时，共生会反映可能在类别定义中有意义的底层关系。当单词显著共生时，将会创建共生规则并将其用作新的子类别的类别描述符。例如，如果许多记录包含单词 `price` 和 `availability`（但少数记录仅包含其中一个），那么这些概念可能会分组到共生规则中 (`price & available`) 并分配到类别（例如 `price`）的子类别。请参阅主题第 100 页的『同现规则』以获取更多信息。

最小 文档。要帮助确定共生的有趣程度，请定义必须包含给定共生才能在类别中用作描述符的最小文档或记录数。

IBM SPSS Modeler Text Analytics 节点

通过随 IBM SPSS Modeler 交付的许多标准节点，您还可以使用文本挖掘角点来将文本分析的能力整合到自己的流中。IBM SPSS Modeler Text Analytics 为您提供了多个文本挖掘节点用于执行此功能。这些节点存储在节点选用板的 IBM SPSS Modeler Text Analytics 选项卡中。

其中包含以下节点:

- **“文件列表”源节点**可生成文档名称列表作为文本挖掘流程的输入。当文本驻留在外部文档中而不是在数据库或其他结构化文件中时，这很有用。节点可输出单个字段，其中针对列出的每个文档或文件夹包含一条记录，可选中这些字段作为后续“文本挖掘”节点中的输入。请参阅第 9 页的『“文件列表”节点』主题以获取更多信息。
- **Web 订阅源源节点**可从 Web 订阅源（例如，RSS 或 HTML 格式的博客或新闻订阅源）中读取文本，并在文本挖掘流程中使用此数据。节点可针对订阅源中找到的每条记录输出一个或多个字段，可选中这些字段作为后续“文本挖掘”节点中的输入。请参阅第 11 页的『Web 订阅源节点』主题以获取更多信息。
- **“文本挖掘”节点**使用语言方法来从文本中抽取关键概念，使您能够利用这些概念和其他数据创建类别，并能够基于已知模式（称为文本链接分析）识别概念之间的关系和关联。此节点可用于探索文本数据内容或者生成概念模型或类别模型。概念和类别可与现有结构化数据（例如人口统计学）进行组合，并且可应用于建模。请参阅第 16 页的『“文本挖掘”建模节点』主题以获取更多信息。
- **“文本链接分析”节点**可抽取概念，并且可基于文本内已知模式识别概念之间的关系。模式抽取可用于发现概念以及随附到这些概念的任何意见或限定符之间的关系。 “文本链接分析”节点提供了更直接的方式来从您的文本中识别和抽取模式，然后将模式结果添加到流中的数据集内。但是您还可以在“文本挖掘”建模节点中使用交互式工作台会话来执行 TLA。请参阅第 39 页的『“文本链接分析”节点』主题以获取更多信息。
- **“翻译”节点**可用于将文本从受支持的语言（例如，阿拉伯语、中文和波斯语）翻译为英语或其他语言用于建模。由此可实现其他方法无法实现的挖掘双字节语言中的文档的能力，并且允许分析人员即使无法使用所涉及的语言，仍可以从这些文档中抽取概念。可从任何文本建模节点调用此功能，但是独立的“翻译”节点可以在多个节点中缓存并复用翻译。请参阅第 47 页的『翻译节点』主题以获取更多信息。
- 从外部文档挖掘文本时，**“文本挖掘输出”节点**可用于生成 HTML 页面，其中包含到从中抽取概念的文档的链接。请参阅第 49 页的『文件查看器节点』主题以获取更多信息。

应用程序

总而言之，需要例行性需要复审大量文档以识别需要进一步探索的关键元素的任何人都可以从 IBM SPSS Modeler Text Analytics 获益。

部分具体应用程序包括:

- **科学和医学研究。**探索辅助研究资料，例如，专利报告、日志文章和协议出版物。识别先前未知的关联（例如与特定产品相关联的博士），展示进一步探索的途径。最大限度缩短药物发现过程中所耗用的时间。用于辅助基因研究。
- **投资的研究。**复审日常分析报告、新文章和公司新闻发布会以确定关键战略要点或市场变化。此类信息的趋势分析会长期不断为企业或行业展示新兴问题或商机。
- **欺诈检测。**使用银行或医疗保健欺诈来检测异常状况和发现大量文本中的危险信号。
- **市场研究。**用于市场研究工作中，以识别开放式调研响应中的关键主题。
- **博客和 Web 订阅源分析。**使用新订阅源、博客等重找到的关键构想来探索和构建模型。
- **CRM。**使用来自所有客户接触方式（例如，电子邮件、事务和调查）的数据构建模型。

第 2 章 读取源文本

文本挖掘的数据可能采用 IBM SPSS Modeler 所使用的任何标准格式（包括数据库或其他表示行和列中的数据的“矩形”格式）或不符合此结构的文档格式（例如，Microsoft Word、Adobe PDF 或 HTML）。

- 要从不符合标准数据结构的文档（例如，Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF、XML、HTML 等）中读取文本，可使用“文件列表”节点来生成文档或文件夹列表作为文本挖掘流程的输入。有关更多信息，请参阅『“文件列表”节点』。
- 要从 Web 订阅源（例如，RSS 或 HTML 格式的博客或新闻订阅源）中读取文本，可使用“Web 订阅源”节点来格式化 Web 订阅源数据作为文本挖掘流程的输入。有关更多信息，请参阅第 11 页的『Web 订阅源节点』。
- 要从 IBM SPSS Modeler 所使用的任何标准数据格式（例如，具有一个或多个文本字段用于添加客户注释的数据库）中读取文本，可使用任何 IBM SPSS Modeler 本机标准源节点。请参阅 IBM SPSS Modeler 节点文档以获取更多信息。

“文件列表”节点

要从以各种格式（例如，Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF、XML、HTML 等）保存的非结构化文档中读取文本，可使用“文件列表”节点来生成文档或文件夹列表作为文本挖掘流程的输入。这是必要的，因为无法以与 IBM SPSS Modeler 所使用的其他数据相同的方式通过字段和记录（行和列）来表示非结构化文本文档。可在“文本挖掘”选用板上找到此节点。

“文本列表”节点充当源节点；但是，除读取和输出源文件的实际数据之外，您还可以使用此节点来读取指定的根目录下的文档或目录的名称并将这些名称生成列表。用于读取文档或目录名称时，输出为单个字段，列出的每个文件含一条记录，可选中这些字段作为后续“文本挖掘”或“文本链接分析”节点的输入。

您可以在 IBM SPSS Modeler 窗口底部的节点选用板的 IBM SPSS Modeler Text Analytics 选项卡上找到此节点。请参阅主题第 7 页的『IBM SPSS Modeler Text Analytics 节点』以获取更多信息。

要点：不支持任何包含机器本地编码中未包含的字符的目录名称和文件名。尝试执行包含年文件列表节点的流时，任何包含这些字符的文件名或目录名称都将导致流执行失败。对于外语目录名称或文件名（如法语语言环境中的日语文件名），可能会发生此情况。

本地数据支持。如果您已连接到 IBM SPSS Modeler Text Analytics Server ，并且具有一个含“文件列表”节点的流，那么数据应驻留在 IBM SPSS Modeler Text Analytics Server 所在机器上，或者确保服务器有权访问存储“文件列表”节点中的源数据的文件夹。

注：您不能将“文件列表”节点用于在 IBM SPSS Collaboration and Deployment Services - Scoring 配置内进行评分。

“文件列表”节点：“设置”选项卡

在该选项卡上，可以定义该节点的目录、文件扩展名和输入。

注：文本挖掘抽取无法处理非 Microsoft Windows 平台下的 Microsoft Office 和 Adobe PDF 文件。但是，始终可处理 XML、HTML 或文本文件。

不支持任何包含机器本地编码中未包含的字符的目录名称和文件名。尝试执行包含年文件列表节点的流时，任何包含这些字符的文件名或目录名称都将导致流执行失败。对于外语目录名称或文件名（如法语语言环境中的日语文件名），可能会发生此情况。

目录 指定包含要列出的文档的根文件夹。

- **包含子目录** 指定还应同时扫描子文件夹。

要在列表中包含的文件类型: 您可以选中或取消选中要使用的文件类型和扩展名。通过取消选中文件扩展名，将忽略含此扩展名的文件。您可以按以下扩展名来过滤:

表 1. 按文件扩展名来进行文件类型过滤.

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

注: 有关更多信息，请参阅第 9 页的『“文件列表”节点』。

如果某些文件无扩展名或者仅含一个尾部句点扩展（例如，`File01` 或 `File01.`），那么请使用**无扩展名**选项来选中这些文件。

输出字段表示 选择输出字段的格式。选项包括:

- **实际文本** 如果该字段将包含实际文本，那么请选择该选项。然后，您可以从以下列表中选择**输入编码**值
  - 自动（欧洲）
  - 自动（日本）
  - UTF-8
  - UTF-16
  - ISO-8859-1
  - ISO-8859-2
  - Windows-1250
  - US ascii
- **文档路径名** 如果输出字段将包含文档所在位置的一个或多个路径名，请选择该选项。

输出显示为 UTF-8 文档文本。

要点: 从 V14 起，“目录列表”选项不再可用，并且仅列出文件列表作为输出。

“文件列表”节点: “其他”选项卡

“类型”选项卡和“注释”选项卡都是 IBM SPSS Modeler 节点中的标准选项卡。

在文本挖掘中使用“文件列表”节点

当驻留在外部非结构化文档中的文本数据的格式为 Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF、XML、HTML 等格式时，会使用“文件列表”节点。除输出实际文本之外，您还可以使用该节点来生成文档或文件夹列表，作为文本挖掘流程（例如，后续“文本挖掘”或“文本链接分析”节点）的输入。

如果使用“文件列表”节点来生成文档列表代替实际文本；那么后续使用“文本挖掘”或“文本链接分析”节点时，请指定**文本字段**表示**到文档的路径名**以指示所选字段包含到文本所在文档的路径，而不是包含要挖掘的实际文本。

例如，假定我们已将“文件列表”节点连接到“文本挖掘”节点以供应驻留在外部文档中的文本:

1. **文件列表节点（“设置”选项卡）。**首先，将此节点添加到流，以指定文本文档的存储位置。选择包含要对其执行文本挖掘的所有文档的目录。
2. **文本挖掘节点（“字段”选项卡）。**接下来，添加文本挖掘节点并将其连接到文件列表节点。在此节点中，定义输入格式、资源模板和输出格式。选择从文件列表节点产生的文件名，并且选择其中文本字段表示**文档路径名**的选项以及其他设置。请参阅主题第 24 页的『在流中使用文本挖掘节点』以获取更多信息。

有关使用“文本挖掘”节点的更多信息，请参阅第 16 页的『“文本挖掘”建模节点』。

Web 订阅源节点

Web 订阅源节点可用于为文本挖掘过程准备 Web 订阅源中的文本数据。此节点接受两种格式的 Web 订阅源:

- RSS 格式。RSS 是 Web 订阅源的基于 XML 的简单标准化格式。此格式的 URL 指向具有链接文章集合（如联合新闻源和博客）的页面。由于 RSS 是标准化格式，因此会自动识别每篇链接文章并在生成的数据流中作为单独的记录进行处理。除非要对文本应用过滤方法，否则无需进一步输入即可识别订阅源中的重要文本数据和记录。
- HTML 格式。可以在“输入”选项卡上定义一个或多个指向 HTML 页面的 URL。然后，在“记录”选项卡中，定义记录开始标记，以及识别用于分隔目标内容的标记并将这些标记分配到所选输出字段（描述、标题、修改日期等）。请参阅主题第 12 页的『Web 订阅源节点: “记录”选项卡』以获取更多信息。

重要！ 如果要尝试通过代理服务器在 Web 上检索信息，那么必须为 IBM SPSS Modeler Text Analytics 客户机和服务器在 `net.properties` 文件中启用代理服务器。执行此文件内详细描述的指示信息。这适用于通过 Web 订阅源节点访问或检索 SDL 软件即服务 (SaaS) 许可证的情况，因为这些连接通过 Java™。缺省情况下，此文件位于 *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties* 中。

此节点的输出是一组用于描述记录的字段。由于**描述**字段包含批量文本内容，因此其最常用。但是，您也可能对其他字段感兴趣，如记录的简短描述（**简短描述**字段）或记录的标题（**标题**字段）。任何输出字段都可选择作为后续文本挖掘节点的输入。

注: 不能使用 Web 订阅源节点在 IBM SPSS Collaboration and Deployment Services - Scoring 配置内进行评分。

您可以在 IBM SPSS Modeler 窗口底部的节点选用板的 IBM SPSS Modeler Text Analytics 选项卡上找到此节点。请参阅主题第 7 页的『IBM SPSS Modeler Text Analytics 节点』以获取更多信息。

Web 订阅源节点: “输入”选项卡

“输入”选项卡用于指定一个或多个 Web 地址或 URL，以便捕获文本数据。在文本挖掘的上下文中，可以指定包含文本数据的订阅源的 URL。

注意！处理非 RSS 数据时，您可能首选使用 Web 脚本编制工具（如 WebQL®）将内容收集自动化，然后使用其他源节点引用来自该工具的输出。

可以设置以下参数:

输入或粘贴 URL。在此字段中，可以输入或粘贴一个或多个 URL。如果是输入多个 URL，请每行仅输入一个 URL，并且使用 **Enter/回车**键分隔各行。请输入文件的完整 URL 路径。这些 URL 可用于以下两种格式之一的订阅源:

- *RSS 格式*。RSS 是 Web 订阅源的基于 XML 的简单标准化格式。此格式的 URL 指向具有链接文章集合（如联合新闻源和博客）的页面。由于 RSS 是标准化格式，因此会自动识别每篇链接文章并在生成的数据流中作为单独的记录进行处理。除非要对文本应用过滤方法，否则无需进一步输入即可识别订阅源中的重要文本数据和记录。
- *HTML 格式*。可以在“输入”选项卡上定义一个或多个指向 HTML 页面的 URL。然后，在“记录”选项卡中，定义记录开始标记，以及识别用于分隔目标内容的标记并将这些标记分配到所选输出字段（描述、标题、修改日期等）。处理非 RSS 数据时，您可能首选使用 Web 脚本编制工具（如 WebQL®）将内容收集自动化，然后使用其他源节点引用来自该工具的输出。请参阅主题『Web 订阅源节点: “记录”选项卡』以获取更多信息。

每个 URL 的要读取的最新条目数。此字段指定以订阅源中找到的第一个记录开头的字段中所列的每个 URL 的要读取的最大记录数。文本量会影响文本挖掘节点或文本链接分析节点中下游抽取期间的处理速度。

保存并复用先前 Web 订阅源（如有可能）。使用此选项时，将会扫描 Web 订阅源并高速缓存处理结果。然后，在后续流执行时，如果给定订阅源的内容尚未更改或者订阅源无法访问（例如，因特网停运），那么会使用高速缓存的版本加快处理时间。在这些订阅源中发现的任何新内容也会高速缓存，以供下次执行节点时使用。

- **标签**。如果选择**保存并复用先前 Web 订阅源（如有可能）**，那么必须指定结果的标签名称。此标签用于描述服务器上的高速缓存订阅源。如果未指定标签或标签无法识别，那么将无法复用。可以在 IBM SPSS Text Analytics Administration Console 的会话表中管理这些 Web 订阅源高速缓存。请参阅《IBM SPSS Text Analytics Administration Console 用户指南》以获取更多信息。

Web 订阅源节点: “记录”选项卡

“记录”选项卡用于通过标识每个新记录的开始位置来指定非 RSS 订阅源的文本内容，以及指定关于每个记录的其他相关信息。如果您知道非 RSS 订阅源 (HTML) 包含位于多个记录中的文本，那么必须在此处标识记录开始标记，否则文本将作为一个记录进行处理。虽然 RSS 订阅源已标准化且无需在此选项卡上指定任何标记，但是仍然可以在“预览”选项卡中预览内容。

注意！处理非 RSS 数据时，您可能首选使用 Web 脚本编制工具（如 WebQL®）将内容收集自动化，然后使用其他源节点引用来自该工具的输出。

URL。此下拉列表包含在“输入”选项卡上输入的 URL 的列表。HTML 和 RSS 格式化订阅源均存在。如果 URL 地址对于下拉列表过长，那么将自动在中间对其进行剪切（使用省略符替换剪切文本，如 *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*）。

- 使用 **HTML 格式化订阅源**时，如果订阅源包含多个记录（或条目），那么可以定义哪些 HTML 标记包含与表中显示的字段对应的数据。例如，可以定义开始标记（用于指示新记录已启动）、修改日期标记或作者名称。
- 使用 **RSS 格式化订阅源**时，由于 RSS 是标准化格式，因此不会提示您输入任何标记。但如果需要，可以在“预览”选项卡上查看样本结果。所有已识别的 RSS 订阅源都前置有 RSS 徽标图标。

“源代码”选项卡。在此选项卡上，可以查看任何 HTML 订阅源的源代码。此代码不可编辑。可以使用“查找”字段在此页面上查找随后可复制并粘贴到下表中的特定标记或信息。“查找”字段不区分大小写，并将与部分字符串相匹配。

“预览”选项卡。在此选项卡上，可以预览 Web 订阅源节点将如何读取记录。这对于 HTML 订阅源特别有用，因为可以通过在“预览”选项卡下的表中定义 HTML 标记来更改将如何读取记录。

非 RSS 记录开始标记。此选项仅适用于非 RSS 订阅源。如果 HTML 订阅源包含要划分为多个记录的多个文本，请在此处指定用于表示记录（如文章或博客条目）开始的 HTML 标记。如果不为非 RSS 订阅源定义开始标记，那么整个页面都作为一个记录进行处理，全部内容都是**描述**字段中的输出，并且节点执行日期同时用作**修改日期**和**发布日期**。

字段表。此选项仅适用于非 RSS 订阅源。在此表中，可以通过为任何预定义输出字段定义开始标记将文本内容划分为多个特定输出字段。请仅输入开始标记。通过解析 HTML 并将表内容与在 HTML 中找到的标记名称和属性相匹配来执行所有匹配。可以使用底部的按钮复制已定义的标记并将其复用于其他订阅源。

表 2. 非 RSS 订阅源的可能输出字段（HTML 格式）

输出字段名称	预期标记内容
标题	用于定界记录标题的标记。（可选）
简短描述	用于定界简短描述或标签的标记。（可选）
描述	用于定界主要文本的标记。如果保留为空白，那么此字段将包含 <body> 标记中的所有其他内容（如果有单个记录）或在当前记录内找到的内容（当已指定记录定界符时）。
作者	用于定界文本作者的标记。（可选）
撰稿者	用于定界撰稿者名称的标记。（可选）
发布日期	用于定界文本发布日期的标记。如果保留为空白，那么此字段将包含节点读取数据的日期。
修改日期	用于标记文本修改日期的标记。如果保留为空白，那么此字段将包含节点读取数据的日期。

在表中输入标记时，将使用此标记作为要匹配的最小标记而非完全匹配来扫描订阅源。也就是说，如果为“标题”字段输入了 ，那么这将会与订阅源中的任何 标记相匹配，包括具有指定属性（如 ）的标记，以便 等同于根标记 ( ) 以及任何包含属性的派生项，并且将该内容用于“标题”输出字段。如果输入根标记，那么还包含任何进一步属性。

表 3. 用于标识输出字段的文本的 HTML 标记示例

如果输入:	它将匹配:	并且还匹配:	但不匹配:
			任何其他标记
<p class="auth">	<p class="auth">	<p color="black" class="auth" id="85643">	<p color="black">

Web 订阅源节点: “内容过滤器”选项卡

“内容过滤器”选项卡用于将过滤器方法应用于 RSS 订阅源内容。此选项卡不适用于 HTML 订阅源。如果订阅源包含许多页眉、页脚、菜单、广告等形式的文本，那么可能要进行过滤。可以使用此选项卡从内容中去除多余的 HTML 标记、JavaScript 和短字或短行。

内容过滤。如果不希望应用清除方法，请选择**无**。否则，选择 **RSS 内容清除器**。

RSS 内容清除器选项。如果选择 **RSS 内容清除器**，那么可以选择基于特定条件来废弃行。行以 HTML 标记（如 <p> 和 ）分隔，但是内嵌标记（如 、<b> 和 <font>）除外。请注意，
 标记作为换行符处理。

- **废弃短行**。此选项会忽略不包含此处定义的**最小字数**的行。

- **废弃包含短字的行**。此选项会忽略具有超过此处定义的**最小平均字长**的行。
- **废弃包含多个单字符字的行**。此选项会忽略包含超过特定**单字符字比例**的行。
- **废弃包含特定标记的行**。此选项会忽略包含字段中指定的任何标记的行中的文本。
- **废弃包含特定文本的行**。此选项会忽略包含字段中指定的任何文本的行。

在文本挖掘中使用 Web 订阅源节点

Web 订阅源节点可用于为文本挖掘过程准备因特网 Web 订阅源中的文本数据。此节点接受 HTML 或 RSS 格式的 Web 订阅源。这些订阅源充当文本挖掘过程的输入（后续文本挖掘节点或文本链接分析节点）。

如果使用 Web 订阅源节点，那么必须确保指定“文本”字段表示文本挖掘节点或文本链接分析节点中的**实际文本**，以指示这些订阅源直接链接到各文章或博客条目。

重要！ 如果要尝试通过代理服务器在 Web 上检索信息，那么必须为 IBM SPSS Modeler Text Analytics 客户机和服务器在 `net.properties` 文件中启用代理服务器。执行此文件内详细描述的指示信息。这适用于通过 Web 订阅源节点访问或检索 SDL 软件即服务 (SaaS) 许可证的情况，因为这些连接通过 Java。缺省情况下，此文件位于 *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties* 中。

示例: 具有文本挖掘建模节点的 Web 订阅源节点（RSS 订阅源）

例如，假设将 Web 订阅源节点连接到文本挖掘节点，以便将 RSS 订阅源中的文本数据提供到文本挖掘过程中。

1. **Web 订阅源节点（“输入”选项卡）**。首先，将此节点添加到流，以指定订阅源内容的所在位置并验证内容结构。在第一个选项卡上，提供 RSS 订阅源的 URL。由于示例针对 RSS 订阅源，因此已定义格式化，并且无需在“记录”选项卡上进行任何更改。可选内容过滤算法可用于 RSS 订阅源，但在本例中未提供该算法。
2. **文本挖掘节点（“字段”选项卡）**。接下来，添加文本挖掘节点并将其添加到 Web 订阅源节点。在此选项卡上，按 Web 订阅源节点定义文本字段输出。在本例中，将要使用**描述**字段。此外，选择表示**实际文本**的选项“文本”字段，以及其他设置。
3. **文本挖掘节点（“模型”选项卡）**。接下来，在“模型”选项卡上，选择构建方式和资源。在本例中，选择使用缺省资源模板直接从此节点构建概念模型。

有关使用文本挖掘节点的更多信息，请参阅第 16 页的『“文本挖掘”建模节点』。

第 3 章 挖掘概念和类别

“文本挖掘”建模节点用于生成以下两种文本挖掘模型块之一:

- *概念模型块*: 显示屏抽取来自结构化和非结构化文本数据的重要概念。
- *类别模型块*: 对文档和记录进行评分并将其分配到由抽取的概念（和模式）组成的类别中。

抽取的概念和类别可与现有结构化数据（例如人口统计学）进行组合，并且可用于借助 IBM SPSS Modeler 的一整套数据挖掘工具来进行建模，以此实现更好更集中的决策。例如，如果客户频繁将登录问题列为完成联机帐户管理任务的主要障碍，那么您可能会希望将“登录问题”合并到模型。

此外，文本挖掘建模节点在 IBM SPSS Modeler 中完全集成，以便您可通过 IBM SPSS Modeler Solution Publisher 部署文本挖掘流，从而实现对应用程序（例如，PredictiveCallCenter）中非结构化数据进行实时评分。能够部署这些流可确保成功对文本挖掘实现进行闭合循环。例如，您的组织现在可通过应用预测模型分析入站或出站调用者的便签式注释，来实时增强市场营销消息的准确性。使用文本挖掘模型使流实现改善预测数据模型的准确性。

注: 要将 IBM SPSS Modeler Text Analytics 与 IBM SPSS Modeler Solution Publisher 一起运行，请将目录 `<install_directory>/ext/bin/spss.TMWBServer` 添加到 `$LD_LIBRARY_PATH` 环境变量。

在 IBM SPSS Modeler Text Analytics 中，我们经常参考抽取的概念和类别。理解概念和类别的意义很重要，因为它们可帮助您在进行说明性工作和构建模型时作出更明智的决策。

概念和概念模型块

抽取过程期间，会扫描和分析文本数据以识别感兴趣或相关单个字（例如 `election` 或 `peace`）和短语（例如，`presidential election`、`election of the president` 或 `peace treaties`）。这些字和短语统一称为*术语*。使用语言资源，抽取相关术语，且会将相似术语分组在称为**概念**的前导术语下。

通过此方式，根据文本和您正使用的一组语言资源，概念可表示多个底层术语。例如，假设具有一份员工满意度调查，且抽取了概念 `salary`。假设您在查看了与 `salary` 关联的记录时，注意到 `salary` 未始终显示在文本中，而是显示了包含某些相似内容的记录，例如，术语 `wage`、`wages` 和 `salaries`。这些术语分组在 `salary` 下，因为抽取引擎根据处理规则或语言资源将其视为相似或确定其为同义词。在此情况下，包含任何这些术语的任何文档或记录将被视为其包含了单词 `salary`。

如果您希望了解哪些术语分组在概念下，那么可浏览交互式工作台中的概念，或查看概念模型中显示了哪些同义词。请参阅第 27 页的『概念模型中的底层术语』主题以获取更多信息。

概念模型块包含一组可用于识别也包含了概念（包括其任何同义词或分组的术语）的记录或文档的概念。可以通过两种方式使用概念模型。第一种方式是浏览和分析在原始源文本中发现的概念，或快速识别相关文档。第二种方式是将此模型应用到新的文本记录或文档，以快速识别新文档/记录中的相同关键概念，例如，实时从呼叫中心发现便签式数据中的关键概念。

请参阅第 25 页的『文本挖掘块: 概念模型』主题以获取更多信息。

类别和类别模型块

您可以创建**类别**，这些类别实质上表示用于捕获以文本表示的关键构想、知识和看法的较高级别概念或主题。类别由一组描述符（例如，*概念*、*类型*和*规则*）组成。这些描述符一起用于识别记录或文档是否属于给定类别。可以扫描文档或记录以查看其任何文本是否匹配描述符。如果找到匹配项，那么会向此类别分配文档/记录。该过程称为**分类**。

可以使用产品的一组成熟自动化技术自动构建类别，或手动使用您可能具有的有关数据的其他了解来构建类别，或同时使用这两种方式。您还可以通过此节点的“模型”选项卡从文本分析软件包装入一组预构建类别。仅可通过交互式工作台手动创建类别或优化类别。请参阅第 19 页的『“文本挖掘”节点: “模型”选项卡』主题以获取更多信息。

类别模型块包含一组类别及其描述符。可使用模型来基于每个文档/记录中的文本分类一组文档或记录。将读取每个文档或记录，然后将其分配到找到了描述符匹配的每个类别。通过此方式，可以将文档或记录分配给多个类别。例如，您可使用类别模型块来查看开放式调查响应或一组博客条目中的关键构想。

请参阅第 33 页的『文本挖掘块: 类别模型』主题以获取更多信息。

“文本挖掘”建模节点

“文本挖掘”节点使用语言和频率技术来从文本中抽取关键概念，并使用这些概念和其他数据创建类别。此节点可用于探索文本数据内容或者生成概念模型块或类别模型块。执行此建模节点时，内部语言抽取引擎会使用自然语言处理方法抽取和组织概念、模式和/或类别。

您可以使用**直接生成**选项执行“文本挖掘”节点并自动生成概念或类别模型块。或者，您可以使用**以交互方式构建**方式来使用更加实践性和探索性的方法，通过这种方式，您可以抽取概念、创建类别和优化您的语言资源，并且可以执行文本链接分析和探索集群。请参阅第 19 页的『“文本挖掘”节点: “模型”选项卡』主题以获取更多信息。

您可以在 IBM SPSS Modeler 窗口底部的节点选用板的 IBM SPSS Modeler Text Analytics 选项卡上找到此节点。请参阅主题第 7 页的『IBM SPSS Modeler Text Analytics 节点』以获取更多信息。

需求。文本挖掘建模节点接受来自 Web 订阅源节点、文本列表节点或者任何标准源代码的文本数据。该节点随 IBM SPSS Modeler Text Analytics 一起安装，可在 IBM SPSS Modeler Text Analytics 选用板上访问此节点。

注: 该节点为所有用户替换“文本抽取”节点，为日语用户替换先前版本的“Clementine 文本挖掘”中提供的旧“文本挖掘”节点。如果您拥有使用这些节点或模型块的较旧的流，那么必须使用新的“文本挖掘”节点来重新构建自己的流。

“文本挖掘”节点: “字段”选项卡

“字段”选项卡用于指定您将从中抽取概念的数据的字段设置。处理较大型的数据集时请考虑使用来自此节点的“样本”节点上游。请参阅第 24 页的『对上游采样以节省时间』主题以获取更多信息。

您可以设置以下参数:

文本字段。选择包含要挖掘的文本的字段、文档路径名或文档的目录路径名。此字段取决于数据源。

文本字段表示。指示在先前设置中指定的文本字段包含的内容。选项包括:

- **实际文本。**如果字段包含应从中抽取概念的确切文本，请选择此选项。
- **文档路径名。**如果字段包含文本文档驻留所在的位置的一个或多个路径名，请选择此选项。

文档类型。仅在指定文本字段表示**文档路径名**的情况下，此选项才可用。文档类型指定文本的结构。请选择以下类型之一:

- **全文本。**用于大多数文档或文本源。系统会扫描整个文本集以进行抽取。与其他选项不同，此选项没有其他设置。
- **结构化文本。**用于书目形式、专利以及任何包含可识别并分析的常规结构的文件。此文档类型用于跳过全部或部分抽取过程。通过它可定义术语分隔符，分配类型和施加最小频率值。如果选择此选项，那么必须单击**设置**按钮并在"文档设置"对话框的**结构化文本格式化**区域中输入文本分隔符。请参阅主题『"字段的文档设置"选项卡』以获取更多信息。
- **XML 文本。**用于指定包含要抽取的文本的 XML 标记。系统会忽略所有其他标记。如果选择此选项，那么必须单击**设置**按钮并在"文档设置"对话框的 **XML 文本格式化**区域中指定包含抽取过程中要读取的文本的 XML 元素。请参阅主题『"字段的文档设置"选项卡』以获取更多信息。

文本统一。仅在指定文本字段表示**文档路径名**并将**全文本**设置为文档类型的情况下，此选项才可用。选择以下抽取方式:

- **文档方式。**用于简短且语义同构的文档，如来自新闻机构的文章。
- **段落方式。**用于 Web 页面和非标记文档。抽取过程按语义划分文档，从而利用诸如内部标记和语法之类的特征。如果选择此方式，那么将逐个段落应用评分。因此，例如，仅在同一段落中找到 apple 和 orange 的情况下，规则 apple & orange 才成立。

注: 由于从 PDF 文档中抽取文本的方式，**段落方式**不适用于这些文档。这是因为抽取会抑制回车符标记。

段落方式设置。仅在指定文本字段表示**文档路径名**并将文本统一选项设置为**段落方式**时，此选项才可用。请指定要在任何抽取中使用的字符阈值。实际大小会四舍五入为最接近的时间段。要确保从文档集合的文本产生的单词关联具有代表性，请避免指定太小的抽取大小。

- **最小值。**指定要在任何抽取中使用的最小字符数。
- **最大值。**指定要在任何抽取中使用的最大字符数。

输入编码。仅在指示文本字段表示**文档路径名**的情况下，此选项才可用。它指定缺省文本编码。对于除日语以外的所有语言，都将从指定编码或已识别编码转换为 ISO-8859-1。因此，即使指定其他编码，抽取引擎也会在对其进行处理之前将其转换为 ISO-8859-1。不符合 ISO-8859-1 编码定义的任何字符都将转换为空格。对于日语文本，可以选择以下若干编码选项之一: SHIFT_JIS、EUC_JP、UTF-8 或 ISO-2022-JP。

分区方式。使用分区方式来选择是否给予类型节点设置进行分区或者选择其他分区。分区会将数据分割为培训和测试样本。

"字段的文档设置"选项卡

结构化文本格式化

如果您因为具有结构化数据或要施加有关如何处理文本的规则而要跳过全部或部分抽取过程，请使用**结构化文本**文档类型选项，并且对包含"文档设置"对话框的**结构化文本格式化**部分中的文本的字段或标记进行声明。抽取的术语仅派生自已声明的字段或标记（和子标记）内包含的文本。将忽略任何未声明的字段或标记。

在特定上下文中，无需语言处理，并且语言抽取引擎可以替换为显式声明。在关键字字段以诸如分号 (;) 或逗号 (,) 之类的分隔符分隔的书目文件中，抽取两个分隔符之间的字符串便已足够。为此，可以暂挂完整抽取过程并改为定义特殊处理规则，以声明术语分隔符，向已抽取的文本分配类型，或者为抽取施加最小频率计数。

请在声明结构化文本元素时使用以下规则:

- 每行只能声明一个字段、标记或元素。它们不必存在于数据中。

- 声明区分大小写。
- 如果声明标记具有属性（如 <title id="1234">）并要包含所有变体（在本例中为所有标识），请添加不带属性或右尖括号 (>) 的标记，例如 <title
- 在字段或标记名称后添加冒号以指示这是结构化文本。直接在字段或标记后但在任何分隔符、类型或频率值之前添加此冒号，例如 author: 或 <place>:。
- 要指示字段或标记中包含多个术语，并且分隔符用于指定个别术语，请声明冒号后的分隔符，例如 author:, 或 <section>:;。
- 要向在标记中找到的内容分配类型，请声明冒号和分隔符后的类型名称，例如 author:,Person 或 <place>:;Location。请按照名称在资源编辑器中的显示方式来使用其对类型进行声明。
- 要为字段或标记定义最小频率计数，请声明行尾的数字，例如 author:,Person1 或 <place>:;Location5。其中 n 是已定义的频率计数，在字段或标记中找到的术语必须在要抽取的文档或记录的整个集合中出现至少 n 次。这也要求定义分隔符。
- 如果您具有包含冒号的标记，那么必须为冒号前置反斜杠字符，以便不会忽略声明。例如，如果具有名为 <topic:source> 的字段，请将其输入为 <topic\:source>。

为说明语法，假设具有以下重现书目字段:

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

针对此示例，如果要使抽取过程重点关注作者和摘要但忽略其余内容，那么将仅声明以下字段:

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

在此示例中，author:,Person1 字段声明表明已暂挂对字段内容的语言处理。相反，它表明作者字段包含多个以逗号分隔符相互分隔的名称，并且这些名称应分配到 Person 类型，如果名称在文档或记录的整个集合中出现至少一次，那么应对其进行抽取。由于所列的字段 abstract: 不带任何其他声明，因此在抽取期间将扫描该字段，并将应用标准语言处理和输入。

XML 文本格式化

如果要将抽取过程仅限于特定 XML 标记内的文本，请使用 **XML 文本**文档类型选项，并且对包含“文档设置”对话框的 **XML 文本格式化**部分中的文本的标记进行声明。抽取的术语仅派生自这些标记或其子标记内包含的文本。

注意！如果要跳过抽取过程并对术语分隔符施加规则，向已抽取的文本分配类型，或者为已抽取的文本施加频率计数，请使用以下描述的**结构化文本**选项。

请在声明 XML 文本格式化的标记时使用以下规则:

- 每行只能声明一个 XML 标记。
- 标记元素区分大小写。
- 如果声明标记具有属性（如 <title id="1234">）并要包含所有变体（在本例中为所有标识），请添加不带属性或右尖括号 (>) 的标记，例如 <title

为说明语法，假设具有以下 XML 文档:

```
<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>
```

针对此示例，将声明以下标记:

```
<section>
<title
```

在此示例中，由于已声明标记 <section>，因此在抽取过程中会扫描此标记及其嵌套标记 Traffic Signals 和 Road signs are helpful 中的文本。但是，由于既未显式声明标记 <p>，也未显式声明嵌套在已声明标记内的标记，因此会忽略 Learning the rules is important。

“文本挖掘”节点: “模型”选项卡

“模型”选项卡用于为节点输出指定构建方法和生成模型设置。

您可以设置以下参数:

模型名称 您可以基于目标或标识字段（或在未指定此类字段的情况下为模型类型）自动生成模型名称，或指定定制名称。

使用分区数据。 如果定义了分区字段，那么此选项可确保仅使用来自培训分区的数据构建模型。

构建方式。指定执行具有此“文本挖掘”节点的流时将如何生成模型块。或者，您可以使用**以交互方式构建**方式来使用更加实践性和探索性的方法，通过这种方式，您可以抽取概念、创建类别和优化您的语言资源，并且可以执行文本链接分析和探索集群。

- **以交互方式构建。**执行流时，该选项会启动一个交互式界面，您可以在其中抽取概念和模式、探索和调整抽取的结果、构建和优化类别、调整语言资源（模板、同义词、类型、库等）以及构建类别模型块。请参阅第 20 页的『以交互方式构建』主题以获取更多信息。
- **直接生成。**该选项指示执行流时，应自动创建模型并将其添加到“模型”选用板中。不同于交互式工作台，在执行时除在节点中定义的设置之外，无需任何其他操作。如果您选中此选项，那么会显示特定于模型的选项，您可以通过这些选项来定义要生成的模型类型。请参阅第 21 页的『直接生成』主题以获取更多信息。

复制以下来源的资源。挖掘文本时，抽取不仅基于“专家”选项卡中的设置，也基于语言资源。这些资源充当抽取期间文本处理方式的基础，从而获取概念、类型和（有时）模式。您可以将资源从资源模板或文本分析包复制到此节点中。选择一项资源，然后单击**装入**以定义将从中抽取资源的包或模板。装入时，会将资源的副本存储在节点中。因此，如果您想要使用经过更新的模板或 TAP，必须在此处或者在交互式工作台会话中重新装入此资源。为方便起见，在节点中会显示复制并装入资源的日期和时间。请参阅第 21 页的『从模板和 TAP 复制资源』主题以获取更多信息。

文本语言。标识进行挖掘的文本的语言。节点中复制的资源控制所呈现的语言选项。可以选择为其调整了资源的语言，或者选择 **ALL** 选项。强烈建议指定文本数据的确切语言；但如果不确定，那么可以选择 **ALL** 选项。**ALL** 不适用于日语文本。此 **ALL** 选项会延长执行时间，因为使用了自动语言识别来扫描所有文档和记录，以便首先标识文本语言。通过此选项，采用受支持且许可的语言的所有记录或文档都由抽取引擎使用相应语言的内部字典进行读取。请参阅主题第 182 页的『语言标识』以获取更多信息。如果您有兴趣为当前不具有访问权的受支持语言购买许可证，请与您的销售代表联系。

以交互方式构建

在文本挖掘建模节点的“模型”选项卡中，可以为模型块选择构建方式。如果选择**以交互方式构建**，那么执行流时会打开一个交互式界面。在此交互式工作台中，您可以:

- 抽取和浏览抽取结果（包括概念和类型）以发现文本数据中的重要构想。
- 使用各种方法来构建和扩展来自概念、类型、TLA 模式和规则的类别，以便您可以将自己的文档和记录存储到这些类别中。
- 优化您的语言资源（资源模板、库、字典、同义词等）以便您可以通过抽取、检验和优化概念的迭代过程来改进自己的结果。
- 执行文本链接分析 (TLA) 并使用发现的 TLA 模式来构建更好的类别模型块。“文本链接分析”节点不提供相同的探索性选项或建模功能。
- 在“可视化”窗格中生成集群以发现新关系并探索概念、类型、模式和类别之间的关系。
- 在 IBM SPSS Modeler 的“模型”选用板中生成经优化的类别模型块并将其用于其他流。

注: 如果要创建 IBM SPSS Collaboration and Deployment Services 作业，那么不能构建交互模型。

使用来自上次节点更新的会话工作（类别、TLA、资源等）。在交互式工作台会话中工作时，可以使用会话数据（抽取参数、资源、类别定义等）更新节点. **使用会话工作**选项允许您使用已保存的会话数据重新启动交互式工作台。首次使用此节点时禁用该选项，因为未曾保存任何会话数据。要了解如何使用会话数据来更新节点以便使用该选项，请参阅第 70 页的『更新建模节点并保存』。

如果*使用*该选项启动会话，那么在您下次启动会话时，抽取设置、类别、资源和上次您从交互式工作台会话执行节点更新时的任何其他工作均可供使用。由于通过该选项来使用保存的会话数据，因此会禁用并忽略某些内容（例如，从以下模板的复制的资源）和其他选项卡。但是如果您*不使用*此选项启动会话，那么仅使用当前定义的节点内容，表示您先前在工作台中执行的任何工作都将不可用。

注: 如果您在使用**使用会话工作...** 选项缓存抽取结果之后更改源节点，那么如果要获得更新的抽取结果，需要在启动交互式工作台会话时运行新的抽取。

跳过抽取并复用已缓存的数据和结果。您可以在交互式工作台会话中复用任何已缓存的抽取结果和数据。当您想要节省时间并复用抽取结果而不是等待启动会话时执行全新的抽取时，该选项特别有用。为使用该选项，您必须先前已从交互式工作台会话内更新了此节点，并且已选中**在抽取结果中保留会话工作和缓存文本数据以供复用**选项。要了解如何使用会话数据来更新节点以便使用该选项，请参阅第 70 页的『更新建模节点并保存』。

会话开始方式。选择指示要在启动交互式工作台会话时首先启动的视图和操作的选项。无论您从任何视图中开始，进入会话后都可以切换至任何视图。

- **使用抽取结果来构建类别。**该选项会在“类别和概念”视图中启动交互式工作台并（如果适用）执行抽取。在此视图中，您可以创建类别和生成类别模型。您还可以切换至其他视图。请参阅第 61 页的第 8 章，『交互式工作台模式』主题以获取更多信息。
- **探索文本链接分析 (TLA) 结果。**该选项会通过抽取并识别文本内概念之间的关系（例如，意见或“文本链接分析”视图中的其他链接）来启动并开始。您必须选择包含 TLA 模式规则的模板或文本分析包才能使用该选项并获取结果。如果您处理较大的数据集，那么 TLA 抽取可能需要一些时间。在此情况下，您可能需要考虑使用“样本”节点上游。请参阅第 127 页的第 12 章，『探索文本链接分析』主题以获取更多信息。
- **分析共现字集群。**该选项在“集群”视图中启动，并更新任何过时的抽取结果。在此视图中，您可以执行共现字集群分析，以生成一组集群。共现字集群是一个过程，首先从基于两个概念在给定记录或文档中的共现来评估这两个概念之间的链接值的强度开始，最后将链接强度较高的概念分组到集群内。请参阅第 61 页的第 8 章，『交互式工作台模式』主题以获取更多信息。

直接生成

在文本挖掘建模节点的“模型”选项卡中，可以为模型块选择构建方式。如果选择**直接生成**，那么可以在节点中设置选项，然后直接执行流即可。输出为概念模型块，此模型块直接放置在“模型”选用板中。不同于交互式工作台，在执行时除在节点中针对此选项定义频率设置之外，无需任何其他操作。

要在模型中包含的概念的最大数量。该选项指示您想要创建概念模型，仅当您自动构建模型（非迭代）时才适用。它还声明此模型包含的概念数量不能超过指定数量。

- **按最高频率选中概念。概念的最大数量。**这是从频率最高的概念开始将选中的概念数量。此处频率表示在文档/记录的整个集合中概念（及其所有底层术语）出现的次数。此数字应高于记录计数，因为在每条记录中每个概念可以多次显示。
- **取消选中在过多记录中出现的概念。记录的百分比。**取消选中记录计数百分比高于您指定的数量的概念。该选项适用于排除文本或每条记录中频繁发生但在分析中不重要的概念。

针对评分速度进行优化。缺省情况下选中该选项，该选项可确保创建的模型体积小并且可快速评分。取消选中该选项会造成大量更大的模型，这些模型评分速度缓慢。但是，更大的模型可确保在生成的概念模型中初始显示的评分与对含模型块的相同文本进行评分时所获得的评分相同。

从模板和 TAP 复制资源

挖掘文本时，抽取不仅基于“专家”选项卡中的设置，也基于语言资源。这些资源充当抽取期间文本处理方式的基础，从而获取概念、类型和（有时）模式。您可以将资源从*资源模板*复制到此节点中，如果您位于“文本挖掘”节点中，那么还可以选择*文本分析包* (TAP)。

缺省情况下，将节点添加到画布中时，会将资源从产品许可语言的基本模板复制到该节点中。如果您拥有多种语言的许可证，那么选中的第一种语言将用于确定要自动装入的模板。

装入时，会将选中的资源的副本存储在节点中。仅复制模板或 TAP 的内容，模板或 TAP 本身不链接至节点。这意味着如果稍后更新次模板或 TAP，那么在节点中这些更新不会自动可用。简而言之，除非您重新装如模板或 TAP 副本，或者除非您更新“文本挖掘”节点并选中**使用会话工作**选项，否则始终使用装入节点的资源。有关**使用会话工作**的更多信息，请参阅本主题中的其他内容。

选择模板或 TAP 时，选择采用与文本数据相同语言的模板或 TAP。您只能使用采用已获得许可的语言的模板或 TAP。如果要执行文本链接分析，必须选择包含 TLA 模式的模板。如果模板包含 TLA 模式，那么会在“装入资源模板”对话框的 TLA 列中显示一个图标。

注：您不能将 TAP 装入“文本链接分析”节点。

资源模板

资源模板是预定义的库和高级语言与非语言资源的集合，这些资源针对特定域或用途经过了微调。在文本挖掘建模节点中，将节点添加到流中时在节点中已装入来自基本模板的资源的副本，但是您可以通过选择**资源模板**或**文本分析包**，然后单击**装入**来更改模板或装入文本分析包。对于模板，您随后可以在“装入资源模板”对话框中选择模板。

注：如果在列表中未显示您所需的模板，但是您的机器上存在已导出的副本，那么现在您可以将其导入。您还可以从该对话框中导出模板以与其他用户共享。请参阅第 149 页的『导入和导出模板』主题以获取更多信息。

文本分析包 (TAP)

文本分析包 (TAP) 是预定义的库和高级语言与非语言资源的集合，这些资源与一个或多个预定义的类别集合绑定。IBM SPSS Modeler Text Analytics 提供了多个英语语言文本和日语语言文本的预构建 TAP，这些 TAP 都

针对特定域经过微调。您不能编辑这些 TAP，但是可以将其用于快速开始类别模型构建。您还可以在交互式会话中创建自己的 TAP。请参阅第 117 页的『装入文本分析包』主题以获取更多信息。

注: 您不能将 TAP 装入“文本链接分析”节点。

使用“使用会话工作”选项（“模型”选项卡）

在“模型”选项卡中将资源复制到节点中时，还可以需要稍后在交互式会话中队资源进行更改，并使用这些最新更改来更新文本挖掘建模节点。在此情况下，您可以在文本挖掘建模节点的“模型”选项卡中选择**使用会话工作**选项。

如果选中**使用会话工作**，那么在节点中会禁用**装入**按钮以指示将使用来自交互式工作台的资源代替先前在此处装入的资源。

要在选中**使用会话工作**选项之后对资源进行更改，可以通过资源编辑器视图在交互式工作台中直接编辑或切换您的资源。请参阅第 148 页的『装入后更新节点资源』主题以获取更多信息。

“文本挖掘”节点: “专家”选项卡

“专家”选项卡包含影响文本抽取和处理方式的高级参数。此对话框中的参数可控制抽取过程的基本行为以及一些高级行为。但是，这些参数仅代表可供您使用的部分选项。还有多种语言资源和选项可影响抽取结果，这些资源和选项由您在“模型”选项卡上所选的资源模板来控制。请参阅第 19 页的『“文本挖掘”节点: “模型”选项卡』主题以获取更多信息。

注: 如果您使用“模型”选项卡上保存的交互式工作台信息选择了**以交互方式进行构建**方式，在此情况下采用上次保存的工作台会话中的抽取设置，因此会禁用这整个选项卡。

针对荷兰语、英语、法语、德语、意大利语、葡萄牙语和西班牙语文本

在抽取除“日语”以外的其他语言（例如，英语、西班牙语、法语、德语等）时，可以设置以下参数:

注: 请参阅本主题以获取有关日语文本的“专家”设置的信息。

将抽取限于全局频率至少为 [n] 的概念。指定单词或短语为进行抽取而必须在文本中出现的最小次数。通过此方式，值为 5 会将抽取限于在记录或文档的整个集合中出现至少五次的单词或短语。

在某些情况下，更改此限制会在产生的抽取结果中造成巨大差异，从而影响类别。假设您处理的是餐厅数据且不会为此选项将限制提高至 1 以上。在此情况下，可能会在抽取结果中找到 *pizza (1)*、*thin pizza (2)*、*spinach pizza (2)* 和 *favorite pizza (2)*。但是，如果要将抽取限于全局频率为 5 或以上并重新抽取，那么将不会再获取其中三个概念。您将改为获取 *pizza (7)*，因为 *pizza* 是最简单的形式，此外该单词已作为可能的候选值存在。根据其余文本，实际频率可能为 7 以上，具体视文本中是否仍有其他包含 pizza 的短语而定。此外，如果 *spinach pizza* 已是类别描述符，那么可能需要添加 *pizza* 作为描述符，而不是捕获所有记录。为此，只要已创建类别，就请谨慎更改此限制。

请注意，这是仅抽取功能；如果模板包含术语（通常如此），并且在文本中找到该模板的术语，那么无论该术语的频率如何，都将对其编制索引。

例如，假设使用在 Core 库中的 `<Location>` 类型下包含“los angeles”的基本资源模板；如果文档仅包含一次 Los Angeles，那么 Los Angeles 将是概念列表的一部分。要防止此情况，将需要设置过滤器，以显示至少出现与**将抽取限于全局频率至少为 [n] 的概念**字段中输入的值相同的次数的概念。

调整标点错误。 此选项可在抽取期间临时标准化包含标点错误的文本（例如，不正确使用），以改善概念的可抽取性。当文本很短且质量不佳（例如，在开放式调查响应、电子邮件和 CRM 数据中）时，或文本包含很多缩略词时，此选项非常有用。

调整拼写错误，最小根字符限制为 [n]。 此选项适用于模糊分组方法，此方法可帮助将普遍拼写有误的单词或拼写接近的单词分组到一个概念下。模糊分组算法临时删除抽取单词中的所有元音（除了第一个元音）和出现的二重/三重辅音，然后进行比较，以查看它们是否相同，以便 modeling 和 modelling 分组到一起。但是，如果每个术语分配给不同类型（除了 <Unknown> 类型），那么不会应用模糊分组方法。

您还可先优化最少数目的所需*根*字符，再使用模糊分组。术语中根字符数通过对所有字符相加减去形成屈折变化后缀的任何字符数以及（使用复合单词术语的情况下）限定词和介词数计算得出。例如，术语 exercises 将计算为 8 个根字符（形式为“exercise”），因为字母单词末尾的 *s* 是屈折变化形式（复数形式）。相似地，apple sauce 将计算为 10 个根字符（“apple sauce”，manufacturing of cars 将计算为 16 个根字符（“manufacturing car”）此计数方法仅用于检查是否应该应用模糊分组，但不会影响匹配单词的方式。

注：如果发现某些单词之后分组不正确，那么可通过在“高级资源”选项卡中的**模糊分组：例外**中显式进行声明来从此方法排除单词对。请参阅主题第 175 页的『模糊分组』，以获取更多信息。

抽取单术语。 此选项用于抽取单个单词（单术语），前提是此单词不属于复合单词的一部分，且其为名词或语音的不可识别部分。

抽取非语言实体。 此选项用于抽取非语言实体，例如，电话号码、社保号、时间、日期、货币、数字、百分比、电子邮件地址和 HTTP 地址。您可以在“高级资源”选项卡中的**非语言实体：配置**部分中包含或排除某些类型的非语言实体。通过禁用任何不需要的实体，抽取引擎不会浪费处理时间。请参阅主题第 179 页的『配置』，以获取更多信息。

大写算法。 此选项用于抽取内置字典中不存在的简单和复合术语，前提是术语的第一个字母为大写。此选项提供了一种很好的方式来抽取大部分正确的名词。

尽可能将部分和完整人员姓名分组在一起。 此选项用于将在文本中显示不同的姓名分组在一起。由于通常在文本开头部分通过全名指代姓名，而之后通过较短的版本指代姓名，因此，此功能会很有帮助。此选项尝试将类型为 <Unknown> 的任何单术语与类型为 <Person> 的任何复合术语的最后一个单词匹配。例如，如果发现了 *doe* 且其最初类型为 <Unknown>，那么抽取引擎会检查以了解 <Person> 类型中的任何复合术语是否将 *doe* 作为最后一个单词包含，例如，*john doe*。此选项不适用于名字，因为大多数名字永不会抽取为单术语。

最大非功能单词排列。 此选项指定应用排列方法时可显示的非功能单词的最大数目。此排列方法将仅包含的非功能单词（例如，of 和 the）不同（不考虑屈折变化）的相似短语分组在一起。例如，假设将此值设置为最多两个单词，且抽取了 company officials 和 officials of the company。在此情况下，这两个抽取的术语将在最终概念列表中分组在一起，因为在忽略 of the 时，这两个术语视为相同。

注：要启用抽取“文本链接分析”结果，必须使用**探索文本链接分析结果**选项开始会话，并选择包含 TLA 定义的资源。您始终可以稍后在交互式工作台会话期间通过“抽取设置”对话框来抽取 TLA 结果。请参阅第 74 页的『抽取数据』主题以获取更多信息。

对于日语文本

此对话框针对“日语”文本包含不同的选项，因为抽取过程存在差异。要处理日语文本，必须在此节点的“模型”选项卡中选择针对日语经过微调的模板或文本分析包。请参阅第 21 页的『从模板和 TAP 复制资源』主题以获取更多信息。

辅助分析。 启动抽取时，将使用一组缺省类型执行基本关键字抽取。 但是，选择辅助分析器时，由于抽取器现在将小品词和助动词作为概念的一部分包含，可获取更多或更丰富的概念。 如果进行观点分析，那么还会包含大量其他类型。此外，选择辅助分析器，还可生成文本链接分析结果。

注: 调用辅助分析器时，需要花费更长时间来完成抽取过程。

- **依赖关系分析**。选择此选项可从基本类型和关键字抽取获取抽取概念的扩展小品词。还可从依赖关系文本链接分析 (TLA) 获取更丰富的模式结果。
- **观点分析**。选择此分析器可获取其他抽取的概念，适用时，会执行 TLA 模式的抽取。除了基本类型，还可从超过 80 种观点类型受益。这些类型用于通过表情、观点和意见说明文本中的概念和模式。具有三个选项，这些选项指示观点分析的焦点: **所有观点**、**仅表示观点**和**仅结论**。
- **无辅助分析器**。该选项会关闭所有辅助分析器。如果在“模型”选项卡上选中了**探索文本链接分析 (TLA) 结果**，那么会隐藏该选项，因为辅助分析器是获取 TLA 结果所必需的。如果您选中该选项，但是稍后选择**探索文本链接分析 (TLA) 结果**，那么在流执行期间会出现错误。

对上游采样以节省时间

当您具有大量数据时，处理时间可能需要几分钟到几小时，尤其是使用交互式工作台会话时。数据大小越大，抽取和分类过程耗时越长。为更有效地进行工作，您可以从自己的“文本挖掘”节点添加一个 IBM SPSS Modeler 样本节点上游。使用此“样本节点”通过使用少量文档或记录子集来执行前几个阶段，以进行随机采样。

少量样本通常完全足以决定如何编辑您的资源，甚至可以创建大部分（甚至全部）类别。在较小的数据集上运行并且对结果满意之后，可以将同样的技术应用于对整个数据集创建类别。然后，您可以寻找不适合您创建的类别的文档或记录，并根据需要进行调整。

注: “样本节点”是一个标准 IBM SPSS Modeler 节点。

在流中使用文本挖掘节点

“文本挖掘”建模节点用于访问数据和抽取流中的概念。您可以使用任何源节点来访问数据，例如数据库节点、变量文件节点、Web 订阅源节点或固定文件节点。对于驻留在外部文档内的文本，可使用“文件列表”节点。

示例 1: 文件列表节点和文本挖掘节点，用于直接构建概念模型块

以下示例显示了如何使用“文件列表”节点和“文本挖掘”建模节点来生成概念模型块。有关使用“文件列表”节点的更多信息，请参阅第 9 页的『“文件列表”节点』。

1. **“文件列表”节点（“设置”选项卡）。**首先，我们将此节点添加到流中以指定文本文件的存储位置。我们选择了包含要对其执行文本挖掘的所有文档的目录。
2. **“文本挖掘”节点（“字段”选项卡）。**接下来，我们将一个“文本挖掘”节点添加并连接到“文件列表”节点。在此节点中，我们定义了输入格式、资源模板和输出格式。我们选择了从“文件列表”节点生成的字段名称，并选择了其中文本字段表示**到文档的路径**的选项以及其他设置。请参阅『在流中使用文本挖掘节点』主题以获取更多信息。
3. **“文本挖掘”节点（“模型”选项卡）。**接下来，在“模型”选项卡上，我们选择了构建方式以从该节点直接生成概念模型块。您可以选择其他资源模板或者保留基本资源。

示例 2: “Excel 文件”节点和“文本挖掘”节点，用于以交互方式构建类别模型

此示例显示了“文本挖掘”节点如何能够另外启动交互式工作台会话。有关交互式工作台的更多信息，请参阅第 61 页的第 8 章，『交互式工作台模式』。

1. **“Excel 源”节点（“数据”选项卡）。**首先，我们将此节点添加到流中以指定文本存储位置。

2. **“文本挖掘”节点（“字段”选项卡）。**接下来，我们添加并连接了一个“文本挖掘”节点。在此第一个选项卡上，我们定义了输入格式。我们从源节点中选择了字段名称，并选择了表示**实际文本**的文本字段选项，因为数据直接来自“Excel 源节点”。
3. **“文本挖掘”节点（“模型”选项卡）。**接下来，在“模型”选项卡上，我们选择了以交互方式构建类别模型块，并使用抽取结果来自动构建类别。在此示例中，我们从文本分析包装入了资源副本和一组类别。
4. **交互式工作台会话。**接下来，我们执行了流并打开了交互式工作台界面。执行抽取之后，我们开始探索数据并改进类别。

文本挖掘块：概念模型

“文本挖掘”概念模型块是在您成功执行“文本挖掘”模型节点时创建的，在执行期间您选中了“模型”选项卡的**直接生成模型**。文本挖掘概念模型块用于实时发现其他文本数据中的关键概念，例如，来自呼叫中心的便笺。

概念模型块本身包含已分配到类型的概念列表。您可以选中此模型中的任意或全部概念以针对其他数据进行评分。执行包含“文本挖掘”模型块的流时，会在构件模型之前根据“文本挖掘”建模节点的“模型”选项卡上所选的构建方式来向数据添加新字段。请参阅『概念模型：“模型”选项卡』主题以获取更多信息。

如果使用已转换文档生成了模型块，那么将在已转换的语言中执行评分。同样，如果使用英语作为语言生成了模型块，那么可以在模型块中指定转换语言，因为文档之后将转换为英语。

文本挖掘模型块在生成后会放置在模型块选用板中（位于 IBM SPSS Modeler 窗口右上方的“模型”选项卡上）。

查看结果

要查看有关模型块的信息，请右键单击模型块选用板中的节点，然后从上下文菜单中选择**浏览**（或针对流中的节点单击**编辑**）。

向流中添加模型

要向流中添加模型块，请单击模型块选用板中的图标，然后单击要将节点放置在的流画布。或者，右键单击图标，然后从上下文菜单中选择**添加到流**。然后，将流连接到节点，即可传递数据以生成预测。

注意：如果要使用评分块来重新生成包含所使用的类别模型和模板建模节点，我们建议您创建 TAP，并在生成评分块之前将其用于交互式会话代替建模节点。

概念模型：“模型”选项卡

在概念模型中，“模型”选项卡显示已抽取的概念集。这些概念以表格式显示，每个概念一行。该选项卡上的目标是选择将用于评分的概念。

注：如果改为生成了类别模型块，那么该选项卡将包含不同信息。请参阅第 33 页的『类别模型块：“模型”选项卡』主题以获取更多信息。

缺省情况下，选中所有概念以进行评分，如最左侧列中的复选框所示。复选框表示此概念将用于评分。未选中的复选框表示将从评分中排除此概念。您可以通过选中行并单击所选的其中一个复选框来选中多个行。

要了解有关每个概念的更多信息，可以查看以下每个列中提供的额外信息:

概念。这是已抽取的引导词或短语。在某些情况下，此概念表示概念名称以及与此概念关联的某些其他底层术语。要查看哪些底层术语属于某个概念，请在此选项卡中显示“底层术语”窗格，并选择概念以在对话框底部查看对应的术语。请参阅第 27 页的『概念模型中的底层术语』主题以获取更多信息。

全局。此处全局（频率）表示在文档/记录的整个集合中概念（及其所有底层术语）出现的次数。

- **条形图。**此概念的全局频率在文本数据中显示为条形图。此条形图采用概念分配到的类型的颜色，以直观区分各种类型。
- **%.** 此概念的全局频率在文本数据中显示为百分比。
- **N。**此概念在文本数据中出现的实际次数。

文档数。此处"文档数"表示文档计数，即其中包含概念（及其所有底层术语）的文档或记录的数量。

- **条形图。**此概念的文档计数显示为条形图。此条形图采用概念分配到的类型的颜色，以直观区分各种类型。
- **%.** 此概念的文档计数显示为百分比。
- **N。**包含此概念的文档或记录的实际数量。

类型。概念分配到的类型。对于每个概念，"全局"和"文档数"列显示为彩色以表示此概念分配到的类型。**类型**是概念的语义分组。请参阅第 161 页的『类型字典』主题以获取更多信息。

处理概念

通过右键单击表中的单元格，可以显示一个上下文菜单，您可以从中:

- **全选。**将选中表中的所有行。
- **复制。**将所选概念复制到剪贴板。
- **随字段复制。**将所选概念与列标题一起复制到剪贴板。
- **选中所选项。**选中表中所选行的所有复选框，从而在评分中包含这些概念。
- **取消选中所选项。**取消选中表中所选行的所有复选框。
- **全部选中。**选中表中的所有复选框。这将导致在最终输出中使用所有概念。
- **全部取消选中。**取消选中表中的所有复选框。取消选中概念表示在最终输出中将不使用此概念。
- **包含概念。**显示"包含概念"对话框。请参阅『包含概念用于评分的选项』主题以获取更多信息。

包含概念用于评分的选项

要快速选中或取消选中将用于评分的选项，请单击**包含概念**的工具栏按钮。

图 1. "包含概念"工具栏按钮

单击此工具栏按钮将打开"包含概念"对话框，以允许您基于规则选择概念。将包含在"模型"选项卡中具有复选标记的所有概念用于评分。在此子对话框中应用规则以更改将用于评分的概念。

您可以从以下选项中进行选择:

按最高频率选中概念。概念的最大数量。这是从全局频率最高的概念开始将选中的概念数量。此处频率表示在文档/记录的整个集合中概念（及其所有底层术语）出现的次数。此数字应高于记录计数，因为在每条记录中每个概念可以多次显示。

按文档计数选中概念。最小计数。这是要选中概念所需的最小文档计数。此处文档计数表示其中显示概念（及其所有底层术语）的文档/记录的数量。

选中分配给类型的概念。从下拉列表中选择类型以选中分配给此类型的所有概念。在抽取过程期间会将概念自动分配给类型。**类型**是概念的语义分组。类型包含较高级别的概念、肯定词、否定词、限定符、上下文限定符、名字、位置、组织等之类的内容。请参阅第 161 页的『类型字典』主题以获取更多信息。

取消选中在过多记录中出现的概念。记录的百分比。取消选中记录计数百分比高于您指定的数量的概念。该选项适用于排除文本或每条记录中频繁发生但在分析中不重要的概念。

取消选中分配给类型的概念。取消选中匹配您从下拉列表中所选类型的概念。

概念模型中的底层术语

您可以查看针对您在表中所选概念定义的底层术语。通过单击工具栏上的底层术语切换按钮，您可以在对话框底部的拆分窗格中显示底层术语表。

这些底层术语包括语言资源中定义的同义词（无论是否在文本中找到这些同义词都是如此）以及文本中用于生成模型块、变换的术语、来自模糊分组的属于等的任何已抽取的复数/单数格式。

图 2. “显示底层术语”工具栏按钮

注: 您不能编辑底层术语列表。此列表是通过语言资源中定义的替换、替换字典中的同义词定义、模糊分组等生成的。要对概念下的术语的分组方式或者其处理方式进行更改，必须直接在资源（在交互式工作台的资源编辑器中或者模板编辑器中可编辑，然后在节点中重新装入）中进行更改，然后重新执行流以获取具有经更新的结果的心模型块。

通过右键单击包含底层术语或概念的单元格，可以显示一个上下文菜单，您可以从中:

- **复制。**将所选单元格复制到剪贴板。
- **随字段复制。**将所选单元格随列标题一起复制到剪贴板。
- **全选。**将选中表中的所有单元格。

概念模型: “设置”选项卡

“设置”选项卡用于定义新输入数据的文本字段值（如果需要）。您也可以在其中为自己的输出定义数据模型（评分方式）。

注: 仅当将模型块放置在画布中时才会显示该选项卡。当您在“模型”选用板中访问此对话框时，该选项卡不存在。

评分方式: 概念作为记录

借助该评分方式，会针对每个概念/文档对创建一条新记录。通常，输出中的记录数量多于输入中的记录数量。

除输入字段之外，会向数据添加以下新字段。

表 4. “概念作为记录”的输出字段.

字段	描述
概念	包含文本数据字段中找到的已抽取的概念名称。
类型	将概念的类型存储为完整的类型名称，例如，*位置*或*人员*。**类型**是概念的语义分组。请参阅第 161 页的『类型字典』主题以获取更多信息。

表 4. "概念作为记录"的输出字段 (续).

字段	描述
计数	显示此概念（及其底层术语）在文本主体（记录/文档）中出现的次数。

选中该选项时，将禁用除**调整标点错误**以外的所有选项。

评分方式: 概念作为字段

在概念模型中，对于每条输入记录，会为给定文档中找到的每个概念创建一条新记录。从而使输出记录数量与输入中的记录数量相同。但是，每条记录（每个行）现在针对"模型"选项卡上使用复选标记选中的每个概念包含一个新字段（一个列）。每个概念字段的值取决于您在此选项卡上选择**标记**还是**计数**作为字段值。

注: 如果使用非常大的数据集，例如，使用 DB2 数据库，那么使用**概念作为字段**可能由于数据量而出现处理问题。在此情况下，我们建议改为使用**概念作为记录**。

字段值。选择每个概念的新字段是否将包含计数或标记值。

- **标记。**该选项用于包含输出中具有两个不同值的标记，例如，*Yes*/*No*、*True*/*False*、*T*/*F* 或 *1* 和 *2*。将自动设置存储类型以反映所选的值。例如，如果针对标记输入数字值，那么将自动将这些数字值作为整数值来处理。标记的存储类型可以是字符串、整数、实数或日期/时间。输入 **True** 和 **False** 的标记值。
- **计数。**用于获取在给定记录中概念发生的次数的计数。

字段名称扩展。指定字段名称的扩展。字段名称是使用概念名称加上此扩展来生成的。

- **添加为。**指定在字段名称中添加扩展的位置。选择**前缀**以将扩展添加到字符串开头。选择**后缀**以将扩展添加到字符串末尾。

调整标点错误。 此选项可在抽取期间临时标准化包含标点错误的文本（例如，不正确使用），以改善概念的可抽取性。当文本很短且质量不佳（例如，在开放式调查响应、电子邮件和 CRM 数据中）时，或文本包含很多缩略词时，此选项非常有用。

注: 处理日语文本时，**调整标点错误**选项不适用。

概念模型: "字段"选项卡

"字段"选项卡用于定义新输入数据的文本字段值（如果需要）。

注: 仅当将模型块放置在流中时才会显示该选项卡。当您在"模型"选用板中访问此输出时，该选项卡不存在。

文本字段。选择包含要挖掘的文本的字段、文档路径名或文档的目录路径名。此字段取决于数据源。

文本字段表示。指示在先前设置中指定的文本字段包含的内容。选项包括:

- **实际文本。**如果字段包含应从中抽取概念的确切文本，请选择此选项。
- **文档路径名。**如果字段包含文本文档驻留所在的位置的一个或多个路径名，请选择此选项。

文档类型。仅在指定文本字段表示**文档路径名**的情况下，此选项才可用。文档类型指定文本的结构。请选择以下类型之一:

- **全文本。**用于大多数文档或文本源。系统会扫描整个文本集以进行抽取。与其他选项不同，此选项没有其他设置。

- **结构化文本。**用于书目形式、专利以及任何包含可识别并分析的常规结构的文件。此文档类型用于跳过全部或部分抽取过程。通过它可定义术语分隔符，分配类型和施加最小频率值。如果选择此选项，那么必须单击**设置**按钮并在“文档设置”对话框的**结构化文本格式化**区域中输入文本分隔符。请参阅主题第 17 页的『“字段的文档设置”选项卡』以获取更多信息。
- **XML 文本。**用于指定包含要抽取的文本的 XML 标记。系统会忽略所有其他标记。如果选择此选项，那么必须单击**设置**按钮并在“文档设置”对话框的 **XML 文本格式化**区域中指定包含抽取过程中要读取的文本的 XML 元素。请参阅主题第 17 页的『“字段的文档设置”选项卡』以获取更多信息。

输入编码。仅在指示文本字段表示**文档路径名**的情况下，此选项才可用。它指定缺省文本编码。对于除日语以外的所有语言，都将从指定编码或已识别编码转换为 ISO-8859-1。因此，即使指定其他编码，抽取引擎也会在对其进行处理之前将其转换为 ISO-8859-1。不符合 ISO-8859-1 编码定义的任何字符都将转换为空格。对于日语文本，可以选择以下若干编码选项之一：SHIFT_JIS、EUC_JP、UTF-8 或 ISO-2022-JP。

文本语言。指示挖掘的文本的语言；这是抽取期间检测到的主要语言。如果您有兴趣购买当前无权访问的受支持的语言的许可证，请与销售代表联系。

概念模型: “摘要”选项卡

“摘要”选项卡提供有关模型本身（*分析*文件夹）、模型中使用的字段（*字段*文件夹）、构建模型时使用的设置（*构建设置*文件夹）和模型培训（*训练摘要*文件夹）的信息。

首次浏览建模节点时，“摘要”选项卡上的文件夹处于折叠状态。要查看感兴趣的结果，可使用项目左侧的展开控件展开项目，或单击**全部展开**按钮显示所有结果。要在查看结果后将其隐藏，请使用展开控件折叠要隐藏的特定文件夹，或者单击**全部折叠**按钮来折叠所有结果。

在流中使用概念模型块

使用“文本挖掘”建模节点时，您可以生成概念模型块或类别模型块（通过交互式工作台会话）。以下示例显示了如何在简单的流中使用概念模型。

示例: 含概念模型块的“Statistics 文件”节点

以下示例显示了如何使用“文本挖掘”概念模型块。

图 3. 示例流: 含“文本挖掘”概念模型块的“Statistics 文件”节点。

1. **“Statistics 文件”节点（“数据”选项卡）。**首先，我们将此节点添加到流中以指定文本文件的存储位置。

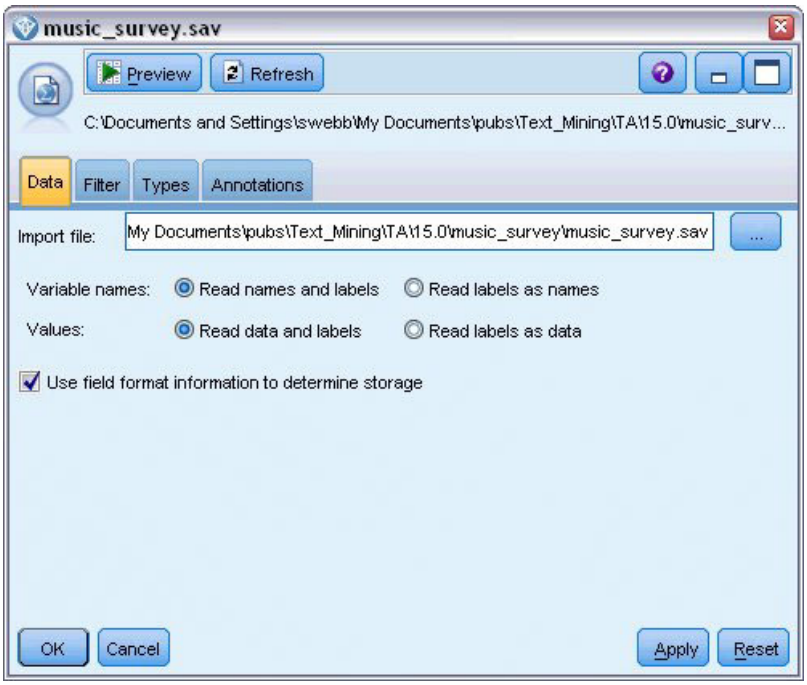


图 4. "Statistics 文件"节点对话框: "数据"选项卡

2. **"文本挖掘"概念模型块（"模型"选项卡）**。接下来，我们将一个概念模型块添加并连接到"Statistics 文件"节点。选中要用于对数据进行评分的概念。

图 5. "文本挖掘"模型块对话框："模型"选项卡

3. **"文本挖掘"概念模型块（"设置"选项卡）。**然后，定义输出格式并选择*概念作为字段*。在"模型"选项卡中将为选中的每个概念创建一个新字段。每个字段名称将由概念名称和前缀"*Concept_*"组成

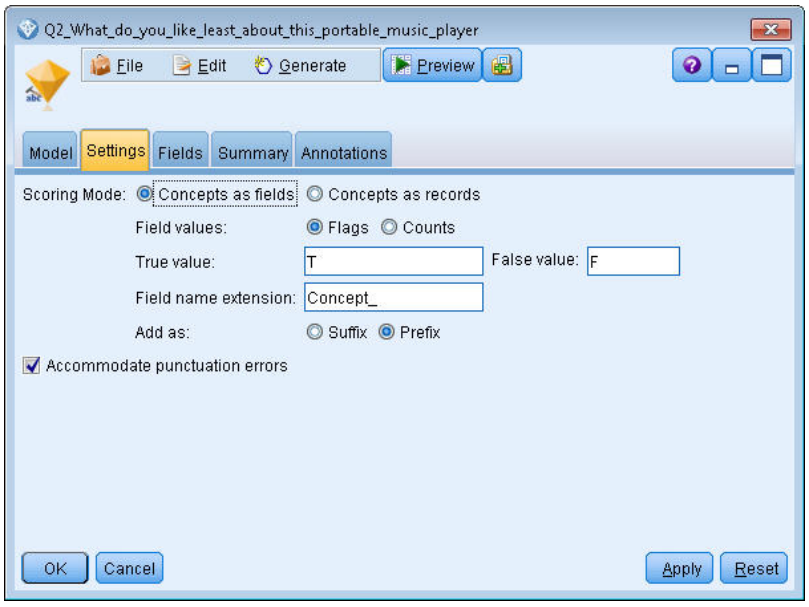


图 6. "文本挖掘"概念模型块对话框："设置"选项卡

4. **“文本挖掘”概念模型块（“字段”选项卡）。**然后，选择文本字段 **Q2_What_do_you_like_least_about_this_portable_music_player**，即来自“Statistics 文件”节点的字段名称。我们还选择了选项**文本字段表示：实际文本**。

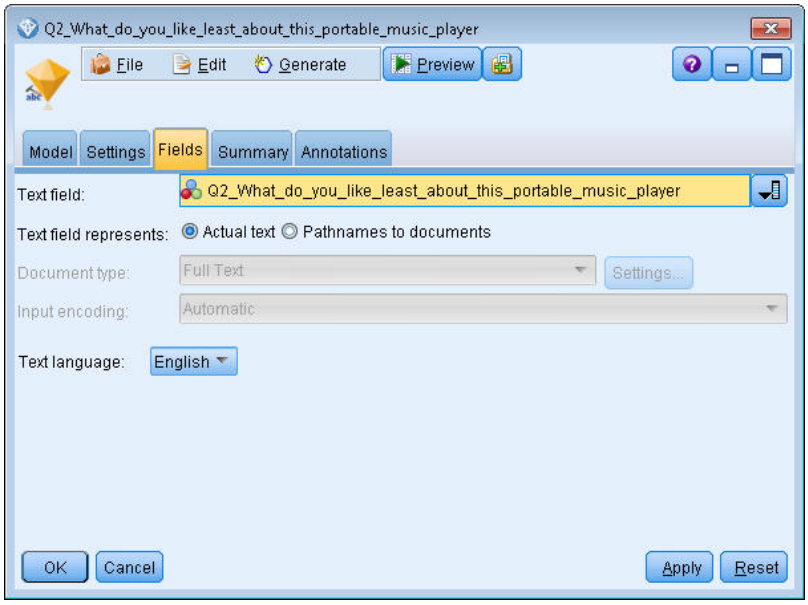
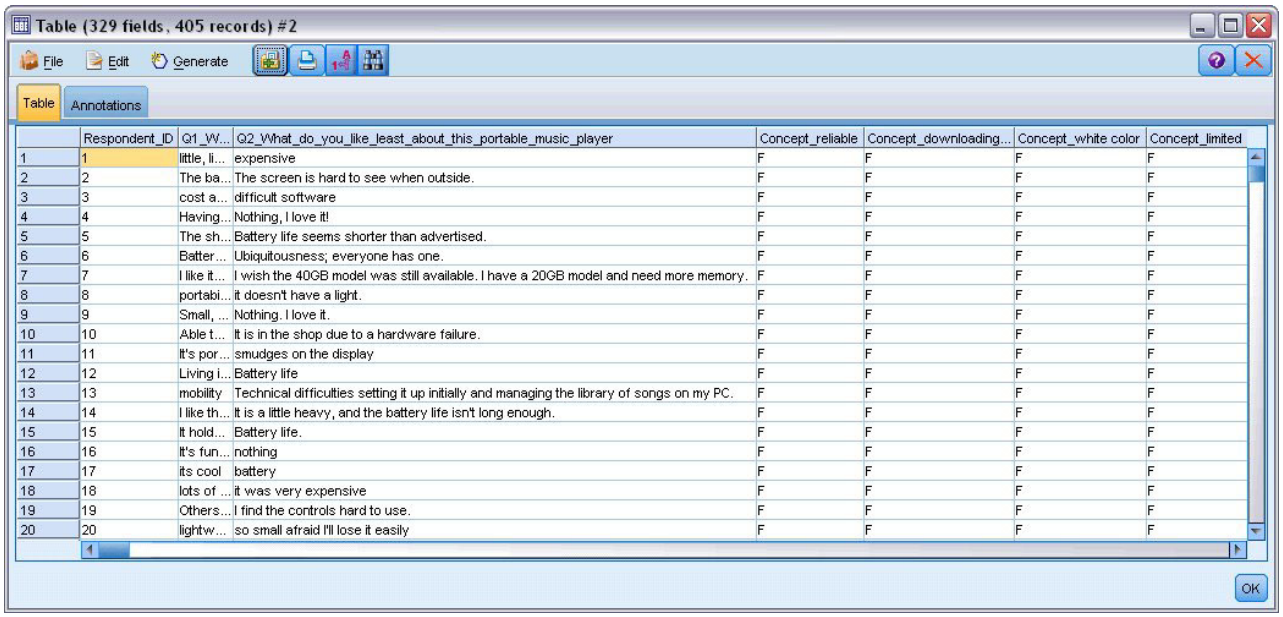


图 7. “文本挖掘”概念模型块对话框：“字段”选项卡

5. **表节点。**然后，附加一个表节点以查看结果并执行流。这样会在屏幕上打开表输出。



	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

图 8. 滚动表输出以显示概念标记

文本挖掘块: 类别模型

从交互式工作台中生成类别模型时就会创建“文本挖掘”类别模型。此建模块包含一组类别，其定义由概念、类型、TLA 模式和/或类别规则组成。此块用于对调查响应、博客条目、其他 Web 订阅源和任何其他文本数据进行分类。

如果在建模块中启动交互式工作台会话，那么可以在生成类别模型之前浏览抽取结果、优化资源、调整类别。执行包含“文本挖掘”模型块的流时，会在构件模型之前根据“文本挖掘”建模节点的“模型”选项卡上所选的构建方式来向数据添加新字段。请参阅『类别模型块: “模型”选项卡』主题以获取更多信息。

如果使用已转换文档生成了模型块，那么将在已转换的语言中执行评分。同样，如果使用英语作为语言生成了模型块，那么可以在模型块中指定转换语言，因为文档之后将转换为英语。

文本挖掘模型块在生成后会放置在模型块选用板中（位于 IBM SPSS Modeler 窗口右上方的“模型”选项卡上）。

查看结果

要查看有关模型块的信息，请右键单击模型块选用板中的节点，然后从上下文菜单中选择**浏览**（或针对流中的节点单击**编辑**）。

向流中添加模型

要向流中添加模型块，请单击模型块选用板中的图标，然后单击要将节点放置在的流画布。或者，右键单击图标，然后从上下文菜单中选择**添加到流**。然后，将流连接到节点，即可传递数据以生成预测。

注意: 如果要使用评分块来重新生成包含所使用的类别模型和模板建模节点，我们建议您创建 TAP，并在生成评分块之前将其用于交互式会话代替建模节点。

类别模型块: “模型”选项卡

对于类别模型，“模型”选项卡会在左侧显示类别模型中的类别列表，在右侧显示所选类别的描述符。每个类别都由多个描述符组成。对于所选的每个类别，会在表中显示关联的描述符。这些描述符可包含概念、类别规则、类型和 TLA 模式。同时也会显示每个描述符的类型以及每个描述符所表示的部分示例。

在该选项卡上，目标是选择要用于评分的类别。对于类别模型，会通过对文档和记录进行评分来划分其类别。如果某个文档或记录在其文本或任何底层术语中包含一个或多个描述符，那么会将此文档或记录分配到此描述符所属的类别中。这些底层术语包括语言资源中定义的同义词（无论是否在文本中找到这些同义词都是如此）以及文本中用于生成模型块、变换的术语、来自模糊分组的属于等的任何已抽取的复数/单数术语。

注: 如果改为生成了概念模型块，那么该选项卡将包含不同结果。请参阅第 25 页的『概念模型: “模型”选项卡』主题以获取更多信息。

类别树

要了解有关每个类别的更多信息，请选中此类别并复审针对此类别中的描述符显示的信息。对于每个描述符，您可以复审以下信息:

- **描述符**名称。该字段包含一个表示此描述符的种类以及描述符名称的图标。

表 5. 描述符图标

	概念		TLA 模式

表 5. 描述符图标 (续)

	类型	*fx*	类别规则

- **类型。**该字段包含描述符的类型名称。类型是相似概念的集合（语义组），例如，组织名称、产品或正面意见。不会将规则分配到类型。
- **详细信息。**该字段包含此描述符中所包含的内容的列表。根据匹配数量，由于对话框中的大小限制，您可能无法看到包含每个描述符的完整列表。

选择和复制类别

缺省情况下，选中所有顶级类别以进行评分，如左侧窗格中的复选框所示。复选框表示此类别将用于评分。未选中的复选框表示将从评分中排除此类别。您可以通过选中行并单击所选的其中一个复选框来选中多个行。并且，如果选中某个类别或子类别但是未选中其中某个子类别，那么复选框会显示蓝色背景以指示仅选中所选类别的部分子类别。

通过右键单击树中的类别，可以显示一个上下文菜单，您可以从中:

- **选中所选项。**选中表中所选行的所有复选框。
- **取消选中所选项。**取消选中表中所选行的所有复选框。
- **全部选中。**选中表中的所有复选框。这将导致在最终输出中使用所有类别。您还可以使用工具栏上对应的复选框图标。
- **全部取消选中。**取消选中表中的所有复选框。取消选中类别表示在最终输出中将不使用此类别。您还可以使用工具栏上对应的空复选框图标。

通过右键单击描述符表中的单元格，可以显示一个上下文菜单，您可以从中:

- **复制。**将所选概念复制到剪贴板。
- **随字段复制。**将所选描述符随列标题一起复制到剪贴板。
- **全选。**将选中表中的所有行。

类别模型块: “设置”选项卡

“设置”选项卡用于定义新输入数据的文本字段值（如果需要）。您也可以在其中为自己的输出定义数据模型（评分方式）。

注: 仅当模型块置于画布上或流中时，该选项卡才会显示在节点对话框中。当您在“模型”选用板中访问此块时，该选项卡不存在。

评分方式: 类别作为字段

借助该选项，输出记录数量与输入中的记录数量相同。但是，每条记录现在针对“模型”选项卡上使用复选标记选中的每个类别包含一个新字段。对于每个字段，针对 **True** 和 **False** 输入标记值，例如，*Yes*/*No*、*True*/*False*、*T*/*F* 或 *1* 和 *2*。将自动设置存储类型以反映所选的值。例如，如果针对标记输入数字值，那么将自动将这些数字值作为整数值来处理。标记的存储类型可以是字符串、整数、实数或日期/时间。

注: 如果使用非常大的数据集，例如，使用 DB2 数据库，那么使用**类别作为字段**可能由于数据量而出现处理问题。在此情况下，我们建议改为使用**类别作为记录**。

字段名称扩展。您可以选择为字段名称指定扩展前缀/后缀，或者可以选择使用类别代码。字段名称是使用类别名称加上此扩展来生成的。

- **添加为。**指定在字段名称中添加扩展的位置。选择**前缀**以将扩展添加到字符串开头。选择**后缀**以将扩展添加到字符串末尾。

如果未选中子类别。通过此选项，可以指定将如何处理属于未选定用于评分的子类别的描述符。存在两个选项。

- 选项**在评分中完全排除其描述符**将导致在评分期间忽略且不使用没有复选标记（未选中）的子类别的描述符。
- 选项**将描述符与父类别中的描述符汇总**将导致没有复选标记（未选中）的子类别的描述符用作父类别（此子类别上方的类别）的描述符。如果有多个子类别级别且未选中，那么将在第一个可用父类别下累积描述符。

调整标点错误。 此选项可在抽取期间临时标准化包含标点错误的文本（例如，不正确使用），以改善概念的可抽取性。当文本很短且质量不佳（例如，在开放式调查响应、电子邮件和 CRM 数据中）时，或文本包含很多缩略词时，此选项非常有用。

注: 处理日语文本时，**调整标点错误**选项不适用。

评分方式: 类别作为记录

借助该选项，会针对每个类别和文档对创建一条新记录。通常，输出中的记录数量多于输入中的记录数量。除输入字段之外，根据模型种类还会向数据添加新字段。

表 6. "类别作为记录"的输出字段.

新输出字段	描述
类别	包含文本文档分配到的类别名称。如果类别属于另一个类别的子类别，那么指向类别名称的完整路径受您在此对话框中所选的值控制。

分层类别的值。该选项控制在输出中显示子类别名称的方式。

- **完整的类别路径。**该选项将输出类别的名称和父类别的完整路径（如果适用），使用斜杠来分隔类别名称与子类别名称。
- **简短的类别路径。**该选项将仅输出类别名称，但是用省略符来显示存在问题的类别的父类别数量。
- **底部级别类别。**该选项将仅输出类别名称，不显示完整路径或父类别。

如果未选中子类别。通过此选项，可以指定将如何处理属于未选定用于评分的子类别的描述符。存在两个选项。

- 选项**在评分中完全排除其描述符**将导致在评分期间忽略且不使用没有复选标记（未选中）的子类别的描述符。
- 选项**将描述符与父类别中的描述符汇总**将导致没有复选标记（未选中）的子类别的描述符用作父类别（此子类别上方的类别）的描述符。如果有多个子类别级别且未选中，那么将在第一个可用父类别下累积描述符。

调整标点错误。 此选项可在抽取期间临时标准化包含标点错误的文本（例如，不正确使用），以改善概念的可抽取性。当文本很短且质量不佳（例如，在开放式调查响应、电子邮件和 CRM 数据中）时，或文本包含很多缩略词时，此选项非常有用。

注: 处理日语文本时，**调整标点错误**选项不适用。

类别模型块: “其他”选项卡

类别模型块的“字段”选项卡和“设置”选项卡与概念模型块的选项卡是相同的。

- “字段”选项卡。请参阅第 28 页的『概念模型: “字段”选项卡』主题以获取更多信息。
- “摘要”选项卡。请参阅第 29 页的『概念模型: “摘要”选项卡』主题以获取更多信息。

在流中使用类别模型块

“文本挖掘”类别模型块是从交互式工作台会话生成的。您可以在流中使用此模型块。

示例: 含类别模型块的“Statistics 文件”节点

以下示例显示了如何使用“文本挖掘”模型块。

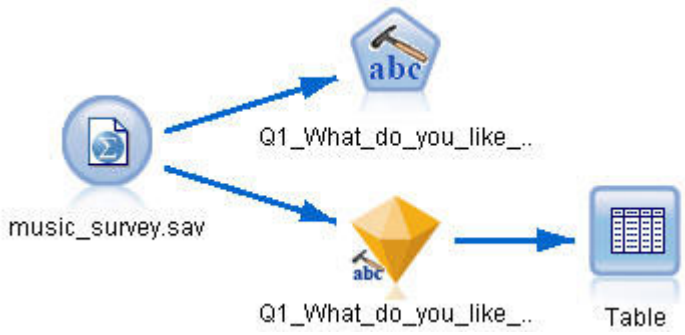


图 9. 示例流: 含“文本挖掘”类别模型块的“Statistics 文件”节点。

1. **“Statistics 文件”节点（“数据”选项卡）。**首先，我们将此节点添加到流中以指定文本文件的存储位置。

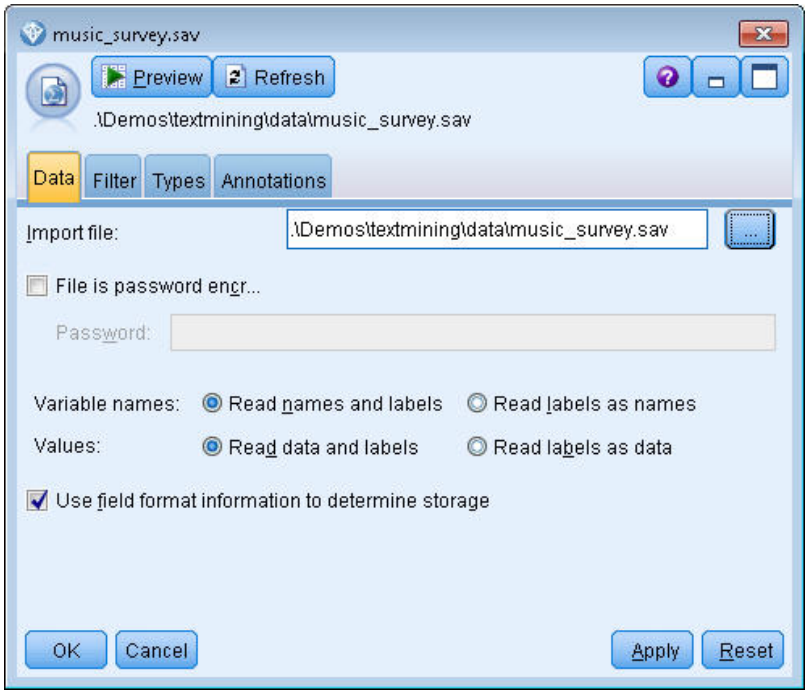


图 10. “Statistics 文件”节点对话框: “数据”选项卡

2. **“文本挖掘”类别模型块（“模型”选项卡）。**接下来，我们将一个类别模型块添加并连接到“Statistics 文件”节点。选中要用于对数据进行评分的类别。

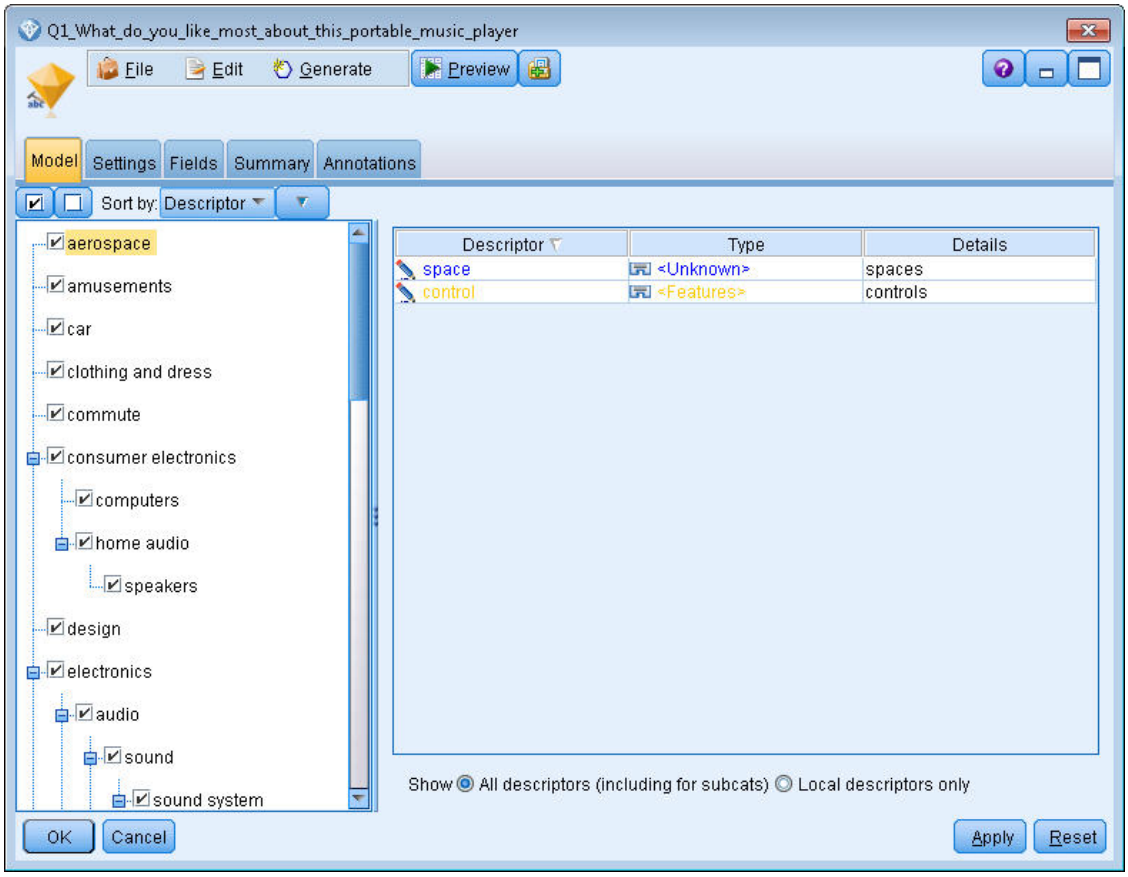


图 11. “文本挖掘”模型块对话框：“模型”选项卡

3. **“文本挖掘”模型块（“设置”选项卡）。**然后，定义输出格式**类别作为字段**。

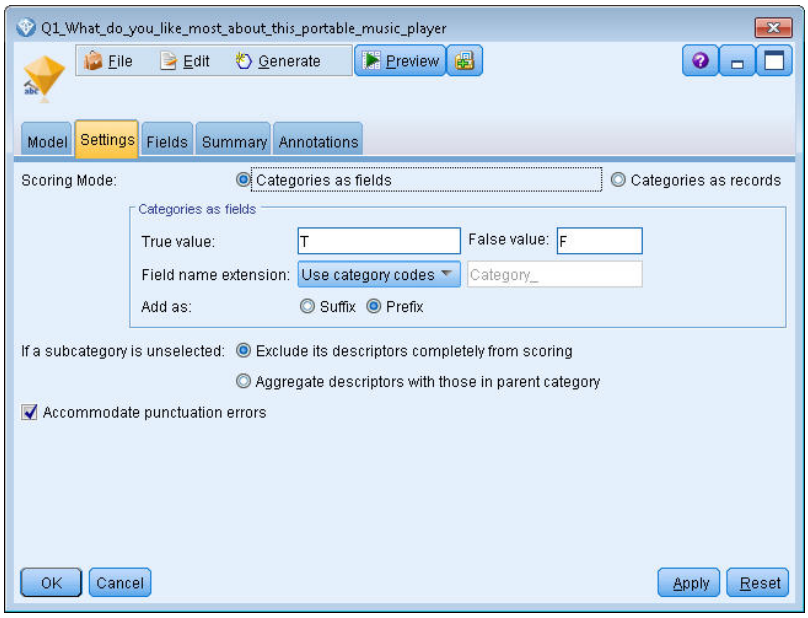


图 12. 类别模型块对话框：“设置”选项卡

4. **“文本挖掘”类别模型块（“字段”选项卡）。**然后，选择文本字段变量（即来自“Statistics 文件”节点的字段名称），并选择选项“文本”字段显示**实际文本**以及其他设置。

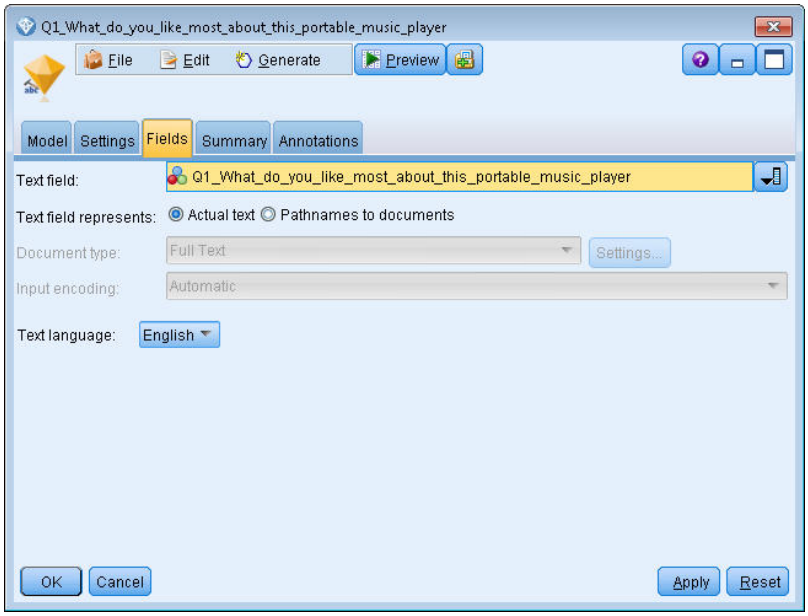


图 13. “文本挖掘”模型块对话框：“字段”选项卡

5. **表节点。**然后，附加一个表节点以查看结果并执行流。

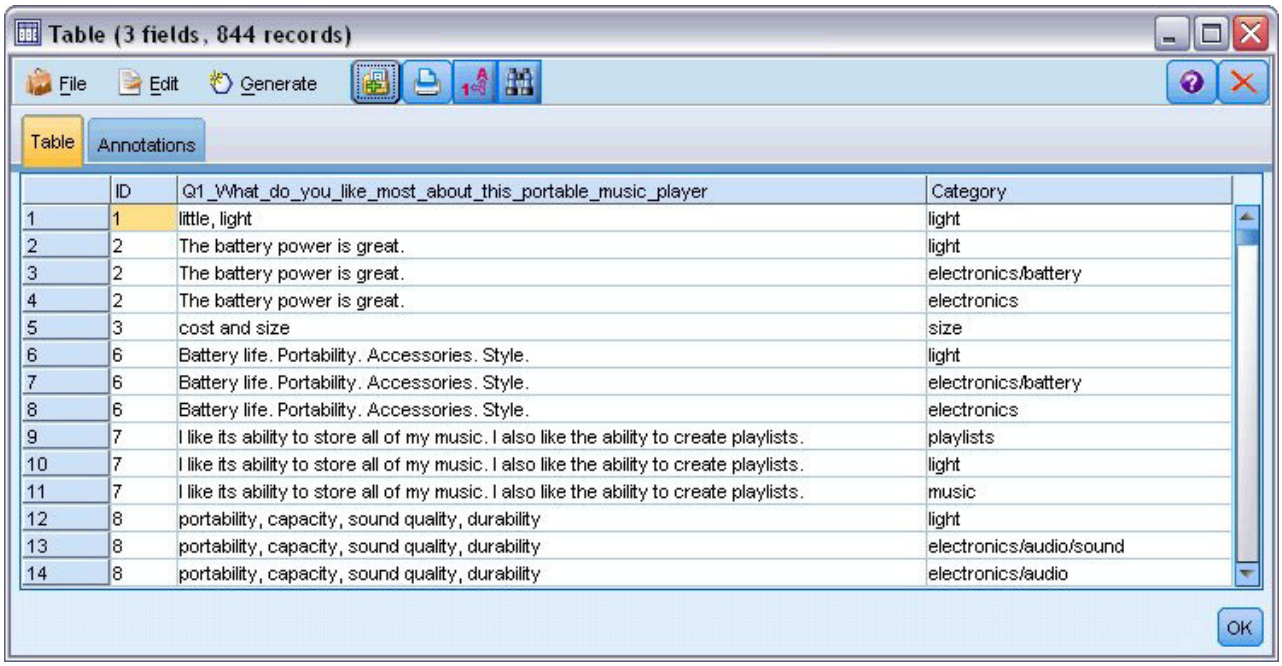


图 14. 表输出

第 4 章 挖掘文本链接

“文本链接分析”节点

“文本链接分析 (TLA)”节点在文本挖掘的概念抽取中添加模式匹配技术以基于已知模式识别文本数据中的概念之间的关系。这些关系可以描述客户对于产品的感受、哪些公司正在合作开展业务，甚至是基因或药品代理之间的关系。

例如，您可能对抽取竞争对手的产品名称兴趣不大。通过使用此节点，您还可以了解人们对于该产品的感受，前提是数据中存在此类意见。通过将已知模式与您的文本数据进行匹配来识别和抽取关系和关联。

您可以在 IBM SPSS Modeler Text Analytics 随附的某些资源模板内使用 TLA 模式规则，或者创建/编辑您自己的 TLA 模式规则。模式规则由宏、文字列表和组成布尔值查询或与您的输入文本比较的规则的字隙组成。当 TLA 模式规则匹配文本时，可通过 TLA 抽取此文本并将其重新构造为输出数据。请参阅第 183 页的第 19 章，『关于文本链接规则』主题以获取更多信息。

“文本链接分析”节点提供了更直接的方式来从您的文本中识别和抽取 TLA 模式，然后将结果添加到流中的数据集内。但是“文本链接分析”节点并非您可执行文本链接分析的唯一方法。您还可以在“文本挖掘”建模节点中使用交互式工作台会话。

在交互式工作台中，您可以探索 TLA 模式规则并将其用作为类别描述符和/或使用向下钻取和图形来了解有关结果的更多信息。请参阅第 127 页的第 12 章，『探索文本链接分析』主题以获取更多信息。事实上，使用“文本挖掘”节点来抽取 TLA 结果是为您的数据探索和调整模板以供稍后在 TLA 节点中直接使用的有效方法。

可在最多 6 个通道或部件中显示输出。日语模式仅作为一个或两个通道输出。请参阅第 42 页的『TLA 节点输出』主题以获取更多信息。

您可以在 IBM SPSS Modeler 窗口底部的节点选用板的 IBM SPSS Modeler Text Analytics 选项卡上找到此节点。请参阅第 7 页的『IBM SPSS Modeler Text Analytics 节点』主题以获取更多信息。

需求。“文本链接分析”节点接受使用任何标准源节点（数据库节点、平面文件节点等）读取到字段中的文本数据。或读取到列出到“文件列表”节点或“Web 订阅源”节点生成的外部文档的路径的字段中的文本数据。

强度。“文本链接分析”节点不仅通过基本概念抽取提供有关概念*之间*的关系的信息，还提供数据中可能揭示的相关意见或限定符。

“文本链接分析”节点: “字段”选项卡

“字段”选项卡用于指定您将从中抽取概念的数据的字段设置。您可以设置以下参数:

“标识”字段。选择包含文本记录的标识的字段。标识必须为整数。“标识”字段充当单个文本记录的索引。如果文本字段表示要挖掘的文本，请使用“标识”字段。如果文本字段表示**到文档的路径名**，请勿使用“标识”字段。

文本字段。选择包含要挖掘的文本的字段、文档路径名或文档的目录路径名。此字段取决于数据源。

文本字段表示。指示在先前设置中指定的文本字段包含的内容。选项包括:

- **实际文本。**如果字段包含应从中抽取概念的确切文本，请选择此选项。
- **文档路径名。**如果字段包含文本文档驻留所在的位置的一个或多个路径名，请选择此选项。

文档类型。仅在指定文本字段表示**文档路径名**的情况下，此选项才可用。文档类型指定文本的结构。请选择以下类型之一:

- **全文本。**用于大多数文档或文本源。系统会扫描整个文本集以进行抽取。与其他选项不同，此选项没有其他设置。
- **结构化文本。**用于书目形式、专利以及任何包含可识别并分析的常规结构的文件。此文档类型用于跳过全部或部分抽取过程。通过它可定义术语分隔符，分配类型和施加最小频率值。如果选择此选项，那么必须单击**设置**按钮并在“文档设置”对话框的**结构化文本格式化**区域中输入文本分隔符。请参阅主题第 17 页的『“字段的文档设置”选项卡』以获取更多信息。
- **XML 文本。**用于指定包含要抽取的文本的 XML 标记。系统会忽略所有其他标记。如果选择此选项，那么必须单击**设置**按钮并在“文档设置”对话框的 **XML 文本格式化**区域中指定包含抽取过程中要读取的文本的 XML 元素。请参阅主题第 17 页的『“字段的文档设置”选项卡』以获取更多信息。

文本统一。仅在指定文本字段表示**文档路径名**并将**全文本**设置为文档类型的情况下，此选项才可用。选择以下抽取方式:

- **文档方式。**用于简短且语义同构的文档，如来自新闻机构的文章。
- **段落方式。**用于 Web 页面和非标记文档。抽取过程按语义划分文档，从而利用诸如内部标记和语法之类的特征。如果选择此方式，那么将逐个段落应用评分。因此，例如，仅在同一段落中找到 `apple` 和 `orange` 的情况下，规则 `apple & orange` 才成立。

  注: 由于从 PDF 文档中抽取文本的方式，**段落方式**不适用于这些文档。这是因为抽取会抑制回车符标记。

段落方式设置。仅在指定文本字段表示**文档路径名**并将文本统一选项设置为**段落方式**时，此选项才可用。请指定要在任何抽取中使用的字符阈值。实际大小会四舍五入为最接近的时间段。要确保从文档集合的文本产生的单词关联具有代表性，请避免指定太小的抽取大小。

- **最小值。**指定要在任何抽取中使用的最小字符数。
- **最大值。**指定要在任何抽取中使用的最大字符数。

输入编码。仅在指示文本字段表示**文档路径名**的情况下，此选项才可用。它指定缺省文本编码。对于除日语以外的所有语言，都将从指定编码或已识别编码转换为 ISO-8859-1。因此，即使指定其他编码，抽取引擎也会在对其进行处理之前将其转换为 ISO-8859-1。不符合 ISO-8859-1 编码定义的任何字符都将转换为空格。对于日语文本，可以选择以下若干编码选项之一：SHIFT_JIS、EUC_JP、UTF-8 或 ISO-2022-JP。

复制以下来源的资源。挖掘文本时，抽取不仅基于“专家”选项卡中的设置，也基于语言资源。这些资源充当抽取期间文本处理方式的基础，从而获取概念、类型和 TLA 模式。您可以将资源从资源模板复制到此节点中。

资源模板是预定义的库和高级语言与非语言资源的集合，这些资源针对特定域或用途经过了微调。这些资源充当抽取期间处理数据的方式的基础。单击**装入**并选择要从中复制您的资源的模板。

模板是在选择模板时而不是在执行流时装入的。装入时，会将资源的副本存储在节点中。因此，如果您想要使用经过更新的模板，必须在此处重新装入这些模板。请参阅第 21 页的『从模板和 TAP 复制资源』主题以获取更多信息。

文本语言。标识进行挖掘的文本的语言。节点中复制的资源控制所呈现的语言选项。可以选择为其调整了资源的语言，或者选择 **ALL** 选项。强烈建议指定文本数据的确切语言；但如果不确定，那么可以选择 **ALL** 选项。**ALL** 不适用于日语文本。此 **ALL** 选项会延长执行时间，因为使用了自动语言识别来扫描所有文档和记录，以便首先标识文本语言。通过此选项，采用受支持且许可的语言的所有记录或文档都由抽取引擎使用相应语言的内部字典进行读取。请参阅主题第 182 页的『语言标识』以获取更多信息。如果您有兴趣为当前不具有访问权的受支持语言购买许可证，请与您的销售代表联系。

“文本链接分析”节点: “模型”选项卡

“模型”选项卡包含单个选项，该选项可影响抽取过程的速度和准确性。

针对评分速度进行优化。缺省情况下选中该选项，该选项可确保创建的模型体积小并且可快速评分。取消选中该选项后所创造的模型评分速度更慢，但可确保完整的概念-类型一致性，即，它可确保仅为给定概念分配一种类型。

“文本链接分析”节点: “专家”选项卡

在此节点中，自动启用文本链接分析 (TLA) 模式结果的抽取。“专家”选项卡包含影响文本抽取和处理方式的额外参数。此对话框中的参数可控制抽取过程的基本行为以及一些高级行为。还有多种语言资源和选项也可能影响抽取结果，这些资源和选项由您所选的资源模板来控制。

针对荷兰语、英语、法语、德语、意大利语、葡萄牙语和西班牙语文本

调整标点错误。 此选项可在抽取期间临时标准化包含标点错误的文本（例如，不正确使用），以改善概念的可抽取性。当文本很短且质量不佳（例如，在开放式调查响应、电子邮件和 CRM 数据中）时，或文本包含很多缩略词时，此选项非常有用。

调整拼写错误，最小根字符限制为 [n]。 此选项适用于模糊分组方法，此方法可帮助将普遍拼写有误的单词或拼写接近的单词分组到一个概念下。模糊分组算法临时删除抽取单词中的所有元音（除了第一个元音）和出现的二重/三重辅音，然后进行比较，以查看它们是否相同，以便 modeling 和 modelling 分组到一起。但是，如果每个术语分配给不同类型（除了 <Unknown> 类型），那么不会应用模糊分组方法。

您还可先优化最少数目的所需*根*字符，再使用模糊分组。术语中根字符数通过对所有字符相加减去形成屈折变化后缀的任何字符数以及（使用复合单词术语的情况下）限定词和介词数计算得出。例如，术语 exercises 将计算为 8 个根字符（形式为“exercise”），因为字母单词末尾的 *s* 是屈折变化形式（复数形式）。相似地，apple sauce 将计算为 10 个根字符（“apple sauce”，manufacturing of cars 将计算为 16 个根字符（“manufacturing car”）此计数方法仅用于检查是否应该应用模糊分组，但不会影响匹配单词的方式。

注: 如果发现某些单词之后分组不正确，那么可通过在“高级资源”选项卡中的**模糊分组: 例外**中显式进行声明来从此方法排除单词对。请参阅主题第 175 页的『模糊分组』，以获取更多信息。

抽取单术语。 此选项用于抽取单个单词（单术语），前提是此单词不属于复合单词的一部分，且其为名词或语音的不可识别部分。

抽取非语言实体。 此选项用于抽取非语言实体，例如，电话号码、社保号、时间、日期、货币、数字、百分比、电子邮件地址和 HTTP 地址。您可以在“高级资源”选项卡中的**非语言实体: 配置**部分中包含或排除某些类型的非语言实体。通过禁用任何不需要的实体，抽取引擎不会浪费处理时间。请参阅主题第 179 页的『配置』，以获取更多信息。

大写算法。 此选项用于抽取内置字典中不存在的简单和复合术语，前提是术语的第一个字母为大写。此选项提供了一种很好的方式来抽取大部分正确的名词。

尽可能将部分和完整人员姓名分组在一起。 此选项用于将在文本中显示不同的姓名分组在一起。由于通常在文本开头部分通过全名指代姓名，而之后通过较短的版本指代姓名，因此，此功能会很有帮助。此选项尝试将类型为 `<Unknown>` 的任何单术语与类型为 `<Person>` 的任何复合术语的最后一个单词匹配。例如，如果发现了 *doe* 且其最初类型为 `<Unknown>`，那么抽取引擎会检查以了解 `<Person>` 类型中的任何复合术语是否将 *doe* 作为最后一个单词包含，例如，*john doe*。此选项不适用于名字，因为大多数名字永不会抽取为单术语。

最大非功能单词排列。 此选项指定应用排列方法时可显示的非功能单词的最大数目。此排列方法将仅包含的非功能单词（例如，of 和 the）不同（不考虑屈折变化）的相似短语分组在一起。例如，假设将此值设置为最多两个单词，且抽取了 `company officials` 和 `officials of the company`。在此情况下，这两个抽取的术语将在最终概念列表中分组在一起，因为在忽略 `of the` 时，这两个术语视为相同。

对于日语文本

对于日语文本，您可以选择要应用的辅助分析器。

辅助分析。 启动抽取时，将使用一组缺省类型执行基本关键字抽取。 但是，选择辅助分析器时，由于抽取器现在将小品词和助动词作为概念的一部分包含，可获取更多或更丰富的概念。 如果进行观点分析，那么还会包含大量其他类型。此外，选择辅助分析器，还可生成文本链接分析结果。

注: 调用辅助分析器时，需要花费更长时间来完成抽取过程。

- **依赖关系分析**。选择此选项可从基本类型和关键字抽取获取抽取概念的扩展小品词。还可从依赖关系文本链接分析 (TLA) 获取更丰富的模式结果。
- **观点分析**。选择此分析器可获取其他抽取的概念，适用时，会执行 TLA 模式的抽取。除了基本类型，还可从超过 80 种观点类型受益。这些类型用于通过表情、观点和意见说明文本中的概念和模式。具有三个选项，这些选项指示观点分析的焦点: **所有观点**、**仅表示观点**和**仅结论**。

TLA 节点输出

运行“文本链接分析”节点之后，会重新构造数据。了解文本挖掘重新构造数据的方式是非常重要的。如果您想要采用不同的数据挖掘结构，那么可以使用“字段操作”选用板上的节点来完成此操作。例如，如果所处理的数据中每个行表示一条文本记录，那么针对源文本数据中揭示的每个模式会创建一个行。对于输出中的每个行，包含 15 个字段:

- 六个字段（**Concept#**，例如，**Concept1**、**Concept2**、...，以及 **Concept6**）表示模式匹配中找到的任何概念。
- 六个字段（**Type#**，例如，**Type1**、**Type2**、...，以及 **Type6**）表示每个概念的类型。
- **规则名称**表示用于匹配文本并声称输出的文本链接规则的名称。

- 一个使用您在节点中指定的“标识”字段的名称并表示输入数据中的记录或文档标识的字段。
- **匹配的文本**表示与 TLA 模式匹配的原始记录或文档中的文本数据部分。

注: 日语文本的文本链接分析模式规则仅生成一个或两个通道模式结果。

注: 任何预先存在的流如果包含来自低于 5.0 的发行版的“文本链接分析”节点，那么都必须在更新节点之后才能完全执行。较高版本的 IBM SPSS Modeler 中的某些改进需要将较低版本的节点替换为更易于部署且更强大的更高版本。

还可以执行某些语言的自动翻译。该功能支持您挖掘采用您可能无法使用（说或读）的语言的文档。如果您要使用翻译功能，必须有权访问 SDL 软件即服务 (SaaS)。请参阅第 48 页的『转换设置』主题以获取更多信息。

缓存 TLA 结果

如果使用高速缓存，文本链接分析结果保存在流中。为避免每次执行流时重复抽取文本链接分析结果，请选择文本链接分析节点，并从菜单中选择**编辑 > 节点 > 高速缓存 > 启用**。下次执行流时，输出会缓存在节点中。节点图标会显示微小的“文档”图形，填充高速缓存时，此图形会从白色更改为绿色。在会话持续时间内会保留高速缓存。要将高速缓存再保留一天（关闭并重新打开流之后），请选中该节点，并从菜单中选择**编辑 > 节点 > 高速缓存 > 保存高速缓存**。下次打开流时，可以重新装入保存的高速缓存，而无须再次运行翻译。

或者，您可以右键单击节点并从上下文菜单中选择**高速缓存**来保存或启用节点高速缓存。

在流中使用“文本链接分析”节点

“文本链接分析”节点用于访问数据和抽取流中的概念。您可以使用任何源节点来访问数据。

示例: 含“文本链接分析”节点的“Statistics 文件”节点

以下示例显示了如何使用“文本链接分析”节点。

图 15. 示例: 含“文本链接分析”节点的“Statistics 文件”节点

1. **“Statistics 文件”节点（“数据”选项卡）**。首先，我们将此节点添加到流中以指定文本存储位置。

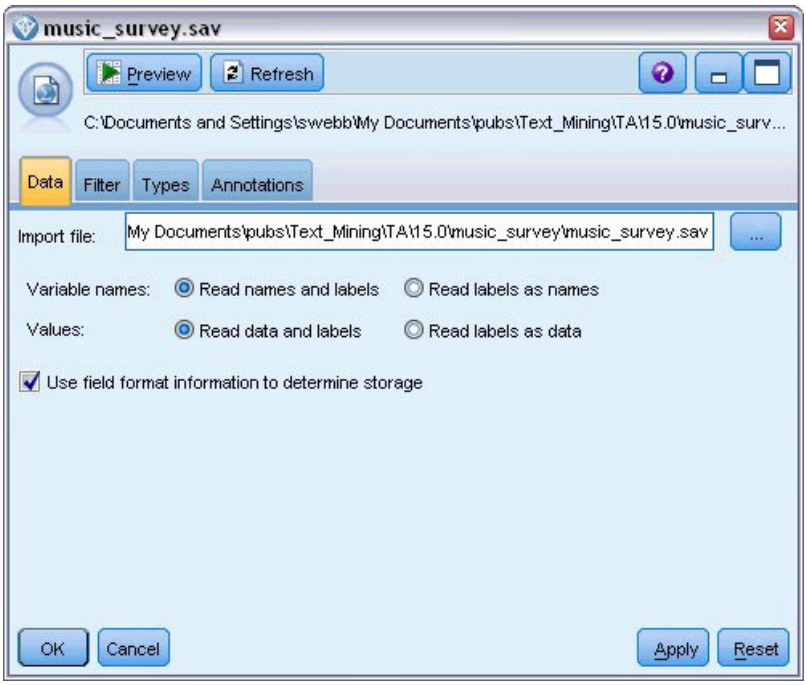


图 16. "Statistics 文件"节点对话框: "数据"选项卡

2. **"文本链接分析"节点（"字段"选项卡）。**接下来，我们将此节点连接到流中以抽取概念用于下游建模或查看。我们指定了包含此数据的"标识"字段和文本字段名称以及其他设置。

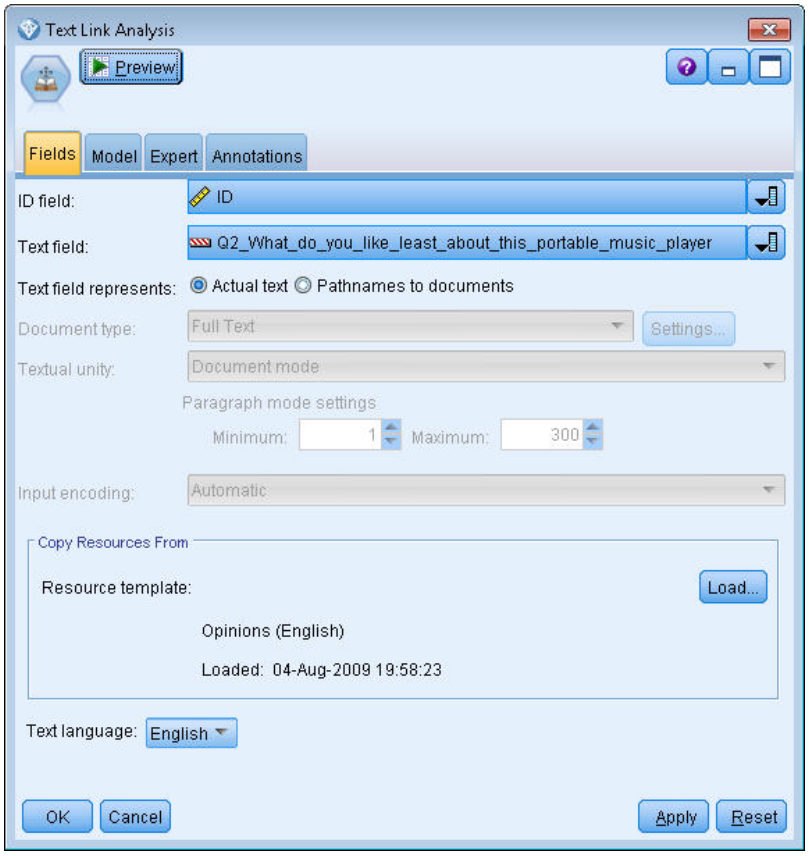
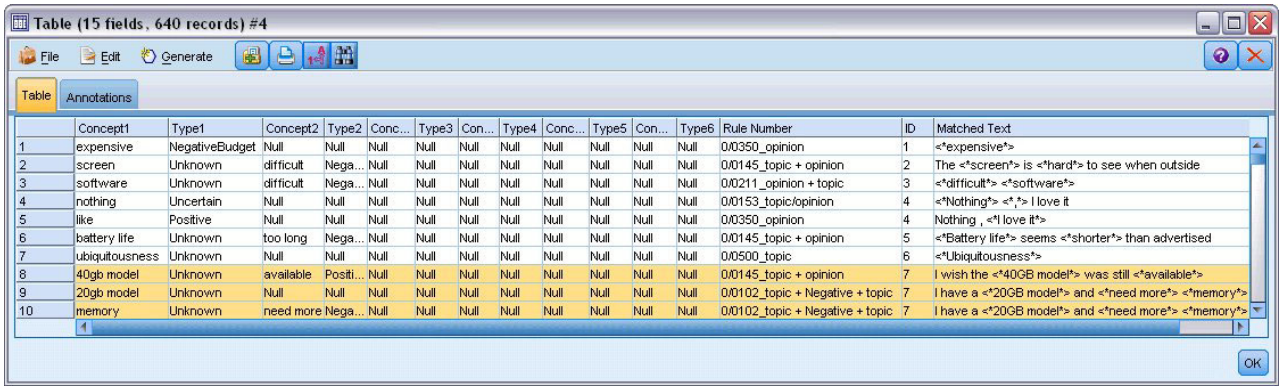


图 17. "文本链接分析"节点对话框："字段"选项卡

3. **表节点。**最后，我们连接了一个"表"节点以查看从文本文档中抽取的概念。在显示的表输出中，您可以看到使用"文本链接分析"节点执行此流之后在数据中找到的 TLA 模式结果。某些结果显示只有一个概念/类型匹配。在其他结果中，结果更为复杂，包含多种类型和多个概念。此外，通过"文本链接分析"节点运行数据和抽取概念导致数据的多个方面发生更改。我们的示例中的原始数据包含 8 个字段和 405 条记录。执行"文本链接分析"节点之后，现在其中包含 15 个字段和 640 条记录。现在针对找到的每个 TLA 模式结果显示一行。例如，ID 7 从原始状态变为三行，因为抽取了三个 TLA 模式记录。如果要将此输出数据重新合并到原始数据中，可以使用"合并"节点。



	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*>

图 18. 表输出节点

第 5 章 为抽取转换文本

翻译节点

“翻译”节点可用于将文本从受支持的语言（例如，阿拉伯语、中文和波斯语）翻译为英语以供使用 IBM SPSS Modeler Text Analytics 对其进行分析。由此可实现其他方法无法实现的挖掘双字节语言中的文档的能力，并且允许分析人员即使无法理解所涉及的语言，仍可以从外语文档中抽取概念。请注意，必须能够连接到 SDL 的软件即服务 (SaaS) 才能使用“翻译”节点。

挖掘采用其中任意语言的文本时，只需在流中的“文本挖掘”建模节点之前添加“翻译”节点即可。您还可以在“翻译”节点中启用高速缓存来避免每次执行此流时都重复进行翻译。

您可以在 IBM SPSS Modeler 窗口底部的节点选用板的 IBM SPSS Modeler Text Analytics 选项卡上找到此节点。请参阅主题第 7 页的『IBM SPSS Modeler Text Analytics 节点』以获取更多信息。

缓存翻译。如果您缓存翻译，那么翻译的文本会存储在流中而不是存储在外部文件中。为避免每次执行流时重复翻译，请选中“翻译”节点，并从菜单中选择**编辑 > 节点 > 高速缓存 > 启用**。下次执行流时，翻译输出会缓存在节点中。节点图标会显示微小的“文档”图形，填充高速缓存时，此图形会从白色更改为绿色。在会话持续时间内会保留高速缓存。要将高速缓存再保留一天（关闭并重新打开流之后），请选中该节点，并从菜单中选择**编辑 > 节点 > 高速缓存 > 保存高速缓存**。下次打开流时，可以重新装入保存的高速缓存，而无须再次运行翻译。

或者，您可以右键单击节点并从上下文菜单中选择**高速缓存**来保存或启用节点高速缓存。

重要！ 如果要尝试通过代理服务器在 Web 上检索信息，那么必须为 IBM SPSS Modeler Text Analytics 客户机和服务器在 `net.properties` 文件中启用代理服务器。执行此文件内详细描述的指示信息。这适用于通过 Web 订阅源节点访问或检索 SDL 软件即服务 (SaaS) 许可证的情况，因为这些连接通过 Java。缺省情况下，此文件位于 *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties* 中。

注：您不能将“翻译”节点用于在 IBM SPSS Collaboration and Deployment Services - Scoring 配置内进行评分。

“翻译”节点：“翻译”选项卡

文本字段：选择包含要挖掘的文本、文档路径名或到文档的目录路径名的字段。此字段取决于数据源。您可以指定任何字符串字段，包括 `Direction=None` 或 `Type=Typeless` 的字段。

文本字段表示。指示先前设置中指定的文本字段包含的内容。选项包括：

- **实际文本**：如果该字段包含应从中抽取概念的准确文本，那么请选择该选项。
- **到文档的路径名**：如果该字段包含到包含要抽取的文本的文档所在位置的一个或多个路径名，那么请选择该选项。例如，如果使用“文件列表”节点在文档列表中读取内容，那么应选中该选项。请参阅第 9 页的『“文件列表”节点』主题以获取更多信息。

输入编码：选择源文本的编码。您可以通过选择**自动**选项开始，但是如果您注意到某些字段未得到正确处理，我们建议您从此处列表中选择实际编码。“自动”选项在处理简短文本（例如，简短的数据库记录时）可能错误识别编码。来自此节点的文本输出以 `UTF-8` 进行编码。

设置：指定流的翻译设置。

- **语言对连接**。选择要使用的语言对；在**翻译设置**对话框中设置到 SDL 服务的链接之后，会在此列表中自动显示可用语言对。请参阅『转换设置』主题以获取更多信息。
- **接触方式**。如果您先前创建了 *SDL 接触方式*，那么请选择要配合翻译使用的接触方式。
- **保存并复用先前翻译的文本（如果可能）**：指定应保存翻译结果，如果下次执行流时存在相同数量的记录/文档，那么假定内容相同，并且将复用翻译结果来节省处理时间。如果在运行时选中该选项，并且记录数与上次保存的数量不匹配，那么会完全翻译此文本，然后将其保存在标签名称下以供下次执行。仅当您选中 SDL 翻译语言时，该选项才可用。

  注：如果文本存储在流中，您还可以在“翻译”节点中启用高速缓存。在此情况下，不仅可复用翻译结果，并且只要高速缓存可用，即可忽略上游任何内容。
- **标签**：如果您选择**保存并复用先前翻译的文本（如果可能）**，那么必须为结果指定标签名称。此标签用于识别先前翻译的文本。如果未指定任何标签，那么执行流并且无可用复用时，会在“流属性”中添加一条警告。

转换设置

在此对话框中，您可以定义和管理能够在转换时随时复用的 SDL 软件即服务 (SaaS) 转换连接。在此处定义连接后，即可在转换时快速选择语言对连接，而不必重新输入所有连接设置。

语言对连接向服务器表明源语言和转换语言以及 URL 连接详细信息。例如，*中文 - 英语*意味着源文本为中文，而产生的转换将是英语。必须手动定义将通过 SDL 联机服务访问的每个连接。

重要！ 如果要尝试通过代理服务器在 Web 上检索信息，那么必须为 IBM SPSS Modeler Text Analytics 客户机和服务器在 `net.properties` 文件中启用代理服务器。执行此文件内详细描述的指示信息。这适用于通过 Web 订阅源节点访问或检索 SDL 软件即服务 (SaaS) 许可证的情况，因为这些连接通过 Java。缺省情况下，此文件位于 *C:\Program Files\IBM\SPSS\Modeler\18\jre\lib\net.properties* 中。

连接 URL 输入 SDL 软件即服务连接的 URL。

API 密钥 输入由 SDL 提供的密钥。

帐户标识 输入由 SDL 提供的唯一标识。

用户标识 输入由 SDL 提供的唯一标识。

测试 单击**测试**以验证是否正确配置连接以及查看在该连接上找到的语言对。

使用转换节点

要从受支持的转换语言（如阿拉伯语、中文或波斯语）抽取概念，请在流中的任何文本挖掘节点前添加转换节点。

如果要转换的文本包含在一个或多个外部文件中，那么可以使用文件列表节点在名称列表中进行读取。在此情况下，将会在文件列表节点与任何后续文本挖掘节点之间添加转换节点，并且输出将是已转换的文本驻留所在的位置。

第 6 章 浏览外部源文本

文件查看器节点

挖掘文档集合时，您可以将文件的完整路径名直接指定到文本挖掘建模节点和转换节点中。但是，当输出到表节点时，将仅显示文档的完整路径名而不是其包含的文本。文件查看器节点可用作表节点的模拟，通过它可访问每个文档中的实际文本，而不必将其全部合并成单个文件。

文件查看器可通过提供对源文本或已转换的文本（从中抽取了概念，否则在流中无法访问）的访问来帮助更好地了解文本抽取结果。此节点添加到流中的文件列表节点后，以获取所有文件的链接列表。

此节点的结果是窗口显示所有已读取并用于抽取概念的文档元素。从此窗口中，可以单击工具栏图标，以在将文档名列为超链接的外部浏览器中启动报告。可以单击链接来打开集合中的对应文档。请参阅主题『使用文件查看器节点』以获取更多信息。

您可以在 IBM SPSS Modeler 窗口底部的节点选用板的 IBM SPSS Modeler Text Analytics 选项卡上找到此节点。请参阅主题第 7 页的『IBM SPSS Modeler Text Analytics 节点』以获取更多信息。

注: 当在客户机/服务器方式下工作且文件查看器节点属于流的一部分时，文档集合必须存储在服务器上的 Web 服务器目录中。由于文本挖掘输出节点产生 Web 服务器目录中存储的文档列表，因此 Web 服务器的安全设置会管理这些文档的许可权。

文件查看器节点设置

您可以为文件查看器节点指定以下设置。

文档字段。从数据中选择包含要显示的文档的全名和路径的字段。

生成的 HTML 页面的标题。创建要在包含文档列表的页面顶部显示的标题。

使用文件查看器节点

以下示例显示如何使用文件查看器节点。

示例: 文件列表节点和文件查看器节点

图 19. 说明文件查看器节点的使用的流

1. **文件列表节点（“设置”选项卡）。**首先，添加此节点以指定文档的所在位置。

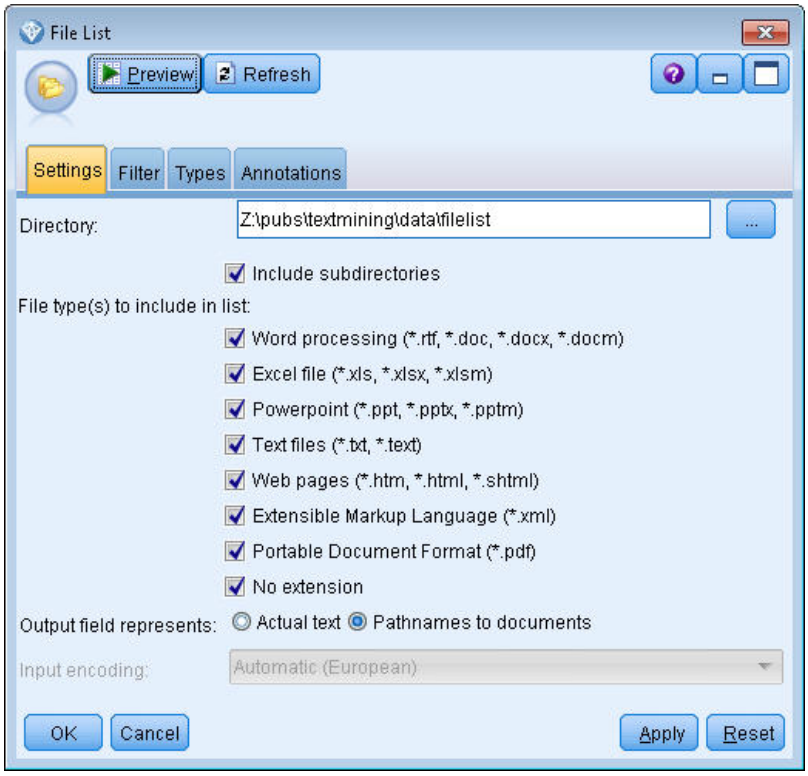


图 20. 文件列表节点对话框：“设置”选项卡

2. **文件查看器节点（“设置”选项卡）。** 接下来，附加文件查看器节点以产生文档的 HTML 列表。

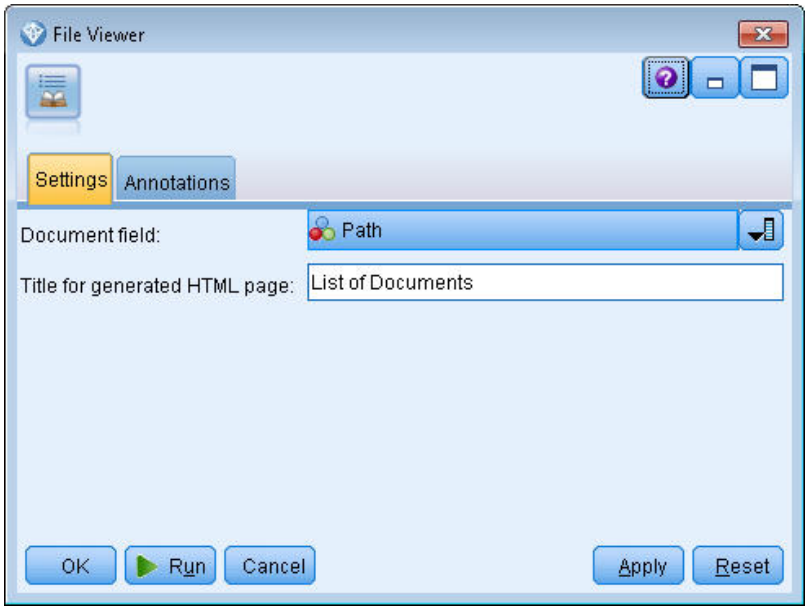


图 21. 文件查看器节点对话框：“设置”选项卡

3. **文件查看器输出对话框。**接下来，执行用于在新窗口中输出文档列表的流。

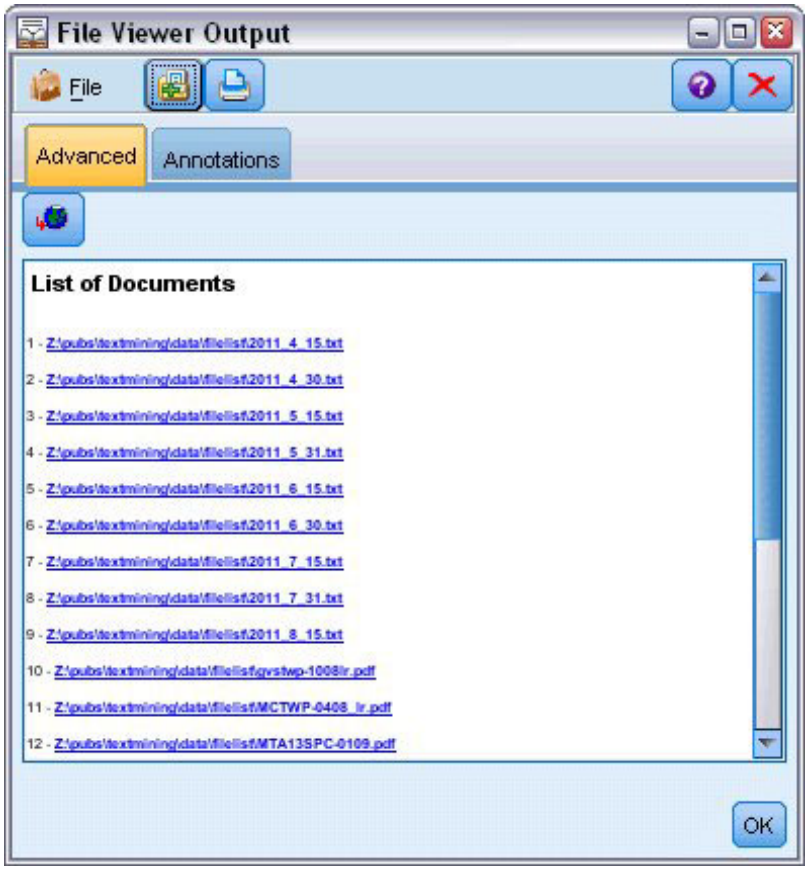


图 22. 文件查看器输出

4. 为查看文档，单击了显示带有红色箭头的地球图标的工具栏按钮。这会在浏览器中打开文档超链接列表。

第 7 章 脚本编制的节点属性

IBM SPSS Modeler 拥有脚本编制语言，允许您从命令行执行流。您可以在此处了解有关特定于随 IBM SPSS Modeler Text Analytics 交付的每个节点的节点属性。有关随 IBM SPSS Modeler 交付的标准节点集合的更多信息，请参阅“脚本编制和自动化指南”。

文件列表节点: filelistnode

您可以使用下表中的属性进行脚本编制。该节点本身称为 filelistnode。

表 7. “文件列表”节点脚本编制属性

脚本编制属性	数据类型
path	*字符串*
recurse	*标记*
word_processing	*标记*
excel_file	*标记*
powerpoint_file	*标记*
text_file	*标记*
web_page	*标记*
xml_file	*标记*
pdf_file	*标记*
no_extension	*标记*

注: “创建列表”参数不再可用，包含该选项的任何脚本都将自动转换为“文件”输出。

Web 订阅源节点: webfeednode

您可以使用下表中的属性进行脚本编制。节点本身称为 webfeednode。

表 8. Web 订阅源节点脚本编制属性

脚本编制属性	数据类型	属性描述
urls	*string1 string2 ...stringn*	每个 URL 都以列表结构进行指定。URL 列表以“\n”分隔
recent_entries	*flag*	
limit_entries	*integer*	每个 URL 的要读取的最新条目数。
use_previous	*flag*	用于保存并复用 Web 订阅源高速缓存。
use_previous_label	*string*	已保存的 Web 高速缓存的名称。
start_record	*string*	非 RSS 开始标记。
url *n* .title	*string*	对于列表中的每个 URL，必须也在此处定义一个 URL。第一个 URL 将是 url1.title，其中数字与其在 URL 列表中的位置相匹配。这是包含内容标题的开始标记。
url *n* .short_description	*string*	对于 url *n* .title 相同。
url *n* .description	*string*	对于 url *n* .title 相同。

表 8. Web 订阅源节点脚本编制属性 (续)

脚本编制属性	数据类型	属性描述
url *n* .authors	*string*	对于 url *n* .title 相同。
url *n* .contributors	*string*	对于 url *n* .title 相同。
url *n* .published_date	*string*	对于 url *n* .title 相同。
url *n* .modified_date	*string*	对于 url *n* .title 相同。
html_alg	无 HTMLCleaner	内容过滤方法。
discard_lines	*flag*	废弃短行。与 min_words 配合使用
min_words	*integer*	最小字数。
discard_words	*flag*	废弃短行。与 min_avg_len 配合使用
min_avg_len	*integer*	
discard_scw	*flag*	废弃包含多个单字符字的行。与 max_scw 配合使用
max_scw	*integer*	一行中单字符字的最大比例 (0-100%)
discard_tags	*flag*	废弃包含特定标记的行。
tags	*string*	特殊字符必须使用反斜杠字符 \ 进行转义。
discard_spec_words	*flag*	废弃包含特定字符串的行
words	*string*	特殊字符必须使用反斜杠字符 \ 进行转义。

文本挖掘节点: TextMiningWorkbench

您可以使用以下参数来通过脚本编制定义或更新节点。该节点本身称为 TextMiningWorkbench。

要点！无法通过脚本编制指定其他资源模板。如果您认为自己需要模板，必须在节点对话框中将其选中。

表 9. "文本挖掘"建模节点脚本编制属性

脚本编制属性	数据类型	属性描述
text	*field*	
method	ReadText ReadPath	
docType	*integer*	可能的值为 (0,1,2)，其中 0 = 全文本，1 = 结构化文本，2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	请注意，具有特殊字符的值（如 "UTF-8"）应加上引号以避免与算术运算符混淆。
unity	*integer*	可能的值为 (0,1)，其中 0 = 段落，1 = 文档
para_min	*integer*	
para_max	*integer*	

表 9. “文本挖掘”建模节点脚本编制属性 (续)

脚本编制属性	数据类型	属性描述
mtag	*string*	包含所有 mtag 设置（来自 XML 文件的“设置”对话框）
mclef	*string*	包含所有 mclef 设置（来自结构化文本文件的“设置”对话框）
partition	*字段*	
custom_field	*标记*	指示是否将指定分区字段。
use_model_name	*标记*	
model_name	*字符串*	
use_partitioned_data	*标记*	如果定义了分区字段，那么仅将培训数据用于模型构建。
model_output_type	Interactive Model	Interactive 会生成类别模型。Model 会生成概念模型。
use_interactive_info	*标记*	针对仅限在工作台会话中以交互方式进行构建。
reuse_extraction_results	*标记*	针对仅限在工作台会话中以交互方式进行构建。
interactive_view	Categories TLA Clusters	针对仅限在工作台会话中以交互方式进行构建。
extract_top	*整数*	当 model_type = Concept 时使用此参数
use_check_top	*标记*	
check_top	*整数*	
use_uncheck_top	*标记*	
uncheck_top	*整数*	
language	de en es fr it ja nl pt	
frequency_limit	*整数*	在 14.0 中不推荐。
concept_count_limit	*integer*	将抽取限于全局频率至少为该值的概念。不适用于日语文本
fix_punctuation	*flag*	不适用于日语文本
fix_spelling	*flag*	不适用于日语文本
spelling_limit	*integer*	不适用于日语文本
extract_uniterm	*flag*	不适用于日语文本
extract_nonlinguistic	*flag*	不适用于日语文本
upper_case	*flag*	不适用于日语文本
group_names	*flag*	不适用于日语文本

表 9. “文本挖掘”建模节点脚本编制属性 (续)

脚本编制属性	数据类型	属性描述
permutation	*integer*	最大非功能单词排列（缺省值为 3）。不适用于日语文本。
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	仅限日语文本抽取。 0 = 情感辅助抽取 1 = 依赖关系抽取 2 = 未设置辅助分析器。
jp_algorithm_sense_mode	0 1 2	仅限日语文本抽取。 0 = 仅包含 2 = 仅表示 3 = 所有情感。

文本挖掘模型块: TMWBModelApplier

您可以使用下表中的属性进行脚本编制。此块本身称为 TMWBModelApplier。

表 10. 文本挖掘模型块属性

脚本编制属性	数据类型	属性描述
scoring_mode	字段 记录	
field_values	标记 计数	该选项在“类别”模型块中不可用。对于标记，请将其设置为 TRUE 或 FALSE
true_value	*字符串*	对于标记，请将其值定义为 true。
false_value	*字符串*	对于标记，请将其值定义为 false。
extension_concept	*字符串*	指定字段名称的扩展。字段名称是使用概念名称加上此扩展来生成的。使用 add_as 值指定此扩展的放置位置。
extension_category	*字符串*	字段名称扩展。您可以选择为字段名称指定扩展前缀/后缀，或者可以选择使用类别代码。字段名称是使用类别名称加上此扩展来生成的。使用 add_as 值指定此扩展的放置位置。
add_as	后缀 前缀	
fix_punctuation	*标记*	
excluded_subcategories_descriptors	RollUpToParent 忽略	仅适用于类别模型。前提是未选中子类别。该选项允许您指定将如何处理属于未选中用于评分的子类别的描述符。存在两个选项。 • 忽略。“从评分中完全排除其描述符”选项将导致在评分期间忽略并且不使用不具有复选标记（未选中）的子类别的描述符。 • RollUpToParent。“将描述符与父类别中的描述符聚合”将导致将不具有复选标记（未选中）的子类别的描述符用作为父类别（高于此子类别的类别）的描述符。如果有多个级别的子类别未选中，那么描述符将上滚至第一个可用父类别下
check_model	*标记*	在 V14 中不推荐

表 10. 文本挖掘模型块属性 (续)

脚本编制属性	数据类型	属性描述
text	*字段*	
method	ReadText ReadPath	
docType	*整数*	可能值为（0、1、2），其中 0 = 完整文本，1 = 结构化文本 和 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	请注意，具有特殊字符的值（如 "UTF-8"）应加上引号以避免与算术运算符混淆。
language	de en es fr it ja nl pt	

“文本链接分析”节点: textlinkanalysis

您可以使用下表中的参数来通过脚本编制定义或更新节点。该节点本身称为 textlinkanalysis。

要点！无法通过脚本编制指定资源模板。要选择模板，必须在节点对话框中将其选中。

表 11. “文本链接分析 (TLA)”节点脚本编制属性

脚本编制属性	数据类型	属性描述
id_field	*字段*	
text	*field*	
method	ReadText ReadPath	
docType	*integer*	可能的值为 (0,1,2)，其中 0 = 全文本，1 = 结构化文本，2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	请注意，具有特殊字符的值（如 "UTF-8"）应加上引号以避免与算术运算符混淆。

表 11. “文本链接分析 (TLA)”节点脚本编制属性 (续)

脚本编制属性	数据类型	属性描述
unity	*integer*	可能的值为 (0,1)，其中 0 = 段落，1 = 文档
para_min	*integer*	
para_max	*integer*	
mtag	*string*	包含所有 mtag 设置（来自 XML 文件的“设置”对话框）
mclef	*string*	包含所有 mclef 设置（来自结构化文本文件的“设置”对话框）
language	de en es fr it ja nl pt	
concept_count_limit	*integer*	将抽取限于全局频率至少为该值的概念。不适用于日语文本
fix_punctuation	*flag*	不适用于日语文本
fix_spelling	*flag*	不适用于日语文本
spelling_limit	*integer*	不适用于日语文本
extract_uniterm	*flag*	不适用于日语文本
extract_nonlinguistic	*flag*	不适用于日语文本
upper_case	*flag*	不适用于日语文本
group_names	*flag*	不适用于日语文本
permutation	*integer*	最大非功能单词排列（缺省值为 3）。不适用于日语文本。
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	仅限日语文本抽取。 0 = 情感辅助抽取 1 = 依赖关系抽取 2 = 未设置辅助分析器。
jp_algorithm_sense_mode	0 1 2	仅限日语文本抽取。 0 = 仅包含 2 = 仅表示 3 = 所有情感。

翻译节点: translatenode

您可以使用下表中的属性进行脚本编制。该节点本身称为 translatenode。

表 12. 翻译节点属性

脚本编制属性	数据类型	属性描述
text	*字段*	
method	ReadText ReadPath	

表 12. 翻译节点属性 (续)

脚本编制属性	数据类型	属性描述
encoding	Automatic "Big5"、"Big5-HKSCS" 、 "UTF-8"、"UTF-16"、 "US-ASCII"、"Latin1"、 "CP850"、"CP874"、 "CP1250"、"CP1251"、 "CP1252"、"CP1253"、 "CP1254"、"CP1255"、 "CP1256"、"CP1257"、 "CP1258"、"GB18030"、 "GB2312"、"GBK"、 "eucJP"、"JIS7"、 "SHIFT_JIS"、"eucKR"、"TSCII"、 "ucs2"、"KOI8-R"、 "KOI8-U"、"ISO8859-1"、 "ISO8859-2"、"ISO8859-3"、 "ISO8859-4"、"ISO8859-5"、 "ISO8859-6"、"ISO8859-7"、 "ISO8859-8"、"ISO8859-8-i"、 "ISO8859-9"、"ISO8859-10"、 "ISO8859-13"、"ISO8859-14"、 "ISO8859-15"、"IBM 850"、 "IBM 866"、"Apple Roman"、 "TIS-620"	请注意含特殊字符的值（例如，"UTF-8"）应以引号括起以避免与数学运算符混淆
lw_server_type	LOC WAN HTTP	
lw_hostname	*字符串*	
lw_port	*整数*	
url	*字符串*	翻译服务器的 URL
apiKey	*字符串*	
user_id	*字符串*	
lpid	*整数*	如果设置了 *language_from* 或 *language_from_id*，那么不使用该值。
translate_from	Arabic、Chinese、 Traditional Chinese、Czech、 Danish、Dutch、 English、French、 German、Greek、 Hindi、Hungarian、 Italian、Japanese、 Korean、Persian、 Polish、Portuguese、 Romanian、Russian、 Spanish、Somali、 Swedish	

表 12. 翻译节点属性 (续)

脚本编制属性	数据类型	属性描述
translate_from_id	ara、chi、 cht、cze、 dan、dut、 eng、fra、 ger、gre、 hin、hun、 ita、jpn、 kor、per、 pol、por、 rum、rus、 som、spa、 swe	
translate_to	English	
translate_to_id	eng	
translation_accuracy	*整数*	指定您希望翻译流程的准确性级别 - 选择 1 到 3 之间的值
use_previous_translation	*标记*	指定翻译结果已存在（来自先前执行）并且可复用
translation_label	*字符串*	输入标签以识别要复用的翻译结果

第 8 章 交互式工作台模式

从文本挖掘建模节点，您可以在流执行期间选择启动交互式工作台会话。在此工作台中，您可以从文本数据提取关键内容、构建类别、探索文本链接分析模式和聚类以及生成类别模型。在本章中，我们从从包含将处理的主要元素的高级透视图讨论工作台界面，包括:

- **提取结果。**在执行提取后，这些是从文本数据标识和提取的关键字和短语，也称为*概念*。这些概念分组到*类型*。您可以使用这些概念和类型来探索数据以及创建类别。在**类别和概念**视图中进行管理。
- **类别。**使用描述符（例如，提取结果、模式和规则）作为定义，您可以手动或自动创建一组类别，将基于文档和类别是否包含类别定义的一部分将它们分配给这些类别。在**类别和概念**视图中进行管理。
- **聚类。** *聚类*是一组概念，已发现这些概念之间存在可标识它们之间的关系的链接。将使用复杂算法分组概念，算法使用两个概念一起出现的频率与单独出现的频率的比较，以及其他因子。在**聚类**视图中进行管理。您也可以将构成聚类的概念添加到类别。
- **文本链接分析模式。**如果您在语言资源中具有文本链接分析 (TLA) 模式规则或者正在使用已具有某些 TLA 规则的资源模板，那么可以从文本数据提取模式。这些模式可帮助发现数据中概念之间感兴趣的关系。您还可以使用这些模式作为类别中的描述符。在**文本链接分析**视图中进行管理。对于日语文本，必须选择辅助分析器并开启 TLA 提取。
- **语言资源。**提取过程依赖于一组参数和语言定义来监管文本提取和处理。在**资源编辑器**视图中以模板和库形式进行管理。

类别和概念视图

应用程序界面由多个视图组成。您可以在“类别和概念”视图窗口中创建和探索类别以及探索和调整提取结果。*类别*指示一组密切相关的构想和模式，将通过评分处理将文档和记录分配给它们。而*概念*指示可用作类别的构建块（称为描述符）的最级别级别的提取结果。

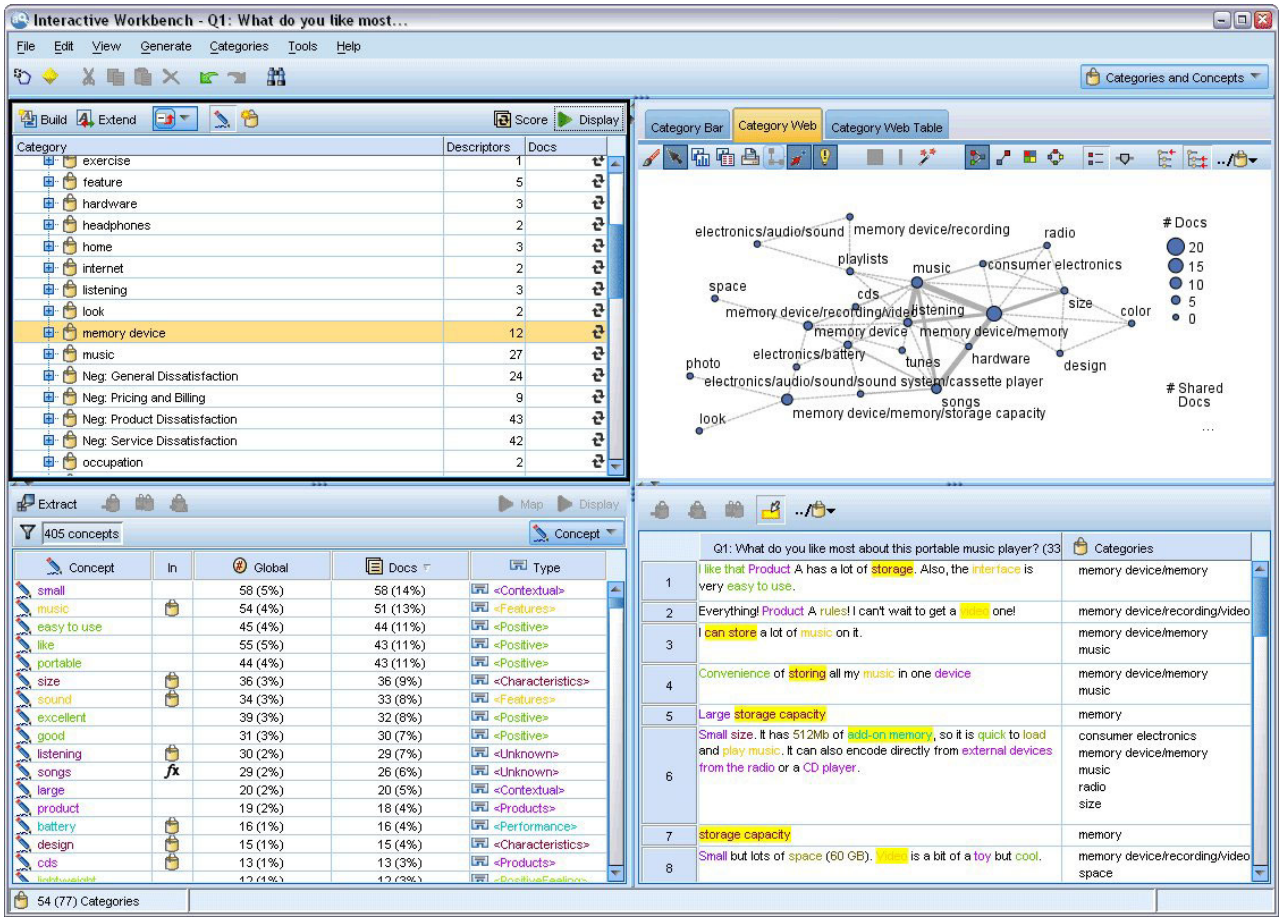


图 23. “类别和概念”视图

“类别和概念”视图由四个窗格组成，可从“视图”菜单中选择其名称以隐藏或显示每个窗格。请参阅主题第 85 页的第 10 章，『对文本数据进行分类』以获取更多信息。

类别窗格

此区域位于左上角，显示一个表，您可以在其中管理生成的任何类别。在从文本数据提取概念和类型后，您可以使用诸如语义网络和概念包含的技术，或者通过手动创建，开始构建类别。如果双击类别名称，那么“类别定义”对话框将打开并显示构成其定义的所有描述符，例如，概念、类型和规则。请参阅主题第 85 页的第 10 章，『对文本数据进行分类』以获取更多信息。并非所有自动化方法都适用于所有语言。

在选择窗格中的行后，您可以在“数据”和“可视化”窗格中显示有关相应的文档/记录或描述符的信息。

提取结果窗格

此区域位于左下角，表示提取结果。在运行提取时，提取引擎读取文本数据、标识相关概念并为每个指定一个类型。**概念**是从文本数据提取的字或短语。**类型**是以类型字典形式存储的概念的语义分组。在提取完成时，概念和类型在“提取结果”窗格中以编码颜色显示。请参阅主题第 73 页的『提取结果：概念和类型』以获取更多信息。

您可通过将鼠标悬停在概念名称上，查看概念的基础术语集合。执行此操作将显示一个工具提示，其中显示概念名称，直到在此概念下分组的多行术语。这些基础术语包括语言资源中定义的同义词（不管是否在文本中发

现了同义词），以及任何抽取的术语复数/单数、轮排术语和进行了模糊分组的术语等。您可通过右键单击概念名称并选择上下文菜单选项，复制这些术语或查看一组完整的基础术语。

文本挖掘是一个迭代式过程，其中根据文本数据的上下文复审提取结果，微调以生成新结果，然后重新评估。可通过修改语言资源来优化提取结果。可以直接在“提取结果”或“数据”窗格中部分完成此微调，但是也可以直接在“资源编辑器”视图中完成。请参阅主题第 67 页的『资源编辑器视图』以获取更多信息。

可视化窗格

此区域位于右上角，显示有关文档/记录分类中共性的多个透视图。每个图形或图表都提供类似的信息，但是以不同的方式表示，或者具有不同级别的详细信息。这些图表和图形可用于分析组织结果，以及帮助微调类别或报告。例如，在图形中，您可能会发现类别太过相似（例如，共享 75% 以上的记录）或者差别过大。图形或图表中的内容对应于其他窗格中的选择。请参阅主题第 133 页的『类别图形和图表』以获取更多信息。

数据窗格

“数据”窗格位于右下角。此窗格显示一个表，其中包含与视图的另一个区域中的选择相对应的文档或记录。根据选择的内容，“数据”窗格中仅显示对应的文本。在进行选择后，单击**显示**按钮以使用相应的文本填充“数据”窗格。

如果在另一个窗格中进行选择，那么相应的文档或记录显示以颜色突出显示的概念，从而便于在文本中识别。您还可以在颜色编码的项上悬停鼠标以显示工具提示，其中显示提取的概念的名称以及指定的类型。请参阅主题第 92 页的『数据窗格』以获取更多信息。

“类别和概念”视图中的搜索和查找

在某些情况下，您可能需要快速查找特定部分中的信息。使用“查找”工具栏，您可以输入想要搜索的字符串，以及定义其他搜索条件，例如，区分大小写或搜索方向。然后，您可以选择想要在其中执行搜索的窗格。

要使用“查找”功能部件

1. 在“类别和概念”视图中，从菜单选择**编辑 > 查找**。“查找”工具栏在“类别”窗格和“可视化”窗格上显示。
2. 在文本框中输入想要搜索的字符串。您可以使用工具栏按钮以控制区分大小写、部分匹配和搜索方向。
3. 在工具栏中，单击想要在其中执行搜索的窗格的名称。如果找到匹配，那么将在窗口中突出显示文本。
4. 要查找文本匹配，请再次单击窗格的名称。

聚类视图

在“聚类”视图中，您可以构建和探索在文本数据中找到的聚类结果。*聚类*是聚类算法基于概念的出现频率以及概念一起出现的频率生成的概念分组。聚类的目标是用来分组在以下情况下一起共同出现的概念：类别的目标是基于包含的文本与每个类别的描述符（概念、规则、模式）的匹配程度分组文档或记录。

在概念在聚类中一起出现的频率越高并且与其他概念一起出现的频率越低的情况下，聚类越容易标识关注的概念关系。当两个概念在相同文档或记录中出现（或者其同义词或术语出现）时，两个概念共现。请参阅主题第 121 页的第 11 章，『分析聚类』以获取更多信息。

您可以构建聚类并在一组图表和图形中进行探索，帮助您发现否则需要消耗大量时间才能找到的概念之间的关系。虽然您无法将整个聚类添加到类别，但是您可以通过“聚类定义”对话框将聚类中的概念添加到类别。请参阅主题第 124 页的『集群定义』以获取更多信息。

您可以对聚类设置进行更改以影响结果。请参阅主题第 122 页的『构建集群』以获取更多信息。

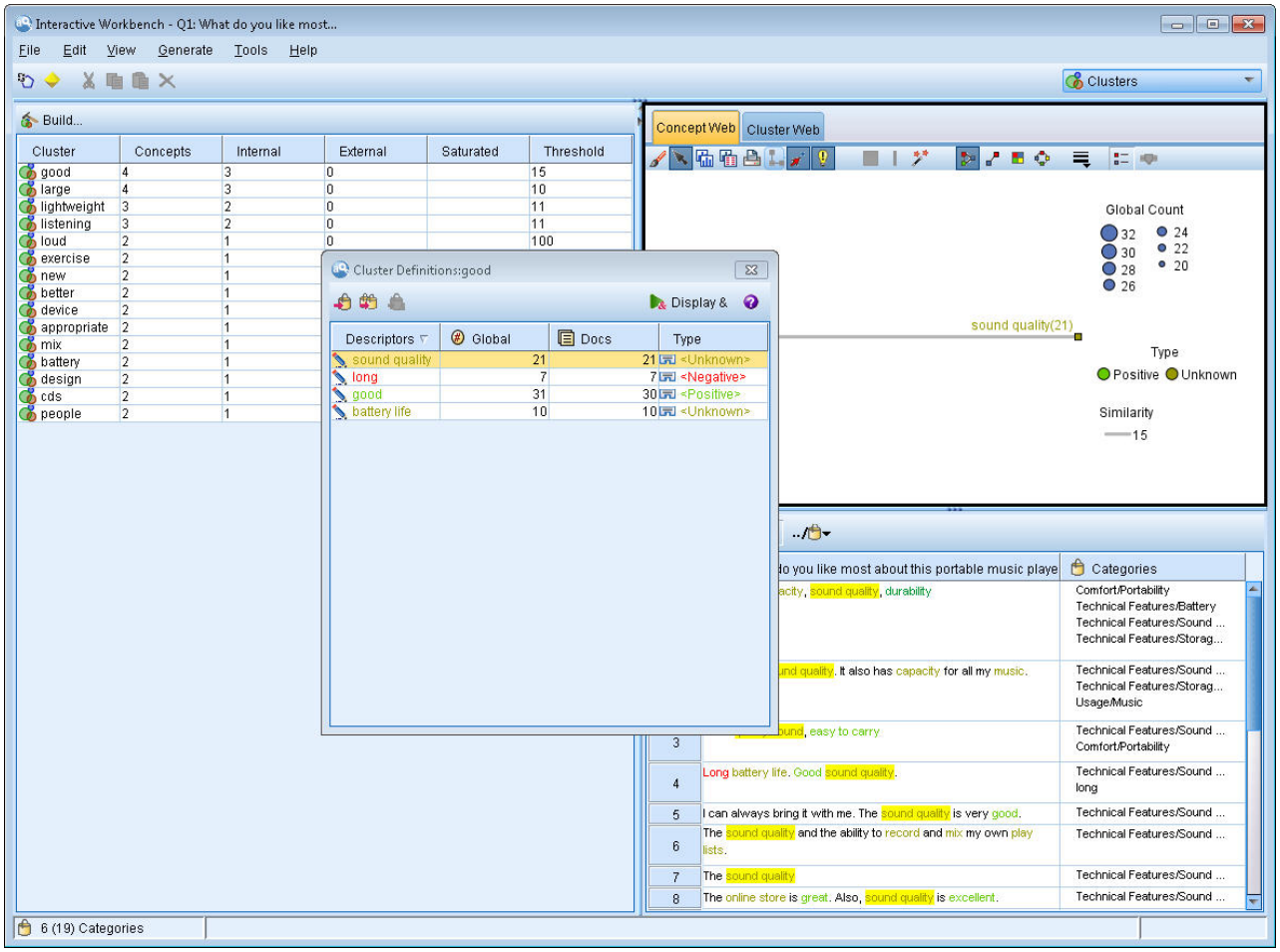


图 24. “聚类”视图

“聚类”视图分为三个窗格，可通过从“视图”菜单中选择其名称来隐藏或显示其中每个窗格。通常，仅显示“聚类”窗格和“可视化”窗格。

聚类窗格

此窗格位于左侧，显示在文本数据中发现的聚类。您可以通过单击**构建**按钮来创建聚类结果。聚类由聚类算法构成，此算法将尝试标识频繁一起发生的概念。

在新提取发生时，将清除聚类结果，并且您必须重建聚类以获取最新结果。在构建聚类时，您可以更改某些设置，例如，要创建的聚类的最大数量、可包含的聚类的最大数量，或者可具有的外部概念的链接的最大数量。请参阅主题第 124 页的『探索集群』以获取更多信息。

可视化窗格

此窗格位于右上角，提供有关聚类的两个透视图：概念 Web 图形和聚类 Web 图形。如果不可见，那么您可以从“视图”菜单访问此窗格（**视图 > 可视化**）。根据聚类窗格中选择的内容，您可以查看聚类之间或其中的相应交互。将以多种格式显示结果：

- **概念 Web**。Web 图形其中显示所选聚类中的所有概念以及聚类之外链接的概念。
- **聚类 Web**。Web 图形显示所选聚类到其他聚类的链接，以及这些其他聚类之间的任何链接。

注: 为了显示“聚类 Web”图形，您必须已构建具有外部链接的聚类。外部链接是不同的聚类中概念对之间的链接（一个聚类中的概念与另一个聚类中的概念）。请参阅主题第 134 页的『聚类图形』以获取更多信息。

数据窗格

“数据”窗格位于右下角，缺省情况下隐藏。您无法在任何“数据”窗格中显示来自于“聚类”窗格的结果，因为这些聚类跨多个文档/记录，使得数据结果无关。但是，您可以在“聚类定义”对话框中查看与选择相对应的数据。根据此对话框中选择的内容，“数据”窗格中仅显示对应的文本。在进行选择后，单击**显示 &** 按钮以使用包含所有概念的文档或记录填充“数据”窗格。

相应的文档或记录显示以颜色突出显示的概念，从而便于在文本中识别。您还可以在颜色编码的项上悬停鼠标以显示提取的概念以及指定的类型。“数据”窗格可包含多个列，但是文本字段列始终显示。其包含提取期间使用的文本字段的名称，或者如果文本数据位于多个不同的文件中，将包含文档名称。其他列可用。请参阅主题第 92 页的『数据窗格』以获取更多信息。

“文本链接分析”视图

在“文本链接分析”视图中，您可以构建和探索在文本数据中找到的文本链接分析模式。文本链接分析 (TLA) 是一种模式匹配技术，支持您定义 TLA 规则并将它们与实际提取的概念以及文本中找到的关系进行比较。

在尝试发现概念之间的关系或者有关特定主题的意见时，模式最有用。某些实例想从调查数据提取有关问题的意见，从医学研究论文中提取基因组关系，或者从情报数据提取人员或位置之间的关系。

在提取 TLA 模式后，您可以在“数据”或“可视化”窗格中进行探索，甚至将它们添加到“类别和概念”视图中的类别。必须在资源模板或库中定义 TLA 规则，这些规则用于提取 TLA 结果。请参阅主题第 183 页的第 19 章，『关于文本链接规则』以获取更多信息。

如果选择提取 TLA 模式结果，那么将在此视图中显示结果。如果未选择执行此操作，那么必须使用**提取**按钮并选择选项以启用模式提取。

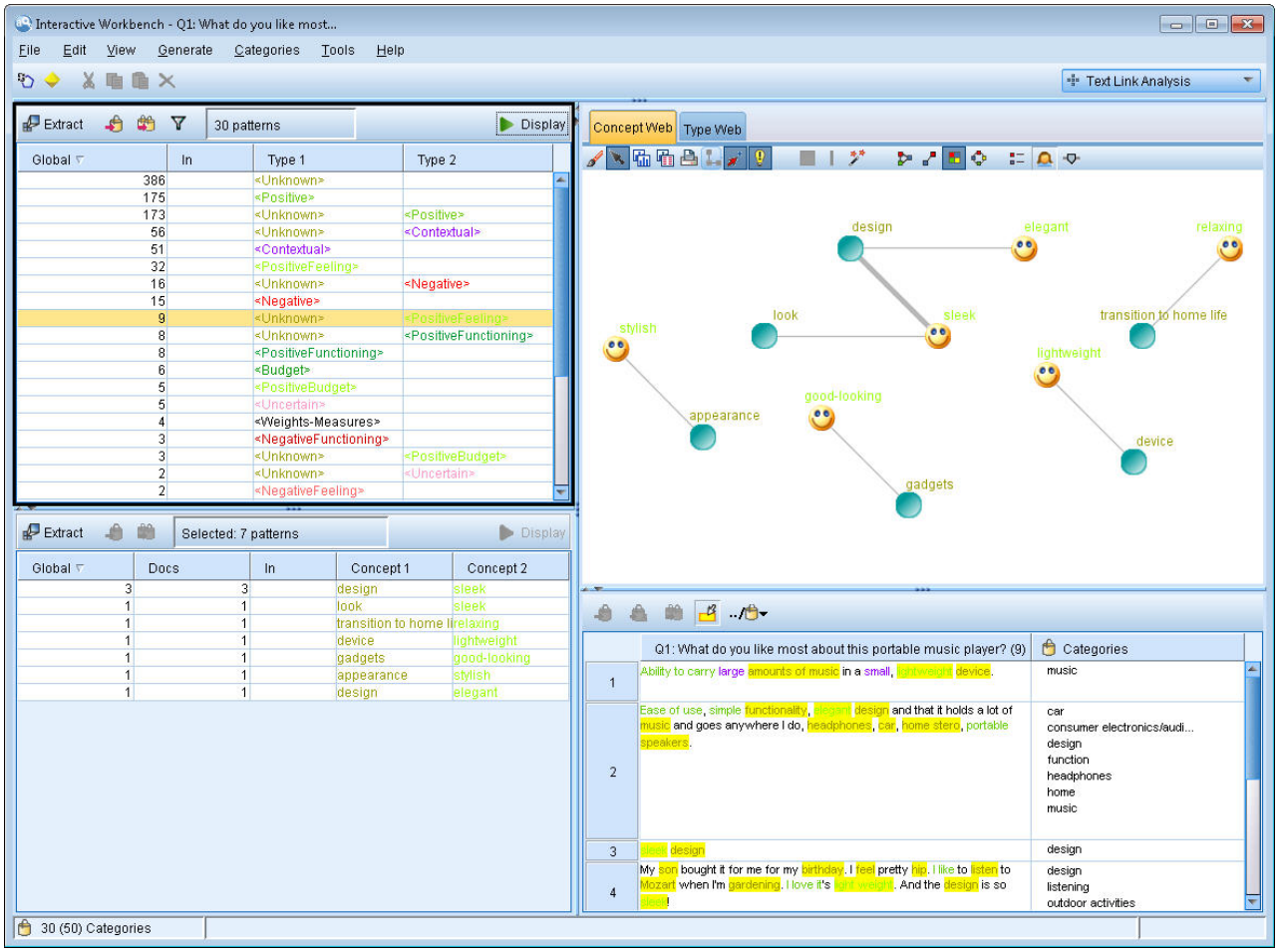


图 25. “文本链接分析”视图

“文本链接分析”视图由四个窗格组成，可从“视图”菜单中选择其名称以隐藏或显示每个窗格。请参阅主题第 127 页的第 12 章，『探索文本链接分析』以获取更多信息。

“类型”和“概念模式”窗格

“类型”和“概念模式”窗格位于左侧，包含两个互连的窗格，您可以在其中探索和选择 TLA 模式结果。模式最多可包含 6 个类型或 6 个概念。请注意，对于日语文本，模式最多只能包含 1 个或 2 个类型或概念。TLA 模式规则（如语言资源中的定义）控制模式规则的复杂性。请参阅主题第 183 页的第 19 章，『关于文本链接规则』以获取更多信息。

模式规则首先在类型级别进行分组，然后划分为概念模式。因此，提供两个不同的结果窗格：“类型模式”（左上方）和“概念模式”（左下方）。

- **类型模式。**“类型模式”窗格显示已提取的模式，其中包含一个或多个匹配 TLA 模式规则的相关类型。类型模式显示为 `<Organization> + <Location> + <Positive>`，这可提供有关组织在特定位置的正反馈。
- **概念模式。**“概念模式”窗格显示在上述“类型模式”窗格中当前选中的所有类型模式的概念级别的提取的模式。概念模式采用以下结构：`hotel + paris + wonderful`。

如“类别和概念”视图中的提取结果一样，您可以在此查看结果。如果看到要对构成这些模式的类型和概念进行任何改进，那么在“类别和概念”视图的“提取结果”窗格中执行更改，或者直接在“资源编辑器”中执行更改，然后重新提取模式。

可视化窗格

此窗格位于“文本链接分析”视图的右上角，将所选模式的 Web 图形显示为类型模式或概念模式。如果不可见，那么您可以从“视图”菜单访问此窗格（**视图 > 可视化**）。根据其他窗格中选择的内容，您可以查看文档/记录以及模式之间的相应交互。

将以多种格式显示结果:

- **概念图形。**此图形表示所选模式中的所有概念。概念图形中的线条宽度和节点大小（如果不显示类型图标）显示所选表中全局出现的数量。
- **类型图形。**此图形表示所选模式中的所有类型。图形中的线条宽度和节点大小（如果不显示类型图标）显示所选表中全局出现的数量。节点由类型颜色或图标表示。

请参阅主题第 135 页的『“文本链接分析”图形』以获取更多信息。

数据窗格

“数据”窗格位于右下角。此窗格显示一个表，其中包含与视图的另一个区域中的选择相对应的文档或记录。根据选择的内容，“数据”窗格中仅显示对应的文本。在进行选择后，单击**显示**按钮以使用相应的文本填充“数据”窗格。

如果在另一个窗格中进行选择，那么相应的文档或记录显示以颜色突出显示的概念，从而便于在文本中识别。您还可以在颜色编码的项上悬停鼠标以显示工具提示，其中显示提取的概念的名称以及指定的类型。请参阅主题第 92 页的『数据窗格』以获取更多信息。

资源编辑器视图

IBM SPSS Modeler Text Analytics 可使用强大的抽取引擎，快速、准确地从文本数据捕获关键概念。此引擎在很大程度上依赖语言资源来指示应该如何分析和解释大量非结构化文本数据。

在资源编辑器视图中，可查看和微调用于抽取概念、按类型对概念进行分组以及在文本数据中发现模式等操作的语言资源。IBM SPSS Modeler Text Analytics 提供了多个预配置资源模板。此外，在一些语言中，还可使用文本分析报中的资源。请参阅主题第 116 页的『使用文本分析包』，以获取更多信息。

由于这些资源可能不会始终完美地适应文本上下文，因此可在资源编辑器中针对特定上下文或域创建、编辑和管理您自己的资源。请参阅主题第 153 页的第 16 章，『处理库』，以获取更多信息。

要简化微调语言资源的过程，可直接通过“抽取结果”和“数据”窗格中的上下文菜单从“类别和概念”视图执行公共字典任务。请参阅主题第 80 页的『优化抽取结果』，以获取更多信息。

注: 针对日语文本进行调整的资源界面会稍有不同。

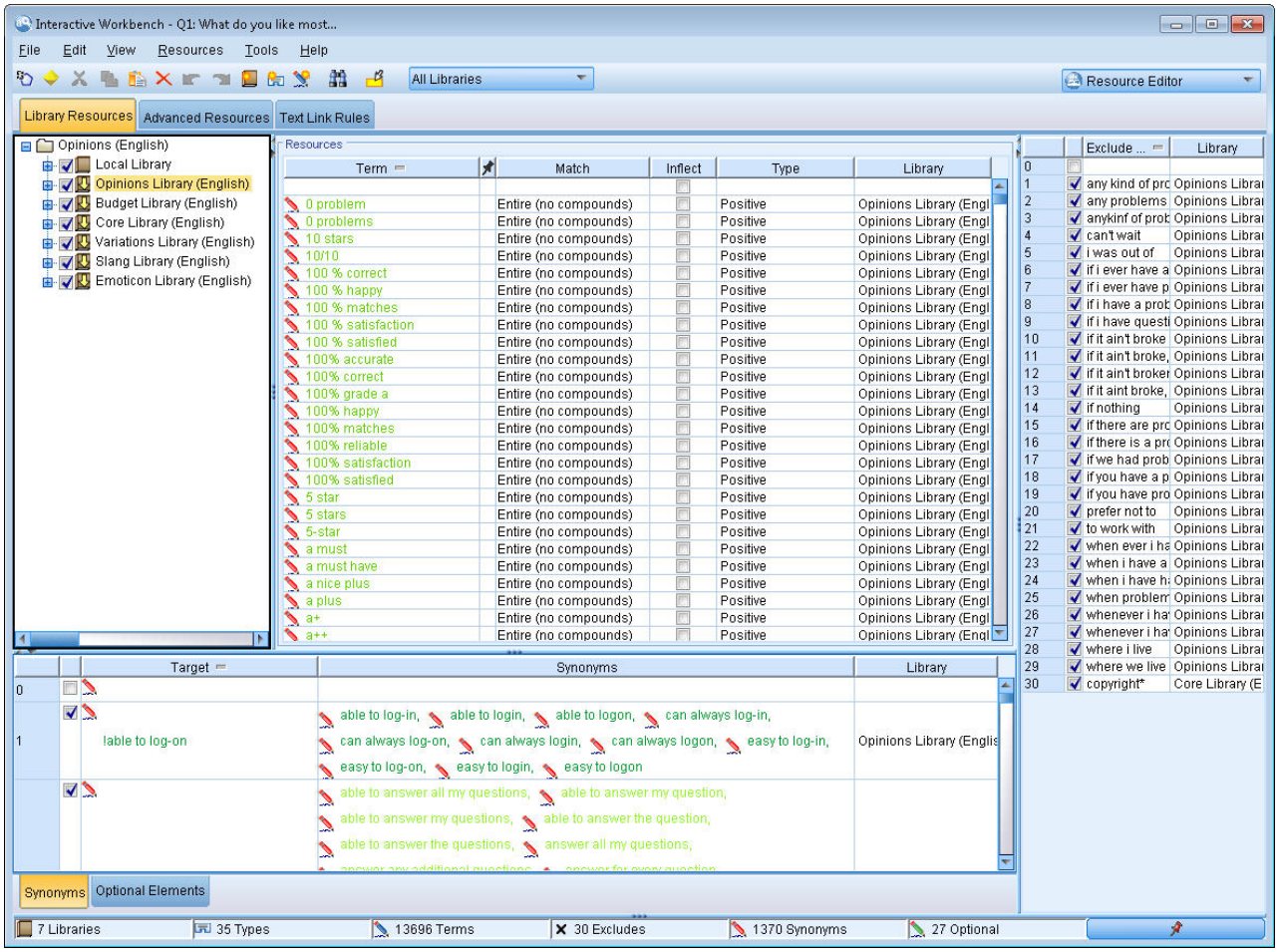


图 26. “资源编辑器”视图

在资源编辑器视图中执行的操作集中于语言资源的管理和微调。这些资源以模板和库的形式进行存储。资源编辑器视图组织为四个部分: “库树”窗格、“类型字典”窗格、“替换字典”窗格和“排除字典”窗格。

注: 请参阅主题第 144 页的『编辑器界面』以获取更多信息。

设置选项

您可以在“选项”对话框中为 IBM SPSS Modeler Text Analytics 设置常规选项。此对话框包含以下选项卡:

- **会话**。此选项卡包含常规选项和定界符。
- **显示**。此选项卡包含界面中使用的颜色选项。
- **声音**。此选项卡包含声音提示选项。

要编辑选项

1. 从菜单，选择**工具 > 选项**。此时“选项”对话框将打开。
2. 选择包含想要更改的信息的选项卡。
3. 更改任何选项。
4. 单击**确定**以保存更改。

选项: “会话”选项卡

在此选项卡上，可以定义一些基本设置。

“数据”窗格和类别图形显示。这些选项会影响数据在“数据”窗格中和在“类别和概念”视图中的“可视化窗格”中的呈现方式。

- **显示“数据”窗格和类别 Web 的限制。**此选项设置要显示的最大文档数，或者用于填充“类别和概念”视图中的“数据”窗格或图形和图表。
- **在显示时显示文档/记录的类别。**如果选定，那么只要单击“显示”便会对文档或记录进行评分，以便可在“数据”窗格中的“类别”列中或在类别图形中显示其所属的任何类别。在某些情况下，尤其对于较大的数据集而言，可能要关闭此选项，以便更快显示数据和图形。

从“数据”窗格添加到类别。这些选项会影响在从“数据”窗格添加文档和记录时添加到类别中的内容。

- **在“类别和概念”视图中，复制。**在此视图中从“数据”窗格添加文档或记录将复制**仅概念**或**概念和模式**。
- **在“文本链接分析”视图中，复制。**在此视图中从“数据”窗格添加文档或记录将复制**仅模式**或**概念和模式**。

资源编辑器定界符。在“资源编辑器”视图中选择要在输入元素（如概念、同义词和可选元素）时用作定界符的字符。

选项: 显示选项卡

在此选项卡上，您可以编辑影响应用程序的整体外观的选项以及用于区分元素的颜色。

注: 要将产品的外观切换到经典外观或先前发行版的外观，请打开 IBM SPSS Modeler 窗口“工具”菜单中的“用户选项”对话框。

定制颜色。编辑屏幕上显示的元素的颜色。对于表中的每个元素，您都可以更改颜色。要指定定制颜色，请单击想要更改的元素右侧的颜色区域，并从下拉颜色列表中选择一种颜色。

- **未提取的文本。**“数据”窗格中还会显示未提取的文本数据。
- **突出显示背景。**在选择窗格中的元素或者“数据”窗格中的文本时文本选择背景颜色。
- **需要提取的背景。**“提取结果”、“模式”和“聚类”窗格的背景色指示已经对库执行的更改并且需要提取。
- **类别反馈背景。**在执行操作后，显示类别背景色。
- **缺省类型。**在“数据”窗格和“提取结果”窗格中显示类型和概念的缺省颜色。此颜色仅应用于在“资源编辑器”中创建的任何定制类型。您可以通过在 资源编辑器 中编辑这些类型字典的属性，针对定制类型字典覆盖此缺省颜色。请参阅主题第 163 页的『创建类型』以获取更多信息。
- **条纹表格 1.** 在“编辑强制”概念对话框的表格中以备用方式使用两种颜色中的第一种，从而区分每一组线。
- **条纹表格 2.** 在“编辑强制”概念对话框的表格中以备用方式使用两种颜色中的第二种，从而区分每一组线。

注: 如果单击**重置为缺省值**按钮，那么此对话框中的所有选项将重置为第一次安装此产品时具有的值。

选项: 声音选项卡

在此选项卡上，您可以编辑影响声音的选项。在“声音事件”下，您可以指定用于在发生事件时通知您的声音。提供大量声音。使用省略号按钮 (**...**) 以浏览并选择声音。用于创建 IBM SPSS Modeler Text Analytics 的声音的 *.wav* 文件存储在安装目录的 *media* 子目录中。如果不想播放声音，请选择**全部静音**。缺省情况下静音。

注: 如果单击**重置为缺省值**按钮，那么此对话框中的所有选项将重置为第一次安装此产品时具有的值。

Microsoft Internet Explorer 帮助设置

Microsoft Internet Explorer 设置

此应用程序中的大多数“帮助”功能都使用基于 Microsoft Internet Explorer 的技术。缺省情况下，某些版本的 Internet Explorer（包括随 Microsoft Windows XP Service Pack 2 一起提供的版本）将拦截本地计算机上 Internet Explorer 窗口中其认为是“活动内容”的项。此缺省设置可能导致拦截“帮助”功能中的某些内容。要访问所有“帮助”内容，您可以更改 Internet Explorer 的缺省行为。

1. 从 Internet Explorer 菜单，选择:

   工具 > Internet 选项...

2. 单击**高级**选项卡。
3. 向下滚动到**安全**部分。
4. 选择（选中）**允许活动内容在“我的电脑”的文件中运行**。

生成模型块和建模节点

在您处于交互式会话中时，您可能想要使用已完成的工作来生成:

- **文本挖掘建模节点。**从交互式工作台会话生成的建模节点是“文本挖掘”节点，其设置和选项反映打开的交互式会话中存储的项。当您不再具有原始“文本挖掘”节点或者想要生成新版本时，这可能非常有用。请参阅主题第 15 页的第 3 章，『挖掘概念和类别』以获取更多信息。
- **类别模型块。**从交互式工作台会话生成的模型块是类别模型块。您必须在“类别和概念”视图中至少具有一个类别才能生成类别模型块。请参阅主题第 33 页的『文本挖掘块: 类别模型』以获取更多信息。

要生成文本挖掘建模节点

1. 从菜单，选择**生成 > 生成建模节点**。使用工作台会话中的所有当前设置，将“文本挖掘”建模节点添加到工作画布。在文本字段后命名节点。

要生成类别模型块

1. 从菜单，选择**生成 > 生成模型**。将使用缺省名称直接在“模型”选用板上生成模型块。

更新建模节点并保存

在使用交互式会话时，建议经常更新建模节点以保存更改。在交互式工作台会话中完成工作和想要保存工作时，也应该更新建模节点。在更新建模节点时，工作台会话内容将保存到启动交互式工作台会话的“文本挖掘”节点。不会关闭输出窗口

重要！此更新不会保存您的流.要保存流,请在更新建模节点后在主 IBM SPSS Modeler 窗口中执行此操作.

要更新建模节点

1. 从菜单，选择**文件 > 更新建模节点**。将使用构建和提取设置以及您具有的任何选项和类别更新建模节点。

关闭和结束会话

在会话中完成工作后，可以通过三种方式离开会话:

- **保存。**通过此选项，可以先将工作保存回到未来会话的起始建模节点中，以及发布任何库以供在其他会话中复用。请参阅主题第 157 页的『共享库』以获取更多信息。保存后，将会关闭会话窗口，并从 IBM SPSS Modeler 窗口中的“输出”管理器删除会话。

- **退出。**此选项将废弃未保存的工作，关闭会话窗口并从 IBM SPSS Modeler 窗口中的“输出”管理器删除会话。要释放内存，建议保存所有重要工作并退出会话。
- **关闭。**此选项将不保存或废弃任何工作。此选项会关闭会话窗口，但是会话将继续运行。可以通过在 IBM SPSS Modeler 窗口中的“输出”管理器中选择此会话再次打开会话窗口。

关闭工作台会话

1. 从菜单中选择**文件 > 关闭**。

键盘辅助功能选项

交互式工作台界面提供键盘快捷键以提高产品的功能可访问性。在最基本的级别，可以按 Alt 键加相应的按键来激活窗口菜单（例如，Alt+F 用于访问“文件”菜单）或按 Tab 间以滚动浏览对话框控件。此部分将覆盖备用导航的键盘快捷键。对于 IBM SPSS Modeler 界面，还存在其他键盘快捷键。

表 13. 通用键盘快捷键

快捷键	功能
Ctrl+1	显示具有多个选项卡的窗格中的第一个选项卡。
Ctrl+2	显示具有多个选项卡的窗格中的第二个选项卡。
Ctrl+A	选择具有焦点的窗格的所有元素。
Ctrl+C	将所选文本复制到剪贴板。
Ctrl+E	在“类别”、“概念”和“文本链接分析”视图中启动抽取。
Ctrl+F	在 资源编辑器/模板编辑器 中显示查找工具栏（如果还不可视）并将焦点置于该处。
Ctrl+I	在“类别”和“概念”视图中，启动所选类别的“类别定义”对话框。在“集群”视图中，启动所选集群的“集群定义”对话框。
Ctrl+R	在 资源编辑器/模板编辑器 中打开“添加术语”对话框。
Ctrl+T	打开“类型属性”对话框以在 资源编辑器/模板编辑器 中创建新类型。
Ctrl+V	粘贴剪贴板内容。
Ctrl+X	从 资源编辑器/模板编辑器 中剪切所选项。
Ctrl+Y	重做视图中的上一项操作。
Ctrl+Z	撤销视图中的上一项操作。
F1	显示帮助，或在处于对话框中时显示项目的上下文帮助。
F2	在表单元格中切换进入和退出编辑方式。
F6	在活动视图中的主要窗格之间移动焦点。
F8	将焦点移至窗格拆分条以调整大小。
F10	展开“文件”主菜单。
向上箭头和向下箭头	在选择拆分条时垂直调整窗格大小。
向左箭头和向右箭头	在选择拆分条时水平调整窗格大小。
Home 和 End	在选择拆分条时将窗格大小调整为最小或最大大小。
Tab	前移经过窗口、窗格或对话框中的项目。
Shift+F10	显示项目的上下文菜单。
Shift+Tab	后移经过窗口或对话框中的项目。
Shift+arrow	在处于编辑方式 (F2) 下时选择编辑字段中的字符。
Ctrl+Tab	将焦点前移至窗口中的下一个主要区域。
Shift+Ctrl+Tab	将焦点后移至窗口中的上一个主要区域。

对话框快捷键

在处理对话框时，多个快捷键和屏幕朗读器按键可有所帮助。在进入对话框时，可能需要按 Tab 键以将焦点置于第一个控件上并启动屏幕朗读器。下表中提供了专用键盘和屏幕朗读器快捷键的完整列表。

表 14. 对话框快捷键

快捷键	功能
Tab	前移经过窗口或对话框中的项目。
Ctrl+Tab	从文本框前移至下一项。
Shift+Tab	后移经过窗口或对话框中的项目。
Shift+Ctrl+Tab	从文本框后移至上一项。
空格条	选择具有焦点的控件或按钮。
Esc	取消更改并关闭对话框。
Enter	验证更改并关闭对话框（与“确定”按钮等效）。如果处于文本框中，那么必须首先按 Ctrl+Tab 以退出该文本框。

第 9 章 提取概念和类型

在执行启动交互式工作台的流时，将自动针对流中的文本数据执行提取。此提取的最终结果是一组概念和类型，而且在此情况下，语言资源模式中存在 TLA 模式。您可以在“提取结果”窗格中查看和处理概念和类型。请参阅主题第 4 页的『抽取的工作方式』以获取更多信息。

如果想要微调提取结果，那么可以修改语言资源并重新提取。请参阅主题第 80 页的『优化抽取结果』以获取更多信息。提取过程依赖于“提取”对话框中的资源和任何参数以指示任何提取和组织结果。如果并非全部，那么您可以使用提取结果来定义类别定义中更适合的部分。

提取结果: 概念和类型

抽取过程期间，将扫描所有文本数据，且会识别和抽取相关概念并将其分配给类型。抽取完成时，会在位于“类别和概念”视图左下角的“抽取结果”窗格中显示结果。首次启动会话时，使用在节点中所选的语言资源模板抽取和组织这些概念和类型。

抽取的概念、类型和 TLA 模型通称为**抽取结果**，它们充当类别的描述符或构建块。还可在类别规则中使用概念、类型和模式。此外，原子方法使用概念和类型来构建类别。

文本挖掘是一个迭代过程，在此过程中，根据文本数据的上下文查看抽取结果，并对其进行微调以生成新结果，然后进行重新评估。抽取后，应该查看结果，并通过修改语言资源进行所需的任何更改。您可以直接从“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框在某种程度上微调资源。请参阅主题第 80 页的『优化抽取结果』，以获取更多信息。 还可直接在资源编辑器视图中执行此操作。请参阅主题第 67 页的『资源编辑器视图』，以获取更多信息。

微调后，可重新抽取以查看新结果。通过从一开始微调抽取结果，可确保每次重新抽取时，将在类别定义中获得相同结果，可完美地适应数据上下文。通过此方式，文档/记录将以更准确和可重复方式分配给类别定义。

概念

抽取过程期间，将扫描和分析文本数据，以便识别文本中感兴趣或相关的单个单词（例如，`election` 或 `peace`）和单词短语（例如，`presidential election`、`election of the president` 或 `peace treaties`）。这些单词和短语通称为*术语*。使用语言资源，将抽取相关术语，然后可将相似术语分组到称为**概念**的引导术语下。

您可通过将鼠标悬停在概念名称上，查看概念的基础术语集合。执行此操作将显示一个工具提示，其中显示概念名称，直到在此概念下分组的多行术语。这些基础术语包括语言资源中定义的同义词（不管是否在文本中发现了同义词），以及任何抽取的术语复数/单数、轮排术语和进行了模糊分组的术语等。您可通过右键单击概念名称并选择上下文菜单选项，复制这些术语或查看一组完整的基础术语。

缺省情况下，概念显示为小写，并根据文档计数（文档列） 按降序排序。抽取概念时，会为其分配类型以帮助分组相似概念。它们将根据此类型进行颜色编码。颜色在资源编辑器中的类型属性中定义。请参阅主题第 161 页的『类型字典』，以获取更多信息。

无论何时在类别定义中使用概念、类型或模式，都会在可排序 **In** 列中显示一个图标。

类型

类型是概念的语义分组。抽取概念时，会为其分配类型以帮助分组相似概念。IBM SPSS Modeler Text Analytics 提供了多种内置类型，例如，<Location>、<Organization>、<Person>、<Positive> 和 <Negative> 等。例如，<Location> 类型对地理关键字和位置进行分组。此类型将分配给诸如 chicago、paris 和 tokyo 的概念。针对多数语言，在任何类型字典中未发现但从文本抽取的概念会自动类型化为<未知>请参阅主题第 162 页的『内置类型』，以获取更多信息。

选择“类型”视图时，缺省情况下，抽取的类型按全局频率以降序显示。 您还可以看到类型进行了颜色编码以帮助区分。颜色属于类型属性。请参阅主题第 163 页的『创建类型』，以获取更多信息。还可创建您自己的类型。

模式

还可从文本数据抽取模式。但是，必须在资源编辑器中具有包含一些文本链接分析 (TLA) 模式规则的库。还必须选择在 IBM SPSS Modeler Text Analytics 节点设置或在“抽取”对话框中使用选项**启用文本链接分析模式抽取**来抽取这些模式。请参阅主题第 127 页的第 12 章，『探索文本链接分析』，以获取更多信息。

抽取数据

无论何时需要执行抽取，“抽取结果”窗格都会变为黄色，且会在此窗格中工具栏下方显示消息**按“抽取”按钮以抽取概念**。

在以下情况下可能会需要抽取：尚不具有任何抽取结果，已更改语言资源，需要更新抽取结果或已打开会话 （在其中未保存抽取结果（**工具 > 选项**）。

注：如果在使用**使用会话工作...** 选项高速缓存抽取结果后更改流的源代码，那么要获取更新的抽取结果，将需要在启动交互式工作台会话后运行新抽取。

运行抽取时，会显示一个进度指示符，以提供抽取状态的反馈。此时间段期间，抽取引擎会读取所有文本数据，并识别和抽取相关术语和模式，将这些术语和模式分配给类型。然后，引擎会尝试在一个称为概念的引导术语下分组同义词术语。完成此过程时，会在“抽取结果”窗格中显示生成的概念、类型和模式。

抽取过程会生成一组概念和类型（例如，启用的文本链接分析 (TLA) 模式）。您可以在“类别和概念”视图的“抽取结果”窗格中查看和使用这些概念和类型。如果已抽取 TLA 模式，那么可在“文本链接分析”视图中看到这些模式。

注：数据集大小与完成抽取过程所需时间相关。您始终可以考虑插入样本节点上游或优化您机器的配置。

抽取数据

1. 从菜单中选择**工具 > 抽取**。或者单击**抽取**工具栏按钮。
2. 如果选择始终显示“抽取设置”对话框，那么会显示此对话框以便您可进行任何更改。请参阅此主题中更多信息以了解每个设置的描述符。
3. 单击**抽取**以启动抽取过程。抽取开始后，会打开进度对话框。抽取后，会在“抽取结果”窗格中显示结果。缺省情况下，概念显示为小写，并根据文档计数（文档列）按降序排序。

您可以使用工具栏选项查看结果，以通过不同方式对结果进行排序、过滤结果或切换到其他视图（概念或类型）。您还可通过使用语言资源优化抽取结果。请参阅主题第 80 页的『优化抽取结果』，以获取更多信息。

针对荷兰语、英语、法语、德语、意大利语、葡萄牙语和西班牙语文本

“抽取设置”对话框包含一些基本抽取选项。

启用“文本链接分析”模式抽取。 指定希望从文本数据抽取 TLA 模式。它还假定您在资源编辑器中其中一个库中具有 TLA 模式规则。此选项可极大地缩短抽取时间。请参阅主题第 127 页的第 12 章，『探索文本链接分析』，以获取更多信息。

调整标点错误。 此选项可在抽取期间临时标准化包含标点错误的文本（例如，不正确使用），以改善概念的可抽取性。当文本很短且质量不佳（例如，在开放式调查响应、电子邮件和 CRM 数据中）时，或文本包含很多缩略词时，此选项非常有用。

调整拼写错误，最小根字符限制为 [n]。 此选项适用于模糊分组方法，此方法可帮助将普遍拼写有误的单词或拼写接近的单词分组到一个概念下。模糊分组算法临时删除抽取单词中的所有元音（除了第一个元音）和出现的二重/三重辅音，然后进行比较，以查看它们是否相同，以便 modeling 和 modelling 分组到一起。但是，如果每个术语分配给不同类型（除了 <Unknown> 类型），那么不会应用模糊分组方法。

您还可先优化最少数目的所需*根*字符，再使用模糊分组。术语中根字符数通过对所有字符相加减去形成屈折变化后缀的任何字符数以及（使用复合单词术语的情况下）限定词和介词数计算得出。例如，术语 exercises 将计算为 8 个根字符（形式为“exercise”），因为字母单词末尾的 *s* 是屈折变化形式（复数形式）。相似地，apple sauce 将计算为 10 个根字符（“apple sauce”，manufacturing of cars 将计算为 16 个根字符（“manufacturing car”）此计数方法仅用于检查是否应该应用模糊分组，但不会影响匹配单词的方式。

注：如果发现某些单词之后分组不正确，那么可通过在“高级资源”选项卡中的**模糊分组：例外**中显式进行声明来从此方法排除单词对。请参阅主题第 175 页的『模糊分组』，以获取更多信息。

抽取单术语。 此选项用于抽取单个单词（单术语），前提是此单词不属于复合单词的一部分，且其为名词或语音的不可识别部分。

抽取非语言实体。 此选项用于抽取非语言实体，例如，电话号码、社保号、时间、日期、货币、数字、百分比、电子邮件地址和 HTTP 地址。您可以在“高级资源”选项卡中的**非语言实体：配置**部分中包含或排除某些类型的非语言实体。通过禁用任何不需要的实体，抽取引擎不会浪费处理时间。请参阅主题第 179 页的『配置』，以获取更多信息。

大写算法。 此选项用于抽取内置字典中不存在的简单和复合术语，前提是术语的第一个字母为大写。此选项提供了一种很好的方式来抽取大部分正确的名词。

尽可能将部分和完整人员姓名分组在一起。 此选项用于将在文本中显示不同的姓名分组在一起。由于通常在文本开头部分通过全名指代姓名，而之后通过较短的版本指代姓名，因此，此功能会很有帮助。此选项尝试将类型为 <Unknown> 的任何单术语与类型为 <Person> 的任何复合术语的最后一个单词匹配。例如，如果发现了 *doe* 且其最初类型为 <Unknown>，那么抽取引擎会检查以了解 <Person> 类型中的任何复合术语是否将 *doe* 作为最后一个单词包含，例如，*john doe*。此选项不适用于名字，因为大多数名字永不会抽取为单术语。

最大非功能单词排列。 此选项指定应用排列方法时可显示的非功能单词的最大数目。此排列方法将仅包含的非功能单词（例如，of 和 the）不同（不考虑屈折变化）的相似短语分组在一起。例如，假设将此值设置为最多两个单词，且抽取了 company officials 和 officials of the company。在此情况下，这两个抽取的术语将在最终概念列表中分组在一起，因为在忽略 of the 时，这两个术语视为相同。

概念地图的索引选项：指定希望在抽取时构建地图索引，以便稍后可以快速绘制概念地图。要编辑索引设置，请单击**设置**。请参阅主题第 79 页的『构建概念映射索引』，以获取更多信息。

开始抽取之前始终显示此对话框。 指定在除了在转至“工具”菜单的情况下未看到“抽取设置”对话框的情况下，是否在每次抽取时希望看到此对话框，或指定是否在每次抽取时希望询问您是否要编辑任何抽取设置。

针对日语文本

“抽取设置”对话框包含一些针对日语文本的基本抽取选项。缺省情况下，此对话框中选择的设置与“文本挖掘”建模节点的“专家”选项卡上选择的设置相同。为了处理日语文本，必须将文本用作输入，并在“文本挖掘”节点的“模型”选项卡中选择日语模板或文本分析包。请参阅主题第 21 页的『从模板和 TAP 复制资源』，以获取更多信息。

辅助分析。 启动抽取时，将使用一组缺省类型执行基本关键字抽取。 但是，选择辅助分析器时，由于抽取器现在将小品词和助动词作为概念的一部分包含，可获取更多或更丰富的概念。 如果进行观点分析，那么还会包含大量其他类型。此外，选择辅助分析器，还可生成文本链接分析结果。

注: 调用辅助分析器时，需要花费更长时间来完成抽取过程。

- **依赖关系分析**。选择此选项可从基本类型和关键字抽取获取抽取概念的扩展小品词。还可从依赖关系文本链接分析 (TLA) 获取更丰富的模式结果。
- **观点分析**。选择此分析器可获取其他抽取的概念，适用时，会执行 TLA 模式的抽取。除了基本类型，还可从超过 80 种观点类型受益。这些类型用于通过表情、观点和意见说明文本中的概念和模式。具有三个选项，这些选项指示观点分析的焦点: **所有观点**、**仅表示观点**和**仅结论**。
- **无辅助分析器**。此选项用于关闭所有辅助分析器。如果选择了选项**启用文本链接分析模式抽取**，那么无法选择此选项，因为需要辅助分析器才能获取 TLA 结果。

启用“文本链接分析”模式抽取。 指定希望从文本数据抽取 TLA 模式。它还假定您在资源编辑器中其中一个库中具有 TLA 模式规则。此选项可极大地缩短抽取时间。此外，必须选择辅助分析器才可抽取 TLA 模式结果。请参阅主题第 127 页的第 12 章，『探索文本链接分析』，以获取更多信息。

过滤提取结果

在处理非常大的数据集时，提取过程可能会生产数百万个结果。对于许多用户，此数量使其难于有效地查看结果。因此，为了放大最关注的内容，您可以通过“提取结果”窗格中提供的“过滤器”对话框来过滤这些结果。

请记住，此“过滤器”对话框中的所有设置将一起用来过滤适用于类别的提取结果。

按频率过滤。您可以过滤以仅显示具有特定全局或文档频率值的结果。

- **全局频率**是概念在整个文档或记录集中出现的总次数，在**全局**列中显示。
- **文档频率**是概念在其中出现的文档或记录的总数，在**文档**列中显示。

例如，如果概念 `nato` 在 500 个记录中出现 800 次，那么可以说此概念的全局频率为 800 并且文档频率为 500。

按类型。您可以进行过滤以仅显示属于特定类型的结果。您可以选择所有类型或仅特定类型。

以及按匹配文本。您还可以进行过滤以仅显示匹配在此处定义的规则的结果。在**匹配文本**字段中输入要匹配的字符集，然后选择要应用匹配的条件。

表 15. 匹配文本条件

条件	描述
包含	如果字符串在任意位置出现，那么文本匹配。（缺省选项）
开始于	仅在概念或类型以指定的文本开始时，文本才匹配。
结束于	仅在概念或类型以指定的文本结束时，文本才匹配。
完全匹配	整个字符串必须匹配概念或类型名称。

按排名。您还可以进行过滤以仅显示根据全局频率（**全局**）或文档频率（**文档**）排名前几位的概念，可以是升序或降序。

提取结果窗格中显示的结果

以下是如何根据过滤器在“提取结果”窗格工具栏中以英语显示结果的一些示例。

表 16. 过滤器反馈的示例

过滤器反馈	描述
	工具栏显示结果数量。因为无文本匹配过滤器并且未满足最大值，不显示其他图标。
	工具栏显示结果限制为过滤器中指定的最大数量，在此情况下为 300。如果显示紫色图标，那么意味着满足最大概念数量。在图标上悬停鼠标以获取更多信息。
	工具栏显示使用匹配文本过滤器限制结果。通过放大镜图标来显示。

要过滤结果

1. 从菜单，选择**工具 > 过滤器**。此时“过滤器”对话框将打开。
2. 选择并优化想要使用的过滤器。
3. 单击**确定**以应用过滤器并在“提取结果”窗格中查看新结果。

探索概念映射

您可以创建一个概念映射以探索概念的相互关联。选择一个概念，然后单击**映射**，这将打开概念映射窗口，可在其中探索与所选概念相关的概念集。您可以通过编辑设置（例如，要包含的类型、要查找的关系的种类等）来过滤显示的概念。

要点：在可以创建映射之前，必须生成索引。这可能需要几分钟时间。但是，在生成索引后，在下一次重新提取前，无需重新生成索引。如果想要在每次提取时自动生成索引，那么可以选提取取设置中的此选项。请参阅主题第 74 页的『 抽取数据 』以获取更多信息。

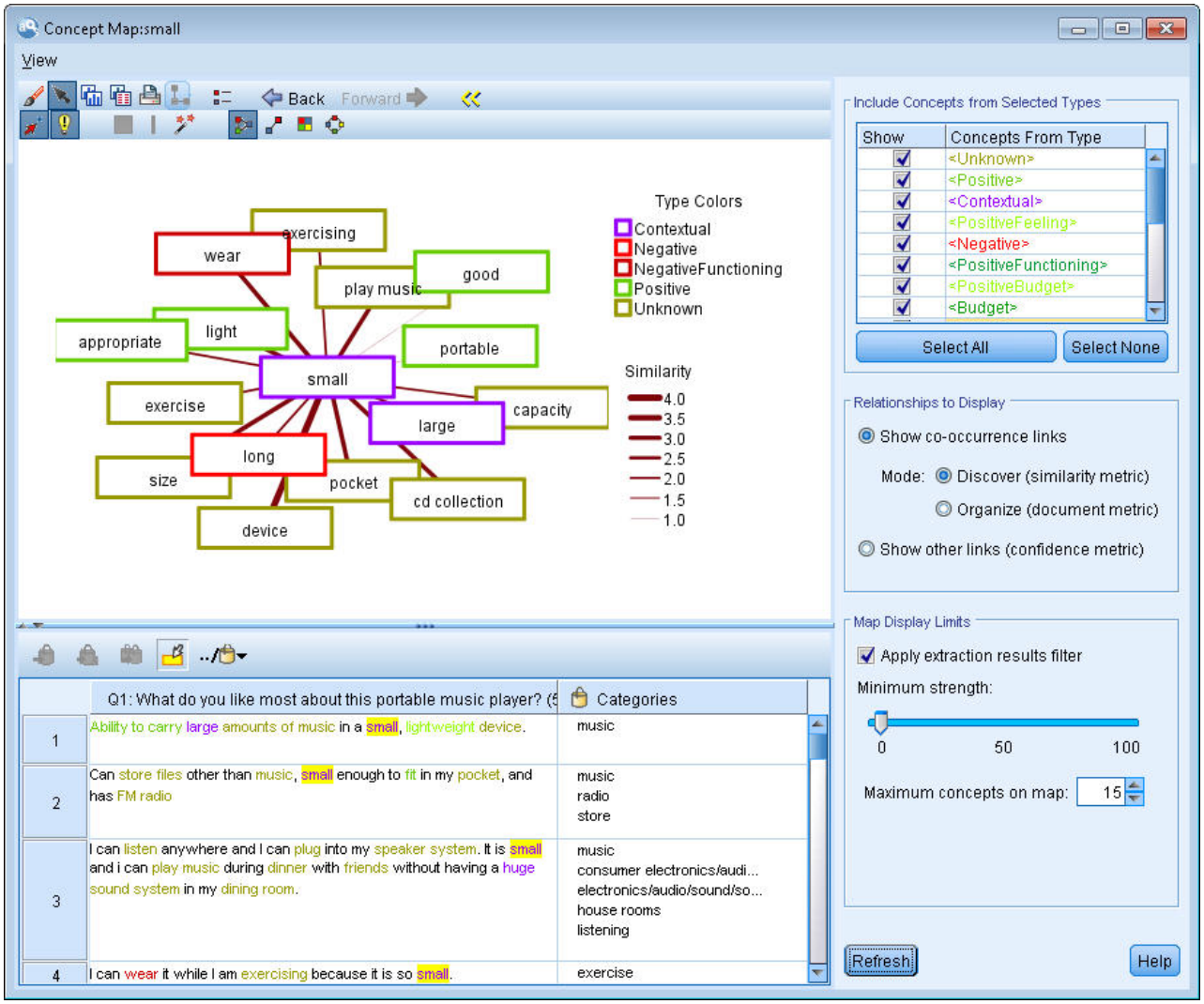


图 27. 选定概念的概念映射

要查看概念映射

1. 在“提取结果”窗格中，选择一个概念。
2. 在此窗格的工具栏中，单击**映射**按钮。如果已生成映射索引，那么概念映射将在一个单独的对话框中打开。如果未生成映射索引或者已过期，那么必须重新构建索引。此过程可能需要几分钟时间。
3. 单击映射周边以进行探索。如果双击一个链接的概念，那么映射将重新绘制自身，并向您显示刚刚双击的概念的链接概念。
4. 顶部工具栏提供一些基本映射工具，例如，移回前一个映射、根据关系强度过滤链接，以及打开过滤器对话框来控制显示的概念的类型和要表示的关系的种类。工具栏的第二行包含图形编辑工具。请参阅主题第 136 页的『使用图形工具栏和选用板』以获取更多信息。
5. 如果不满意找到的链接的种类，那么查看映射右侧显示的此映射的设置。

映射设置：包含来自所选类型的概念

映射中显示仅属于此表中所选类型的概念。要隐藏特定类型的概念，请在表中取消选择此类型。

映射设置: 要显示的关系

显示共现链接 如果想要显示共现链接，请选择模式。模式影响链接强度的计算方式。

- *发现（相似性度量值）*。通过此度量值，使用更复杂的计算来计算链接的强度，考虑两个概念分别出现的频率以及一起出现的频率。高强度值意味着概念对更多趋向于一起出现，而不是分别出现。通过以下公式，将任何浮点值转换为整数。

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

图 28. 相似性系数公式

在此公式中，C_I 是出现概念 I 的文档或记录的数量。

C_J 是出现概念 J 的文档或记录的数量。

C_{IJ} 是文档集中概念对 I 和 J 共现的文档或记录的数量。

- *组织（文档度量值）*。按照共现的原始计数来确定包含此度量值的链接的强度。通常，两个概念的频率越高，那么它们同时一起出现的可能性也越高。高强度值意味着一对概念频率一起出现。

显示其他链接（置信度度量值）。您可以选择要显示的其他链接；这些可以是语义、派生（形态）或包含（句法），而且涉及从链接的概念除去某个概念的步骤数。这些帮助您优化资源，尤其是同义或消除歧义。有关其中每个分组方法的简短描述，请参阅第 95 页的『高级语言设置』

注: 请记住，如果如果在构建索引时未选择这些或者找不到关系，那么不会显示任何项。请参阅主题『构建概念映射索引』以获取更多信息。

映射设置: 映射显示限制

应用提取结果过滤器。如果您不想使用所有概念，那么可以使用提取结果窗格中的过滤器来显示显示的项。然后，选择此选项，并且 IBM SPSS Modeler Text Analytics 将使用此过滤集查找相关概念。请参阅主题第 76 页的『过滤提取结果』以获取更多信息。

最小强度。在此设置最小链接强度。映射中隐藏关系强度低于此限制的任何相关概念。

映射上的最大概念数。指定要在映射上显示的关系的最大数量。

构建概念映射索引

在可以创建映射之前，必须生成概念关系的索引。在创建概念映射时，IBM SPSS Modeler Text Analytics 引用此索引。您可以通过选择此对话框中的方法来选择要建立索引的关系。

分组方法。选择一个或多个方法。有关其中每个方法的简要描述，请参阅第 97 页的『有关语言方法』。并非所有方法都适用于所有文本语言。

防止特定概念的配对。选择此复选框以停止在输出中将两个概念分组在一起或配对的过程。要创建或管理概念对，请单击**管理对**。请参阅主题第 97 页的『管理链接异常对』以获取更多信息。

构建索引可能需要几分钟时间。但是，在生成索引后，在下一次重新提取前，无需重新生成索引，或者除非想要更改设置以包含更多关系。如果想要在提取时生成索引，您可以在提取设置中选择此选项。请参阅主题第 74 页的『抽取数据』以获取更多信息。

优化抽取结果

抽取是迭代式过程，您可以在其中抽取、复审结果、对其进行修改，然后重新抽取以更新结果。由于准确性和连续性是成功的文本挖掘和分类的基本要素，请从开始时就优化您的抽取结果，以确保您每次重新抽取时，都将在类别定义中获得完全相同的结果。通过这种方式，将可以更准确且可重复的方式将记录和文档分配到您的目录中。

抽取结果充当类别的构建块。使用这些抽取结果创建类别时，如果记录和文档包含与一个或多个类别描述符匹配的文本，那么会自动将其分配到这些类别。虽然您可以在对语言资源进行任何优化之前开始分类，但在开始之前至少复查一次您的抽取结果是很有帮助的。

在您复查自己的结果时，可能会发现您希望抽取引擎以其他方式来处理的某些元素。请考虑以下示例:

- **无法识别的同义词。**假定您找到多个您视为同义词的概念（例如，smart（智慧）、intelligent（智能）、bright（机智）和 knowledgeable（博学）），并且这些概念在抽取结果中均显示为单独的概念。那么您可以创建一个同义词定义，在其中将 intelligent、bright 和 knowledgeable 都分组到目标概念 smart 下。这样将把所有这些概念与 smart 分组在一起，全局频率计数也将更高。请参阅『添加同义词』主题以获取更多信息。
- **错误输入的概念。**假定您的抽取结果中的概念显示在某个类型中，而您想要将其分配到另一个类型。在另一个示例中，假设您在自己的抽取结果中找到 15 个蔬菜概念，您想要将其全部添加到称为 <Vegetable>（蔬菜）的新类型中。针对多数语言，在任何类型字典中未发现但从文本抽取的概念会自动类型化为<未知>您可以将概念添加到类型。请参阅第 81 页的『将概念添加到类型』主题以获取更多信息。
- **无关紧要的概念。**假设您发现抽取的某个概念的频率计数较高 - 即在许多记录或文档中均可找到此概念。但是，您认为此概念对您的分析无关紧要。您可以从抽取中将其排除。请参阅第 82 页的『从抽取中排除概念』主题以获取更多信息。
- **错误匹配。**假定在复查包含某个概念的记录或文档时，您发现有两个词被错误分组在一起，例如，faculty（教职员工）和 facility（设施）。此匹配可能是由于称为模糊分组的内部算法所导致的，此算法会临时忽略双辅音或三辅音和元音，以将常见拼写错误分组在一起。您可以将这些词添加到不应分组在一起的词对列表中。请参阅第 175 页的『模糊分组』主题以获取更多信息。模糊分组不可用于日语文本。
- **未抽取的概念。**假设您期望查找某些已抽取的概念，但是在复查记录或文档文本时注意到某些词或短语未抽取。通常这些文字是您不感兴趣的动词或形容词。但有时您希望使用未抽取的一个文字或短语作为类别定义的一部分。要抽取此概念，您可以强制将术语添加到类型字典中。请参阅第 83 页的『强制将文字添加到抽取中』主题以获取更多信息。

可以从“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框中选择一个或多个元素并右键单击以访问上下文菜单来直接执行其中许多更改。

完成更改之后，窗格背景色会更改以显示您需要重新抽取以查看更改。请参阅第 74 页的『抽取数据』主题以获取更多信息。如果您处理较大的数据集，在进行多项更改之后重新抽取可能比在每次更改之后进行重新抽取更有效。

注: 您可以在资源编辑器视图（视图 > 资源编辑器）中查看用于生成抽取结果的可编辑语言资源的完整集合。这些资源会以库和字典的形式显示在此视图中。您可以在库和字典中直接定制这些概念和类型。请参阅第 153 页的第 16 章，『处理库』主题以获取更多信息。

添加同义词

*同义词*用于将具有相同含义的两个或更多个词关联在一起。同义词通常还用于将术语与其缩写分组在一起，或者将经常拼错的词与其正确拼写分组在一起。通过使用同义词，目标概念的频率更高，使其更易于发现文本数据中以不同方式显示的类似信息。

随本产品交付的语言资源模板和库包含许多预定义的同义词。但是，如果您发现未识别的同义词，可以对其进行定义以便在下次抽取时可以识别这些同义词。

第一步是确定目标概念或线索概念。*目标概念*是一个文字或短语，您可将最终结果中所有同义词术语分组到目标概念下。抽取期间，同义词会分组到此目标概念下。第二步是识别此概念的所有同义词。在最终抽取中将使用所有同义词替换此目标概念。必须抽取术语才能使其成为同义词。但是，无需抽取目标概念即可发生替换。例如，如果希望将 `intelligent`（智能）替换为 `smart`（智慧），那么 `intelligent` 是同义词，`smart` 是目标概念。

如果您创建新的同义词定义，那么会将新的目标概念添加到字典中。随后，您必须将同义词添加到此目标概念。只要您创建或编辑同义词，就会在资源编辑器的同义词字典中记录这些更改。如果您要查看这些同义词字典的完整内容，或者进行大量更改，可能更愿意在资源编辑器中直接完成这些工作。请参阅第 167 页的『替换/同义词字典』主题以获取更多信息。

任何新的同义词都将自动存储在资源编辑器视图的库树中列出的第一个库中，缺省情况下为*本地库*。

注: 如果您查找同义词定义但是通过上下文菜单或者在资源编辑器中直接查找时无法找到此定义，那么通过内部模糊分组技术可能已生成匹配。请参阅第 175 页的『模糊分组』主题以获取更多信息。

要创建新同义词

1. 在“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框中，选择要为其创建新同义词的概念。
2. 从菜单中选择**编辑 > 添加到同义词 > 新建**。这样会打开“创建同义词”对话框。
3. 在“目标文本”框中输入目标概念。所有同义词都将分组到此概念下。
4. 如果要添加更多同义词，请在“同义词”列表框下输入这些同义词。使用全局分隔符来分隔每个同义词术语。请参阅第 69 页的『选项: “会话”选项卡』主题以获取更多信息。
5. 如果处理日语文本，请通过选择**来自类型的同义词**字段中的类型名称来指定这些同义词的类型。但是，目标采用抽取期间分配的类型。但是，如果目标未抽取为概念，那么此列中所列出的类型会分配给抽取结果中的目标。
6. 单击**确定**以应用更改。这样会关闭此对话框，“抽取结果”窗格背景色会发生更改，以指示您需要重新抽取才能查看更改。如果有几项更改需要完成，请在重新抽取之前完成更改。

要添加到同义词

1. 在“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框中，选择要添加到现有同义词定义的概念。
2. 从菜单中选择**编辑 > 添加到同义词**。此菜单会显示一组同义词，最近创建的同义词列在列表顶部。选择要将所选概念添加到的同义词名称。如果您看到正在查找的同义词，请将其选中，这样会将所选概念添加到此同义词定义。如果您未看到此同义词，请选择**更多**以显示“所有同义词”对话框。
3. 在“所有同义词”对话框中，您可以按自然排序顺序（创建顺序）或者按升序或降序来对列表进行排序。选择要将所选概念添加到其中的同义词的名称，然后单击**确定**。这样会关闭此对话框，并将概念添加到同义词定义。

将概念添加到类型

运行抽取时，抽取的概念会分配到类型中以将具有共通点的术语分组在一起。 IBM SPSS Modeler Text Analytics 交付了许多内置类型。请参阅第 162 页的『内置类型』主题以获取更多信息。针对多数语言，在任何类型字典中未发现但从文本抽取的概念会自动类型化为<未知>

复查结果时，您可能会发现某些概念显示在某一个类型中，而您想将这些概念分配到另一个类型中，或者您可能会发现一组文字实际上本身属于一个新的类型。在这些情况下，您可能想要将这些概念重新分配到另一个类型或者彻底创建一个新类型。您无法为日语文本创建新类型。

例如，假定您正在处理与汽车有关的调查数据，并且您想要按不同领域的车辆来进行分类。您可以创建一个称为 <Dashboard>（仪表盘）的类型以将与车辆仪表盘上找到的仪表和旋钮相关的所有概念分组在一起。然后，可以将诸如 gas gauge（油量表）、heater（暖气）、radio（无线电）和 odometer（里程表）添加到此新类别。

在另一个示例中，假定您正在处理与高等院校相关的调查数据，并且抽取类型为 Johns Hopkins（大学名称）作为 <Person>（人员）类型而不是作为 <Organization>（组织）类型。在此情况下，您可以将此概念添加到 <Organization>（组织）类型中。

当您创建类型或者将概念添加到类型的术语列表中时，在资源编辑器中您的语言资源库中的类型字典中会记录这些更改。如果您要查看这些库的内容，或者进行大量更改，可能更愿意在资源编辑器中直接完成这些工作。请参阅第 164 页的『添加术语』主题以获取更多信息。

要将概念添加到类型

1. 在“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框中，选择要添加到现有类型的概念。
2. 右键单击以打开上下文菜单。
3. 从菜单中选择**编辑 > 添加到类型**。此菜单会显示一组类型，最近创建的类型列在列表顶部。选择要将所选概念添加到的类型名称。如果您看到正在查找的类型名称，请将其选中，这样会将所选概念添加到此类型。如果您未看到类型名称，请选择**更多**以显示“所有类型”对话框。
4. 在“所有类型”对话框中，您可以按自然排序（创建顺序）或者按升序或降序来对列表进行排序。选择要将所选概念添加到其中的类型的名称，然后单击**确定**。这样会关闭此对话框，并将其作为术语添加到此类型中。

注: 对于日语文本，在某些实例中，更改术语类型不会更改最终抽取列表中最终要将此术语分配到的类型。这是由于对于某些基本术语，在抽取期间内部字典的优先顺序更高。

要创建新类型

1. 在“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框中，选择要为其创建新类型的概念。
2. 从菜单中选择**编辑 > 添加到类型 > 新建**。这样会打开“类型属性”对话框。
3. 在“名称”文本框中输入此类型的新名词，并对其他字段进行任何更改。请参阅第 163 页的『创建类型』主题以获取更多信息。
4. 单击**确定**以应用更改。这样会关闭此对话框，“抽取结果”窗格背景色会发生更改，以指示您需要重新抽取才能查看更改。如果有几项更改需要完成，请在重新抽取之前完成更改。

从抽取中排除概念

查看结果时，您可能偶尔会发现您不想要抽取的概念或者供任何自动类别构建技术所使用的概念。在某些情况下，这些概念具有极高的频率计数，完全对您的分析无关紧要。在此情况下，您可以将概念标记为从最终抽取中排除。通常，您添加到此列表中的概念为文本中用于连续性的填充词或短语，但是不增加任何重要内容，并且可能造成抽取结果混乱。通过将概念添加到排除字典中，可以确保从不抽取这些概念。

通过排除概念，在下次抽取时，抽取结果中将不显示所有已派出的概念的变体。如果在类别中此概念已显示为一个描述符，那么重新抽取之后它将保留在类别中并且计数为零。

排除时，会在资源编辑器的排除字典中记录这些更改。如果您要查看所有排除定义并直接对其进行编辑，可能更愿意在资源编辑器中直接完成这些工作。请参阅第 170 页的『排除字典』主题以获取更多信息。

注: 对于日语文本，在某些实例中排除术语或类型将不会导致将其排除。这是由于对于日语资源的某些基本术语，在抽取期间内部字典的优先顺序更高。

要排除概念

1. 在“抽取结果”窗格、“数据”窗格、“类别定义”对话框或“集群定义”对话框中，选择要从抽取中排除的概念。
2. 右键单击以打开上下文菜单。
3. 选择**从抽取中排除**。这样会将此概念添加到资源编辑器的排除字典中，“抽取结果”窗格背景色会发生更改，以指示您需要重新抽取才能查看更改。如果有几项更改需要完成，请在重新抽取之前完成更改。

注: 您排除的任何文字都将自动存储在资源编辑器的库树中列出的第一个库中，缺省情况下为*本地库*。

强制将文字添加到抽取中

抽取后在“数据”窗格中查看文本数据时，可能会发现未抽取某些文字或短语。通常这些文字是您不感兴趣的动词或形容词。但有时您希望使用未抽取的一个文字或短语作为类别定义的一部分。

如果您要抽取这些文字和短语，可以强制将某个术语添加到类型库中。请参阅第 166 页的『强制术语』主题以获取更多信息。

要点！将字典中的某个术语标记为强制并非万无一失。这表示即使您已显式将某个术语添加到字典中，有时在重新抽取后的“抽取结果”窗格中仍不显示此术语，或者虽然显示此术语但是与您声明的术语不完全相同。虽然这种情况很罕见，但当已将某个文字或短语作为较长的短语的一部分来抽取时，可能会发生这种情况。为防止出现这种情况，请在类型字典中将**完整（非复合）**匹配选项添加到此术语。请参阅第 164 页的『添加术语』主题以获取更多信息。

第 10 章 对文本数据进行分类

在“类别和概念”视图中，您可以创建**类别**，这些类别实质上表示用于捕获以文本表示的关键构想、知识和看法的较高级别概念或主题。

作为 IBM SPSS Modeler Text Analytics 14 发行版的一部分，类别还可包含分层结构，这表示其中可包含子类别，这些子类别还可包含其自己的子类别，以此类推。您可以将预定义的类别结构（原称为代码帧）导入分层式类别，并在该产品内构建这些分层式类别。

实际上，分层式类别支持您利用一个或多个子类别来构建树结构，用于更准确地对项目（例如，不同概念或主题领域）进行分组。以下是一个与休闲活动相关的简单示例；如果要回答诸如*如果您有更多时间，那么会想要参加哪些活动？*，您可能会列出的首要类别包括*体育*、*艺术与工艺品*、*钓鱼*等；向下一级，在*体育*下，可能包括*球类运动*、*水上相关运动*等子类别。

类别由一组描述符组成，例如，*概念*、*类型*、*模式*和*类别规则*。通过将这些描述符结合在一起可用于识别某个文档或记录是否属于给定类别。可通过扫描文档或记录中的文本来查看是否有任何文本与描述符匹配。如果找到匹配，那么会将此文档/记录分配至此类别。该过程称为**分类**。

您可以使用“类别和概念”视图的四个窗格中显示的数据来处理、构建和直观探索您的类别，可通过选择“视图”菜单中各窗格的名称来隐藏或显示这些窗格。

- **“类别”窗格。**在此窗格中构建和管理您的类别。请参阅第 86 页的『“类别”窗格』主题以获取更多信息。
- **“抽取结果”窗格。**在此窗格中探索和处理抽取的概念和类型。请参阅第 73 页的『提取结果：概念和类型』主题以获取更多信息。
- **“可视化”窗格。**在此窗格中直观探索您的类别及其交互方式。请参阅第 133 页的『类别图形和图表』主题以获取更多信息。
- **“数据”窗格。**在此窗格中探索和查看对应于选项的文档和记录中包含的文本。请参阅第 92 页的『数据窗格』主题以获取更多信息。

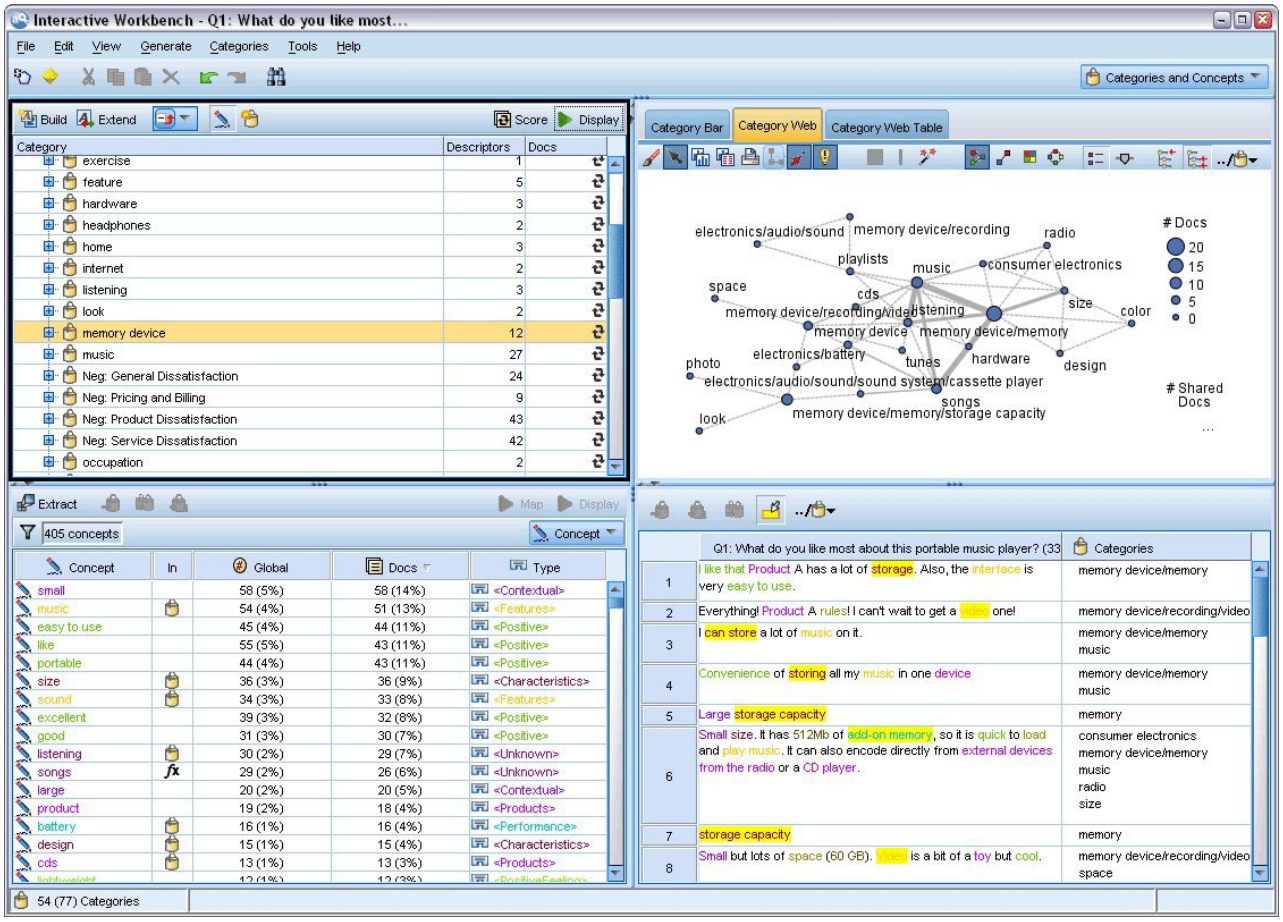


图 29. “类别和概念”视图

虽然您可以首先使用来自文本分析包 (TAP) 的一组类别或者从预定义的类别文件导入类别，但您还可能需要创建自己的类别。可以使用本产品健全的自动化技术集来自动创建类别，这些技术使用抽取结果（概念、类型和模式）来生成类别及其描述符。还可以使用您所具有的有关数据的其他洞察来手动创建类别。但是，只能手动创建类别或者通过交互式工作台来对类别进行微调。请参阅第 19 页的『“文本挖掘”节点：“模型”选项卡』主题以获取更多信息。您可以通过将抽取结果拖放到类别中来手动创建类别定义。您可以通过以下方式来丰富这些类别或任何空类别：将类别规则添加到类别中、使用自己的预定义类别或采用上述方式的组合。

每一种技术和方法都适合某种类型的数据和情境，但在同一份分析中组合多种技术以捕获完整范围的文档或记录是很有帮助的。在分类过程中，您可能会发现需要对语言资源进行其他更改。

“类别”窗格

“类别”窗格是您在其中构建和管理自己的类别的区域。此窗格位于“类别和概念”视图左上角。在从文本数据中抽取概念和类型之后，可以开始使用各种技术（例如，概念包含、共现等）自动或手动构建类别。请参阅第 93 页的『构建类别』主题以获取更多信息。

每次创建或更新类别时，都可通过单击**评分**按钮来评分文档或记录，以查看是否有任何文本与给定类别中的描述符匹配。如果找到匹配，那么会将此文档或记录分配至此类别。最终结果是基于类别中的描述符将大部分（或全部）文档或记录分配到各类别中。

类别树表格

此窗格中的树表格显示了一组类别、子类别和描述符。此树还包含多个列，用于显示每个树项的信息。以下列可供显示:

- **代码。**列出每个类别的代码值。缺省情况下隐藏该列。您可以通过以下菜单显示该列: **视图 > 类别窗格**。
- **类别。**包含类别树，其中显示了类别和子类别的名称。此外，如果单击描述符工具栏图标，那么还将显示一组描述符。
- **描述符。**提供组成描述符定义的描述符数量。此计数不包括子类别中的描述符数量。当在**类别**列中显示描述符名称时不提供计数。 您可以通过菜单在树中显示或隐藏描述符本身: **视图 > 类别窗格 > 所有描述符**。
- **文档。**评分后，此列会提供分类到某个类别及其所有子类别中的文档或记录的数量。如果基于记录描述符有 5 条记录与其主要类别相匹配，并且基于记录描述符另有 7 条记录与子目录相匹配，那么主要类别的文档总数为两者之和 - 在此情况下为 12。但是，如果同一条记录与主要目录及其子目录都匹配，那么计数为 11。

不存在类别时，该表格仍包含两个行。主要的行称为**所有文档**，它是文档或记录总数。第二个行称为**未分类**，其中显示了尚未分类的文档/记录数量。

对于窗格中每个类别，在类别名称之前会带有一个小型黄色桶状图标。如果您双击某个类别，或选择菜单中的**视图 > 类别定义**，那么会打开“类别定义”对话框并显示组成其定义的所有元素（称为**描述符**），例如概念、类型、模式和类别规则。请参阅第 91 页的『有关类别』主题以获取更多信息。缺省情况下，类别树表格不显示类别中的描述符。如果要在树种直接查看描述符而不是在“类别定义”对话框中查看，请单击工具栏中画笔图标所表示的切换按钮。选中此切换按钮时，您也可以展开自己的树以查看描述符。

评分类别

文档。类别树表格中的列会显示已分类到特定类别中的文档或记录的数量。如果此数量已过期或者尚未计算，那么会在该列中显示一个图标。您可以单击窗格工具栏上的**评分**来重新计算文档数量。切记，处理较大的数据集时，评分过程需要一些时间。

选择树中的类别

在树中进行选择时，只能选择同代类别，即如果选择顶级类别，那么不能同时选择子类别。或者，如果选择给定类别的 2 个子类别，那么不能同时选择另一个类别的子类别。选择不连续的类别将导致丢失先前的选择。

在“数据”和“可视化”窗格中显示

选中表格中的某个行时，可以单击**显示**按钮以使用对应于您的选择的信息来刷新“可视化”和“数据”窗格。如果某个窗格不可见，那么单击**显示**将导致显示此窗格。

优化您的类别

首次尝试时，分类可能无法为您的数据生成完美的结果，您可能想要删除某些类别或者将某些类别与其他类别加以组合。通过查看抽取结果，您还会发现某些未创建的类别可能很有用。如果是这样，您可以对结果进行手动更改，以针对您的特定上下文对结果进行微调。请参阅第 119 页的『编辑和优化类别』主题以获取更多信息。

用于创建类别的方法和策略

如果尚未抽取或抽取结果到期，那么使用其中一个类别构建或扩展方法都会提示您进行自动抽取。应用方法后，分组到一个类别的概念和类型仍可用于使用其他方法构建的类别。这表示您可能会看到一个概念处于多个类别中，除非选择不复用这些类别。

为了帮助创建最佳类别，请查看以下内容:

- 用于创建类别的方法
- 用于创建类别的策略
- 用于创建类别的提示

用于创建类别的方法

由于每个数据集唯一，类别创建方法数和应用这些方法的顺序会随着时间发生变化。此外，由于文本挖掘目标在不同数据组中可能会不同，因此可能需要试验不同方法，以了解哪种方法可针对给定文本数据产生最佳效果。自动方法都无法对数据进行完美分类；因此，建议寻求和应用非常适用于您的数据的一种或多种自动方法。

除了将文本分析包（TAP，**.tap*）用于预构建类别集，还可使用以下任何方法组合对响应进行分类:

- **自动构建方法。** 提供了多个基于语言和基于频率的类别选项来自动为您构建类别。请参阅主题第 93 页的『构建类别』，以获取更多信息。
- **自动扩展方法。** 提供了多个语言方法，通过添加和增强描述符以便其可捕获更多记录，来扩展现有类别。请参阅主题第 101 页的『扩展类别』，以获取更多信息。
- **手动方法。** 提供了多种手动方法，例如，拖放。请参阅主题第 104 页的『手动创建类别』，以获取更多信息。

用于创建类别的策略

以下策略列表并不详尽，但可为您提供一些关于如何开始构建类别的构想。

- 定义“文本挖掘”节点时，从文本分析包 (TAP) 选择类别集，以便使用一些预构建类别开始分析。这些类别可能足以从一开始对文本进行分类。但是，如果要添加更多类别，那么可编辑“构建类别”设置（**类别 > 构建设置**）。打开**高级设置**: **语言**对话框，并选择“类别”输入选项**未使用的抽取结果**，并构建其他类别。
- 定义节点时，从 TAP 选择类别集（在交互式工作台中的“类别和概念”视图中）。下一步，将未使用概念或模式拖放到您认为适当的类别中。然后，展开您刚刚编辑的现有类别（**类别 > 扩展类别**），以获取与现有类别描述符相关的更多描述符。
- 使用高级语言设置自动构建类别（**类别 > 构建类别**）。然后通过删除描述符、删除类别或合并相似类别手动优化类别，直到您满意生成的类别。此外，如果最初在**不**使用**尽可能使用通配符一般化**选项的情况下构建类别，还可尝试使用**一般化**选项，使用“扩展类别”自动简化类别。
- 使用非常具有描述性的类别名称和/或注释导入预定义类别文件。如果最初在**不**选择此选项的情况下从类别名称导入或生成描述符，那么稍后可使用“扩展类别”对话框，并选择**使用从类别名称生成的描述符扩展空类别**选项。然后，再次扩展这些类别，但这次使用分组方法。
- 按频率对概念或概念模式进行排序，然后将最相关的概念或概念模式拖放到“类别”窗格中，手动创建第一组类别。具有此组初始类别后，使用“扩展”功能（**类别 > 扩展类别**）扩展并优化所有所选类别，以便其将包含其他相关描述符，从而匹配更多记录。

应用这些方法后，建议查看生成的类别，并使用手动方法来执行较小的调整，除去任何错误分类，或添加可能已缺少的记录或单词。此外，由于使用其他方法可能会生成冗余类别，还可根据需要合并或删除类别。请参阅主题第 119 页的『编辑和优化类别』，以获取更多信息。

创建类别的提示

为帮助您创建更好的类别，您可以查看有助于您决定自己的方法的部分提示。

有关类别与文档比率的提示

文档和记录分配到的类别在定性文本分析中通常不会互斥，这至少存在两个原因:

- 第一，一般经验法则表明文本文档或记录越长，表达的构想和意见越清晰。因此，将文档或记录分配到多个类别的可能性显著增加。
- 其次，通常存在多种方式来分组和解释逻辑上相关的文本文档或记录。对于有关受访者政治信念（例如，*自由党*和*保守党*或*共和党*和*民主党*）以及更具体的类别（例如，*社会自由主义*、*财政保守*等）的开放式问题调查。这些类别无需互斥并详尽。

有关要创建的类别数的提示

类别创建应直接流自数据 - 当您看到与您的数据有关的有趣内容时，即可创建一个类别来表示此信息。总而言之，对于可以创建的类别数量没有建议的上限。但是，创建类别过多当然可能造成难以管理。在此适用两大原则:

- **类别频率。**要使类别实用，其中必须包含最低数量的文档或记录。一个或两个文档可能包含非常有趣的内容，但如果是来自 1,000 个文档之一或之二，那么其中包含的信息在总数中的频率可能不足以使其成为实用信息。
- **复杂性。**您创建的类别越多，在完成分析后必须审查和总结的信息越多。但是，类别过多在增加复杂性的同时可能无法添加实用的详细信息。

不幸的是，对于确定多少数量的类别为过多或者确定每个类别的最低记录数量不存在任何规则。您必须基于特定情况的需求来自行决定。

但是我们可以提供有关从何处开始的建议。虽然类别数量不应过多，但是在分析的早期阶段，类别数量多比数量少更好。这样更易于将相对类似的类别分组在一起，而不是将案例拆分为新的类别，因此从更多类别开始工作并逐渐减少类别数量的策略通常是最佳实践。考虑到文本挖掘的迭代性质以及通过此软件程序可以轻松完成这项工作，开始时构建更多类别是可以接受的。

选择最佳描述符

以下信息包含为您的类别选择或生成最佳描述符（概念、类型、TLA 模式和类别规则）的部分准则。描述符是类别的构建块。当文档或记录中的部分或全部文本与某个描述符匹配时，即表示此文档或记录与类别匹配。

除非描述符包含或对应于已抽取的概念或模式，否则将不会与任何文档或记录匹配。因此，请按以下段落中的描述来使用概念、类型、模式和类别规则。

由于概念不仅表示其本身还表示一组底层术语，这些术语范围广泛包括复数/单数形式、同义词到拼写变体等，因此仅限概念本身才能用作为描述符或描述符的一部分。要了解有关任何给定概念的底层术语的更多信息，请单击“类别和概念”视图的“抽取结果”窗格中的概念名称。悬停于概念名称上时，会显示一个工具提示，其中显示了在上次抽取期间在您的文本中找到的任何底层术语。并非所有概念都包含底层术语。例如，如果 `car`（汽车）和 `vehicle`（车辆）属于同义词，但是已抽取 `car` 作为概念，并使用 `vehicle` 作为底层术语，那么在描述符中只能使用 `car`，因为它将自动匹配带有 `vehicle` 的文档或记录。

使用概念和类型作为描述符

要查找包含概念（或其任何底层术语）的所有文档或记录时，请使用概念作为描述符。在此情况下，无需使用更复杂的类别规则，因为准确的概念名称已足够。切记，使用抽取意见的资源时，有时在 TLA 模式抽取期间概念会发生更改以捕获句子更真实的含义（请参阅下一部分中有关 TLA 的示例）。

例如，某项调查响应指示每个人最喜欢的水果为“*Apple and pineapple are the best（最喜欢苹果和菠萝）*”，这可能导致抽取 `apple`（苹果）和 `pineapple`（菠萝）。通过在您的类别中添加 `apple`（苹果）概念作为描述符，可使包含 `apple` 概念（或者其任何底层术语）的所有响应都与此类别相匹配。

但是，如果您只想知道哪些响应以任何方式提到了 *apple*，那么可以编写诸如 * apple * 之类的类别规则，这样还将捕获包含诸如 apple（苹果）、apple sauce（苹果酱）或 french apple tart（法式苹果挞）之类的概念的响应。

您还可以通过直接使用诸如 <Fruit（水果）> 之类的类型作为描述符，来捕获包含相同类型的概念的所有文档或记录。请注意，对于类型不能使用 *。

请参阅第 73 页的『提取结果：概念和类型』主题以获取更多信息。

使用文本链接分析 (TLA) 模式作为描述符

要捕获更精确的、差别更细微的构想时，请使用 TLA 模式结果作为描述符。在 TLA 抽取期间会对文本进行分析，每次对文本中的一个句子或一个子句进行处理，而不是查看整个文本（文档或记录）。通过考虑将单个句子的所有部分整合在一起，TLA 可以识别意见、两个元素之间的关系或否定，从而更好地了解真正的含义。您可以使用概念模式或类型模式作为描述符。请参阅第 129 页的『类型和概念模式』主题以获取更多信息。

例如，如果文本为“*the room was not that clean（房间不太干净）*”，那么可以抽取以下概念：room（房间）和 clean（干净）。但是，如果在抽取设置中已启用 TLA 抽取，那么 TLA 可能检测到 clean（干净）带有负面含义，实际对应于 not clean（不干净），是 dirty（脏）概念的同义词。您可以在此处看到单独使用 clean（干净）概念作为描述符将与此文本匹配，但还会捕获其他提到干净情况的文档或记录。因此，使用以 dirty 作为输出概念的 TLA 概念模式更好，因为它会匹配此文本，可能成为更适合的描述符。

使用类别业务规则作为描述符

类别规则是基于使用抽取的概念、类型、模式和布尔运算符的逻辑表达式自动将文档或记录分类到某个类别中的语句。例如，您可以编写如下含义的表达式：*include all records that contain the extracted concept* embassy *but not* argentina *in this category*（在此类别中包含含有已抽取的“大使馆”概念但不含“阿根廷”概念的所有记录）。

您可以使用 &、| 和 !() 布尔值在自己的类别中编写和使用类别规则作为描述符，来表达多种不同的构。有关这些规则的语法以及如何编写和编辑这些规则的详细信息，请参阅第 105 页的『使用类别规则』。

- 使用带有 & (AND) 布尔运算符的类别规则来帮助您查找其中发生了 2 个或更多个概念的文档或记录。以 & 运算符连接的 2 个或更多个概念无需发生在同一个句子或短语内，而是可发生在要视为与类别匹配的相同文档或记录中的任何位置。例如，如果创建规则 food & cheap（食物和便宜）作为描述符，那么将于包含文本“*the food was pretty expensive, but the rooms were cheap*（食物很贵，但房间和便宜）”的记录相匹配，尽管 food（食物）并不是被称为 cheap（便宜）的那个名词，因为此文本同时包含 food（食物）和 cheap（便宜）。
- 使用采用 !() (NOT) 布尔运算符作为描述符的类别规则来帮助您查找其中某些事物已发生而某些未发生的文档或记录。这样可帮助避免基于文字看似相关但基于上下文则不相关的分组信息。例如，如果创建类别规则 <组织> & !(ibm) 作为描述符，那么将匹配文本 *SPSS Inc. was a company founded in 1967*（SPSS Inc. 是一家创建于 1967 年的公司），但不匹配文本 *the software company was acquired by IBM.*（该关键公司被 IBM 收购）。
- 使用带有 | (OR) 布尔运算符作为描述符的类别规则来帮助您查找包含某一个概念或类型的文档或记录。例如，如果创建类别规则 (personnel|staff|team|coworkers) & bad 作为描述符，那么将匹配其中可找到这些名词并带有概念 bad（糟糕）的所有文档或记录。
- 在类别规则中使用类型来使其更通用并且可能更易于部署。例如，如果处理酒店数据，那么您可能非常有兴趣了解客户对于酒店工作人员的想法。相关术语可能包括 receptionist（接待员）、waiter（男性侍者）、waitress（女性侍者）、reception desk（服务台）、front desk（前台）等。在此情况下，您可以创建称为 <HotelStaff>（酒店员工）的新类型以及此类型的所有前述术语。虽然可以为每一种类型的员工创建一条

类别规则（例如，[* waitress * & nice]，[* desk * & friendly]，[* receptionist * & accommodating]），但是可以使用 <HotelStaff> 类型来创建更通用的类别规则以捕获对酒店员工表现赞许意见的所有响应，其形式为：[<HotelStaff> & <Positive>]。

注： 在类别规则中包含 TLA 模式时，可以在这些类别规则中使用 + 和 &。请参阅第 106 页的『在类别规则中使用 TLA 模式』主题以获取更多信息。

使用概念、TLA 或类别规则作为描述符的匹配差异示例

以下示例演示了如何使用概念作为描述符、使用类别规则作为描述符或者使用 TLA 模式作为描述符对于文档或记录的分类方式的影响。假设您拥有以下 5 条记录。

- A: "*awesome restaurant staff, excellent food and rooms comfortable and clean.*（出色的餐厅员工，美味的食物，房间舒适又干净）"
- B: "*restaurant personnel was awful, but rooms were clean.*（餐厅员工太糟糕，但房间很干净）"
- C: "*Comfortable, clean rooms.*（房间舒适而且干净）"
- D: "*My room was not that clean.*（我的房间不太干净）"
- E: "*Clean.*（干净）"

由于这些记录包含文字 *clean*（干净），并且您想要捕获此信息，因此您可以创建下表中所显示的描述符之一。由于本质上您正在尝试捕获，您可以看到使用某一种类型的描述符相比于使用其他描述符所生成的不同结果。

表 17. 示例记录与描述符匹配的方式.

描述符	A	B	C	D	E	说明
clean	*匹配*	*匹配*	*匹配*	*匹配*	*匹配*	描述符是抽取的概念。包含 clean 概念的每条记录（包括记录 D），如果没有 TLA，那么 TLA 规则无法自动获悉"*not clean*（不干净）"意为 dirty（脏）。
clean + .	-	-	-	-	*匹配*	描述符是代表 clean 本身的 TLA 模式。仅匹配已抽取 clean 并且在 TLA 抽取期间没有关联概念的记录。
[clean]	*匹配*	*匹配*	*匹配*	-	*匹配*	描述符是类别规则，查找本身包含 clean 或者同时包含其他内容的 TLA 规则。匹配找到包含 clean 的 TLA 输出的所有记录，无论 clean 是否在任何位置与诸如 room（房间）之类的任何其他概念相链接都是如此。

有关类别

类别指一组紧密相关的概念、意见或态度。为实现其目的，还可通过阐述其基本含义的短语或标签轻松描述类别。

例如，如果分析使用者提供的有关新上市洗衣皂的调查响应，那么可创建标签为 *odor* 的类别，此类别包含描述产品味道的所有响应。但是，此类别不会区分气味好闻的产品和气味不好闻的产品。由于在使用相应资源时 IBM SPSS Modeler Text Analytics 可以抽取意见，那么可创建两个其他类别以标识*喜欢 odor* 的响应者以及*不喜欢 odor* 的响应者。

您可在"类别和概念"视图窗口左上窗格的"类别"窗格中创建和使用类别。每个类别通过一个或多个描述符定义。**描述符**为概念、类型和模式以及已用于定义类别的类别规则。

如果要查看组成给定类别的描述符，那么可在“类别”窗格工具栏中单击铅笔图标，然后展开树以查看描述符。或者，选择类别并打开“类别定义”对话框（**视图 > 类别定义**）。

使用类别构建方法（例如，概念包含）自动构建类别时，这些方法将使用概念和类型作为描述符来创建类别。如果抽取 TLA 模式，那么您还可以添加模式或这些模式的部分作为类别描述符。 请参阅主题第 127 页的第 12 章，『探索文本链接分析』，以获取更多信息。如果构建集群，那么可将集群中的概念添加到新类别或现有类别。 最后，您可以手动创建类别规则，以用作类别中的描述符。请参阅主题第 105 页的『使用类别规则』，以获取更多信息。

类别属性

除了描述符，类别还具有属性，可编辑这些属性来重命名类别、添加标签或添加注释。

存在以下属性:

- **名称。**缺省情况下，此名称显示在树中。使用自动方法创建类别时，会自动为其指定名称。
- **标签。** 创建更有意义的类别描述以在其他产品或者其他表或图形中使用时，使用标签会很有帮助。如果选择选项以显示标签，那么会在界面中使用标签来标识类别。
- **代码。** 代码数字与此类别的代码值对应。.
- **注释。** 您可以在此字段中添加每个类别的简短描述。通过“构建类别”对话框生成类别时，会自动向此注释添加说明。您还可直接通过选择文本并从菜单选择**类别 > 添加到注释**，将样本文本添加到注释。

数据窗格

您可以在创建类别时检查所处理的某些文本数据。例如，如果创建一个分类 640 个文档的类别，那么可能想要查看部分或全部这些文档以查看实际写入的文本。您可以在右下方的“数据”窗格中查看记录或文档。如果缺省情况下未显示，那么从菜单中选择**视图 > 窗格 > 数据**。

“数据”窗格根据“类别”窗格、“提取结果”窗格或“类别定义”对话框中的选择，每个 文档或记录显示一行，直至特定显示限制。缺省情况下，限制“数据”窗格中显示的 文档或记录的数量，以便更快地查看数据。但是，您可以在“选项”对话框中进行调整。请参阅主题第 69 页的『选项: “会话”选项卡』以获取更多信息。

显示和刷新数据窗格

“数据”窗格不会自动刷新其显示，因为对于较大的数据集，自动数据刷新可能需要一些时间才能完成。因此，在此视图的另一个窗格中或“类别定义”对话框中进行选择时，单击**显示**以刷新“数据”窗格的内容。

文本文档或记录

如果文本数据采用记录格式，并且文本长度相对较短，那么“数据”窗格中的文本字段整体显示文本数据。但是，在处理记录和较大的数据集时，文本字段列显示文本的一小部分，打开右侧的“文本预览”窗格将显示表中选择的记录的更多文本或全部文本。如果文本数据采用单个文档形式，那么“数据”窗格显示文档的文件名。在选择文档时，“文本预览”窗格打开，并包含选中的文档文本。

颜色和突出显示

在显示数据时，在这些文档或记录中找到的数据、概念和描述符将使用显示突出显示，从而便于在文本中识别。颜色编码对应于概念所属的类型。您还可以在颜色编码的项上悬停鼠标以显示提取的概念以及指定的类型。未提取的任何文本将显示为空白。通常，这些未提取的字通常是连接词（*and* 或 *with*）、代词（*me* 或 *they*）和动词（*is*、*have* 或 *take*）。

数据窗格列

文本字段列始终显示，您还可以显示其他列。要显示其他列，请从菜单中选择**视图 >数据窗格**，然后选择想要在“数据”窗格中显示的列。以下列可供显示:

- **“文本字段名称”(#)/文档。**添加从中提取概念和类型的文本数据集的列。如果数据位于文档中，列称为“文档”，并且仅显示文档文件名或完整路径。要查看这些文档的文本，必须在“文本预览”窗格中进行查看。在列名称后的括号中显示“数据”窗格中的行数。有时候，由于“选项”对话框中为提高装入速度而采取的限制，不显示完整文档或记录。如果到达最大数量，那么数量将后跟 **- Max**。请参阅主题第 69 页的『选项: “会话”选项卡』以获取更多信息。
- **类别。**列出记录所属的每个类别。在显示此列时，刷新“数据”窗格可能需要较长时间才能显示最新信息。
- **相关性排名。**针对单个列别中的每条记录提供一个排名。此排名显示与此类别中的其他记录相比，此记录适合列别的程度。选择“类别”窗格（左上方窗格）中的记录以查看排名。请参阅主题『类别相关性』以获取更多信息。
- **类别计数。**列出记录所属的类别的数量。

类别相关性

为帮助您构建更好的类别，您可以查看每个目录中文档或记录的相关性以及文档或记录所属的所有类别的相关性。

类别与记录的相关性

只要在“数据”窗格中显示文档或记录，就会在“类别”列中列出其所属的所有类别。当某个文档或记录属于多个类别时，此列中的类别会按相关性匹配从高到低的顺序显示。最先列出的类别被视作为与此文档或记录对应的程度最高。请参阅第 92 页的『数据窗格』主题以获取更多信息。

记录与类别的相关性

选择类别时，可以在“数据”窗格的“相关性等级”列中查看其每条记录的相关性。此相关性等级指示文档或记录相比于所选类别内其他记录与此类别的适合程度。要查看单个类别的记录等级，请在“类别”窗格（左上窗格）中选中此类别，这样会在该列中显示此文档或记录的等级。缺省情况下此列不可见，但是您可以选择显示此列。请参阅第 92 页的『数据窗格』主题以获取更多信息。

记录等级数越低表示越适合或者此记录与所选类别相关程度越高（例如，1 表示最适合）。如果多条记录的相关性相同，那么每条记录显示等级时都会后接一个等号 (=) 以表示其相关性相等。例如，可能存在以下等级: 1=、1=、3、4 等，这表示存在两条记录针对此类别被视为同样属于最佳匹配。

提升: 您可以将相关性最高的记录的文本添加到类别注释中以帮助提供对类别的更好的描述。通过选择文本并从菜单中选择**类别 > 添加到注释**来从“数据”窗格直接添加文本。

构建类别

虽然您可能具有文本分析包中的类别，但还可使用一些语言和频率方法自动构建类别。通过“构建类别设置”对话框，可应用自动化语言和频率方法，以从概念或概念模式生成类别。

通常，类别由多种描述符（类型、概念、TLA 模式和类别规则）组成。使用自动化类别构建方法构建类别时，生成的类别将根据概念或概念模式（取决于所选输入）进行命名，每个类别包含一组描述符。这些描述符形式可以为类别规则或概念，并包含方法发现的所有相关概念。

构建类别后，可通过在“类别”窗格中进行查看或通过图形和图表进行研究来了解有关类别的很多信息。然后可使用手动方法来执行较小的调整，除去任何错误分类，或添加可能已缺少的记录或单词。应用方法后，分组到

一个类别的概念、类型和模式仍可用于使用其他方法。此外，由于使用其他方法可能会生成冗余或不合适的类别，还可合并或删除类别。请参阅主题第 119 页的『编辑和优化类别』，以获取更多信息。

重要！ 在先前发行版中，应该针对同现和同义词规则使用方括号。在此发行版中，方括号现在指示文本链接分析模式结果。而同现和同义词规则将使用括号引起来 (speaker systems|speakers)。

构建类别

1. 从菜单中选择**类别 > 构建类别**。除非选择永不提示，否则会显示一个消息框。
2. 选择希望立即构建还是首先编辑设置。

- 单击**立即构建**以使用当前设置开始构建类别。缺省情况下选择的设置通常足以开始分类过程。类别构建过程将开始，且会显示进度框。
- 单击**编辑**以查看和修改构建设置。

注: 可显示的最大类别数为 10,000。如果达到或超过此数目，那么会显示警告。如果发生此情况，那么应该更改“构建或扩展类别”选项以减少构建的类别数。

输入

将根据派生自类型模式或类型的描述符构建类别。 在表中，可选择要在类别构建过程中包含的单个类型或模式。

类型模式。如果选择类型模式，那么将根据模式而不是类型或自己的概念构建类别。在此方式中，将对包含属于所选类型模式的概念模式的任何记录或文档进行分类。因此，如果选择表中的 <Budget> 和 <Positive> 类型模式，那么可能会生成诸如 cost & <Positive> 或 rates & excellent 的类别。

将类型模式用作构建自动化类别的输入时，有时方法可确定多种方式来形成类别结构。实际上，不仅仅只有一种正确方法用于生成类别；但是可能会发现一种结构比另一种结构更适用于您的分析。要在此情况下帮助定制输出，可将类型指定为首选焦点。所有生成的顶级类别将来自此处所选的类型（而不是其他类型）概念。每个子类别将包含来自此类型的文本链接模式。在**按模式类型构造类别**: 字段中选择此类型，表将更新以仅显示包含所选类型的适用模式。通常会为您预先选择 <Unknown>。这会选择包含类型 <Unknown>（适用于非日语文本）的所有模式。. 表将按降序顺序显示类型，以具有最多数目记录或文档 (**文档**计数）的类型开始。

类型。如果选择类型，那么将从属于所选类型的概念构建类别。因此，如果选择表中的 <Budget> 类型，那么可以生成诸如 cost 或 price 的类别，因为 cost 和 price 是分配给 <Budget> 类型的概念。

缺省情况下，仅会选择捕获最多纪录或文档的类型。此预先选择使您可快速关注最相关的类型，并避免构建不相关的类别。表将按降序顺序显示类型，以具有最多数目记录或文档 (**文档**计数）的类型开始。缺省情况下，会在类型表中取消选择 Opinions 库中的类型。

所选的输入会影响获取的类别。选择将类型用作输入时，可更容易地看到明确相关的概念。例如，如果在使用类型作为输入的情况下构建类别，那么可获取具有概念（例如，apple、pear、citrus fruits 和 orange 等）的类别 Fruit。如果改为选择类型模式作为输入，例如，选择模式 <Unknown> + <Positive>，那么可能会获得一个类别 fruit + <Positive>，此类别具有一种或多种水果，例如，fruit + tasty 和 apple + good。第二种结果仅显示 2 个概念模式，因为出现的其他水果不一定合格。虽然这可能足够用于当前文本数据，但是在使用不同文档集的纵向研究中，可能需要手动添加诸如 citrus fruit + positive 的其他描述符或使用类型。将类型单独用作输入可帮助您查找所有可能的水果。

方法

由于每个数据集唯一，方法数和应用这些方法的顺序会随着时间发生变化。由于文本挖掘目标在不同数据组中可能会不同，因此可能需要试验不同方法，以了解哪种方法可针对给定文本数据产生最佳效果。

您不需要非常了解这些设置也可使用这些设置。缺省情况下，已选择最常见的普通设置。因此，可跳过高级设置对话框，直接构建类别。同样地，如果在此处执行更改，那么每次不必返回设置对话框，因为会始终保留最新设置。

选择语言或频率方法，并单击“高级设置”按钮以显示所选方法的设置。自动方法都无法对数据进行完美分类；因此，建议寻求和应用非常适用于您的数据的一种或多种自动方法。无法同时使用语言和频率方法进行构建。

- **高级语言方法**。有关更多信息，请参阅『高级语言设置』。
- **高级频率方法**。有关更多信息，请参阅第 101 页的『高级频率设置』。

高级语言设置

构建类别时，可从多个高级语言类别构建方法选择，包括*概念根派生*（不可用于日语）、*概念包含*、*语义网络*（仅针对英语文本）以及*同现规则*。这些方法可单独使用，也可与其他方法组合使用来创建类别。

请记住，由于每个数据集唯一，方法数和应用这些方法的顺序会随着时间发生变化。由于文本挖掘目标在不同数据组中可能会不同，因此可能需要试验不同方法，以了解哪种方法可针对给定文本数据产生最佳效果。自动方法都无法对数据进行完美分类；因此，建议寻求和应用非常适用于您的数据的一种或多种自动方法。

“高级设置：语言”对话框中提供了以下区域和字段：

输入和输出

类别输入。 选择将从其构建类别的输入：

- **未使用的抽取结果**。此选项可从任何现有类别中未使用的抽取结果构建类别。这可最大程度降低记录匹配多个类别的趋势，并限制生成的类别数。
- **所有抽取结果**。此选项可使用任何抽取结果构建类别。不存在任何分类或存在很少的分类时，这会很有帮助。

类别输出。 选择将构建的类别的常规结构：

- **分层，具有子类别**。此选项可用于创建子类别及其子类别。您可以通过选择可创建的最大级别数（**创建的最大级别数**字段）来设置类别的深度。如果选择 3，那么类别可包含子类别，这些子类别还可具有子类别。
- **平面类别（仅单个级别）**。此选项仅启用要构建的一个类别级别，这表示不会生成任何子类别。

分组方法

每个可用方法都明确适合于特定类型的数据和情况，但通常有助于组合同一分析中的方法以捕获完整范围的文档或记录。您可能会看到属于多个类别的概念或找到冗余类别。

概念根派生。此方法采用某个概念并通过分析任何概念组件是词法相关还是共享根来查找与其相关的其他概念，从而创建类别。此方法对于识别同义复合词概念非常有用，因为所生成的每个类别中的概念是同义词或含义紧密相关。它处理变长数据并生成更小数量的紧凑类别。例如，概念 `opportunities to advance` 将与概念 `opportunity for advancement` 和 `advancement opportunity` 分组在一起。请参阅主题第 97 页的『概念根派生』以获取更多信息。此选项不适用于日语文本。

语义网络。此方法首先从每个概念的广泛单词关系索引确定其可能含义，然后通过将相关概念分组来创建类别。此方法在概念对于语义网络已知且不太模糊时最适用。在文本包含对于网络未知的专用术语或行话时帮助不大。在一个示例中，概念 `granny smith apple` 可能会与 `gala apple` 和 `winesap apple` 分组在一起，因为它们是 granny smith 的同代。在另一个示例中，概念 `animal` 可能与 `cat` 和 `kangaroo` 分组在一起，因为它们是 `animal` 的下义词。此方法仅在本发行版中适用于英语文本。请参阅主题第 99 页的『语义网络』以获取更多信息。

概念包含。此方法通过基于多术语概念（复合词）包含属于另一个类别中单词的子集还是超集的单词将其分组来构建类别。例如，概念 seat 将会与 safety seat、seat belt 和 seat belt buckle 分组在一起。请参阅主题第 98 页的『概念包含』以获取更多信息。

共生。此方法根据在文本中找到的共生来创建类别。构想是当通常在文档和记录中同时找到概念和概念模式时，共生会反映可能在类别定义中有意义的底层关系。当单词显著共生时，将会创建共生规则并将其用作新的子类别的类别描述符。例如，如果许多记录包含单词 price 和 availability（但少数记录仅包含其中一个），那么这些概念可能会分组到共生规则中 (price & available) 并分配到类别（例如 price）的子类别。请参阅主题第 100 页的『同现规则』以获取更多信息。

最小 文档。要帮助确定共生的有趣程度，请定义必须包含给定共生才能在类别中用作描述符的最小文档或记录数。

最大搜索距离。 选择生成类别之前希望通过方法搜索的最大范围。值越小，获得的结果越少，但是，这些结果将更为简单，且更可能互相紧密链接或关联。值越大，获得的结果可能越多，但是，这些结果可靠性和相关性将降低。此选项全局应用于所有方法时，影响最大的是同现和语义网络。

防止对特定概念进行配对。 选中此复选框，以停止在输出中将两个概念分组在一起或配对的过程。要创建或管理概念对，请单击**管理对**。请参阅主题第 97 页的『管理链接异常对』，以获取更多信息。

尽可能使用通配符一般化。 选择此选项，使产品使用星号通配符在类别中生成通用规则。例如，不是生成多个描述符（例如，[apple tart + .] 和 [apple sauce + .]），而是使用通配符可生成 [apple * + .]. 如果使用通配符一般化，那么通常将获得之前所获得的相同数据的记录或文档。但是，此选项具有减少数目和简化类别描述符的优势。此外，此选项还通过针对新文本数据（例如，在纵波研究中）使用这些类别，提高对更多记录或文档进行分类的能力。

其他用于构建类别的选项

除了可选择要应用的分组方法，还可编辑多个其他构建选项，如下所示:

创建的顶级类别最大数。 使用此选项，可限制在接下来单击“构建类别”按钮时生成的类别数。在某些情况下，如果将此值设置为较高，然后删除任何不相关的类别，那么可能会获得较好的效果。

每个类别的最大描述符和/或子类别数。 使用此选项，定义创建类别时必须包含的最小描述符和子类别数。此选项可帮助限制未捕获到大量记录或文档的类别的创建。

允许描述符在多个类别中显示。 选择此项时，允许在接下来将构建的多个类别中使用描述符。通常，会选择此选项，因为项通常或自然地属于两个或更多类别，通常，允许其实现此操作会生成较高质量的类别。如果未选择此选项，那么减少多个类别中重叠记录，根据所具有的数据类型，可能需要执行此操作。但是，针对多数数据类型，将描述符限制为单个类别通常会导致降低质量或类别覆盖范围。例如，假设具有概念 car seat manufacturer。使用此选项，此概念可能会显示在基于文本 car seat 的一个类别中，以及基于 manufacturer 的另一个类别中。如果未选择此选项，尽管您仍可能会获得这两个类别，概念 car seat manufacturer 仍仅作为描述符显示在其基于多个因素最佳匹配的类别中，这些因素包含在其中出现 car seat 和 manufacturer 的记录数。

解决重复类别名称方法。 选择如何处理器名称将与现有类别相同的任何新类别或子类别。您可以将新项（及其描述符）与具有相同名称的现有类别合并。或者，如果在现有类别中发现重复名称，可以选择跳过创建任何类别。

管理链接异常对

类别构建、集群和概念映射期间，内部算法按已知关键对单词进行分组。为防止对两个概念进行配对或将其链接到一起，可以在**构建类别高级设置**对话框、**构建集群**对话框和**概念映射索引设置**对话框中打开此功能部件，并单击**管理对**按钮。

在生成的**管理链接异常**对话框中，可添加、编辑或删除概念对。每行输入一个对。在此处输入对将放置在构建或扩展类别、集群和概念映射时发生配对。根据您的需要精确输入单词，例如，单词的重音版本不等于单词的非重音版本。

例如，如果需要确保不对 `hot dog` 和 `dog` 进行分组，那么可将对作为表中的单独行添加。

有关语言方法

构建或扩展类别时，可从多个高级语言类别构建方法选择，包括*概念根派生*（不可用于日语）、*概念包含*、*语义网络*（仅针对英语文本）以及*同现规则*。这些方法可单独使用，也可与其他方法组合使用来创建类别。

您不需要非常了解这些设置也可使用这些设置。缺省情况下，已选择最常见的普通设置。如果需要，可跳过此高级设置对话框，直接构建或扩展类别。同样地，如果在此处执行更改，那么每次不必返回设置对话框，因为会始终记住上次使用设置。

但是，请记住，由于每个数据集唯一，方法数和应用这些方法的顺序会随着时间发生变化。由于文本挖掘目标在不同数据组中可能会不同，因此可能需要试验不同方法，以了解哪种方法可针对给定文本数据产生最佳效果。自动方法都无法对数据进行完美分类；因此，建议寻求和应用非常适用于您的数据的一种或多种自动方法。

用于构建类别的主要自动化语言方法包括:

- **概念根派生。** 此方法通过以下方式创建类别: 获取概念并通过分析任何概念组件是否在词法上相关来查找与其相关的其他概念。请参阅主题『概念根派生』，以获取更多信息。此选项不适用于日语文本。
- **概念包含。** 此方法通过获取概念并查找包含此概念的其他概念来创建类别。请参阅主题第 98 页的『概念包含』，以获取更多信息。
- **语义网络。** 此方法通过从每个概念的单词关系的扩展索引来识别概念的可能含义开始，然后通过分组相关概念来创建类别。请参阅主题第 99 页的『语义网络』，以获取更多信息。此选项仅适用于英语文本。
- **同现。** 此方法可创建可用于创建新类别、扩展类别或作为其他类别方法的输入的同现规则。请参阅主题第 100 页的『同现规则』，以获取更多信息。

概念根派生

注: 此方法不适用于日语文本。

概念根派生方法通过以下方式创建类别: 获取概念并通过分析任何概念组件是否在词法上相关来查找与其相关的其他概念。组成部分为单词。方法尝试通过查看概念中每个组件的结尾（后缀）并查找可从其检索的其他概念来对概念进行分组。此构想为，相互之间派生单词时，这些单词可能会具有相同或相近含义。为了识别结尾部分，使用内部特定于语言的规则。例如，将使用 `opportunity for advancement` 和`advancement opportunity` 对概念 `opportunities to advance` 进行分组。

您可以针对任何种类的文本使用概念根派生。概念根派生本身生成的类别很少，每个类别将包含的概念也很少。每个类别中的概念是同义词或情境相关。即使您正在手动构建类别，可能会发现使用此算法很有帮助；其发现的同义词可能为您特别关注的那些概念的同义词。

注: 您可以通过显式指定来阻止将概念分组在一起。请参阅主题『管理链接异常对』，以获取更多信息。

术语成分化和反屈折变化

应用概念根派生或概念包含方法时，会首先将术语划分为组成部分（单词），然后会对组成部分执行反屈折变化操作。应用方法时，概念及其关联术语将转入并基于分隔符（例如，空格、连字符和撇号）拆分为组成部分。例如，术语 system administrator 将拆分为组成部分，例如，{administrator，system}。

但是，原始术语的一些部分可能不会使用，这些部分称为停用词。在英语中，一些可忽略组成部分可能包含 a、and、as、by、for、from、in、of、on、or、the、to 和 with。

例如，术语 examination of the data 具有组成部分集 {data，examination}，of 和 the 均被视为可忽略。此外，组成部分顺序不在组成部分集中。通过此方法，以下三个术语可能等效：cough relief for child、child relief from a cough 和 relief of child cough，因为这三个术语都具有相同组成部分集 {child，cough，relief}。每次术语对识别为等效时，会将相应概念合并以形成引用所有术语的新概念。

此外，由于术语的组成部分可能已进行屈折变化，因此特定于语言的规则在内部应用以识别等效术语，与可屈折变化的变体（复数形式）无关。通过此方式，可以将术语 level of support 和 support levels 识别为等效，因为反屈折变化的单数形式将为 level。

概念根派生工作方式

术语已成分化和反屈折变化（参见先前部分）后，概念根派生算法会分析组成部分结尾或后缀，以查找组成部分根，然后将概念与具有相同或相似根的其他概念分组在一起。将使用特定于文本语言的一组语言派生规则识别结尾。例如，针对英语文本，具有派生规则说明了后缀为 ical 的概念组成部分结尾可能派生自具有相同根词干和后缀为 ic 的结尾的概念。使用此规则（以及反屈折变化）时，算法可以将概念 epidemiologic study 和 epidemiological studies 分组在一起。

由于术语已成分化，且识别了可忽略的组成部分（例如，in 和 of），因此概念根派生算法还可以将概念 studies in epidemiology 和 epidemiological studies 分组在一起。

已选择一组组成部分派生规则，以便通过此算法分组的多数概念为同义词：概念 epidemiologic studies、epidemiologicalstudies 和 studies in epidemiology 都是等效术语。为了改善完整性，提供了一些派生规则，允许算法对情境相关的概念进行分组。例如，算法可对诸如 empire builder 和 empire building 的概念进行分组。

概念包含

概念包含方法通过获取概念并使用词汇系列算法构建类别，识别其他概念中包含的概念。此构想在于，概念中的单词为其他概念的一部分时，它会反映底层语义关系。包含是一种可用于任何文本类型的强大方法。

此方法非常适合与语义网络组合使用，但也可单独使用。概念包含还可在文档或记录包含很多特定于域的术语或行话时获取更好的效果。这尤其适用于您事先调整了字典以便抽取特殊术语并相应对这些术语进行分组（使用同义词）的情况。

概念包含工作方式

应用概念包含算法之前，会对术语进行成分化和反屈折变化。请参阅主题第 97 页的『概念根派生』，以获取更多信息。之后，概念包含算法会分析组成部分集。针对每个组成部分集，算法会查找属于第一个组成部分集的另一个组成部分集。

例如，如果具有概念 continental breakfast，此概念具有组成部分集 {breakfast，continental}，且具有概念 breakfast，此概念具有组成部分集 {breakfast}，那么算法得出 continental breakfast 属于 breakfast 这一结论且会将其分组在一起。

在较大的示例中，如果在“抽取结果”窗格中具有概念 seat，且应用了此算法，那么还会在此类别中对诸如 safety seat、leather seat、seat belt、seat belt buckle、infant seat carrier 和 car seat laws 的概念进行分组。

由于术语已成分化且识别了可忽略组成部分（例如，in 和 of），那么概念包含算法将识别到概念 advanced spanish course 包含概念 course in spanish。

注：您可以通过显式指定来阻止将概念分组在一起。请参阅主题第 97 页的『管理链接异常对』，以获取更多信息。

语义网络

在此版本中，语义网络技术仅限以英语文本提供。

此技术会使用文字关系的内置网络来构建类别。出于此原因，当术语非常确切并且不模糊时，此技术可生成非常良好的结果。但是，您不应期望此技术可在高度技术性概念/专业化概念之间找到任何联系。处理此类概念时，您可能会发现概念包含和概念根派生技术更实用。

语义网络的工作方式

语义网络技术背后的构想在于利用已知的文字关系来创建同义词和下义词的类别。**下义词**即某种概念属于次级概念，因此存在分层关系，也称为 ISA 关系。例如，如果 animal（动物）是一个概念，那么 cat（猫）和 kangaroo（袋鼠）即 animal（动物）的下义词，因为它们都是某种动物。

除同义词和下义词关系之外，语义网络技术还会从 <位置> 类型来检验任何概念之间的部分和完整链接。例如，此技术会将 normandy（诺曼底）、provence（普罗旺斯）和 france（法国）概念分组为一个类别，因为诺曼底和普罗旺斯都属于法国的一部分。

语义网络首先识别语义网络中每个概念之间可能的理解。当概念识别为同义词或下义词时，会将其分组到单个类别中。例如，此技术会创建包含以下三个概念的单个类别：eating apple（吃苹果）、dessert apple（苹果甜点）和 granny smith（绿苹果），因为语义网络包含以下信息：1) dessert apple（苹果甜点）是 eating apple（吃苹果）的同义词，并且 2) granny smith（绿苹果）是 eating apple（吃苹果）的一种（表示它是 eating apple（吃苹果）的下义词）。

如果逐个考量，许多概念（尤其是单元词）是较模糊的。例如，buffet（自助餐或饮食餐车）可表示一种餐饮和一件家具。如果一组概念包含 meal（餐饮）、furniture（家具）和 buffet（自助餐或饮食餐车），那么算法会在将 buffet（自助餐或饮食餐车）与 meal（餐饮）分组在一起或者将其与 furniture（家具）分组在一起之间强制进行选择。请注意，在某些情况下，此算法所做的选择可能不适用于特定记录或文档集的上下文。

语义网络技术在某些数据类型中比概念包含更适用。虽然语义网络和概念包含都将 apple pie（苹果派）识别为某种 pie（派），但只有语义网络将 tart（挞）也识别为一种 pie（派）。

语义网络将配合其他技术运作。例如，假定您同时选中了语义网络和包含技术，并且语义网络将概念 teacher（教师）与概念 tutor（导师）分组在一起（因为导师也是一种类型的教师）。包含算法可将概念 graduate tutor（研究生导师）与 tutor（导师）分组在一起，由此这两种算法课协作生成包含全部三个概念的输出类别：tutor（导师）、graduate tutor（研究生导师）和 teacher（教师）。

语义网络的选项

另有多种设置适用于此技术。

- 更改**最大搜索距离**。选择生成类别之前希望通过方法搜索的最大范围。值越小，生成的结果越少，但是，这些结果将更为简单，且更可能互相紧密链接或关联。值越大，获得的结果将越多，但是，这些结果可靠性和相关性将降低。

例如，根据距离，此算法会搜索 Danish pastry（丹麦点心）至 coffee roll（咖啡卷）（其父级），然后搜索 bun（小圆甜点）（其祖父级）上至 bread（面包）。

如果您认为生成的类别过大或者将太多事物分组在一起，那么通过缩短搜索距离，此技术可生成范围较小的类别，更易于处理。

要点！此外，我们建议您在使用此技术时不要将**调整拼写错误，调整的最小根字符限制为**（在节点的"专家"选项卡上 或"抽取"对话框中定义）应用于模糊分组，因为某些错误分组可能对结果产生较大的负面影响。

同现规则

同现规则使您可发现和分组在一组文档或记录中很大程度上相关的概念。此构想为，在文档和记录中频繁发现概念时，此同现会反映可能在进行类别定义时提供价值的底层关系。此方法可创建可用于创建新类别、扩展类别或作为其他类别方法的输入的同现规则。如果两个概念经常在一组记录中一起显示且很少单独在任何其他记录中显示，那么这两个概念在很大程度上具有同现性。此方法可生成具有较大型数据集（带至少几百个文档或条记录）的良好结果。

例如，如果很多记录包含单词 price 和 availability，那么这些概念可分组到一个同现规则 (price & available)。在另一个示例中，如果概念 peanut butter、jelly 和 sandwich 在多数情况下一起显示而不是单独显示，那么会将其分组为一个概念同现规则 (peanut butter & jelly & sandwich)。

重要！ 在先前发行版中，应该针对同现和同义词规则使用方括号。在此发行版中，方括号现在指示文本链接分析模式结果。而同现和同义词规则将使用括号引起来 (speaker systems|speakers)。

同现规则工作方式

此方法可扫描文档或记录，查找两个或更多将一起显示的概念。如果两个或更多概念经常在一组文档或记录中一起显示且很少单独在任何其他文档或记录中显示，那么这些概念在很大程度上具有同现性。

发现同现概念时，会形成类别规则。这些规则包含使用 & 布尔运算符连接的两个或更多概念。这些规则为逻辑语句，在规则中此组概念在文档或记录中一起同现时，这些逻辑语句会自动将该文档或记录分类到一个类别。

同现规则的选项

如果要使用同现规则方法，那么可微调影响生成的规则的多个设置:

- 更改**最大搜索距离**。选择希望通过此方法搜索同现的最大范围。随着搜索距离变远，每个同现所需的最小相似性值将降低；因此，可能会生成很多同现规则，但具有较低相似值的规则通常并不重要。搜索距离越短，所需的最小相似性值越大；因此，生成的同现规则较少，但这些规则将更为重要（强大）。
- **最少 文档数。** 必须包含给定概念对才能将其视为同现的最小记录或文档数；为此选项设置的值越小，越容易找到同现。增大此值将生成更少但更重要的同现。作为示例，假设在 2 个记录中一起发现了概念"apple"和"pear"（这两个概念均未出现在任何其他记录中）。**最少 文档数。** 设置为 2（缺省值）的情况下，同现方法将创建类别规则（apple 和 pear）。如果此值增加到 3，那么不再创建规则。

注: 针对较小的数据集（< 1000 个响应），无法使用缺省设置找到任何同现。如果是这样，请尝试增大搜索距离值。

注: 您可以通过显式指定来阻止将概念分组在一起。请参阅主题第 97 页的『管理链接异常对』，以获取更多信息。

高级频率设置

您可以基于直接和机械频率方法构建类别。使用此方法，可以为基于给定记录或文档计数发现的每个项（类型、概念或模式）构建一个类别。此外，还可以针对出现频率较低的所有项构建单个类别。计数时，我们指的是包含抽取概念（及其任何同义词）、类型或问题模式）的记录或文档数，而不是整个文本中出现总次数。

通常，对出现的项进行分组可获取相关结果，因为这可能会指示常见或重要响应。在应用其他方法后，此方法对于未使用的抽取结果很有用。其他应用程序在不存在任何其他类别的情况下，在抽取后将立即运行此方法，编辑结果以删除不相关的类别，然后扩展这些类别以便其匹配更多记录或文档。请参阅主题『扩展类别』，以获取更多信息。

您可以通过在“抽取结果”窗格中减少记录或文档数，对概念或概念模式进行排序，然后将顶级项拖放到“类别”窗格中以创建相应类别，而不是使用此方法。

“高级设置：频率”对话框中提供了以下字段:

常规类别描述符。选择描述符的输入类型。请参阅主题第 93 页的『构建类别』，以获取更多信息。

- **概念级别**。选择此选项表示将使用概念或概念模式频率。如果将类型选作构建类别的输入，那么将使用概念，如果选择了类型模式，那么将使用概念模式。通常，向概念级别应用此方法将产生更具体的结果，这是因为概念和概念模式表示较低级别的度量。
- **类型级别**。选择此选项表示将使用类型或类型模式频率。如果将类型选作构建类别的输入，那么将使用类型，如果选择了类型模式，那么将使用类型模式。向类型级别应用此方法，可快速了解有关给定的信息表示的类型的信息。

最小文档数 计入将具有其类别的项。 使用此选项，可从频繁出现的项构建类别。此选项将输出限制为仅包含描述符的类别，此描述符至少在 X 个记录或文档中出现，其中 X 是为此选项输入的值。

将所有剩余项分组到命名的类别。 使用此选项，可将所有出现不频繁的概念或类型分组到单个具有所选名称的“捕获全部”类别。缺省情况下，此类别名为*其他*。

类别输入。 选择要向其应用方法的组:

- **未使用的抽取结果**。此选项可从任何现有类别中未使用的抽取结果构建类别。这可最大程度降低记录匹配多个类别的趋势，并限制生成的类别数。
- **所有抽取结果**。此选项可使用任何抽取结果构建类别。不存在任何分类或存在很少的分类时，这会很有帮助。

解决重复类别名称方法。 选择如何处理其名称将与现有类别相同的任何新类别或子类别。您可以将新项（及其描述符）与具有相同名称的现有类别合并。或者，如果在现有类别中发现重复名称，可以选择跳过创建任何类别。

扩展类别

扩展是一个用于自动添加或增强描述符以“扩展”现有类别的过程。目标是生成更好的类别，用于捕获最初未分配给此类别的相关记录或文档。

选择的自动分组方法将尝试识别与现有类别描述符相关的概念、TLA 模式和类别规则。然后，会将这些新概念、模式和类别规则作为新描述符添加或添加到现有现有描述符。扩展的分组方法包括*概念根派生*（不可用于日语）、*概念包含*、*语义网络*（仅针对英语文本）以及*同现规则*。**使用从类别名称生成的描述符扩展空类别**方法可使用类别名称中的单词生成描述符，因此，类别名称描述性越强，获取的结果也越好。

注: 扩展类别时，频率方法不可用。

扩展是一种强大的可通过交互方式改善类别的方法。具有一些可能需要扩展类别的示例:

- 拖放概念模式以在“类别”窗格中创建类别后
- 手动创建类别并添加简单类别规则和描述符后
- 导入预定义类别文件（其中类别具有描述性很强的名称）后
- 优化来自期间所选的 TAP 中的类别后

您可以多次扩展类别。例如，如果导入具有描述性很强的名称的预定义类别文件，那么可使用**使用从类别名称生成的描述符扩展空类别**选项进行扩展，以获取第一组描述符，然后再次扩展这些类别。但是，在其他情况中，如果描述符扩展范围越来越广，那么多次扩展可能会生成过于通用的类别。由于构建和扩展分组方法使用相似底层算法，因此在构建类别后直接扩展可能不会生成更相关的结果。

提示:

- 如果尝试扩展且不希望使用结果，那么在扩展后始终可立即撤销操作（**编辑 > 撤销**）。
- 扩展可在一个类别中生成精确匹配一组相同文档的两个或更多类别规则，因为在此过程期间单独构建了规则。如果需要，可通过手动编辑类别描述，查看类别和除去冗余项。请参阅主题第 119 页的『编辑类别描述符』，以获取更多信息。

扩展类别

1. 在“类别”窗格中，选择要扩展的类别。
2. 从菜单中选择**类别 > 扩展类别**。除非选择永不提示，否则会显示一个消息框。
3. 选择希望立即构建还是首先编辑设置。

- 单击**立即扩展**以使用当前设置开始扩展类别。过程将开始，且会显示进度框。
- 单击**编辑**以查看和修改设置。

尝试扩展后，将在“类别”窗格中通过单词**已扩展**对发现新描述符的任何类别添加标记，以便可快速识别这些类别。“已扩展”文本将保留，直到您再次扩展、通过其他方式编辑此类别或通过上下文菜单清除这些类别。

注: 可显示的最大类别数为 10,000。如果达到或超过此数目，那么会显示警告。如果发生此情况，那么应该更改“构建或扩展类别”选项以减少构建的类别数。

构建或扩展类别时可采用的每种方法均适用于特定类型的数据和情况，但通常在相同分析中组合方法对于捕获完整的文档或记录而言很有用。在交互式工作台中，在下次构建类别时，分组到一个类别的概念和类型仍供可用。这表示您可能会看到一个概念处于多个类别中，或找到冗余的类别。

“扩展类别: 设置”对话框中提供了以下区域和字段:

扩展方式。 选择将用于扩展类别的输入:

- **未使用的抽取结果**。此选项可从任何现有类别中未使用的抽取结果构建类别。这可最大程度降低记录匹配多个类别的趋势，并限制生成的类别数。
- **所有抽取结果**。此选项可使用任何抽取结果构建类别。不存在任何分类或存在很少的分类时，这会很有帮助。

分组方法

有关这些方法的简短描述，请参阅第 95 页的『高级语言设置』。这些方法包括:

- **概念根派生**（*不适用于日语*）
- **语义网络**（*仅适用于英语文本，如果未选择“仅一般化”选项，那么不会使用此选项。*）

- **概念包含**
- **同现**和**最小文档数**子选项。

多个类型将从语义网络方法永久排除，因为这些类型不会生成相关结果。它们包含<Positive>、<Negative>、<IP> 和其他非语言类型等。

最大搜索距离。 选择生成类别之前希望通过方法搜索的最大范围。值越小，获得的结果越少，但是，这些结果将更为简单，且更可能互相紧密链接或关联。值越大，获得的结果可能越多，但是，这些结果可靠性和相关性将降低。此选项全局应用于所有方法时，影响最大的是同现和语义网络。

防止对特定概念进行配对。 选中此复选框，以停止在输出中将两个概念分组在一起或配对的过程。要创建或管理概念对，请单击**管理对**。请参阅主题第 97 页的『管理链接异常对』，以获取更多信息。

可能时: 选择是仅扩展描述符、使用通配符一般化描述符还是同时执行这两项操作。

- **扩展并一般化。** 此选项将扩展所选类别，然后一般化描述符。选择一般化时，产品将使用星号通配符在类别中创建通用类别规则。例如，不是生成多个描述符（例如，[apple tart + .] 和 [apple sauce + .]），而是使用通配符可生成 [apple * + .]. 如果使用通配符一般化，那么通常将获得之前所获得的相同数据的记录或文档。但是，此选项具有减少数目和简化类别描述符的优势。此外，此选项还通过针对新文本数据（例如，在纵波研究中）使用这些类别，提高对更多记录或文档进行分类的能力。
- **仅扩展**。此选项将扩展类别而不进行一般化。针对手动创建的类别首选选择**仅扩展**选项，然后使用**扩展并一般化**选项再次扩展相同类别，这会很有帮助。
- **仅一般化。** 此选项将一般化描述符，而不以任何其他方式扩展类别。

  注: 选择此选项将禁用**语义网络**选项；这是因为**语义网络**选项仅在将扩展描述时可用。

其他用于扩展类别的选项

除了可选择要应用的方法，还可编辑以下任一选项:

要通过其扩展描述符的最大项数。 使用项（概念、类型和其他表达）扩展描述符时，定义可添加到单个描述符的最大项数。如果将此限制设置为 10，那么可向现有描述符添加最多 10 个附加项。如果要添加 10 个以上的项，那么方法会在添加第 10 个项后停止添加新项。执行此操作可使描述符列表保持较短，但不保证会首先是要最为相关的项。您可能希望使用**尽可能使用通配符一般化**选项，减小扩展大小而不降低质量。此选项仅适用于包含布尔值 & (AND) 或 ! (NOT) 的描述符。

还扩展子类别。 此选项还将扩展所选类别下的任何子类别。

使用从类别名称生成的描述符扩展空类别。 此方法仅适用于空类别（具有 0 个描述符）。如果类别已包含描述符，那么不会通过此方式对其进行扩展。此选项尝试根据组成类别名称的单词为每个类别自动创建描述符。将扫描类别名称以查看名称中的单词是否匹配任何抽取的概念。如果识别了某个概念，那么将使用此概念查找匹配的概念模式，概念和概念模式将用于形成类别的描述符。当类别名称较长且具有描述性时，此选项会产生最佳效果。通过此方法，可快速生成类别描述符，这些描述符可使目录捕获包含这些描述符的记录。从其他任何位置导入类别时，或手动创建具有较长描述性名称的类别时，此选项最为有用。

一般描述符。 仅当选择先前选项时，此选项才适用。

- **概念。** 选择此选项以生成采用概念形式的描述符，而不管是否已从源文本抽取了这些概念。
- **模式。** 选择此选项以生成采用模式形式的描述符，而不管是否已抽取了生成的模式或任何模式。

手动创建类别

除了使用自动化类别构建方法和规则编辑器创建类别外，还可以手动创建类别。提供了以下手动方法:

- 创建空类别，将向其逐个添加元素。请参阅主题『新建或重命名类别』，以获取更多信息。
- 将术语、类型和模式拖动到类别窗格中。请参阅主题『通过拖放创建类别』，以获取更多信息。

新建或重命名类别

您可以创建空类别以向其添加概念和类型。还可重命名类别。

新建空类别

1. 转至“类别”窗格。
2. 从菜单中选择**类别 > 创建空类别**。将打开“类别属性”对话框。
3. 在“名称”字段中输入此类别的名称。
4. 单击**确定**以接受名称并关闭对话框。对话框将关闭，且会在窗格中显示新类别名称。

现在，您可开始添加到此类别。请参阅主题第 119 页的『向类别添加描述符』，以获取更多信息。

重命名类别

1. 选择类别并选择**类别 > 重命名类别**。将打开“类别属性”对话框。
2. 在“名称”字段中输入此类别的新名称。
3. 单击**确定**以接受名称并关闭对话框。对话框将关闭，且会在窗格中显示新类别名称。

通过拖放创建类别

拖放方法是手动方法，不基于算法。您可以通过拖动以下项在“类别”窗格中创建类别:

- 将“抽取结果”窗格中的抽取概念、类型或模式拖动到“类别”窗格。
- 将“数据”窗格中的抽取概念拖动到“类别”窗格。
- 将“数据”窗格中的整行拖动到“类别”窗格。这会创建一个类别，其中包含此行中具有的所有抽取概念和模式。

注: “抽取结果”窗格支持选择多项以加速多个元素的拖放。

重要！ 无法从“数据”窗格拖放不是抽取自文本的概念。如果要强制抽取数据中发现的概念，那么必须将此概念添加到类型。然后，再次运行抽取。新的抽取结果将包含刚添加的概念。之后，可在类别中进行使用。请参阅主题第 81 页的『将概念添加到类型』，以获取更多信息。

要通过拖放创建类别:

1. 从“抽取结果”窗格或“数据”窗格，选择一个或多个概念、模式、类型、记录或部分记录。
2. 按住鼠标按钮的同时，将元素添加到现有类别或拖动到窗格区域以创建新类别。
3. 到达希望放下元素的区域时，释放鼠标按钮。元素将添加到“类别”窗格。修改的类别显示时带有特殊背景色。此颜色称为**类别反馈背景**。请参阅主题第 68 页的『设置选项』，以获取更多信息。

注: 将自动命名生成的类别。如果希望更改名称，那么可对其进行重命名。请参阅主题『新建或重命名类别』，以获取更多信息。

如果要了解哪些记录分配给类别，请在“类别”窗格中选择此类别。“数据”窗格将自动刷新，并显示此类别的所有记录。

使用类别规则

您可以通过很多方式创建类别。其中一种方式是定义类别规则以表示构想。类别规则是用于使用抽取概念、类型、模式以及布尔运算符基于逻辑表达式，将文档或记录自动分类为类别的语句。例如，可以编写含义为 *include all records that contain the extracted concept* `embassy` *而不是* `argentina` *in this category* 的表达式。

虽然在使用分组方法构建类别时会自动生成一些类别规则，例如，*同现*和*概念根派生*（**类别 > 构建设置 > 高级设置**：**语言**），仍可通过有关在数据和上下文的类别理解在规则编辑器中手动创建类别规则。每个规则附加到一个类别，以便之后匹配规则的每个文档或记录可进行评分以划入此类别。

类别规则通过允许您对具有更大特异性的响应进行分类，帮助增强文本挖掘结果的质量和效率，进一步量化分析。您具有的丰富经验和业务知识可能会为您提供对数据和上下文的具体理解。通过理解数据和上下文，将知识转换为类别规则以通过将抽取元素与布尔值逻辑组合，更有效、准确地对文档或记录进行分类。

具有创建这些规则的能力后，您可将业务知识与产品的抽取方法结合使用，从而增强编码精度和效率。

注：有关规则如何匹配文本的示例，请参阅第 110 页的『类别规则示例』

类别规则语法

虽然在使用分组方法构建类别时会自动生成一些类别规则，例如，*同现*和*概念根派生*（**类别 > 构建设置 > 高级设置**：**语言**），仍可在规则编辑器中手动创建类别规则。每个规则为一个类别的描述符，因此，匹配规则的每个文档或记录可自动进行评分以划入此类别。

注：有关规则如何匹配文本的示例，请参阅第 110 页的『类别规则示例』

创建或编辑规则时，必须在规则编辑器中打开了规则。您可以添加概念、类型或模式，还可使用通配符扩展匹配项。使用抽取的概念、类型和模式时，其将查找所有相关概念，对您会有所帮助。

重要！ 要避免出现常见错误，建议直接从“抽取结果”窗格、“文本链接分析”窗格或“数据”窗格，将概念拖放到规则编辑器中，或可能时通过上下文菜单进行添加。

识别了概念、类型和模式时，会在文本旁显示一个图标。

表 18. 抽取图标

图标	描述
	抽取的概念
	抽取的类型
	抽取的模式

规则语法和运算符

下表包含将用于定义规则语法的字符。使用这些字符以及概念、类型和模式创建规则。

表 19. 支持的语法

字符	描述
&	“and”布尔值。例如，a & b 同时包含 a *和* b ，例如: - invasion & united states - 2016 & olympics - good & apple
|	“or”布尔值是一个包含运算符，表示如何发现任何或所有元素，将执行匹配。例如，a | b 包含 a *或* b，如: - attack | france - condominium | apartment
!()	“not”布尔值。例如，!(a) 不包含 a。如， !(good & hotel) , assassination & !(austria) 或 !(gold) & !(copper)
*	根据使用情况表示单个字符或整个单词的通配符。请参阅主题第 108 页的『在类别规则中使用通配符』，以获取更多信息。
()	表达式定界符。将首先评估括号内的任何表达式。
+	用于形成特定于顺序的模式的模式连接符。如果存在，那么必须使用方括号。请参阅主题『在类别规则中使用 TLA 模式』，以获取更多信息。
[]	如果希望根据类别规则内的抽取的 TLA 模式执行匹配，那么需要模式定界符。方括号中的内容指 TLA 模式，用不会根据简单同现匹配概念或类型。如果未抽取此 TLA 模式，那么不会发现任何匹配项。请参阅主题『在类别规则中使用 TLA 模式』，以获取更多信息。如果希望匹配概念和类型而不是模式，请勿使用方括号。 *注*: 在较旧的版本中，通过类别构建方法生成的同现和同义词规则通常带有方括号。在所有新版本中，方括号现在指示 TLA 模式。而通过同现方法和同义词生成的规则将使用括号，例如，(speaker systems|speakers)。

& 和 | 运算符表示可交换，如，a & b = b & a 以及 a | b = b | a。

使用反斜杠对字符进行转义

如果您所具有的概念包含的任何字符也是语法字符，那么必须在此字符前放置一个反斜杠，以便正确解释此规则。反斜杠 (\) 字符用于对具有特殊含义的字符进行转义。拖放到编辑器后，会自动为您添加反斜杠。

必须在以下规则语法字符前添加反斜扛，以便将其视为不仅仅是规则语法:

```
& ! | + < > ( ) [ ] *
```

例如，由于概念 r&d 包含“and”运算符 (&)，因此在规则编辑器中输入时必须添加反斜杠，例如: r\&d。

在类别规则中使用 TLA 模式

可以在类别规则中显式定义文本链接分析模式，以允许您获取更为具体、上下文更清楚的结果。在类别规则中定义模式时，将绕过较为简单的概念抽取结果和仅基于抽取的文本链接分析模式结果匹配的文档和记录。

重要！ 为了在类别规则中使用 TLA 模式匹配文档，必须在启用了文本链接分析的情况下运行抽取。类别规则将查找此过程期间发现的匹配项。如果未在“文本挖掘”节点的“模型”选项卡中选择浏览 TLA 结果，那么可选择在交互式会话内的抽取设置中启用 TLA 抽取，然后重新抽取。请参阅主题第 74 页的『抽取数据』，以获取更多信息。

使用方括号定界。 如果在类别规则中使用 TLA 模式，那么必须对其使用方括号 [ ]。如果希望根据抽取的 TLA 模式执行匹配，那么需要模式定界符。由于类别规则可包含类型、概念或模式，因此针对此规则，带括号

的内容清除说明了其指代抽取的 TLA 模式。如果未抽取此 TLA 模式，那么不会发现任何匹配项。如果在“类别”窗格中看到某个模式不带方括号（例如，apple + good），那么这可能表示模式已在类别规则编辑器外部直接添加到类别。例如，如果从文本链接分析视图直接向类别添加概念模式，那么显示的词模式不带方括号。但是，在类别规则中使用模式时，必须针对在类别规则内针对此模式使用方括号，例如，[banana + !(good)]。

在模式中使用 + 符号。 在 IBM SPSS Modeler Text Analytics 中，可具有多达 6 部分或插槽的模式。要指示顺序很重要，请使用 + 符号连接每个元素，例如，[company1 + acquired + company2]。此处，顺序很重要，因为顺序将更改所获取的是哪个公司这一含义。顺序不由句子结构确定，而是由构造 TLA 模式输出的方式确定。例如，如果具有文本“*I love Paris*”，且希望抽取此构想，那么 TLA 模式可能为 [paris + like] 或 [<Location> + <Positive>] 而不是 [<Positive> + <Location>]，因为缺省意见资源通常会在带 2 个部分的模式的第二个位置中放置意见。因此，在类别中直接将模式用作描述符可帮助避免出现问题。但是，如果需要将模式用作更复杂语句的一部分，请特别注意“文本链接分析”视图中显示的模式内元素的顺序，因为顺序对于是否可发现匹配项非常关键。

例如，假设具有以下两个样本文本，表达式为“I *like pineapple*""和“I hate *pineapple*. However, I *like* strawberries”。表达式 like & pineapple 将匹配这两个文本，因为它是概念表达式而不是文本链接规则（未带方括号）。表达式 pineapple + like 仅匹配“I *like pineapple*”，因为在第二个文本中，单词 *like* 改为与 *strawberries* 关联。

使用模式分组。 您可以使用自己的模式简化规则。假设您希望捕获以下三个表达式：cayenne peppers + like、chili peppers + like 和 peppers + like。您可以将其分组为单个类别规则，例如，[* peppers & like]。如果具有另一个表达式 hot peppers + good，那么可使用诸如 [* peppers + <Positive>] 的规则对这四项进行分组。

模式中的顺序。 为了更好地组织输出，随产品安装的模板中提供的文本链接分析规则尝试以相同顺序输出基本模式，而不管句子中单词顺序如何。例如，如果具有包含文本“*Good presentations.*”的记录和另一个包含“*the presentations were good*”的记录，那么这两个文本都会通过相同规则进行匹配，且以概念模式结果中 presentation + good 的顺序输出，而不是以 presentation + good 和 good + presentation 的顺序输出。而且在带两个插槽的模式（例如，示例中的模式）中，缺省情况下，分配给“意见”库中类型的概念将在输出中最后部分显示，例如，apple + bad。

表 20. 模式语法和布尔值使用

表达式	匹配文档或记录（
[ ]	包含任何 TLA 模式）。如果希望根据抽取的 TLA 模式执行匹配，那么*类别规则中*需要模式定界符。方括号中的内容指 TLA 模式，而不是简单概念和类型。如果未抽取此 TLA 模式，那么不会发现任何匹配项。 如果要创建未包含任何模式的规则，那么可使用 !([ ]).
[a]	包含的模式中至少其中一个元素为 a，而与其在模式中的位置无关。例如，[deal] 可匹配 [deal + good] 而不仅仅是 [deal + .]
[a + b]	包含概念模式。例如，[deal + good]。 *注：* 如果仅希望捕获此模式，而不添加任何其他元素，建议直接向类别添加模式，而不是使用其制定规则。
[a + b + c]	包含概念模式。+ 符号指示匹配元素的顺序很重要。例如，[company1 + acquired + company2]。
[<A> + <B>]	包含第一个插槽中类型为 <A> 和第二个插槽类型为 <B> 的任何模式，且仅具有两个插槽。+ 符号指示匹配元素的顺序很重要。例如，[<Budget> + <Negative>]。 *注：* 如果仅希望捕获此模式，而不添加任何其他元素，建议直接向类别添加模式，而不是使用其制定规则。

表 20. 模式语法和布尔值使用 (续)

表达式	匹配文档或记录（
[<A> & <B>]	包含带类型 <A> 和类型 <B> 的任何类型模式。例如，[<Budget> & <Negative>]。永不会抽取此 TLA 模式；但是编写时使其实际等于 [<Budget> + <Negative>]|[<Negative> + <Budget>]。匹配元素的顺序不重要。此外，其他元素可能位于模式中，但其必须至少具有 <Budget> 和 <Negative>。
[a + .]	包含一个模式，其中 a 为唯一概念，且此模式的任何其他插槽中没有内容。例如，[deal + .] 匹配其唯一输出为概念 deal 的概念模式。如果将概念 deal 作为类别描述符添加，那么将获得 deal 为概念的所有记录，且包含有关 deal 的正面描述。但是，使用 [deal + .] 将仅匹配表示 deal 而不表示任何其他关系或意见且不匹配 deal + fantastic 的记录模式结果。*注:* 如果仅希望捕获此模式，而不添加任何其他元素，建议直接向类别添加模式，而不是使用其制定规则。
[<A> + <>]	包含 <A> 为唯一类型的模式。例如，[<Budget> + <>] 表示其唯一输出为类型 <Budget> 的概念的模式。*注:* 仅当将 <> 放置到类型模式中的模式 + 符号之后（例如，[<Budget> + <>]，不包括 [price + <>]）时，可使用此项来指示空类型。*注:* 如果仅希望捕获此模式，而不添加任何其他元素，建议直接向类别添加模式，而不是使用其制定规则。
[a + !(b)]	至少包含一个模式，此模式包含概念 a 而不包含概念 b。必须至少包含一个模式。例如，[price + !(high)] 或针对类型，[!(<Fruit>|<Vegetable>) + <Positive>]
!([<A> & <B>])	不包含特定模式。例如，!([<Budget> & <Negative>])。

注: 有关规则如何匹配文本的示例，请参阅第 110 页的『类别规则示例』

在类别规则中使用通配符

可以在规则中向概念添加通配符以扩展匹配功能。可以在单词之前和/或之后放置星号 * 通配符，以指示可如何匹配概念。可使用两种类型的通配符:

- **附加通配符**。这些通配符将立即用作前缀或后缀，字符串和星号之间没有任何空格进行分隔。例如，operat* 可匹配 *operat*、*operate*、*operates*、*operations* 和 *operational* 等。
- **单词通配符**。这些通配符将用作概念的前缀或后缀，同时在概念和星号之间使用空格。例如，* operation 可匹配 *operation*、*surgical operation* 和 *post operation* 等。此外，可将单词通配符与附加通配符一起使用，例如，* operat* *，可匹配 *operation*、*surgical operation*、*telephone operator* 和 *operatic aria* 等。正如您在以上示例中所看到的，建议谨慎使用通配符以便不会使范围太广而捕获到不需要的匹配项。

例外！

- 通配符永远不可单独使用。例如，(apple | *) 是不可接受的。
- 通配符永远不用于匹配类型名称。<Negative*> 不会匹配任何类型名称。
- 无法从通过通配符发现的概念匹配项过滤特定类型。将自动使用向其分配了概念的类型。
- 通配符不能位于单词顺序的中间，不管其是否为某个单词的结尾或开始 (open* account) 还是单独组成部分 (open * account)。也无法在类型名称中使用通配符。例如，word* word，如 apple* recipe 不会匹配 applesauce recipe 也不会匹配其他任何项。但是，apple* * 将匹配 *applesauce recipe*、*apple pie* 和 *apple* 等。在另一个示例中，word * word，如 apple * toast 不会匹配 *apple cinnamon toast* 或其他任何项，因为星号出现在两个单词之间。但是 apple * 将匹配 *apple cinnamon toast*、*apple* 和 *apple pie* 等。

表 21. 通配符使用

表达式	匹配文档或记录（
*apple	包含结尾为写入字母但可能具有任何数目的字符作为前缀的概念。例如: *apple 结尾为字母 *apple* 但可采用前缀，例如: - apple - pineapple - crabapple
apple*	包含开头为写入字母但可能具有任何数目的字符作为后缀的概念。例如，apple* 开头为字母 *apple*，但可采用后缀或不采用后缀，例如: - apple - applesauce - applejack 例如，apple* & !(pear* | quince)，包含开头为字母 apple 的概念，不包含开头为字母 *pear* 概念或概念 quince，将不匹配: apple & quince 但可匹配: - applesauce - apple & orange
product	包含具有写入字母 product 但可具有任何数目的字符作为前缀和/或后缀的概念。 例如: *product* 可匹配: - product - byproduct - unproductive
* loan	包含具有单词 loan 但可能在其之前放置了另一个单词的复合词的概念。例如，* loan 可匹配: - loan - car loan - home equity loan 例如，[* delivery + <Negative>] 包含在第一个位置以单词 delivery 结尾的概念，在第二个位置包含类型 <Negative> 的概念，可匹配以下概念模式: - package delivery + slow - overnight delivery + late
event *	包含具有单词 event 但可能为后跟另一个单词的复合词的概念。例如，event * 可匹配: - event - event location - event planning committee
* apple *	包含可能以任何单词开始后跟单词 apple 且可能后跟另一个单词的概念。* 表示 0 或 n，因此它还匹配 apple。例如，* apple * 可匹配: - gala applesauce - granny smith apple crumble - famous apple pie - apple 例如，[* reservation* * + <Positive>]，包含在一个位置中具有单词 reservation（与其在概念中的位置无关）以及在第二个位置中包含类型 <Positive> 的概念，可匹配概念模式: - reservation system + good - online reservation + good

注: 有关规则如何匹配文本的示例，请参阅第 110 页的『类别规则示例』

类别规则示例

考虑通过以下示例，帮助说明规则如何根据用于进行表示的语法通过不同方式与记录匹配。

示例记录

假设具有两个记录:

- **记录 A**: *"when I checked my wallet, I saw I was missing 5 dollars."*
- **记录 B**: *"$5 was found at the picnic area, but the blanket was missing."*

以下两个表说明可为概念和类型以及概念模式和类型模式抽取的内容。

从示例抽取的概念和类型

表 22. 抽取的概念和类型示例

抽取的概念	输入的概念为
wallet	<Unknown>
missing	<Negative>
USD5	<Currency>
blanket	<Unknown>
picnic area	<Unknown>

从示例抽取的 TLA 模式

表 23. 抽取的 TLA 模式输出示例

抽取的概念模式	抽取的类型模式	从记录
picnic area + .	<Unknown> + <>	记录 B
wallet + .	<Unknown> + <>	记录 A
blanket + missing	<Unknown> + <Negative>	记录 B
USD5 + .	<Currency> + <>	记录 B
USD5 + missing	<Currency> + <Negative>	记录 A

类别规则如何匹配

下表包含可在类别规则编辑器中输入的某个语法。并非此处的所有规则都适用，也并非所有规则匹配相同记录。了解不同语法如何影响匹配的记录。

表 24. 样本规则

规则语法	结果
USD5 & missing	同时匹配记录 A 和 B，因为它们都包含抽取的概念 missing 和 USD5。这等效于: (USD5 & missing)
missing & USD5	同时匹配记录 A 和 B，因为它们都包含抽取的概念 missing 和 USD5。这等效于: (missing & USD5)

表 24. 样本规则 (续)

规则语法	结果
missing & <Currency>	同时匹配记录 A 和 B，因为它们都包含抽取的概念 missing 和匹配类型 <Currency> 的概念。这等效于: (missing & <Currency>)
<Currency> & missing	同时匹配记录 A 和 B，因为它们都包含抽取的概念 missing 和匹配类型 <Currency> 的概念。这等效于: (<Currency> & missing)
[USD5 + missing]	匹配 A 但**不**匹配 B，因为记录 B 未生成包含 USD5 + missing 的任何 TLA 模式输出（请参见上表）。这等效于 TLA 模式输出: USD5 + missing
[missing + USD5]	既不匹配记录 A 也不匹配记录 B，因为抽取的 TLA 模式（请参见上表）匹配此处表示的顺序，且 missing 处于第一个位置中。这等效于 TLA 模式输出: USD5 + missing
[missing & USD5]	匹配 A 但**不**匹配 B，因为此类 TLA 模式都不是抽取自 B。使用字符 & 指示在匹配时顺序不重要；因此，此规则查找 [missing + USD5] 或 [USD5 + missing] 的模式匹配项。仅来自记录 A 的 [USD5 + missing] 具有匹配项。
[missing + <Currency>]	既不匹配记录 A 也不匹配记录 B，因为抽取的 TLA 模式都不匹配此顺序。这没有等效项，因为 TLA 输出仅基于术语 (USD5 + missing) 或类型 (<Currency> + <Negative>)，但不具有混合概念和类型。
[<Currency> + <Negative>]	匹配记录 A 但不匹配记录 B，因为没有 TLA 模式抽取自记录 B。这等效于 TLA 输出: <Currency> + <Negative>
[<Negative> + <Currency>]	既不匹配记录 A 也不匹配记录 B，因为抽取的 TLA 模式都不匹配此顺序。在 Opinions 模板中，缺省情况下，当 *topic* 发现带有 *opinion* 时，*topic* (<Currency>) 会占用第一个插槽位置，*opinion* (<Negative>) 会占用第二个插槽位置。

创建类别规则

创建或编辑规则时，必须在规则编辑器中打开了规则。您可以添加概念、类型或模式，还可使用通配符扩展匹配项。使用识别的概念、类型和模式时，由于其将查找所有相关概念，对您会有所帮助。例如，使用概念时，其所有关联术语、复数形式和同义词也会匹配规则。同样地，使用类型时，规则还会捕获其所有概念。

您可通过编辑现有规则或右键单击类别名称并选择**创建规则**，来打开规则编辑器。

您可使用上下文菜单将概念、类型和模式拖放到编辑器中，或手动在编辑器中输入。然后，将这些对象与布尔运算符（&、!()，|）和方括号组合使用以形成规则表达式。要避免出现常见错误，建议直接从“抽取结果”窗格或“数据”窗格，将概念拖放到规则编辑器中。务必注意规则的语法以避免错误。请参阅主题第 105 页的『类别规则语法』，以获取更多信息。

注: 有关规则如何匹配文本的示例，请参阅第 110 页的『类别规则示例』。

创建规则

1. 如果尚未抽取任何数据或您的抽取到期，请立即执行此操作。请参阅主题第 74 页的『抽取数据』，以获取更多信息。

注: 如果过滤抽取时使所有概念不再可见，那么在尝试创建或编辑类别规则时，会显示错误消息。为防止出现此情况，请修改抽取过滤器以使概念可用。

2. 在“类别”窗格中，选择要在其中添加规则的类别。
3. 从菜单中选择**类别 > 创建规则**。将在窗口中打开类别规则编辑器窗格。
4. 在“规则名称”字段中，输入规则的名称。如果未提供名称，那么会自动将表达式用作名称。稍后可重命名此规则。
5. 在较大的表达式文本字段中，可执行以下操作:
   - 直接在字段中输入文本或从其他窗格执行拖放操作。仅使用抽取的概念、类型和模式。例如，如果输入单词 `cats`，但“抽取结果”窗格中仅显示单数形式 `cats`，那么编辑器将无法识别 `cats`。在此最后情况中，单数形式可能会自动包含复数，否则可使用通配符。请参阅主题第 105 页的『类别规则语法』，以获取更多信息。
   - 选择要添加到规则的概念、类型或模式，并使用菜单。
   - 添加布尔运算符以将规则中的元素链接在一起。使用工具栏按钮向规则添加“and”布尔值 **&**、“or”布尔值 **|** 和“not”布尔值 **!()**、括号 **()** 以及针对模式的方括号 **[ ]**。
6. 单击**测试规则**按钮以验证规则格式是否正确。请参阅主题第 105 页的『类别规则语法』，以获取更多信息。发现的文档或记录数显示在文本**测试结果**旁的括号中。在此文本的右侧，可看到规则中已识别的元素或任何错误消息。如果类型、模式或概念旁的图形显示时带有一个红色问号，那么这指示元素不匹配任何已知抽取。如果元素不匹配，那么规则不会找到任何记录。
7. 要测试规则的一部分，请选择此部分并单击**测试所选项**。
8. 如果发现问题，请进行任何必需更改并重新测试规则。
9. 完成时，单击**保存并关闭**，以再次保存规则并关闭编辑器。新规则名称显示在类别中。

编辑和删除规则

创建并保存规则后，可随时编辑此规则。请参阅主题第 105 页的『类别规则语法』，以获取更多信息。

如果不再需要某个规则，那么可将其删除。

编辑规则

1. 在“类别定义”对话框的“描述符”表中，选择规则。
2. 从菜单选择**类别 > 编辑规则**，或双击规则名称。将打开编辑器并显示所选规则。
3. 使用抽取结果和工具栏按钮对规则执行任何更改。
4. 重新测试规则，以确保其返回预期结果。
5. 单击**保存并关闭**，以再次保存规则并关闭编辑器。

删除规则

1. 在“类别定义”对话框的“描述符”表中，选择规则。
2. 从菜单中选择**编辑并删除**。从类别删除了规则。

导入和导出预定义类别

如果在 Microsoft Excel (*.*xls*, *.*xlsx*) 文件中存储了您自己的类别，那么可将其导入 IBM SPSS Modeler Text Analytics 。

还可将在开放式交互式工作台会话中具有的类别导出为 Microsoft Excel（*.xls 和 *.xlsx）文件格式。导出类别时，可选择包含或排除诸如描述符和评分的一些附加信息。请参阅主题第 116 页的『导出类别』，以获取更多信息。

如果预定义类别不具有代码或需要新代码，那么可通过从菜单选择**类别 > 管理类别 > 自动生成代码**，在“类别”窗格中为一组类别生成一组新的代码。这将除去任何现有代码并对其进行自动重新编号。

导入预定义类别

您可以将预定义类别导入 IBM SPSS Modeler Text Analytics 。导入之前，确保预定义类别文件是 Microsoft Excel (*.xls, *.xlsx) 文件，且使用其中一种支持格式进行构造。还可选择让产品自动检测格式。支持以下格式:

- **平面列表格式**: 请参阅主题第 114 页的『平面列表格式』，以获取更多信息。
- **压缩格式**: 请参阅主题第 114 页的『压缩格式』，以获取更多信息。
- **缩进格式**: 请参阅主题第 115 页的『缩进格式』，以获取更多信息。

导入预定义类别

1. 从交互式工作台菜单，选择**类别 > 管理类别 > 导入预定义类别**。将显示“导入预定义类别”向导。
2. 从“查找范围”下拉列表，选择驱动器和文件所在的文件夹。
3. 从列表中选择文件。将在“文件名”文本框中显示文件的名称。
4. 从列表中选择包含预定义类别的工作表。将在“工作表”字段中显示工作表名称。
5. 要开始选择数据格式，请单击**下一步**。
6. 选择文件的格式，并选择选项以允许产品尝试自动检测格式。对最常见的格式执行自动检测效果最佳。
   - **平面列表格式**: 请参阅主题第 114 页的『平面列表格式』，以获取更多信息。
   - **压缩格式**: 请参阅主题第 114 页的『压缩格式』，以获取更多信息。
   - **缩进格式**: 请参阅主题第 115 页的『缩进格式』，以获取更多信息。
7. 要定义其他导入选项，请单击**下一步**。如果选择自动检测格式，那么会将您定向到最后一个步骤。
8. 如果一个或多个行包含列标题或其他多余信息，请选择要在**开始导入行**选项中从其开始导入的行号。例如，如果类别名称以行 7 开始，那么针对此选项必须输入数字 7 以正确导入文件。
9. 如果文件包含类别代码，请选择选项**包含类别代码**。执行此操作可帮助向导正确识别数据。
10. 查看颜色编码的单元和图注，以确保正确识别了数据。文件中检测到的任何错误显示为红色，并在格式预览表下引用。如果选择了错误格式，请返回并选择其他格式。如果需要纠正文件，请通过再次选择此文件，执行这些更改并重新启动向导。必须先纠正所有错误，才可完成向导。
11. 要查看将导入的一组类别和子类别，以及定义为这些类别创建描述符的方式，请单击**下一步**。
12. 查看将在表中导入的一组类别。如果未看到预期作为描述符的关键字，那么可能是在导入期间未识别这些关键字。确保这些关键字带有正确前缀且显示在正确单元中。
13. 选择希望如何处理会话中的任何预先存在的类别。
    - **替换所有现有类别**。此选项将清除所有现有类别，并在其位置单独使用新导入的类别。
    - **追加到现有类别**。此选项将导入类别，并将任何公共类别与现有类别合并。添加到现有类别时，需要确定希望如何处理任何重复项。一个选择（选项: **合并**）是在任何要导入的类别与现有类别共享一个类别名称时将其合并。另一个选择（选项: **从导入排除**）是禁止导入同名类别。
14. **将关键字作为描述符导入**选项用于将数据中识别的关键字作为关联类别的描述符导入。

15. **通过驱动描述符扩展类别**选项将根据表示类别或子类别名称的单词和/或组成注释的单词来生成描述符。如果单词匹配抽取的结果，那么会将其作为描述符添加到类别。当类别名称或注释都较长且具有描述性时，此选项会产生最佳效果。通过此方法，可快速生成类别描述符，这些描述符可使类别捕获包含这些描述符的记录。
    - 使用**自**字段，可选择派生描述符的文本、类别和子类别的名称和/或注释中的单词。
    - 使用**作为**字段，可选择以概念或 TLA 模式形式创建这些描述符。如果未执行 TLA 抽取，那么会在此向导中禁用**模式**的选项。
16. 要将预定义类别导入“类别”窗格，请单击**完成**。

平面列表格式

在平面列表格式中，仅具有一个不包含任何层次结构的顶级类别，这表示不具有任何子类别或子网。类别名称在单列中显示。

可在此格式的文件中包含以下信息:

- 可选**代码**列包含用于唯一标识每个类别的数字值。如果指定数据文件包含代码（**内容设置**步骤中的**包含类别代码**选项），那么包含每个类别的唯一代码的列必须在类别名称正左侧的单元中。如果数据不包含代码，但希望稍后创建一些代码，那么稍后可始终生成代码（**类别 > 管理类别 > 自动生成代码**）。
- *必需***类别名称**列包含所有类别名称。需要此列以使用此格式导入。
- 类别名称正右侧的单元中的可选**注释**。此注释包含用于描述类别/子类别的文本。
- 可以将可选**关键字**作为类别的描述符导入。为了进行识别，这些关键字必须存在于关联类别/子类别名称正下方单元中，且必须针对关键字列表添加下划线 (_) 字符作为前缀，例如，`_firearms, weapons / guns`。关键字单元可包含一个或多个用于描述每个类别的单词。根据您在向导中最后一个步骤中指定的内容，将这些单词作为描述符导入或忽略这些单词。之后，描述符与文本中的抽取结果进行比较。如果找到匹配项，那么此记录或文档会根据评分划分到包含此描述符的类别。

表 25. 具有代码、关键字和注释的平面列表格式

列 A	列 B	列 C
类别代码（*可选*）	类别名称	注释
	_描述符/关键字列表（*可选*）	

压缩格式

压缩格式的构造与平面列表格式的构造相似，除了压缩格式用于分层类别。因此，需要代码级别列来定义每个类别和子类别的分层级别。

可在此格式的文件中包含以下信息:

- *必需***代码级别**列包含指示此行中后续信息的分层位置的数字。例如，如果指定了值 1、2 或 3 且同时具有类别和子类别，那么 1 用于类别、2 用于子类别，3 用于子类别的子类别。如果仅具有类别和子类别，那么 1 用于类别，2 用于子类别。依次类推，直到所需类别深度。
- 可选**代码**列包含用于唯一标识每个类别的值。如果指定数据文件包含代码（**内容设置**步骤中的**包含类别代码**选项），那么包含每个类别的唯一代码的列必须在类别名称正左侧的单元中。如果数据不包含代码，但希望稍后创建一些代码，那么稍后可始终生成代码（**类别 > 管理类别 > 自动生成代码**）。
- *必需***类别名称**列包含所有类别名称和子类别名称。需要此列以使用此格式导入。
- 类别名称正右侧的单元中的可选**注释**。此注释包含用于描述类别/子类别的文本。
- 可以将可选**关键字**作为类别的描述符导入。为了进行识别，这些关键字必须存在于关联类别/子类别名称正下方单元中，且必须针对关键字列表添加下划线 (_) 字符作为前缀，例如，`_firearms, weapons / guns`。关

键字单元可包含一个或多个用于描述每个类别的单词。根据您在向导中最后一个步骤中指定的内容，将这些单词作为描述符导入或忽略这些单词。之后，描述符与文本中的抽取结果进行比较。如果找到匹配项，那么此记录或文档会根据评分划分到包含此描述符的类别。

表 26. 具有代码的压缩格式示例

列 A	**列 B**	**列 C**
分层代码级别	类别代码（*可选*）	类别名称
分层代码级别	子类别代码（*可选*）	子类别名称

表 27. 不具有代码的压缩格式示例

列 A	**列 B**
分层代码级别	类别名称
分层代码级别	子类别名称

缩进格式

在缩进文件格式中，内容是分层的，这表示它包含类别和一个或多个子类别级别。此外，其结构进行缩进以说明此层次结构。文件中每个行包含一个类别或子类别，但子类别从类别缩进，任何子类别的子类别从子类别缩进，依此类推。您可以而在 Microsoft Excel 中手动创建此结构，或使用从其他产品导出并保存为 Microsoft Excel 格式的结构。

- **顶级类别节点和类别名称**分别位于列 A 和列 B 中。或者，如果不存在任何节点，那么类别名称位于列 A 中。
- **子类别节点和子类别名称**分别位于列 B 和列 C 中。或者，如果不存在任何节点，那么子类别名称位于列 B 中。子类别属于类别。如果不具有顶级类别，那么无法具有子类别。

表 28. 具有代码的缩进结构

列 A	**列 B**	**列 C**	**列 D**
类别代码（*可选*）	类别名称		
	子类别代码（*可选*）	子类别名称	
		子类别代码（*可选*）	子类别名称

表 29. 不具有代码的缩进结构

列 A	**列 B**	**列 C**
类别名称		
	子类别名称	
		子类别名称

可在此格式的文件中包含以下信息:

- 可选**代码**必须为用于唯一标识每个类别或子类别的值。如果指定数据文件包含代码（**内容设置**步骤中的**包含类别代码**选项），那么每个类别或子类别的唯一代码必须在类别/子类别名称正左侧的单元中。如果数据不包含代码，但希望稍后创建一些代码，那么稍后可始终生成代码（**类别 > 管理类别 > 自动生成代码**）。
- 每个类别和子类别*必需***名称**。子类别必须在单独行中从类别向右缩进一个单元。
- 类别名称正右侧的单元中的可选**注释**。此注释包含用于描述类别/子类别的文本。
- 可以将可选**关键字**作为类别的描述符导入。为了进行识别，这些关键字必须存在于关联类别/子类别名称正下方单元中，且必须针对关键字列表添加下划线 (_) 字符作为前缀，例如，`_firearms, weapons / guns`。关

键字单元可包含一个或多个用于描述每个类别的单词。根据您在向导中最后一个步骤中指定的内容，将这些单词作为描述符导入或忽略这些单词。之后，描述符与文本中的抽取结果进行比较。如果找到匹配项，那么此记录或文档会根据评分划分到包含此描述符的类别。

重要！ 如果使用一个级别的代码，那么针对每个类别和子类别必须包含一个代码。否则，导入过程将失败。

导出类别

还可将在开放式交互式工作台会话中具有的类别导出为 Microsoft Excel（*.xls 和 *.xlsx）文件格式。大部分将导出的数据来自“类别”窗格的当前内容或来自类别属性。因此，如果计划还导出**文档**评分值，建议再次评分。

表 30. 类别导出选项

始终导出...	可选择性地导出...
• 类别代码（如果存在）	• **文档**评分
• 类别（和子类别）名称	• 类别注释
• 存在的代码级别（*平面/压缩*格式）	• 描述符名称
• 列标题（*平面/压缩*格式）	• 描述符计数

重要！ 导出描述符时，会将其转换为文本字符串，并添加下划线作为前缀。如果重新导入到此产品，那么将失去区分作为模式的描述符哪些是类别规则，哪些是纯概念的能力。如果要在此产品中复用这些类别，我们强烈建议改为生成文本分析包 (TAP) 文件，因为 TAP 格式会保留当前定义的所有描述符以及使用的所有类别、代码和语言资源。可同时在 IBM SPSS Modeler Text Analytics 和 IBM SPSS Text Analytics for Surveys 中使用 TAP 文件。请参阅主题『使用文本分析包』，以获取更多信息。

导出预定义类别

1. 从交互式工作台菜单，选择**类别 > 管理类别 > 导出类别**。将显示“导出类别”向导。
2. 选择位置，并输入将导出的文件的名称。
3. 在“文件名”文本框中输入输出文件的名称。
4. 要选择类别数据的导出格式，请单击**下一步**。
5. 从以下格式进行选择:
   - **平面或压缩列表格式**: 请参阅主题第 114 页的『平面列表格式』，以获取更多信息。平面列表不包含子类别。请参阅主题第 114 页的『压缩格式』，以获取更多信息。压缩列表格式包含分层类别。
   - **缩进格式**: 请参阅主题第 115 页的『缩进格式』，以获取更多信息。
6. 要开始选择要导出的内容以及查看建议的数据，请单击**下一步**。
7. 查看导出文件的内容。
8. 选择或取消选择要导出的其他内容设置（例如，**注释**或**描述符名称**）。
9. 要导出类别，请单击**完成**。

使用文本分析包

文本分析包（也称为 TAP）用作文本响应分类的模板，使用 TAP 可轻松对文本数据进行分类，同时最大程度减少干预，这是因为 TAP 包含快速、自动对大量记录进行编码所需的预构建类别集和语言资源。使用语言资源，文本数据可得以分析和挖掘，从而抽取关键概念。根据文本中发现的关键概念和模式，可以将记录分类到在 TAP 中所选的类别集。您可以创建自己的 TAP 或更新 TAP。

TAP 包含以下元素:

- **类别集**。类别集基本上包含预定义类别、类别代码、每个类别的描述符以及整个类别集的名称。描述符为语言元素（概念、类型、模式和规则），例如，术语 *cheap* 或模式 *good price*。描述符用于定义类别，以便在文本匹配任何类别描述符时，将文档或记录放入类别中。
- **语言资源**。语言资源是一组库和高级资源，这些库和高级资源经过调整以抽取关键概念和模式。而这些抽取概念和模式用作将记录放入类别集中类别的描述符。

创建您自己的 TAP，更新 TAP 或装入文本分析包。

选择 TAP 并选择类别集后，IBM SPSS Modeler Text Analytics 可抽取和分类记录。

注: 可创建 TAP 并可在 IBM SPSS Text Analytics for Surveys 和 IBM SPSS Modeler Text Analytics 之间交换使用。

生成文本分析包

只要您具有带至少一个类别和一些资源的会话，都可从打开的交互式工作台会话的内容生成文本分析包 (TAP)。可以在 TAP 中生成此组类别和描述符（概念、类型、规则或 TLA 模式输出），同时在资源编辑器中打开所有语言资源。

您可以查看创建资源的语言。在模板编辑器或资源编辑器的“高级资源”选项卡中设置语言。

生成文本分析包

1. 从菜单选择**文件 > 文本分析包 > 生成包**。将显示“生成包”对话框。
2. 浏览到将用于保存 TAP 的目录。缺省情况下，TAP 将保存到产品安装目录的 \TAP 子目录。
3. 在**文件名**字段中输入 TAP 的名称。
4. 在**包标签**字段中输入标签。输入文件名时，会名称会自动显示为标签，但此标签可更改。
5. 要从 TAP 执行类别集，请取消选中**包含**复选框。执行此操作将确保不会将其添加到包。缺省情况下，针对每个问题，会在 TAP 中包含一个类别集。TAP 中必须始终至少具有一个类别集。
6. 重命名任何类别集。缺省情况下，**新建类别集**列包含通用名称，这些名称通过向文本变量名称添加 Cat_ 前缀生成。在单元中单击可编辑此名称。在其他位置输入或单击可应用重命名。如果重命名类别集，那么名称仅会在 TAP 中更改，不会在打开的会话中更改变量名称。
7. 如果需要，使用类别集表右侧的方向键对类别集进行重新排序。
8. 单击**保存**以生成文本分析包。将关闭此对话框。

装入文本分析包

配置文本挖掘建模节点时，必须指定将在抽取期间使用的资源。您可以选择文本分析包 (TAP) 来将其资源以及类别集复制到节点，而不是选择资源模板。

以交互方式创建类别模型时，TAP 最为相关，这是因为您可将类别集用作分类的起始点。执行流时，会启动交互式工作台会话，此组类别将显示在“类别”窗格中。通过此方式，可直接使用这些类别对文档和记录进行评分，然后继续优化、构建和扩展这些类别，直到其满足您的需求。请参阅主题第 87 页的『用于创建类别的方法和策略』，以获取更多信息。

从 V14 开始，还可以在单击**装入**并选择 TAP 时，查看定义此 TAP 中资源所采用的语言。

装入文本分析包

1. 编辑文本挖掘建模节点。
2. 在“模型”选项卡中，在**复制资源源位置**部分中选择*文本分析包*。

3. 单击**装入**。将打开“装入文本分析包”对话框。
4. 浏览到包含您要复制到节点的资源和类别集的 TAP 的位置。缺省情况下，TAP 将保存到产品安装目录的 \TAP 子目录。
5. 在**文件名**字段中输入 TAP 的名称。将自动显示标签。
6. 选择要使用的类别集。此组类别将显示在交互式工作台会话中。然后，可手动或使用“构建类别”或“扩展类别”选项来调整和改进这些类别。
7. 单击**装入**以将文本分析包的内容复制到节点。将关闭此对话框。装入 TAP 时，会将 TAP 的副本复制到节点；因此，在显式更新和重新装入 TAP 后，您对资源和类别执行的任何更改才会反映到 TAP 中。

更新文本分析包

如果改进类别集和语言资源或生成全新类别集，那么可更新文本分析包 (TAP)，以便在以后更易于复用这些改进。要执行此操作，必须处于包含要放入 TAP 中的信息的打开会话中。更新时，可选择追加类别集、替换资源、更改包标签或对类别集进行重命名/重新排序。

更新文本分析包

1. 从菜单选择**文件 > 文本分析包 > 更新包**。将显示“更新包”对话框。
2. 浏览到包含要更新的文本分析包的目录。
3. 在**文件名**字段中输入 TAP 的名称。
4. 要将 TAP 中语言资源替换为当前会话中的语言资源，请选择**将此包中的资源替换为打开项目中的资源**选项。通常，由于语言资源用于抽取用于创建类别定义的关键概念和模式，因此更新这些语言资源会很有意义。具有最新语言资源可确保在对记录进行分类时获取最佳效果。如果未选择此选项，那么包中已具有的语言资源会保持不变。
5. 要仅更新语言资源，请确保选择**将此包中的资源替换为打开会话中的资源**选项，并仅选择 TAP 中已具有的当前类别集。
6. 要将打开会话中的新类别集包含到 TAP 中，请选中要添加的每个类别集的复选框。您可以添加一个或多个类别集，也可以不添加。
7. 要从 TAP 除去类别集，请取消选中相应**包含**复选框。由于您要添加改进的类别集，可以选择除去 TAP 中已具有的类别集。要执行此操作，请取消选中“当前类别集”列中相应类别集的**包含**复选框。TAP 中必须始终至少具有一个类别集。
8. 如果需要，重命名类别集。在单元中单击可编辑此名称。在其他位置输入或单击可应用重命名。如果重命名类别集，那么名称仅会在 TAP 中更改，不会在打开的会话中更改变量名称。如果两个类别集具有相同名称，那么名称会显示为红色，直到您纠正重复项。
9. 要创建新包并将会话内容与所选 TAP 的内容合并，请单击**保存为新项**。将显示“保存为文本分析包”对话框。请参阅以下指示信息。
10. 单击**更新**以保存对所选 TAP 执行的更改。

保存文本分析包

1. 浏览到将用于保存 TAP 文件的目录。缺省情况下，TAP 文件保存到安装目录的 TAP 子目录。
2. 在“文件名”字段中输入 TAP 文件的名称。
3. 在“包标签”字段中输入标签。输入文件名时，会自动将此名称用作标签。但是，可重命名此标签。必须具有标签。
4. 单击**保存**以创建新包。

编辑和优化类别

创建一些类别后，将总是希望检查并执行一些调整。除了优化语言资源，应通过查找组合或清除其定义的方式以及检查部分分类文档或记录，查看类别。还可查看类别中的文档或记录，并进行调整，以便通过捕获细微差别定义类别。

您可以使用内置的自动化类别构建方法来创建类别；但是，可能希望对这些类别执行一些调整。使用一种或多种方法后，将在窗口中显示多个新类别。然后，可查看类别中的数据并进行调整，直到您满意类别定义。请参阅主题第 91 页的『有关类别』，以获取更多信息。

以下提供了一些用于优化类别的选项，其中多数选项在以下页面中进行了描述:

向类别添加描述符

使用自动化方法后，很可能仍会具有在任何类别定义中未使用的抽取结果。您应该在“抽取结果”窗格中查看此列表。如果找到希望将其移动到类别的元素，那么可将其添加到现有类别或新类别。

将概念或类型添加到类别

1. 从“抽取结果”窗格和“数据”窗格，选择要添加到新类别或现有类别的元素。
2. 从菜单中选择**类别 > 添加到类别**。 “所有类别”对话框将显示一组类别。选择要向其添加所选元素的类别。如果要向新类别添加元素，请选择**新类别**。新类别将显示“类别”窗格中，同时使用第一个所选元素的名称。

编辑类别描述符

创建一些类别后，可打开每个类别以查看组成其定义的所有描述符。在“类别定义”对话框中，可对类别描述符进行多次编辑。此外，如果类别显示在类别树中，那么还可从此树使用这些类别。

编辑类别

1. 选择要在“类别”窗格中编辑的类别。
2. 从菜单中选择**视图 > 类别定义**。将打开“类别定义”对话框。
3. 选择要编辑的描述符并单击相应工具栏按钮。

下表描述了可用于编辑类别定义的每个工具栏按钮。

表 31. 工具栏按钮和描述.

图标	描述
	从类别删除所选描述符
	将所选描述符移至新类别或现有类别。
	将采用 & 类别规则形式的所选描述符移至类别。请参阅主题第 105 页的『使用类别规则』，以获取更多信息。
	将每个所选描述符作为其自己的新类别移动
 显示	根据所选描述符更新“数据”窗格和“可视化”窗格中显示的内容。

移动类别

如果要将类别放到其他现有类别中，或将描述符移动到其他类别，可进行移动。

移动类别

1. 在“类别”窗格中，选择要移动到其他类别中的类别。
2. 从菜单中选择**类别 > 移动到类别**。菜单显示一组类别，且会在列表顶部显示最新创建的类别。选择要将所选概念移动到的类别的名称。

- 如果看到要查找的名称，请选择此名称，且会向此类别添加所选元素。
- 如果未看到此名称，请选择**更多**以显示“所有类别”对话框，并从列表选择类别。

序列化类别

具有带类别和子类别的分层类别结构时，可序列化结构。序列化类别时，会将此类别的子类别中所有描述符移动到所选类别，且会删除目前的子类别。通过此方式，用于匹配子类别的所有文档现在分类为所选类别。

序列化类别

1. 在“类别”窗格中，选择要平铺的类别（顶级或子类别）。
2. 从菜单中选择**类别 > 序列化类别**。将除去子类别，且会将描述符合并到所选类别。

合并或组合类别

如果要将两个或更多现有类别组合到新类别，那么可进行合并。合并类别时，那么会使用通用名称创建新类别。类别描述符中使用的所有概念、类型和模式将移动到此新类别。稍后，可以通过编辑类别属性重命名此类别。

合并类别或部分类别

1. 在“类别”窗格中，选择要合并在一起的元素。
2. 从菜单中选择**类别 > 合并类别**。将显示“类别属性”对话框，可在其中输入新创建类别的名称。所选类别作为子类别合并到新类别。

删除类别

如果不再需要保留某个类别，那么可将其删除。

删除类别

1. 在“类别”窗格中，选择要删除的一个或多个类别。
2. 从菜单中选择**编辑并删除**。

第 11 章 分析聚类

您可以在“聚类”视图（**视图 > 聚类**）中构建和探索概念聚类。**聚类**是聚类算法基于这些概念在文档/记录集中出现的频率以及它们一起在相同的文档出现（也称为**共现**）的频率生成的一组相关概念。聚类中的每个概念都随聚类中的至少一个其他概念一起出现。聚类的目标是用来分组在以下情况下一起共同出现的概念：类别的目标是基于包含的文本与每个类别的描述符（概念、规则、模式）的匹配程度分组文档或记录。

优秀的聚类包含具有强链接且频繁出现的概念，而且这些与其他聚类中的概念的链接较少。在处理大型数据集时，此技术可能导致较长的处理时间。

注：使用“构建聚类”对话框中的**用于计算聚类的最大文档数**选项以仅使用所有文档或记录的一部分来构建聚类。

聚类过程首先从分析一组概念并查找文档中经常一起出现的概念。在文档中一起出现的两个概念被视为概念对。接下来，聚类过程通过将概念对一起出现的文档数与每个概念出现的文档数进行比较，来评估每个概念对的**相似性值**。请参阅主题第 123 页的『计算相似性链接值』以获取更多信息。

最后，聚类过程通过汇总并考虑其链接值以及“构建聚类”对话框中定义的设置，将类似概念分组为聚类。汇总是指添加概念，或者将较小的聚类合并到较大的聚类，直至聚类饱和。当再合并概念或较小的聚类将导致聚类超过“构建聚类”对话框中的设置（概念数量、内部链接或外部链接）时，聚类**饱和**。聚类可使用聚类中到其他概念的链接总数最高的概念的名称。

最后，并非所有概念对最终位于同一聚类中，因为可能在其他聚类中存在更强的链接，或者饱和度可能阻止合并它们所在的聚类。对于此原因，提供内部链接和外部链接。

- **内部链接**是聚类中概念对之间的链接。在聚类中，并非所有概念都相互链接。但是，在聚类内，每个概念都至少链接到一个其他概念。
- **外部链接**是不同的聚类中概念对之间的链接（一个聚类中的概念与另一个聚类中的概念）。

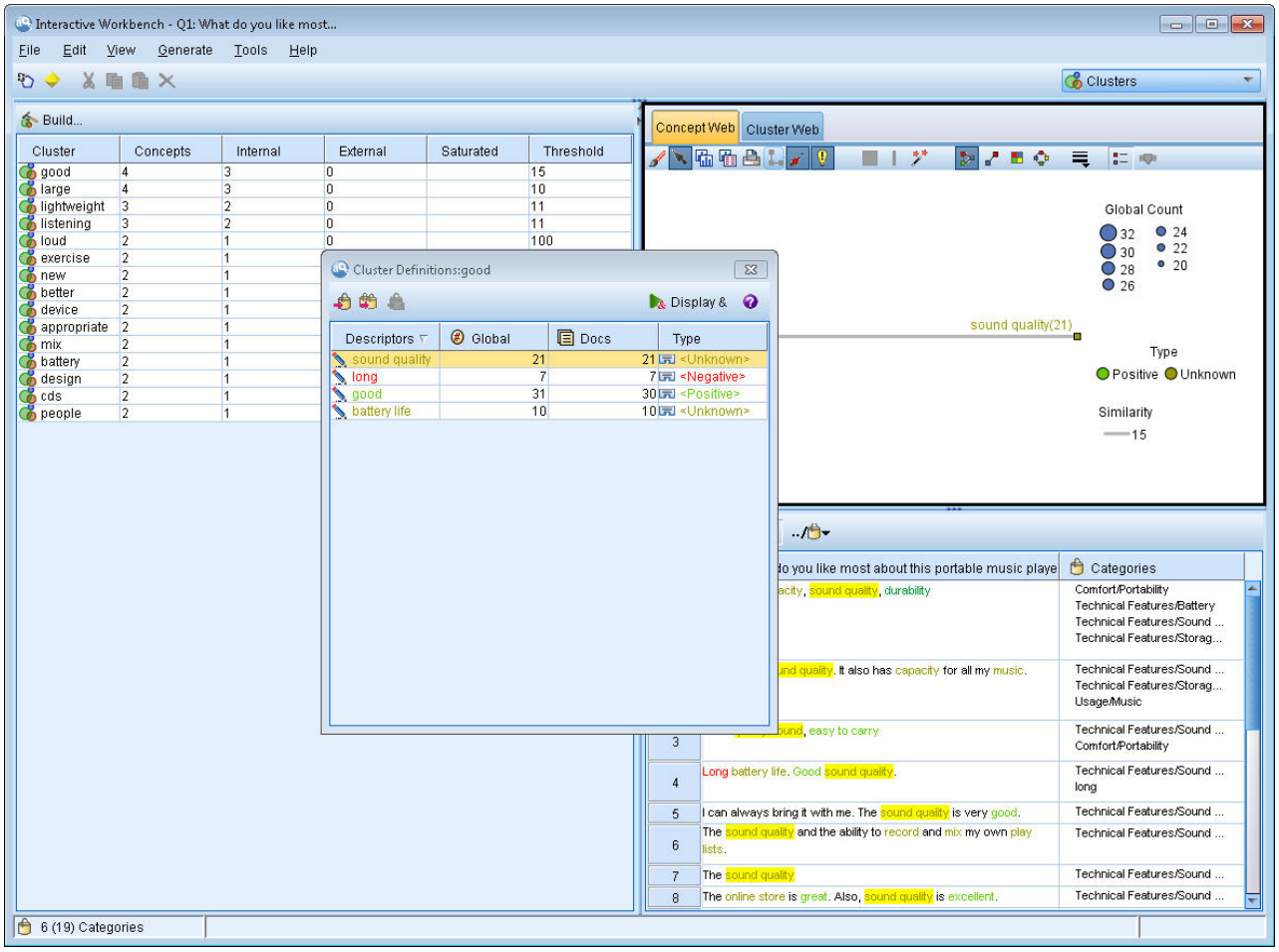


图 30. “聚类”视图

“聚类”视图分为三个窗格，可通过从“视图”菜单中选择其名称来隐藏或显示其中每个窗格:

- **“聚类”窗格。**您可以在此窗格中构建和管理聚类。请参阅主题第 124 页的『探索集群』以获取更多信息。
- **“可视化”窗格。**您可以在此窗格中直观地探索聚类及其交互。请参阅主题第 134 页的『聚类图形』以获取更多信息。
- **“数据”窗格。**您可以探索和复审与“聚类定义”对话框中的选择相对应的文档和记录中包含的文本。请参阅主题第 124 页的『集群定义』以获取更多信息。

构建集群

当您首先访问“集群”视图时，无任何集群可视。您可以通过菜单（**工具 > 构建集群**）或通过单击工具栏上的**构建...** 按钮来构建集群。此操作会打开“构建集群”对话框，可在其中为构建集群定义设置和限制。

注: 只要抽取结果不再与资源匹配，此窗格就会如同“抽取结果”窗格一样变为黄色。您可以重新抽取来获取最新抽取结果，并且黄色将消失。但是，每次执行抽取时都会清除“集群”窗格，并且您将必须重新构建集群。同样，集群不会从一个会话保存到另一个会话。

以下区域和字段在“构建集群”对话框中可用:

输入

输入表。集群是从派生自特定类型的描述符构建而成。在该表中，可以选择要在构建过程中包含的类型。缺省情况下，将会预先选择捕获最多记录或文档的类型。

要集群的概念: 选择选取要用于集群的概念的方法。通过减少概念数，可以加快集群过程。您可以使用顶级概念的数量、顶级概念的百分比或使用所有概念进行集群:

- **基于文档计数的数量。**当选择**最大概念数**时，输入要考虑进行集群的概念的数量。基于具有最高文档计数值的概念来选择概念。Doc 计数是其中出现概念的文档或记录的数量。
- **基于文档计数的百分比。**当选择**最大概念百分比**时，输入要考虑进行集群的概念的百分比。基于具有最高文档计数值的概念的此百分比来选择概念。

要用于计算集群的最大文档数。缺省情况下，使用整个文档集或记录集来计算链接值。但在某些情况下，可能要通过限制用于计算链接的文档或记录的数量来加快集群过程。限制文档数可能会降低集群的质量。要使用此选项，请选中其左侧的复选框并输入要使用的文档或记录的最大数量。

输出限制

要创建的最大集群数。该值是要在“集群”窗格中生成并显示的集群的最大数量。在集群过程中，饱和集群呈现在不饱和集群之前，因此产生的许多集群都将饱和。为查看更多不饱和集群，可以将此设置更改为大于饱和集群数的值。

集群中的最大概念数。该值是集群可以包含的概念的最大数量。

集群中的最小概念数。该值是必须链接才能创建集群的概念的最小数量。

最大内部链接数。该值是集群可以包含的内部链接的最大数量。内部链接是集群内的概念对之间的链接。

最大外部链接数。该值是集群外的概念的链接的最大数量。外部链接是不同集群中的概念对之间的链接。

最小链接值。该值是为考虑将概念对进行集群而接受的最小链接值。使用相似性公式来计算链接值。请参阅主题『计算相似性链接值』以获取更多信息。

防止特定概念的配对。选择此复选框以停止在输出中将两个概念分组在一起或配对的过程。要创建或管理概念对，请单击**管理对**。请参阅主题第 97 页的『管理链接异常对』以获取更多信息。

计算相似性链接值

只是知道概念对出现的文档的数量并不会自行说明两个概念之间的相似程度。在这些情况下，相似性值非常有用。通过将出现文档计数与关系中每个概念单独的文档计数进行比较，来度量相似性链接值。在计算相似性时，计量单位是在其中找到概念或概念对的文档数量（文档计数）。如果概念或概念对在文档中至少出现一次，那么在文档中“找到”此概念或概念对。您可以在概念图中选择线条宽度来表示图中的相似性链接值。

算法揭示这些关系最强，意味着这些概念在文本数据中一起显示的趋势高于其独立出现的趋势。在内部，算法生成 0 和 1 之间的相似性系数，其中值 1 表示两个概念总是一起出现，并且从不分离。然后，相似性系数结果乘以 100 并舍入最近的整数。使用下图中显示的公式计算相似性系数。

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

图 31. 相似性系数公式

其中:

- C_I 是出现概念 I 的文档或记录的数量。
- C_J 是出现概念 J 的文档或记录的数量。
- C_{IJ} 是文档集中概念对 I 和 J 共现的文档或记录的数量。

例如，假定你有 5,000 个文档。 I 和 J 为提取的概念，而 IJ 为 I 和 J 的概念对共现。下表假定两个场景以演示如何计算系数和链接值。

表 32. 概念频率示例

概念/对	场景 A	场景 B
概念: I	在 20 个文档中出现	在 30 个文档中出现
概念: J	在 20 个文档中出现	在 60 个文档中出现
概念对: IJ	在 20 个文档中共现	在 20 个文档中共现
相似性系数	1	0.22222
相似性链接值	100	22

在场景 A 中，概念 I 和 J 以及概念对 IJ 在 20 个文档中出现，生成相似性系数 1，意味着概念总是一起出现。该对的相似性链接值将为 100。

在场景 B 中，I 在 30 个文档中初期，而概念 J 在 60 个文档中出现，但是概念对 IJ 仅在 20 个文档中出现。因此，相似性系数为 0.22222。该对的相似性链接值将舍入为 22。

探索集群

构建集群后，您可以在“集群”窗格中查看结果集。对于每个集群，表中会提供以下信息:

- **集群。**这是集群的名称。集群以具有最高内部链接数的概念命名。
- **概念。**这是集群中的概念的数量。请参阅主题『集群定义』以获取更多信息。
- **内部。**这是集群中的内部链接的数量。内部链接是集群内的概念对之间的链接。
- **外部。**这是集群中的外部链接的数量。外部链接是在一个概念位于一个集群中而另一个概念位于另一个集群中时概念对之间的链接。
- **饱和。**如果存在符号，那么这指示此集群本可更大，但会超过一个或多个限制，因此该集群的集群过程结束并视为*饱和*。在集群过程结束时，饱和集群呈现在不饱和集群之前，因此产生的许多集群都将饱和。为查看更多不饱和集群，可以将**要创建的最大集群数**设置更改为大于饱和集群数的值或减小**最小链接值**。请参阅主题第 122 页的『构建集群』以获取更多信息。
- **阈值。**对于集群中的所有共生概念对，这是集群中其中最小的相似性链接值。请参阅主题第 123 页的『计算相似性链接值』以获取更多信息。具有高阈值的集群表示该集群中的概念比阈值更低的集群中的概念具有跟你广告的整体相似性且更紧密相关。

要了解有关给定集群的信息，可以选择该集群，并且右侧的“可视化”窗格将显示两个图形来帮助探索集群。请参阅主题第 134 页的『聚类图形』以获取更多信息。您还可以将表的内容剪切并粘贴到其他应用程序中。

只要抽取结果不再与资源匹配，此窗格就会如同“抽取结果”窗格一样变为黄色。您可以重新抽取来获取最新抽取结果，并且黄色将消失。但是，每次执行抽取时都会清除“集群”窗格，并且您将必须重新构建集群。同样，集群不会从一个会话保存到另一个会话。

集群定义

您可以通过在“集群”窗格中选择集群并打开“集群定义”对话框来查看该集群内的所有概念（**查看 > 集群定义**）。

所选集群中的所有概念都会显示在“集群定义”对话框中。如果在“集群定义”对话框中选择一个或多个概念并单击**显示 &**，那么“数据”窗格将显示所有记录或文档，其中*所有选定概念都显示在一起*。但是，在“集群”窗格中选择集群时，“数据”窗格不显示任何文本记录或文档。有关“数据”窗格的常规信息，请参阅中。

在此对话框中选择概念还会更改概念 Web 图形。请参阅主题第 134 页的『聚类图形』以获取更多信息。同样，当在“集群定义”对话框中选择一个或多个概念时，“可视化”窗格将显示来自这些概念的所有外部和内部链接。

列描述

系统会显示图标，以便您可以轻松标识每个描述符。

表 33. 列和描述符图标

列	描述
描述符	概念的名称。
 **全局**	显示此描述符在整个数据集中出现的次数，也称为全局频率。
 **文档**	显示其中出现此描述符的文档或记录的数量，也称为文档频率。
类型	显示描述符所属的一个或多个类型。如果描述符是类别规则，那么此列中不显示任何类型名称。

工具栏操作

从此对话框中，还可以选择要在类别中使用的一个或多个概念。执行此操作有多种方法，但最有趣的是选择在集群中共生的概念并将其添加为类别规则。请参阅主题第 100 页的『同现规则』以获取更多信息。可以使用工具栏按钮将概念添加到类别。

表 34. 用于将概念添加到类别的工具栏按钮

图标	描述
	将所选概念添加到新类别或现有类别
	将 & 类别规则形式的所选概念添加到新类别或现有类别。请参阅主题第 105 页的『使用类别规则』以获取更多信息。
	将每个所选概念添加为各自的新类别
	根据所选描述符来更新“数据”窗格和“可视化”窗格中显示的内容

注：您也可以使用上下文菜单将概念作为同义词或作为排除项添加到某个类型。

第 12 章 探索文本链接分析

在文本链接分析 (TLA) 视图中，您可以探索文本链接分析模式结果。文本链接分析是一种模式匹配技术，支持您定义模式规则并将这些与实际提取的概念以及文本中找到的关系进行比较。

例如，提取有关组织的构想可能不足以吸引您。使用 TLA，您还可以了解此组织与其他组织或组织中的人员之间的链接。您还可以使用 TLA 来提取有关产品的意见，或者对于某些语言，种族之间的关系。

在提取某些 TLA 模式结果后，您可以在“文本链接分析”视图的“类型”和“概念模式”窗格中查看它们。请参阅主题第 129 页的『类型和概念模式』以获取更多信息。您可以在此视图的“数据”或“可视化”窗格中进一步探索。最重要的是，您可以将它们添加到类别。

如果尚未选择执行此操作，那么可以单击**提取**，并在“提取设置”对话框中选择**启用文本链接分析模式提取**。请参阅主题第 128 页的『提取 TLA 模式结果』以获取更多信息。

必须在资源模板或库中定义 TLA 模式规则，这些规则用于提取 TLA 模式结果。您可以使用 IBM SPSS Modeler Text Analytics 随附的特定资源模板中的 TLA 模式。提取的关系和模式的种类完全取决于资源中定义的 TLA 规则。您可以针对除日语之外的所有文本语言定义自己的 TLA 规则。模式由宏、字列表和字间距（用于构成布尔值查询）或规则（与输入文本进行比较）组成。请参阅主题第 183 页的第 19 章，『关于文本链接规则』以获取更多信息。

在 TLA 模式规则匹配文本时，可以提取此文本作为模式并重构为输出数据。然后，可以在“文本链接分析”视图窗格中显示结果。可通过从“视图”菜单中选择名称来隐藏或显示每个窗格:

- **“类型”和“概念模式”窗格。**您可以在这两个窗格中构建和探索模式。请参阅主题第 129 页的『类型和概念模式』以获取更多信息。
- **“可视化”窗格。**您可以在此窗格中直观地探索模式中概念和类型的交互。请参阅主题第 135 页的『“文本链接分析”图形』以获取更多信息。
- **“数据”窗格。**您可以探索和复审与另一个窗格中的选择相对应的文档和就路中包含的文本。请参阅主题第 130 页的『数据窗格』以获取更多信息。

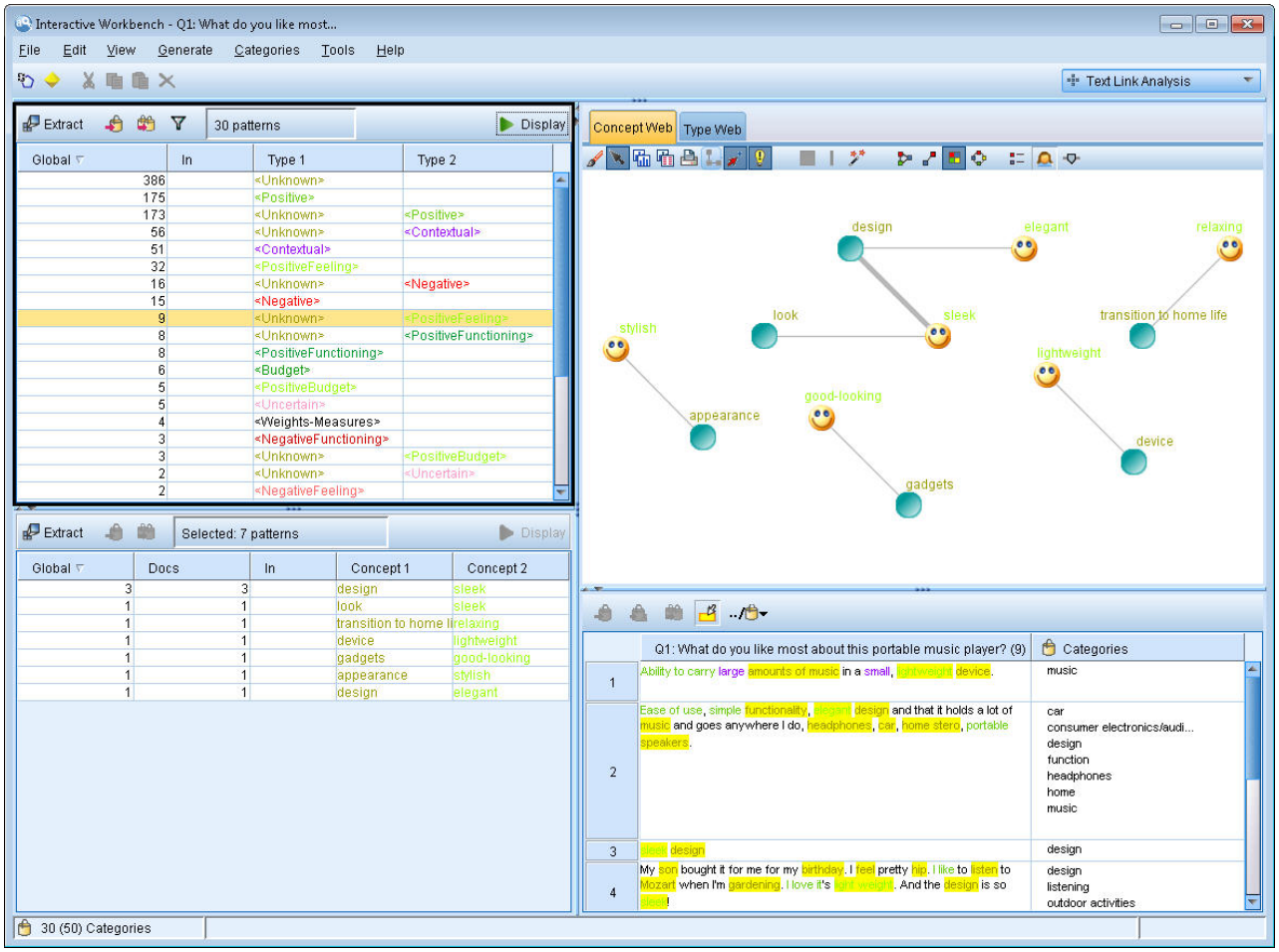


图 32. “文本链接分析”视图

提取 TLA 模式结果

提取过程生成一组概念和类型，以及文本链接分析 (TLA) 模式（如果启用）。如果已提取 TLA 模式，那么您会在“文本链接分析视图中看到这些。在提取结果与资源不同步时，“模式”窗格颜色将变为黄色，指示重新提取将生成不同的结果。

您必须使用**启用文本链接分析模式提取**在节点设置或“提取”对话框中选择提取这些模式。请参阅主题第 74 页的『抽取数据』以获取更多信息。

注: 数据集大小与完成提取过程所用时间有关。请参阅安装指示信息以获取性能统计信息和建议。您可以总是考虑插入样本节点上游或优化机器配置。

要提取数据

1. 从菜单，选择**工具 > 提取**。或者，单击**提取**工具栏按钮。
2. 更改想要使用的任何选项。请记住，必须在此选项卡上选择**启用文本链接分析模式提取**以及在模板中使用 TLA 规则以提取 TLA 模式结果。请参阅主题第 74 页的『抽取数据』以获取更多信息。
3. 单击**提取**以启动提取过程。

在提取开始时，进度对话框将打开。如果想要异常中止提取，请单击**取消**。在提取完成时，对话框关闭并在窗格中显示结果。请参阅主题『类型和概念模式』以获取更多信息。

类型和概念模式

模式由两部分组成，概念和类型的组合。在尝试发现有关特定主题的意见或概念之间的关系时，模式最有用。例如，提取竞争对手的产品名称可能不足以吸引您。在此情况下，您可以查看提取的模式以查看是否可以找到文档或记录包含表示此产品好、坏或昂贵的文本的示例。

模式最多可包含 6 个类型或 6 个概念。因此，两个模式窗格中的行最多包含 6 个槽或者位置。每个槽对应于 TLA 模式规则中一个元素的特定位置，如语言资源定义中所示。在交互式工作台中，如果槽不包含值，那么不会在表格中显示。例如，如果最长的模式结果包含不超过 4 个槽，那么不显示最后两个。请参阅主题第 183 页的第 19 章，『关于文本链接规则』以获取更多信息。

在提取模式结果时，首先将在类型级别进行分组，然后划分为概念模式。因此，提供两个不同的结果窗格：**类型模式**（左上方）和**概念模式**（左下方）。要查看返回的所有概念模式，请选择全部类型模式。底部概念模式窗格然后将显示所有概念模式，直至最大排名值（“过滤器”对话框中定义）。

类型模式。此窗格显示包含一个或多个匹配 TLA 模式规则的相关类型的模式结果。类型模式显示为 `<Organization> + <Location> + <Positive>`，这可提供有关组织在特定位置的正反馈。语法如下所示:

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

概念模式。此窗格显示在上述“类型模式”窗格中当前选中的所有类型模式的概念级别的模式结果。概念模式采用以下结构: `hotel + paris + wonderful`。语法如下所示:

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

在模式结果使用的槽数小于最大值 6 时，仅显示必需的槽（或列）数。将丢弃两个填充槽之间找到的任何空槽，例如，模式 `<Type1>+<>+<Type2>+<>+<>+<>` 可表示为 `<Type1>+<Type3>`。对于概念模式，这可以是 `concept1+.+concept2`（其中，`.` 表示空值）。

如“类别和概念”视图中的提取结果一样，您可以在此查看结果。如果看到要对构成这些模式的类型和概念进行任何改进，那么在“类别和概念”视图的“提取结果”窗格中执行更改，或者直接在“资源编辑器”中执行更改，然后重新提取模式。在作为规则或规则的一部分在类别定义中使用概念、类型或模式时，将在“模式或提取结果”表的**输入**列中显示类别或规则图标。

过滤 TLA 结果

在处理非常大的数据集时，提取过程可能会生产数百万个结果。对于许多用户，此数量使其难于有效地查看结果。然后，您可以过滤这些结果以放大最感兴趣的内容。您可以在“过滤器”对话框中更改设置以显示显示的模式。所有这些设置一起使用。

在 TLA 视图中，“过滤器”对话框包含以下区域和字段。

按频率过滤。您可以过滤以仅显示具有特定全局或文档频率值的结果。

- **全局频率**是模式在整个文档或记录集中出现的总次数，在**全局**列中显示。
- **文档频率**是模式在其中出现的文档或记录的总数，在**文档**列中显示。

例如，如果模式在 500 个记录中出现 300 次，那么可以说此模式的全局频率为 300 并且文档频率为 500。

以及按匹配文本。您还可以进行过滤以仅显示匹配在此处定义的规则的结果。在**匹配文本**字段中输入要匹配的字符集，然后选择是在概念中查找此文本，还是通过标识槽数或全部来输入名称。然后，选择要在其中应用匹配的条件（您无需使用尖括号来表示类型名称的开始或结束）。从下拉列表中选择**与**或**或**，从而使规则匹配两个语句或仅其中一个，并按照与第一个语句相同的方式定义第二个文本匹配语句。

表 35. 匹配文本条件

条件	描述
包含	如果字符串在任意位置出现，那么文本匹配。（缺省选项）
开始于	仅在概念或类型以指定的文本开始时，文本才匹配。
结束于	仅在概念或类型以指定的文本结束时，文本才匹配。
完全匹配	整个字符串必须匹配概念或类型名称。

按排名。您还可以进行过滤以仅显示根据全局频率（**全局**）或文档频率（**文档**）排名前几位的模式，可以是升序或降序。此最大排名值显示针对显示返回的模式总数。

在应用过滤器时，产品添加类型模式，直至到达概念模式的最大总数（排名最大值）。首先查找排名最前的模式类型，然后计算相应的概念模式之和。如果和未超过排名最大值，那么将在视图中显示模式。然后，将对下一个类型模式的概念模式数量求和。如果此数字加上先前类型模式中概念模式的总数小于排名最大值，那么也将在视图中显示这些模式。继续，直至显示不超过排名最大值的尽可能多的模式。

模式窗格中显示的结果

假定您使用英语版本的软件；以下是基于过滤器在“模式”窗格上显示的结果的示例。

图 33. 过滤结果示例 1

在此示例中，工具栏显示由于在过滤器中指定排名最大值，返回的模式结果数受限。如果显示紫色图标，那么意味着满足最大模式数量。在图标上悬停鼠标以获取更多信息。请参阅**以及按排名**过滤器的前置解释。

图 34. 过滤结果示例 2

在此示例中，工具栏显示使用匹配文本过滤器限制结果（请查看放大镜图标）。您可以在图标上悬停鼠标以查看匹配文本。

要过滤结果

1. 从菜单，选择**工具 > 过滤器**。此时“过滤器”对话框将打开。
2. 选择并优化想要使用的过滤器。
3. 单击**确定**以应用过滤器并查看新结果。

数据窗格

在提取和探索文本链接分析模式时，您可能想要查看正在处理的某些数据。例如，您可能想要查看在其中发现一组模式的实际记录。您可以在右下方的“数据”窗格中查看记录或文档。如果缺省情况下未显示，那么从菜单中选择**视图 > 窗格 > 数据**。

“数据”窗格对应于视图中的选择每个文档或记录显示一行，直至特定显示限制。缺省情况下，限制“数据”窗格中显示的文档或记录数量，以便您更快地查看自己的数据。但是，您可以在“选项”对话框中进行调整。请参阅主题第 69 页的『选项: “会话”选项卡』以获取更多信息。

显示和刷新数据窗格

“数据”窗格不会自动刷新其显示，因为对于较大的数据集，自动数据刷新可能需要一些时间才能完成。因此，在此视图中选择类型或概念模式时，您可以单击**显示**以刷新“数据”窗格的内容。

文本文档或记录

如果文本数据采用记录格式，并且文本长度相对较短，那么“数据”窗格中的文本字段整体显示文本数据。但是，在处理记录和较大的数据集时，文本字段列显示文本的一小部分，打开右侧的“文本预览”窗格将显示表中选择的记录的更多文本或全部文本。如果文本数据采用单个文档形式，那么“数据”窗格显示文档的文件名。在选择文档时，“文本预览”窗格打开，并包含选中的文档文本。

颜色和突出显示

在显示数据时，在这些文档或记录中找到的数据、概念和描述符将使用显示突出显示，从而便于在文本中识别。颜色编码对应于概念所属的类型。您还可以在颜色编码的项上悬停鼠标以显示提取的概念以及指定的类型。未提取的任何文本将显示为空白。通常，这些未提取的字通常是连接词（*and* 或 *with*）、代词（*me* 或 *they*）和动词（*is*、*have* 或 *take*）。

数据窗格列

文本字段列始终显示，您还可以显示其他列。要显示其他列，请从菜单中选择**视图 >数据窗格**，然后选择想要在“数据”窗格中显示的列。以下列可供显示:

- **“文本字段名称”(#)/文档。**添加从中提取概念和类型的文本数据集的列。如果数据位于文档中，列称为“文档”，并且仅显示文档文件名或完整路径。要查看这些文档的文本，必须在“文本预览”窗格中进行查看。在列名称后的括号中显示“数据”窗格中的行数。有时候，由于“选项”对话框中为提高装入速度而采取的限制，不显示完整文档或记录。如果到达最大数量，那么数量将后跟 **- Max**。请参阅主题第 69 页的『选项: “会话”选项卡』以获取更多信息。
- **类别。**列出记录所属的每个类别。在显示此列时，刷新“数据”窗格可能需要较长时间才能显示最新信息。
- **相关性排名。**针对单个列别中的每条记录提供一个排名。此排名显示与此类别中的其他记录相比，此记录适合列别的程度。选择“类别”窗格（左上方窗格）中的记录以查看排名。请参阅主题第 93 页的『类别相关性』以获取更多信息。
- **类别计数。**列出记录所属的类别的数量。

第 13 章 可视化图形

“类别和概念”视图、“聚类”视图和“文本链接分析”视图全都在窗口的右上角包含一个可视化窗格。您可以使用此窗格来直观地探索数据。以下图形和图表可用。

- **“类别和概念”视图。**此视图具有三个图形和图表: *类别栏*、*类别 Web* 和*类别 Web 表*。在此视图中，仅在单击**显示**时更新图形。请参阅主题『类别图形和图表』以获取更多信息。
- **“聚类”视图。**此视图具有两个 Web 图形: *概念 Web 图形*和*聚类 Web 图形*。请参阅主题第 134 页的『聚类图形』以获取更多信息。
- **“文本链接分析”视图。**此视图具有两个 Web 图形: *概念 Web 图形*和*类型 Web 图形*。请参阅主题第 135 页的『“文本链接分析”图形』以获取更多信息。

有关用于图形编辑的所有常规工具栏和选用板的更多信息，请参阅联机帮助或 *ModelerSPOnodes.pdf* 文件（作为部分产品下载提供）中的“编辑图形”部分。

类别图形和图表

构建类别时，请务必花时间查看类别定义、其包含的文档或记录以及了解类别如何重叠。可视化窗格提供有关类别的多个透视图。“可视化”窗格位于“类别和概念”视图的右上角。如果它尚不可见，那么可从“视图”菜单（**视图 > 窗格 > 可视化**）访问此窗格。

在此视图中，可视化窗格提供有关文档或记录类别共性的三个透视图。此窗格中的图表和图形可用于分析类别结果，并帮助微调类别或报告。优化类别时，可使用此窗格查看类别定义以发现特别相似（例如，共享超过 75% 的文档或记录）或特别不同的类别。如果两个类别很相似，那么可帮助您决定组合这两个类别。或者，可以通过从一个类别或另一个类别除去特定描述符，决定优化类别定义。

根据在“抽取结果”窗格“分类”窗格或“类别定义”对话框中所选的内容，可在此窗格中每个选项卡上查看文档/记录和类别之间的相应交互。每项表示相似信息，但采用不同方式或使用不同详细信息级别。但是，为了刷新当前所选项的图形，请单击您在其中进行选择的窗格或对话框的工具栏上的**显示**。

“可视化”窗格（“类别和概念”视图中）提供了以下图形和图表:

- **类别条形图。** 表和条形图表示和所选项对应的文档/记录与关联类别之间的重叠。条形图还表示类别中文档/记录数占文档/记录 请参阅主题『类别条形图』，以获取更多信息。
- **类别 Web 图形。** 此图形表示根据在其他窗格中所选项，文档/记录所属的类别的文档/记录重叠。请参阅主题第 134 页的『类别 Web 图』，以获取更多信息。
- **类别 Web 表。** 此表表示的信息与类别 Web 表相同，但格式不同。此表包含三列，可通过单击列标题对这些列进行排序。请参阅主题第 134 页的『类别 Web 表格』，以获取更多信息。

请参阅主题第 85 页的第 10 章，『对文本数据进行分类』，以获取更多信息。

类别条形图

该选项卡显示了一个表格和一个条形图，其中显示了对应于您的选项和关联的类别的文档/记录之间的重叠部分。此条形图还显示了类别中的文档/记录与文档或记录总数的比例。您不能编辑此图表的布局。但是，您可以通过单击列标题来对列进行排序。

此图表包含以下列:

- **类别。**此列用于显示您的选项中的类别名称。缺省情况下，首先列出您的选项中最常用的类别。
- **条形。**此列以可视化方式显示给定类别中的文档或记录数与文档或记录总数的比例。
- **选项比例 (%)。**此列基于某个类别的文档或记录总数逾选项中提供的文档或记录总数的比例显示一个百分比值。
- **文档。** ： 此列显示给定类别中某个选项的文档或记录的数量。

类别 Web 图

该选项卡显示了类别 Web 图。此 Web 根据其他窗格中的选项显示了文档或记录所属类别的文档或记录重叠部分。如果存在类别标签，那么会在图中显示这些标签 。您可以使用此窗格中的工具栏按钮来选择图形布局（网络布局、圆形布局、定向布局或网格布局）。

在 Web 中，每个节点表示一个类别。通过使用鼠标，可以在窗格内选择和移动节点。节点大小表示基于您的选项中该类别的文档或记录数的相对大小。两个类别之间的线粗细和颜色标识两个类别包含的公用文档或记录数。如果在探索模式下将鼠标悬停在节点上，会显示一个工具提示，其中显示此类别的名称（或标签）和此类别中文档或记录的总数。

注: 缺省情况下，针对您可以在其中移动节点的图形已启用探索模式。但是，您可以切换至编辑模式以编辑您的图形布局，包括颜色、字体、图注等。请参阅第 136 页的『使用图形工具栏和选用板』主题以获取更多信息。

类别 Web 表格

此选项卡显示的信息与“类别 Web”选项卡中显示的信息相同，但采用表格格式。表格包含三个列，可通过单击列标题来对其进行排序。

- **计数。**此列显示两个类别之间共享或公用的文档或记录数。
- **类别 1。**此列显示第一个类别的名称，后接其中包含的文档或记录总数（显示在括号内）。
- **类别 2。**此列显示第二个类别的名称，后接其中包含的文档或记录总数（显示在括号内）。

聚类图形

在构建聚类后，您可以在“可视化”窗格的 Web 图形中以直观方式进行探索。可视化窗格提供有关聚类的两个透视图: 概念 Web 图形和聚类 Web 图形。此窗格中的 Web 图形可用于分析聚类结果，并帮助发现可能想要添加到类别的概念和规则。“可视化”窗格位于“聚类”视图右上角。如果不可见，那么您可以从“视图”菜单访问此窗格（**视图 > 窗格 > 可视化**）。通过在“聚类”窗格中选择聚类，您可以自动在“可视化”窗格中显示相应的图形。

注: 缺省情况下，图形采用交互式/选择方式，您可以在其中移动节点。您可以在“编辑”方式中编辑图形布局，包括颜色、字体、图例等。请参阅主题第 136 页的『使用图形工具栏和选用板』以获取更多信息。

“聚类”视图具有两个 Web 图形。

- **概念 Web 图形**。此图形表示所选聚类中的所有概念以及聚类之外链接的概念。该图可帮助查看聚类中概念的链接情况以及任何外部链接。请参阅主题第 135 页的『概念 Web 图形』以获取更多信息。
- **聚类 Web 图形**。该图表示所选聚类以及显示为虚线的所选聚类之间的所有外部链接。请参阅主题第 135 页的『聚类 Web 图形』以获取更多信息。

请参阅主题第 121 页的第 11 章，『分析聚类』以获取更多信息。

概念 Web 图形

此选项卡显示一个 Web 图形，其中显示所选聚类中的所有概念以及聚类之外链接的概念。该图可帮助查看聚类中概念的链接情况以及任何外部链接。聚类中的每个概念都表示为一个节点，这是根据类型颜色编码的颜色。请参阅主题第 163 页的『创建类型』以获取更多信息。

将绘制聚类中概念之间的内部链接，并且根据图形工具栏上选择，每个链接的线条宽度直接与每个概念对共现的文档计数或相似性链接值相关。也将显示聚类概的概念与聚类外部的概念之间的外部链接。

如果在“聚类定义”对话框中选中概念，那么“概念 Web”图形将显示这些概念以及这些概念的任何关联的内部和外部链接。图形上不会显示不包含一个选中的概念的其他概念之间的任何链接。

注：缺省情况下，图形采用交互式/选择方式，您可以在其中移动节点。您可以在“编辑”方式中编辑图形布局，包括颜色、字体、图例等。请参阅主题第 136 页的『使用图形工具栏和选用板』以获取更多信息。

聚类 Web 图形

此选项卡显示一个 Web 图形，其中显示选中的聚类。所选聚类之间的外部链接以及其他聚类之间的任何链接都将显示为虚线。在“聚类 Web”图形中，每个节点都表示整个聚类，而它们之间绘制的线条宽度表示两个聚类之间的外部链接的数量。

重要！为了显示“聚类 Web”图形，您必须已构建具有外部链接的聚类。外部链接是不同的聚类中概念对之间的链接（一个聚类中的概念与另一个聚类中的概念）。

例如，比如说我们有两个聚类。`Cluster A` 具有三个概念：`A1`、`A2` 和 `A3`。`Cluster B` 具有两个概念：`B1` 和 `B2`。以下概念相链接：`A1-A2`、`A1-A3`、`A2-B1`（外部）、`A2-B2`（外部）、`A1-B2`（外部）和 `B1-B2`。这意味着在“聚类 Web”图形中，线条宽度将表示三个外部链接。

注：缺省情况下，图形采用交互式/选择方式，您可以在其中移动节点。您可以在“编辑”方式中编辑图形布局，包括颜色、字体、图例等。请参阅主题第 136 页的『使用图形工具栏和选用板』以获取更多信息。

“文本链接分析”图形

在提取文本链接分析 (TLA) 模式后，您可以在“可视化”窗格的 Web 图形中以直观方式进行探索。可视化窗格提供有关 TLA 模式的两个透视图：概念（模式）Web 图形和类型（模式）Web 图形。此窗格中的 Web 图形可用于直观地表示模式。“可视化”窗格位于“文本链接分析”的右上角。如果不可见，那么您可以从“视图”菜单访问此窗格（**视图 > 窗格 > 可视化**）。如果无选择，那么图形区域为空。

注：缺省情况下，图形采用交互式/选择方式，您可以在其中移动节点。您可以在“编辑”方式中编辑图形布局，包括颜色、字体、图例等。请参阅主题第 136 页的『使用图形工具栏和选用板』以获取更多信息。

“文本链接分析”视图具有两个 Web 图形。

- **概念 Web 图形**。此图形表示所选模式中的所有概念。概念图形中的线条宽度和节点大小（如果不显示类型图标）显示所选表中全局出现的数量。请参阅主题第 136 页的『概念 Web 图形』以获取更多信息。
- **类型 Web 图形**。此图形表示所选模式中的所有类型。图形中的线条宽度和节点大小（如果不显示类型图标）显示所选表中全局出现的数量。节点由类型颜色或图标表示。请参阅主题第 136 页的『类型 Web 图形』以获取更多信息。

请参阅主题第 127 页的第 12 章，『探索文本链接分析』以获取更多信息。

概念 Web 图形

此 Web 图形显示当前选择中的所有概念。例如，如果选择具有三个匹配概念模式的类型模式，那么此图将显示三组链接的概念。概念图形中的线条宽度和节点大小表示全局频率计数。图形直观地表示与模式窗格中选择的内容相同的信息。每个概念的类型由颜色或图标表示，这取决于图形工具栏上的选择。请参阅主题『使用图形工具栏和选用板』以获取更多信息。

类型 Web 图形

此 Web 图形显示当前选择的每个类型模式。例如，如果选择了两个概念模式，那么该图将在选中的模式中每个类型显示一个节点，并显示在相同模式中找到的这些项之间的链接。线条宽度和节点大小表示集合的全局频率计数。图形直观地表示与模式窗格中选择的内容相同的信息。除了图形中显示的类型名称，还通过颜色或类型图标来表示类型，这取决于图形工具栏上的选择。请参阅主题『使用图形工具栏和选用板』以获取更多信息。

使用图形工具栏和选用板

对于每个图形，存在一个工具栏为您提供对某些常用选用板的快速访问，您可以从这些选用板对图形执行多个操作。每个视图（“类别和概念”、“集群”和“文本链接分析”）都包含一个工具栏，这些工具栏之间稍有不同。您可以在*探索*视图模式或*编辑*视图模式之间进行选择。

“探索模式”允许您以分析方式来探索可视化所提供的数据和值，“编辑模式”允许您更改可视化的布局和外观。例如，您可以更改字体和颜色以匹配组织的样式指南。要选择此模式，请从菜单中选择**视图 > 可视化窗格 > 编辑模式**（或单击工具栏图标）。

在编辑模式中，存在多个影响可视化布局不同方面的工具栏。如果您发现有任何不需要使用的工具栏，可以将其隐藏以增加对话框中显示图形的空间。要选择或取消选择工具栏，请单击“视图”菜单上的相关工具栏或者选用板名称。

有关用于图形编辑的所有常规工具栏和选用板的更多信息，请参阅联机帮助或 *SourceProcessOutputNodes.pdf* 文件（作为部分产品下载提供）中的“编辑图形”部分。

表 36. 文本分析工具栏按钮.

按钮/列表	描述
	启用编辑模式。切换至编辑模式以更改图形外观，例如，放大字体、更改颜色以匹配公司样式指南或者除去标签和图注。
	启用探索模式。缺省情况下开启“探索模式”，这表示您可以围绕图形移动和拖动节点，并且可以悬停在图形对象上以显示额外的工具提示信息。
	为“类别和概念”视图以及“文本链接分析”视图中的图形选择 Web 显示类型。 • **圆形布局。**可应用于任何图形的常规布局。它会按链接未定向方式来布局图形，以相同方式处理所有节点。仅围绕圆圈外围放置节点。 • **网络布局。**可应用于任何图形的常规布局。它会按链接未定向方式来布局图形，以相同方式处理所有节点。在布局内自由放置节点。 • **定向布局。**仅应用于定向图形的布局。此布局会生成类树结构，从根节点起直至叶节点，按颜色组织。分层数据按此布局显示较为美观。 • **网格布局。**可应用于任何图形的常规布局。它会按链接未定向方式来布局图形，以相同方式处理所有节点。节点仅置于此空间内的网格点。

表 36. 文本分析工具栏按钮 (续).

按钮/列表	描述
	链接大小表示法。选择图形中表示的线条厚度。这仅适用于“集群”视图。“集群”Web 图形仅显示集群之间外部链接的数量。您可以选择: • **相似性。**厚度指示两个集群之间的外部链接数量。 • **共现性。**厚度指示其中发生描述符共现的文档数量。
	切换按钮，按下此按钮时会显示图注。不按下此按钮时，不显示图注。
	切换按钮，按下此按钮时会在图形中显示类型图标而不是类型颜色。这仅适用于“文本链接分析”视图。
	切换按钮，按下此按钮时会在图形下显示链接滑块。您可以通过滑动箭头来过滤结果。
	将显示所选类别的最高级别的类别图形，而不是其子类别。
	将显示所选最低级别的类别图形。
	该选项控制在输出中显示子类别名称的方式。 • **完整的类别路径。**该选项将输出类别的名称和父类别的完整路径（如果适用），使用斜杠来分隔类别名称与子类别名称。 • **简短的类别路径。**该选项将仅输出类别名称，但是用省略符来显示存在问题的类别的父类别数量。 • **底部级别类别。**该选项将仅输出类别名称，不显示完整路径或父类别。

第 14 章 会话资源编辑器

IBM SPSS Modeler Text Analytics 快速且准确地从文本数据捕获和提取关键字。此提取过程主要依赖于语言资源以指定如何从文本数据提取信息。缺省情况下，这些资源来自于资源模板。

IBM SPSS Modeler Text Analytics 随附一组专用的**资源模板**，其中包含语言和非语言资源，采用库和高级资源形式，旨在帮助定义如何处理和提取数据。请参阅主题第 143 页的第 15 章，『模板和资源』以获取更多信息。

在节点对话框中，您可以将模板资源的副本装入节点。在交互式工作台会话中，您可以专门针对节点数据定制这些资源（如果想要）。在交互式工作台会话期间，您可以在资源编辑器视图中处理资源。在启动交互式会话时，将使用节点对话框中装入的资源来执行提取，除非已在节点中高速缓存数据和提取结果。

在资源编辑器中编辑资源

资源编辑器 针对用于在交互式工作台会话中生成提取结果（概念、类型和模式）的资源提供资源集访问。此编辑器非常类似于模板编辑器，但是在资源编辑器中除外，您可以在其中定义此会话的资源。您完成处理资源以及任何其他工作后，您可以更新建模节点以保存此工作，从而可在后续交互式工作台会话中复原。请参阅主题第 70 页的『更新建模节点并保存』以获取更多信息。

如果想要直接处理用于将资源装入节点的模板，建议使用模板编辑器。您可以在资源编辑器中执行的大多数任务与在模板编辑器中执行类似，例如:

- **处理库**。请参阅主题第 153 页的第 16 章，『处理库』以获取更多信息。
- **创建类型字典**。请参阅主题第 163 页的『创建类型』以获取更多信息。
- **向字典添加术语**。请参阅主题第 164 页的『添加术语』以获取更多信息。
- **创建同义词**。请参阅主题第 168 页的『定义同义词』以获取更多信息。
- **导入和导出模板**。请参阅主题第 149 页的『导入和导出模板』以获取更多信息。
- **发布库**。请参阅主题第 158 页的『发布库』以获取更多信息。

针对荷兰语、英语、法语、德语、意大利语、葡萄牙语和西班牙语文本

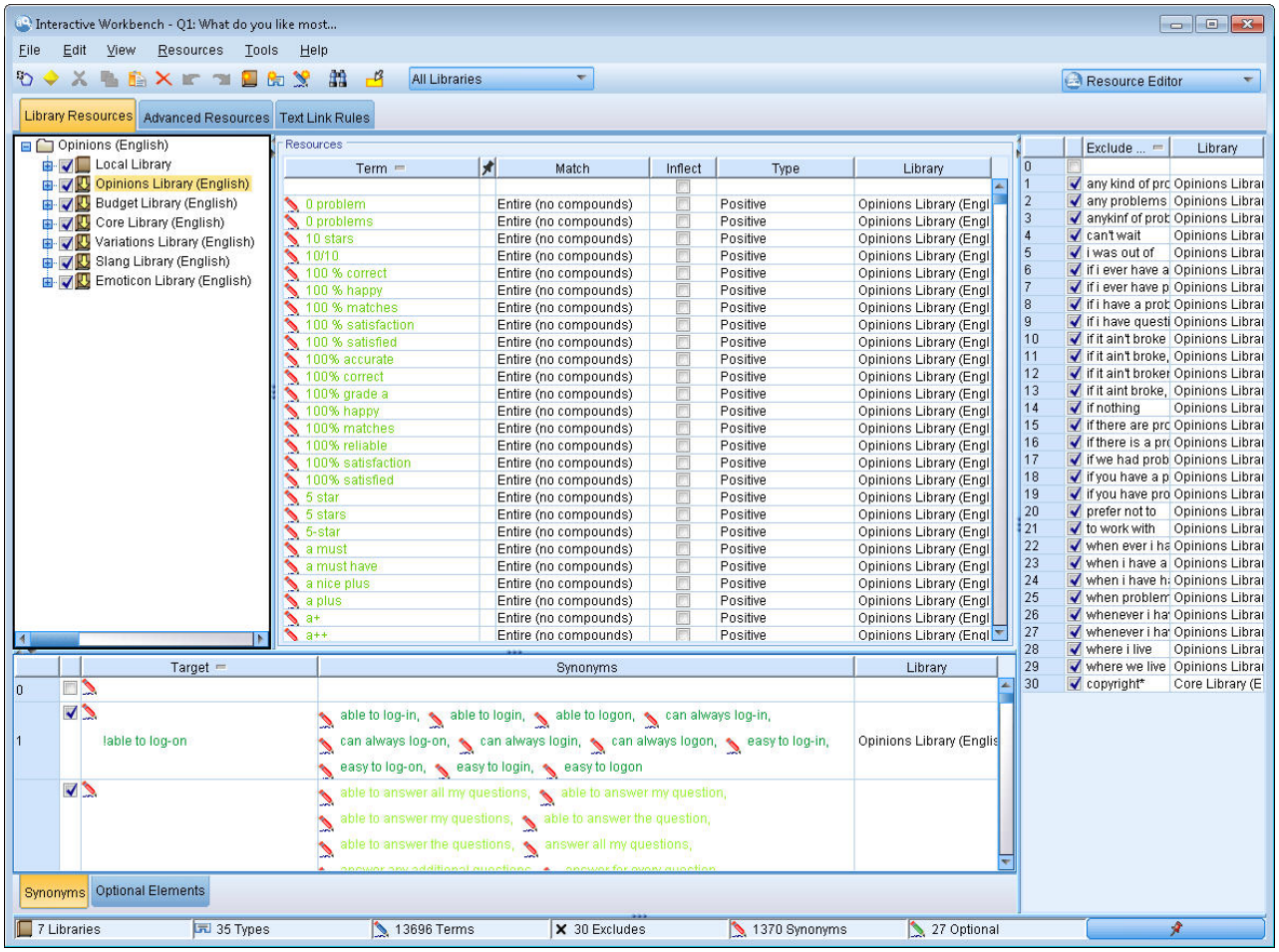


图 35. 非日语语言的“资源编辑器”视图

对于日语文本

日语文本语言的编辑器界面与其他文本语言不同。

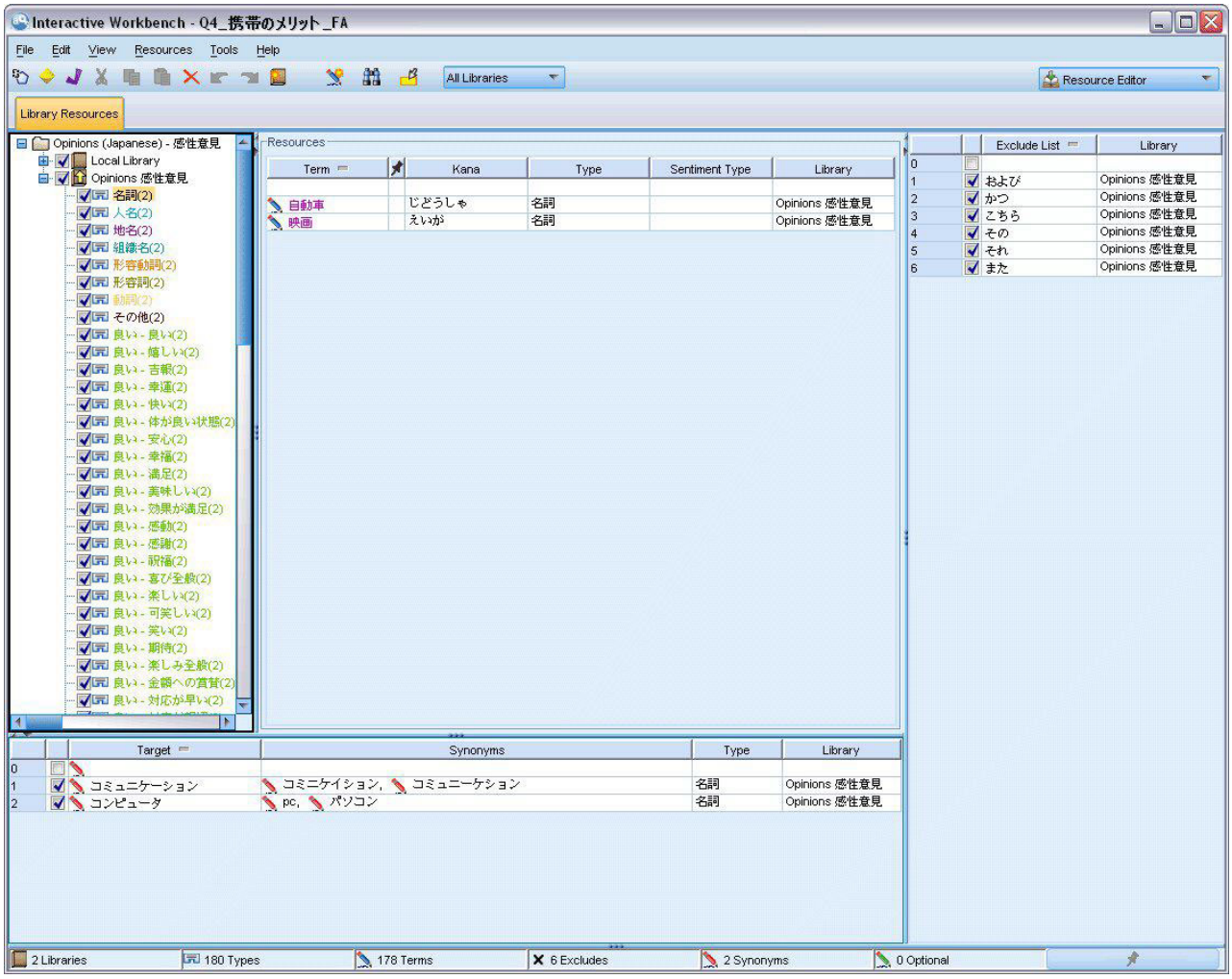


图 36. 针对日语文本的“资源编辑器”视图

创建和更新模板

对资源进行更改并且想要在将来复用这些资源时，可以将这些资源保存为模板。执行此操作时，可以选择使用现有模板名称进行保存或者提供新名称。然后，将来装入此模板时，将能够获取相同的资源。请参阅第 21 页的『从模板和 TAP 复制资源』主题以获取更多信息。

注：您还可以发布和共享自己的库。请参阅第 157 页的『共享库』主题以获取更多信息。

要创建（或更新）模板

1. 从资源编辑器视图的菜单中选择**资源 > 创建资源模板**。这样会打开“创建资源模板”对话框。
2. 如果要创建新模板，请在“模板名称”字段中输入新名称。如果要使用当前装入的资源覆盖现有模板，请选择表中的模板。
3. 单击**保存**以创建模板。

要点！由于是在节点中选择模板时而不是在执行流时装入模板的，如果要获取最新更新，请确保在使用此资源模板的任何其他节点中重新装如此资源模板。请参阅第 148 页的『装入后更新节点资源』主题以获取更多信息。

切换资源模板

如果要将会话中当前装入的资源替换为来自其他模板的资源副本，可以切换至这些资源。这样将覆盖会话中当前装入的所有资源。如果要切换资源以使用部分预定义的文本链接分析 (TLA) 模式规则，请确保选中模板以在 TLA 列中标记这些模板。.

要点！无法从日语模板切换至非日语模板，反之亦然。

当您想要恢复会话工作（类别、模式和资源）但是希望从模板装入经更新的资源副本并且不丢失其他会话工作时，切换资源特别有用。您可以选择想要将其中内容复制到资源编辑器中的模板，然后单击**确定**。这样会替换您在此会话中所拥有的资源。如果要在下次启动交互式工作台会话时保留这些更改，请确保在会话结束时更新建模节点。

注: 如果在交互式会话期间切换至另一个模板的内容，那么节点中列出的模板的名称仍将是上次装入和复制的模板的名称。为从这些资源或其他会话工作中获益，请在退出会话之前更新您的建模节点，并选中节点中的**使用会话工作**选项。请参阅第 70 页的『更新建模节点并保存』主题以获取更多信息。

要切换资源

1. 从资源编辑器视图的菜单中选择**资源 > 切换资源模板**。这样会打开“切换资源”对话框。
2. 从表中显示的模板中选择您要使用的模板。
3. 单击**确定**以放弃当前已装入的资源，从选中模板中装入这些资源的副本以代替这些资源。如果对资源进行了更改，并且想要保存库以供将来使用，那么可以在切换之前发布、更新和共享这些库。请参阅第 157 页的『共享库』主题以获取更多信息。

第 15 章 模板和资源

IBM SPSS Modeler Text Analytics 快速且精确地从文本数据捕获和提取关键概念。此提取过程主要依赖于语言资源以指定如何从文本数据提取信息。请参阅主题第 4 页的『抽取的工作方式』以获取更多信息。您可以在资源编辑器视图中微调这些资源。

在安装软件时，还会获取一组专用资源。这些随附资源使您能够利用数年的研究，并且它们针对特定语言和特定应用程序进行微调。因为随附的资源并非总是完美地适合您的数据上下文，您可以编辑这些资源模板，甚至是创建和使用专门针对您的组织数据进行过微调的定制库。这些资源以各种形式提供，并且可用于您的会话。可以在以下位置找到资源:

- **资源模板**。模板由一组库、类型和一些高级资源构成，它们一起构成一组适合特定域或上下文（例如，产品意见）的专用资源。
- **文本分析包 (TAP)**。除了模板中存储的资源，TAP 还捆绑了一个或多个使用这些资源生成的专用类别集，从而使类别和资源存储在一起并且可复用。请参阅主题第 116 页的『使用文本分析包』以获取更多信息。
- **库**。库用作 TAP 和模板的构建块。也可以单独添加到会话中的资源。每个库都由多个字典组成，这些字典用于定义和管理类型、同义词和排除列表。在还单独交付库时，库将与模板和 TAP 预先打包在一起。请参阅主题第 153 页的第 16 章，『处理库』以获取更多信息。

注: 在提取期间，还会使用某些已编译的内部资源。这些已编译的资源包含大量定义，用于补充核心库中的类型。这些已编译的资源不可编辑。

资源编辑器 针对用于生成提取结果（概念、类型和模式）的资源提供资源集访问。您可以在资源编辑器中执行大量任务，包括:

- **处理库**。请参阅主题第 153 页的第 16 章，『处理库』以获取更多信息。
- **创建类型字典**。请参阅主题第 163 页的『创建类型』以获取更多信息。
- **向字典添加术语**。请参阅主题第 164 页的『添加术语』以获取更多信息。
- **创建同义词**。请参阅主题第 168 页的『定义同义词』以获取更多信息。
- **更新 TAP 中的资源**。请参阅主题第 118 页的『更新文本分析包』以获取更多信息。
- **生成模板**。请参阅主题第 141 页的『创建和更新模板』以获取更多信息。
- **导入和导出模板**。请参阅主题第 149 页的『导入和导出模板』以获取更多信息。
- **发布库**。请参阅主题第 158 页的『发布库』以获取更多信息。

模板编辑器与资源编辑器

可通过两个主要方法来处理和编辑模板、库及其资源。您可以在模板编辑器或资源编辑器中处理语言资源。

模板编辑器

模板编辑器允许您创建和资源资源模板，而无需工作台会话和独立的特定节点或流。您可以使用此编辑器来创建或编辑资源模板，然后将它们装入“文本链接分析”节点和“文本挖掘”建模节点。

模板编辑器可通过 IBM SPSS Modeler 主工具栏的**工具 > 文本分析模板编辑**菜单进行访问。

资源编辑器

资源编辑器可通过交互式工作台会话进行访问，允许您在特定节点和数据库的上下文中处理资源。在将“文本挖掘”建模节点添加到流时，您可以装入资源模板的内容或文本分析包（类别集*和*资源）的副本，来控制针对文本挖掘提取文本的方式。在启动交互式工作台会话时，除了创建类别、提取文本链接分析模式以及创建类别模型，您还可以在集成的 资源编辑器 视图中微调此会话数据的资源。请参阅主题第 139 页的『在资源编辑器中编辑资源』以获取更多信息。

在交互式工作台会话中处理资源时，这些更改仅适用于此会话。如果想要保存工作（资源、类别、模式等）以便在后续会话中继续，必须更新建模节点。请参阅主题第 70 页的『更新建模节点并保存』以获取更多信息。

如果想要将更改保存回原始模板（其内容已复制到建模节点），从而可以将此更新的模板装入其他节点，您可以从资源生成模板。请参阅主题第 141 页的『创建和更新模板』以获取更多信息。

编辑器界面

在模板编辑器或资源编辑器中执行的操作涉及管理和微调语言资源。这些资源存储为模板和库形式。请参阅主题第 161 页的『类型字典』以获取更多信息。

“库资源”选项卡

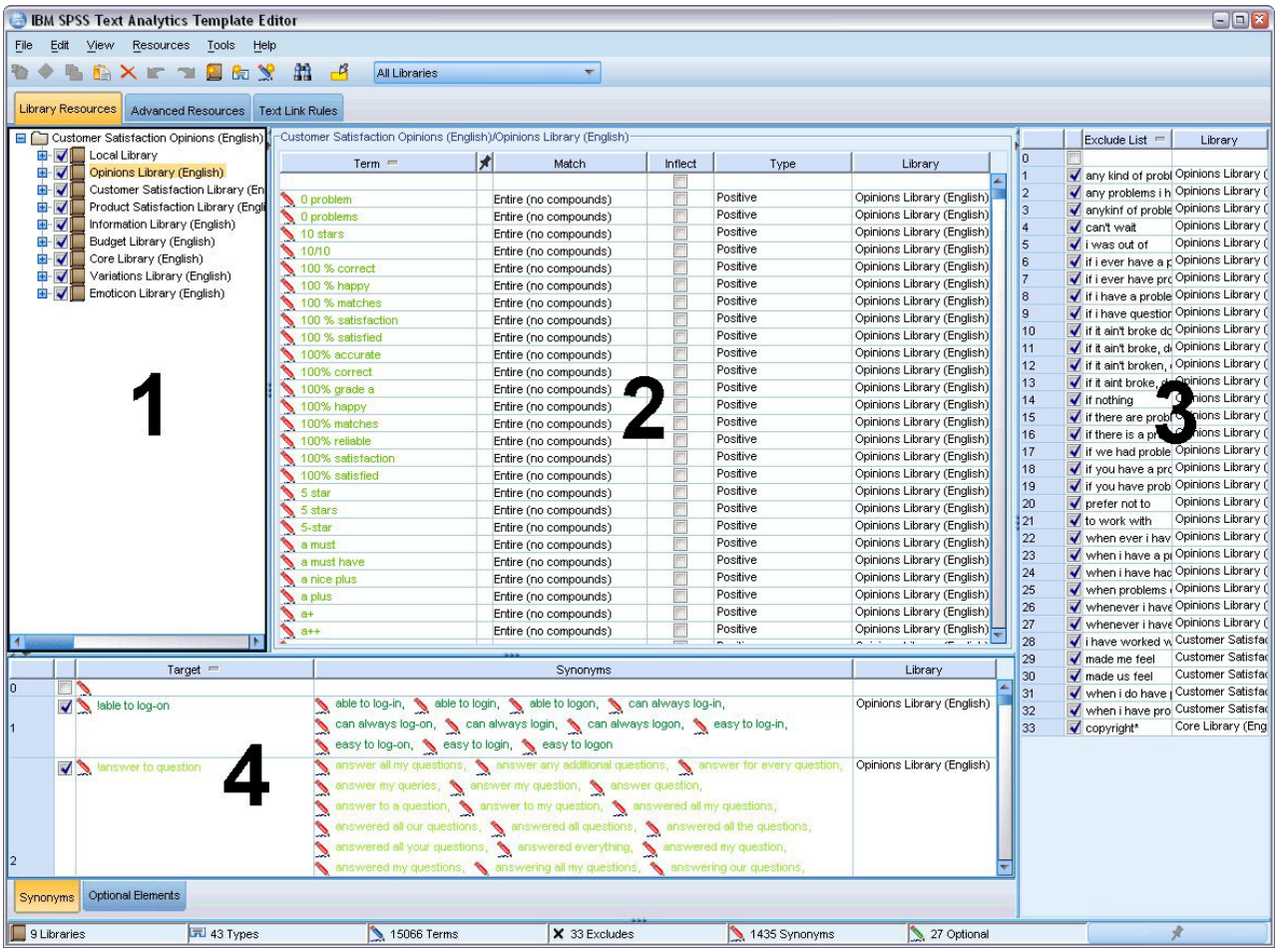


图 37. 文本挖掘模板编辑器

界面组织为四个部分，如下所示:

1.“库树”窗格。 此窗格位于左上角，显示了库的树。您可以在此树中启用和禁用库，还可通过在树中选择库来在其他窗格中过滤视图。您可以使用上下文菜单，在此树中执行很多操作。如果在树中展开某个库，那么可查看其包含的一组类型。如果希望仅关注特定库，那么还可通过**视图**菜单过滤此列表。

2. “类型字典”窗格中的“术语列表”。 此窗格位于库树右侧，显示了树中所选库的类型字典的术语列表。**类型字典**是要在一个标签、类型和名称下分组的一组术语。抽取引擎读取文本数据时，会将文本中发现的单词与类型字典中的术语进行比较。如果抽取的概念在类型字典中显示为术语，那么会分配此类型名称。您可以将类型字典视为包含具有公共内容的术语的不同字典。例如，核心库中 `<Location>` 类型包含诸如 `new orleans`、`great britain` 和 `new york` 的概念。这些术语全部表示地理位置。库可包含一个或多个类型字典。请参阅主题第 161 页的『类型字典』，以获取更多信息。

3. “排除字典”窗格。 此窗格位于右侧，显示将从最终抽取结果排除的术语集合。在此排除字典中显示的术语不会显示在“抽取结果”窗格中。排除术语可以存储在所选的库中。但是，“排除字典”窗格显示库树中可见的所有库的所有排除术语。请参阅第 170 页的『排除字典』，以获取更多信息。

4.“替换字典”窗格。 此窗格位于左下方，在其选项卡中显示同义词和可选元素。在最终抽取结果中，同义词和可选元素帮助将相似术语分组到一个引导术语、目标或概念下。此字典可包含已知同义词和用户定义的同义词和元素以及与正确拼写配对的常见错误拼写。可以在所选库中存储同义词定义和可选元素。但是，替换字典窗格显示库树中可见的所有库的所有内容。虽然此窗格显示所有库中所有同义词或可选元素，但是树中所有库的替换项会在此窗格中显示在一起。一个库仅可包含一个替换字典。请参阅主题第 167 页的『替换/同义词字典』，以获取更多信息。 请注意，“可选元素”选项卡不适用于日语文本语言资源。

注:

- 如果要过滤以便仅看到与单个库相关的信息，那么可使用工具栏上的下拉列表更改库视图。它包含名为**所有库**的顶级条目，针对每个单独库包含一个额外的条目。请参阅主题第 155 页的『查看库』，以获取更多信息。
- 日语文本语言的编辑器界面与其他文本语言不同。

“高级资源”选项卡

可从编辑器视图的第二个选项卡使用高级资源。您可以在此选项卡中复审和编辑高级资源。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。

重要！此选项卡不适用于针对日语文本优化的资源。

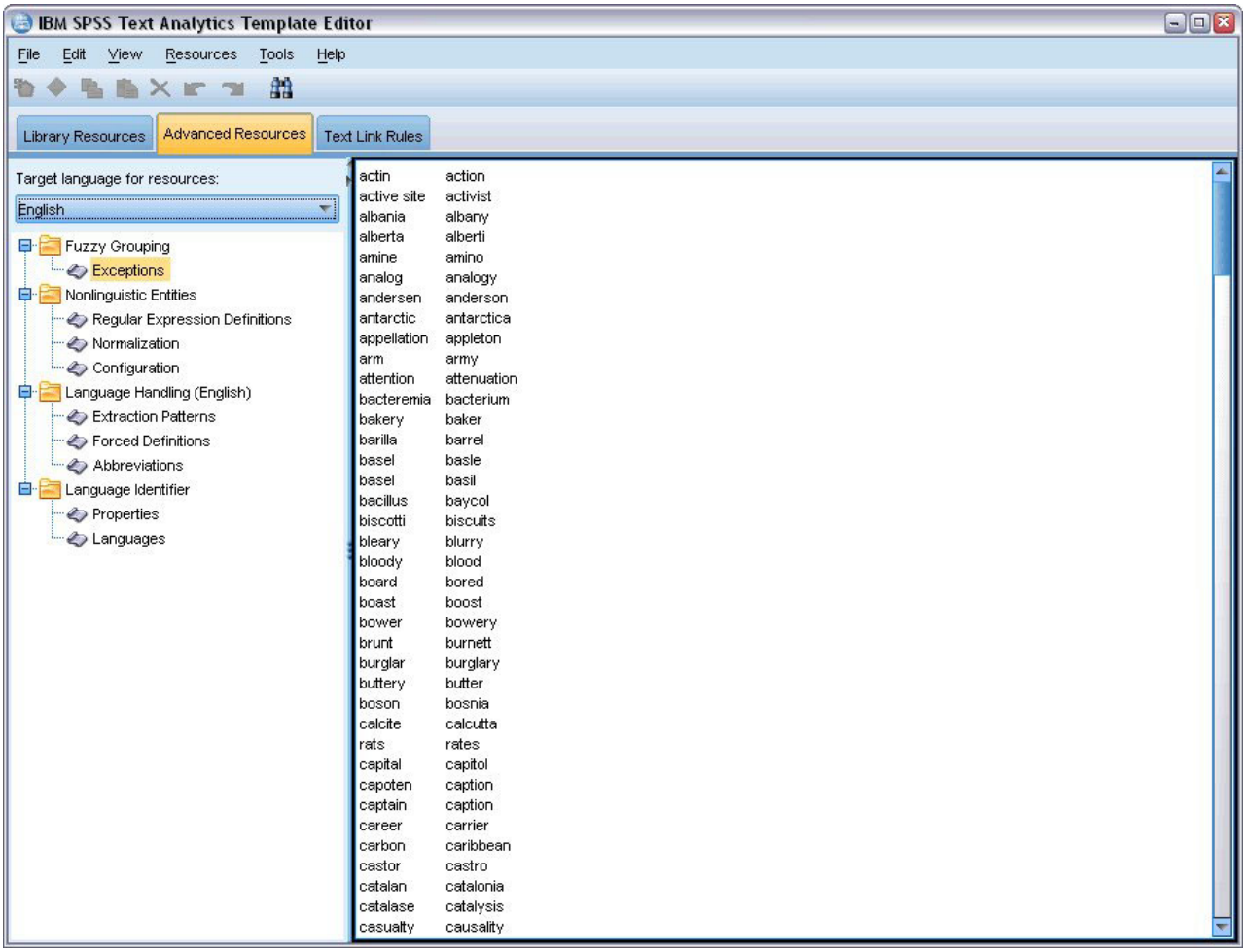


图 38. “文本挖掘模板编辑器 - 高级资源”选项卡

“文本链接规则”选项卡

自 V14 开始，可以在编辑器视图自己的选项卡中编辑文本链接分析规则。您可以使用规则编辑器，创建自己的规则，甚至运行模拟以查看规则对于 TLA 结果的影响。请参阅主题第 183 页的第 19 章，『关于文本链接规则』以获取更多信息。

重要！此选项卡不适用于针对日语文本优化的资源。

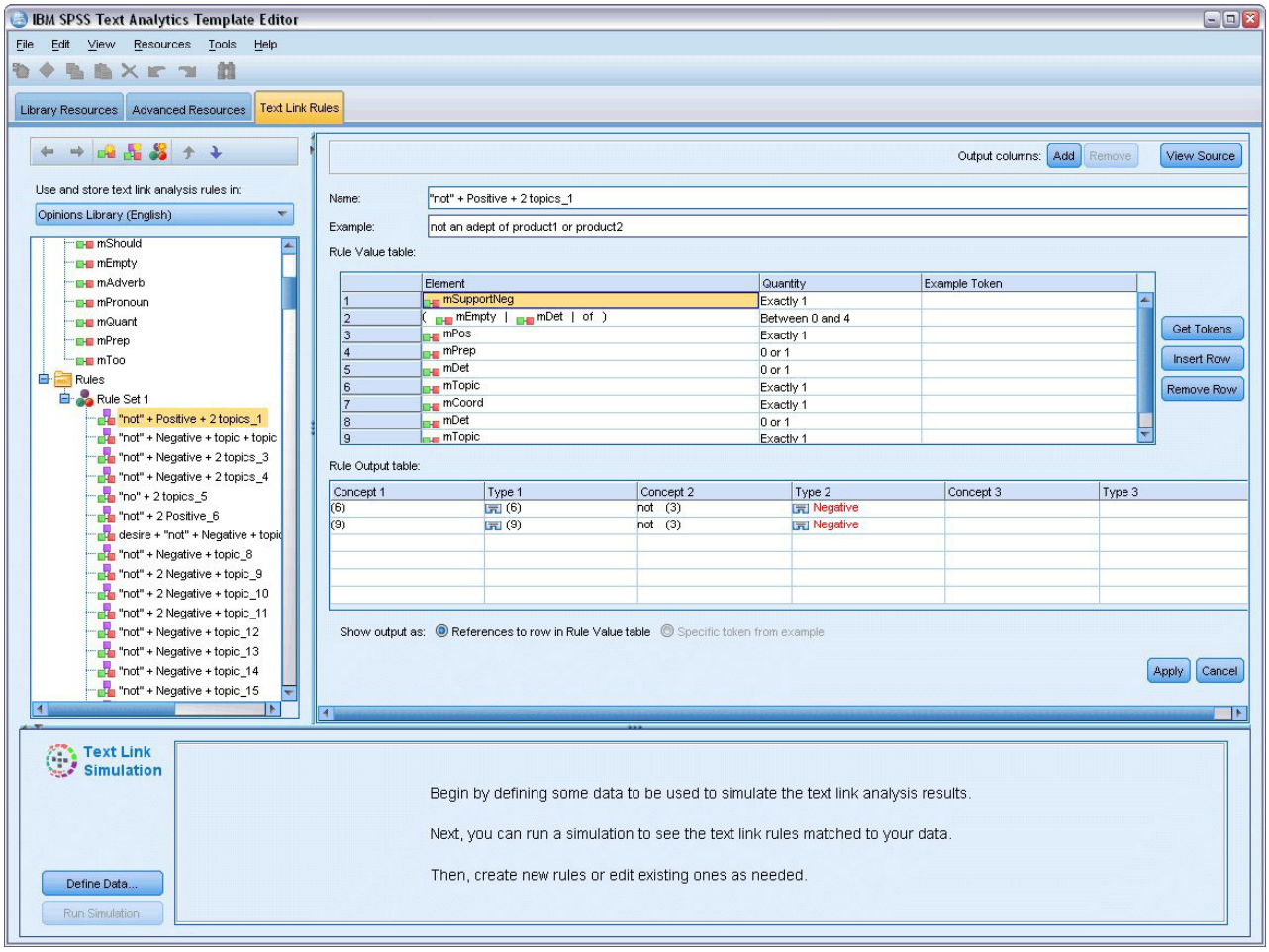


图 39. “文本挖掘模板编辑器 - 文本链接规则”选项卡

打开模板

在启动模板编辑器时，将提示您打开模板。同样，您可以从文件菜单打开模板。如果想要包含某些“文本链接分析”(TLA) 规则的模板，确保选择在 TLA 列具有图标的模板。“语言”列中显示创建模板的语言。

如果想要导入表中未显示的模板，或者想要导出模板，您可以使用“打开模板”对话框中的按钮。请参阅主题第 149 页的『导入和导出模板』以获取更多信息。

要打开模板

1. 从模板编辑器的菜单中，选择**文件 > 打开资源模板**。此时“打开资源模板”对话框将打开。
2. 从表中显示的模板中选择想要使用的模板:
3. 单击**确定**以打开此模板。如果当前在编辑器中打开其他模板，单击“确定”将放弃此模板，并显示此处选中的模板。如果对资源进行了更改并且想要保存库以供以后使用，那么可以在打开其他模板前发布、更新和共享模板。请参阅主题第 157 页的『共享库』以获取更多信息。

保存模板

在模板编辑器中，您可以所做更改保存到模板。您可以选择使用现有模板名称或提供新名称来保存。

如果对先前已装入到节点的模板进行更改，那么必须重新将模板内容装入到节点以获取最新更改。请参阅主题第 21 页的『从模板和 TAP 复制资源』以获取更多信息。

否则，如果在“文本挖掘”节点的“模型”选项卡中使用选项**使用保存的交互式工作**，那么需要从交互式工作台会话切换到此模板资源。请参阅主题第 142 页的『切换资源模板』以获取更多信息。

注：您还可以发布和共享库。请参阅主题第 157 页的『共享库』以获取更多信息。

要保存模板

1. 从模板编辑器的菜单中，选择**文件 > 保存资源模板**。此时“保存资源模板”对话框将打开。
2. 如果想要将此模板保存为新模板，请在“模板名称”字段中输入新名称。如果现有用当前装入的资源覆盖现有模板，请选择表中的模板。
3. 如果期望，输入描述以在表中显评论或注释。
4. 单击**保存**以保存模板。

重要！因为将资源从模板或 TAP 装入/复制到节点，如果对模板进行更改并且想要使用现有系统中的这些更改，那么必须重新装入来更新资源。请参阅主题『装入后更新节点资源』以获取更多信息。

装入后更新节点资源

缺省情况下，在向流添加节点时，将从缺省模板装入一组资源并将其嵌入节点。而且，如果更改模板或使用 TAP，那么在装入时，这些资源的副本将覆盖资源。因为模板和 TAP 未直接链接到节点，对模板或 TAP 进行的任何更改都不会自动用于预先存在的节点。为了使用这些更改，您必须更新此节点中的资源。可通过以下两种方式之一来更新资源。

方法 1：在“模型选项卡”中重新装入资源

如果想要使用新的或更新的模板或 TAP 更新资源，那么可以在节点的“模型”选项卡中重新装入。通过重新装入，将节点中资源的副本替换为更新的副本。为方便起见，更新时间和日期将随原始模板名称一起显示在“模型”选项卡上。请参阅主题第 21 页的『从模板和 TAP 复制资源』以获取更多信息。

但是，如果在“文本挖掘”建模节点中处理交互式会话数据，并且已在“模型”选项卡上选择**使用会话工作**选项，那么将使用保存的会话工作和资源，并且禁用**装入**按钮。因为在交互式工作台会话期间，选择**更新建模节点**选项，并保持类别、资源和其他会话工作，所以禁用。在此情况下，如果想要更改或更新这些资源，那么可以在资源编辑器中尝试下一个方法“切换资源”。

方法 2：在资源编辑器中切换资源

在交互式会话期间想要使用不同资源时，您可以使用“切换资源”对话框来交换这些资源。在想要复用类别工作但不替换资源时，这特别有用。在此情况下，您可以选择“文本挖掘”建模节点的“模型”选项卡上的**使用会话工作**选项。执行此操作将禁用通过节点对话框重新装入模板的能力，而是保留会话期间所做的设置和更改。然后，您可以通过执行流启动交互式工作台会话，并在资源编辑器中切换资源。请参阅主题第 142 页的『切换资源模板』以获取更多信息。

为了保留会话工作后续后续会话，包括资源，需要在交互式工作台会话中更新建模节点，从而将资源（和其他数据）保存回节点。请参阅主题第 70 页的『更新建模节点并保存』以获取更多信息。

注：如果在交互式会话期间切换到另一个模板的内容，那么节点中列出的模板的名称仍是装入和复制的最新模板的名称。为了使用资源或其他会话工作，先更新建模节点，然后再退出会话。

管理模板

您可以随时对模板执行一些基本管理任务，例如，重命名模板、导入和导出模板，或者删除废弃的模板。在“管理模板”对话框中执行这些任务。导入和导出模板支持您与其他用户共享模板。请参阅主题『导入和导出模板』以获取更多信息。

注: 您无法重命名或删除在此产品中安装（或随附）的模板。而是，如果想要重命名，那么可以打开已安装的模板，使用选择的名称生成一个新模板。您可以删除定制模板；但是，如果尝试删除随附的模板，那么将重置为初始安装的版本。

要重命名模板

1. 从菜单，选择**资源 > 管理资源模板**。此时“管理模板”对话框将打开。
2. 选择想要重命名的模板，然后单击**重命名**。表中的名称框将变为可编辑字段。
3. 输入新名称，然后按 Enter 键。将打开一个确认对话框。
4. 如果满意名称更改，请单击**是**。如果不满意，请单击**否**。

要删除模板

1. 从菜单，选择**资源 > 管理资源模板**。此时“管理模板”对话框将打开。
2. 在“管理模板”对话框中，选择想要删除的模板。
3. 单击**删除**。将打开一个确认对话框。
4. 单击**是**以删除，或者单击**否**以取消请求。如果单击**是**，那么将删除模板。

导入和导出模板

您可以通过导入和导出来与其他用户或机器共享模板。模板存储在内部数据库中，但是可以作为 **.lrt* 文件导出到硬盘驱动器。

因为存在可能想要导入或导出模板的情况，所以有多个对话框提供这些功能。

- 在模板编辑器中打开“模板”对话框
- 在“文本挖掘”建模节点和“文本链接分析”节点中装入“资源”对话框。
- 在模板编辑器和资源编辑器中管理“模板”对话框。

要导入模板

1. 在对话框中，单击**导入**。此时“导入模板”对话框将打开。
2. 选择要导入的资源模板文件 (**.lrt*)，然后单击**导入**。您可以使用其他名称保存导入的模板，或者覆盖现有模板。对话框将关闭，并且模板现在在表中显示。

要导出模板

1. 在对话框中，选择想要导出的模板，然后单击**导出**。此时“选择目录”对话框将打开。
2. 选择想要导出到的目录，然后单击**导出**。此对话框关闭，并且模板将导出并具有文件扩展名 (**.lrt*)

退出模板编辑器

在模板编辑器中完成工作后，您可以保存工作并退出编辑器。

要退出模板编辑器

1. 从菜单，选择**文件 > 关闭**。此时“保存并关闭”对话框将打开。
2. 选择**将更改保存到模板**以在关闭编辑器前保存打开的模板。
3. 如果想要在关闭编辑器前发布打开的模板中的任何库，请选择**发布库**。如果选择此选项，那么将提示选择要发布的库。请参阅主题第 158 页的『发布库』以获取更多信息。

备份资源

出于安全考虑，您可能想要不定时备份资源。

重要！在复原时，将全部清除资源的全部内容，并且在产品中只能访问备份文件的内容。这包括任何打开的工作。

注: 只能备份并复原到软件的相同主版本。例如，如果从 V15 备份，那么无法将备份复原到 V16。

要备份资源

1. 从菜单，选择**资源 > 备份工具 > 备份资源**。此时“备份”对话框将打开。
2. 输入备份文件的名称，然后单击**保存**。对话框关闭，并且将创建备份文件。

要复原资源

1. 从菜单，选择**资源 > 备份工具 > 复原资源**。将显示一个警告，提醒您复原将覆盖数据库的当前内容。
2. 单击**是**以继续。将打开一个对话框。
3. 选择想要复原的备份文件，然后单击**打开**。对话框关闭，并在应用程序复原资源。

导入资源文件

如果您已直接在本产品外的资源文件中进行更改，那么可以通过选择已选定的库并继续导入来将其导入到该库中。在导入目录时，也可以将所有受支持文件导入到特定开放库中。只能导入 *.txt 文件。

注意！对于日语语言文件，要导入的 .txt 文件必须使用 UTF8 进行编码。此外，不能为日语导入排除列表。

每个已导入文件必须每行仅包含一个条目，并且如果内容构造为:

- 单词或短语列表（每行一个），那么会将文件导入为类型字典的术语列表，其中类型字典采用不带扩展名的文件名。
- 诸如 *term1* <TAB> *term2* 之类的条目列表，那么会将其导入为同义词列表，其中 *term1* 是底层术语集，*term2* 是目标术语。

导入单个资源文件

1. 从菜单中选择**资源 > 导入文件 > 导入单个文件**。这会打开“导入文件”对话框。
2. 选择要导入的文件，然后单击**导入**。这会将文件内容转换为内部格式并添加到库中。

导入目录中的所有文件

1. 从菜单中选择**资源 > 导入文件 > 导入整个目录**。这会打开“导入目录”对话框。

2. 选择要在其中从**导入**列表导入所有资源文件的库。如果选择**缺省**选项，那么将使用目录的名称作为库名称来创建新库。
3. 选择要从中导入文件的目录。将不读取子目录。
4. 单击**导入**。这会关闭对话框，并且这些已导入资源文件中的内容现在会以字典和高级资源文件形式显示在编辑器中。

第 16 章 处理库

供抽取引擎用于从文本数据中抽取术语并进行分组的资源始终包含一个或多个库。可以在位于模板编辑器和资源编辑器的左上部分的库树中查看库集。库由三种字典组成: 类型、替换和排除。请参阅主题第 161 页的第 17 章，『关于库字典』以获取更多信息。

资源模板或来自所选 TAP 的资源包含若干库，通过这些库可立即开始从文本数据中抽取概念。不过，您也可以创建自己的库并将其另外发布，从而可对其进行复用。请参阅主题第 158 页的『发布库』以获取更多信息。

例如，假设您频率处理与汽车行业相关的文本数据。在分析数据后，您决定将要创建一些定制资源来处理特定于行业的词汇或行话。通过使用模板编辑器，可以创建新模板，并且其中包含用于抽取汽车术语并进行分组的库。由于您将再次需要此库中的信息，请将库发布到在**管理库**对话框中可访问的中心库，以便可在不同流会话中独立对其进行复用。

假设您还对将特定于不同子行业（如电子设备、引擎、冷却系统甚至特定制造商或市场）的术语分组感兴趣。您可以为每个组创建一个库，然后发布这些库，以便可将其用于多个文本数据集。通过此方式，可以添加与文本数据的上下文最佳对应的库。

注: 可以在“高级资源”选项卡中配置和管理其他资源。一些资源适用于所有库，并会管理非语言实体和模糊分组异常等。此外，也可以在“文本链接规则”选项卡中编辑特定于库的文本链接分析模式规则。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。

随附库

缺省情况下，随 IBM SPSS Modeler Text Analytics 一起安装多个库。 您可以使用这些预先格式化的库来访问数千个预先定义的术语和同义词以及多种不同的类型。这些随附库针对多个不同的域进行微调，并且适用于多种不同的语言。

提供大量库，但是大多数的常用方式如下所示:

- **本地库。**用于存储用户定义的字典。这是缺省情况下添加到所有资源的空库。其中也包含空类型的字典。在通过“类别和概念”视图、“聚类”视图和“文本链接分析”视图直接对资源进行更改或改进（例如，向某个类型添加字）时最有用。在此情况下，这些更改和改进将自动存储在 资源编辑器 的库树中列出的第一个库中；缺省情况下，这是*本地库*。您无法发布该库，因为器特定于会话数据。如果想要发布其内容，首先必须重命名库。
- **核心库。**在大多数情况下使用，因为其包含基本的 5 个内置类型，分别表示人员、位置、组织、产品和未知。当您看到在其中一个类型字典中仅列出少量几个术语时，核心库中表示的类型实际上是文本挖掘产品随附的内部已编译资源中找到的强类型的补充。对于每种类型，这些内部、已编译的资源包含数千个术语。因此，如果在字典术语列表中看不到术语时，仍可以使用核心类型提取和输入。这阐述在核心库的 `<Person>` 类型字典中仅显示 *John* 时，如何提取并输入 *George* 之类的名称作为 `<Person>`。类似，如果未包含核心库，那么可能仍会在提取结果中看到这些类型，因为提取引擎仍使用包含这些类型的已编译的资源。
- **意见库。**通常用于从文本数据提取意见和观点。此库包含数千个表示属性、限定词和首选项（在与其他术语一起使用时）的字，指示有关某个主题的意见。该库包含大量内置类型、同义词和排除。还包含一大组用于文本链接分析的父规则。为利用该库中的文本链接分析规则及其生成的模式结果，必须在“文本链接规则”选项卡中指定该库。请参阅主题第 183 页的第 19 章，『关于文本链接规则』以获取更多信息。
- **预算库。**用于提取涉及成本之类的术语。该库包含多个字和短语，表示形容词、限定词以及有关价格或数量之类的判断。

- **变体库。**用于包含特定语言变体需要同义词定义以正确分组的情况。该库仅包含同义词定义。

虽然在模板之外随附的某些库类似某些模板中的内容，但是模板专门针对特定应用程序进行调整，并且包含额外的高级资源。建议您尝试使用针对正在处理的文本数据种类设计的模板并对这些资源进行更改，而不仅仅是向更通用的模板添加单个库。

IBM SPSS Modeler Text Analytics 还随附已编译的资源。 在提取过程中总是使用它们，它们包含针对缺省库中内置类型字典的大量补充定义。因此已编译这些资源，无法进行查看或编辑。但是，您可以强制这些已编译的资源在任何其他字典中输入术语。请参阅主题第 166 页的『强制术语』以获取更多信息。

创建库

您可以创建任意数量的库。在创建新库后，您可以开始在该库中创建类型字典，并输入术语、同义词和排除。

要创建库

1. 从菜单，选择**资源 > 新建库**。此时库属性对话框将打开。
2. 在“名称”文本框中输入库的名称。
3. 如果想要，在“注释”文本框中输入注释。
4. 如果想要在库中输入任何项之前立即发布该库，请单击**发布**。请参阅主题第 157 页的『共享库』以获取更多信息。您还可以在以后随时发布。
5. 单击**确定**以创建库。对话框关闭，并且在树形视图中显示库。如果在树中展开库，那么您将看到库中已自动包含一个空类型字典。您可以在其中直接开始添加术语。请参阅主题第 164 页的『添加术语』以获取更多信息。

添加公用库

如果想要从其他会话数据复用某个库，那么可以将库添加到当前资源，前提是该库是公用库。**公用库**是已发布的库。请参阅主题第 158 页的『发布库』以获取更多信息。

重要！您无法将日语库添加到非日语资源，反之亦然。

在添加公用库时，会将**本地**副本嵌入您的会话数据中。您可以对该库进行更改；但是，如果想要共享更改，必须重新发布库的公用版本。

在添加公用库时，如果在一个库和其他本地库的术语和类型之间发现任何冲突，可能会显示“解决冲突”对话框。您必须解决这些冲突或接受提议的解决方法才能完成此操作。请参阅主题第 158 页的『解决冲突』以获取更多信息。

注: 如果在启动交互式工作台会话时总是更新库，或者在关闭一个 ，那么您很少会遇到不同步的库。请参阅主题第 157 页的『共享库』以获取更多信息。

要添加库

1. 从菜单，选择**资源 > 添加库**。此时“添加库”对话框将打开。
2. 选择列表中的一个或多个库。
3. 单击**添加**。如果在新添加的库和任何已有库之间发生任何冲突，那么将要求您验证冲突解决办法或者进行更改，然后才能完成操作。请参阅主题第 158 页的『解决冲突』以获取更多信息。

查找术语和类型

您可以使用“查找”功能在编辑器的各个窗格中执行搜索。在编辑器中，您可以从菜单中选择**编辑 > 查找**，此时将显示“查找”工具栏。您可以使用此工具栏一次查找一个出现。通过再次单击**查找**，您可以查找搜索项的后续出现。

在搜索时，编辑器仅搜索“查找”工具栏上的下拉列表中列出的库。如果选择**所有库**，那么程序将搜索编辑器中的所有库。

在启动搜索时，将从焦点区域中开始。搜索继续至各个部分，循环，直至返回到活动单元。您可以使用方向箭头来更改搜索顺序。您还可以选择搜索是否区分大小写。

要在视图中查找字符串

1. 从菜单，选择**编辑 > 查找**。此时将显示“查找”工具栏。
2. 输入想要搜索的字符串。
3. 单击**查找**按钮以启动搜索。然后，将突出显示术语或类型的下一个出现。
4. 再次单击按钮以切换出现。

查看库

您可以显示一个特定库或所有库的内容。在处理多个库时或者想要在发布前复审特定库的内容时非常有用。更改视图仅影响在此“库资源”选项卡中看的内容，而不会在提取期间禁用任何库。请参阅主题第 156 页的『禁用本地库』以获取更多信息。

缺省视图为**所有库**，这显示树中的所有库以及它们在其他窗格中的内容。您可以使用工具栏上的下拉列表或者通过菜单选择（**视图 > 库**）来更改此选择。在查看单个库时，其他库中的所有项不在视图中显示，但是提取期间仍会读取。

要更改库视图

1. 从“库资源”选项卡中的菜单，选择**视图 > 库**。此时将打开一个包含所有本地库的菜单。
2. 选择想要查看的库，或者选择**所有库**选项以查看所有库的内容。将根据选择过滤视图的内容。

管理本地库

本地库是交互式工作台会话中或者模板中的库，与公用库相反。请参阅主题第 156 页的『管理公用库』以获取更多信息。您可以执行一些基本本地库管理任务，包括：重命名、禁用或删除本地库。

重命名本地库

您可以重命名本地库。在重命名本地库时，如果存在公用版本，那么会解除与公用版本的关联。这意味着不再与公用版本共享后续更改。您可以使用新名称重新发布本地库。这还意味着，您无法使用对本地版本执行的任何更改来更新原始公用版本。

注：您无法重命名公用库。

1. 从菜单，选择**编辑 > 库属性**。此时“库属性”对话框将打开。

要重命名本地库

1. 在树形视图中，选择想要重命名的库。
2. 在“名称”文本框中输入库的新名称。

3. 单击**确定**以接受库的新名称。对话框关闭，并且在树形视图中更新库名。

禁用本地库

如果想要临时从提取过程中排除某个库，那么可以取消选中树视图中库名左侧的复选框。这指示您想要保留库，但是在检查冲突和提取期间忽略其内容。

要禁用库

1. 在库树窗格中，选择想要禁用的库。
2. 单击空格键。此时将清除名称左边的复选框。

删除本地库

您可以除去库而不删除库的公用版本，反之亦然。删除本地库仅从会话中删除库及其所有内容。删除库的本地版本不会从其他会话或公用版本中除去该库。请参阅主题『管理公用库』以获取更多信息。

要删除本地库

1. 在树形视图中，选择想要删除的库。
2. 从菜单中，选择**编辑 > 删除**以删除库。这将除去库。
3. 如果之前从未发布过该库，那么消息将询问是要删除还是保留该库。单击**删除**以继续，或者如果要保留该库，请单击**保留**。

注: 必须始终保留一个库。

管理公用库

为复用本地库，您可以通过“管理库”对话框发布、处理和查看库（**资源 > 管理库**）。请参阅主题第 157 页的『共享库』以获取更多信息。您可以执行的基本公用库管理任务包括导入、导出或删除公用库。您无法重命名公用库。

导入公用库

1. 在“管理库”对话框中，单击**导入...**。此时“导入库”对话框将打开。
2. 选择想要导入的库文件（*.*lib*），并且如果还想要本地添加该库，请选择**将库添加到当前项目**。
3. 单击**导入**。对话框将关闭。如果已存在使用相同名称的公用库，那么将要求重命名导入的库，或者覆盖当前公用库。

导出公用库

您可以将公用库导出为 *.lib* 格式以便共享。

1. 在“管理库”对话框中，从列表中选择想要导出的库。
2. 单击**导出**。此时“选择目录”对话框将打开。
3. 选择想要导出到的目录，然后单击**导出**。对话框将关闭，并导出库文件（*.*lib*）。

删除公用库

您可以除去本地库而不删除库的公用版本，反之亦然。但是，如果从此对话框中删除库，那么在重新发布本地版本之前，无法再将其添加到任何会话资源。

如果删除随产品安装的库，那么将复原原始安装的版本。

1. 在“管理库”对话框中，选择想要删除的库。您可以单击相应的标题对列表进行排序。
2. 单击**删除**以删除库。IBM SPSS Modeler Text Analytics 验证库的本地版本是否与公用库相同。如果是，那么将除去库而无警报。如果库版本不同，那么将打开一个警报，询问保留还是除去发布的公用版本。

共享库

库允许您通过易于在多个交互式工作台会话中共享的方式处理资源。库可以有两个状态或版本。在编辑器中可编辑且属于交互式工作台会话的库被称为**本地库**。例如，在使用交互式工作台会话时，您可以对*蔬菜*库进行大量更改。如果更改对于其他数据有用，那么可以通过创建**蔬菜**库的*公用库*版本来提供这些资源。公用库，顾名思义，可用于任何其他任何交互式工作台会话中的资源。

您可以在“管理库”对话框中查看公用库。在此公用库版本存在时，您可以将其添加到其他上下文中的资源，从而可共享这些定制的语言资源。

随附的库是初始公用库。可编辑这些库中的资源，然后创建新的公用版本。然后，可以在其他交互式工作台会话中访问这些新版本。

在继续处理新库并进行更改时，库版本将变成不同步。在某些情况下，本地版本可能比公用版本更新，而在另一些情况下，公用版本可能比本地版本更新。如果在另一个交互式工作台对话中更新了公用版本，那么公用版本和本地版本可能都包含对方没有的更改。如果库版本变成为不同步，您可以重新进行同步。同步库版本包括重新发布和/或更新本地库。

在启动交互式工作台会话或关闭时，将提示同步需要更新或重新发布的任何库。此外，您可以使用树形视图库名旁边显示的图标或者查看“库属性”对话框，方便地发现本地库的同步状态。您还可以选择通过菜单选择随时执行此操作。下表描述 5 个可能的状态及其关联的图标。

表 37. 本地库同步状态.

图标	本地库状态描述
	未发布 - 从未发布本地库。
	已同步 - 本地和公用库版本相同。这也适用于无法发布的*本地库*，因为其旨在包含特定于会话的资源。
	已过期 - 公用库版本比本地版本新。您可以使用更改更新本地版本。
	更新 - 本地库版本比公用版本新。您可以将本地版本重新发布到公用版本。
	未同步 - 本地和公用库都包含对方没有的更改。您必须决定是更新还是发布本地库。如果更新，那么将丢失上次更新或发布以来所做的更改。如果选择发布，那么将覆盖公用版本中的更改。

注: 如果在启动交互式工作台会话时总是更新库，或者在关闭一个，那么您很少会遇到不同步的库。

您可以在认为库中的更改有益于其他也包含该库的流时重新发布库。然而，如果您的更改受益于其他流，那么可更新这些流中的本地版本。通过此方式，您可以通过创建新库和/或向资源添加任意数量的公用库，针对应用于您的数据的每个上下文或域创建流。

如果共享库的公用版本，那么本地版本和公用版本不一样的可能性将增加。在启动或关闭交互式工作台会话以及从其发布，或者从 模板编辑器 打开或关闭模板时，将显示一条消息，支持您发布和/或更新版本与“管理库”对话框中的版本不同步的任何库。如果公用库版本比本地版本新，那么对话框询问是否要更新打开的版本。您可以选择保留本地版本代替使用公用版本更新，或者将更新合并到本地库。

发布库

如果从未发布某个特定库，那么发布将会在数据库中创建本地库的公用副本。如果重新发布库，那么本地库的内容将替换现有公用版本的内容。在重新发布后，您可以更新任何其他流会话中的该库，从而使其本地版本与公用版本保持同步。即使您可以发布库，也总是在会话中存储本地版本。

重要！如果对本地库进行更改，同时库的公用版本也发生更改，那么库被视为不同步。建议首先用公共更改更新本地版本，然后执行想要的任何更改，最后重新发布本地版本以使两个版本相同。如果首先进行更改并发布，那么将覆盖公用版本中的任何更改。

要将本地库发布到数据库

1. 从菜单，选择**资源 > 发布库**。此时"发布库"对话框将打开，缺省情况下选中需要发布的所有库。
2. 选择想要发布或重新发布的每个库左侧的复选框。
3. 单击**发布**以将库发布到"管理库"数据库。

更新库

在启动或关闭交互式工作台会话时，您可以更新或发布不再与公用版本同步的任何库。如果公用库版本比本地版本新，那么对话框询问是否要更新打开的库。您可以选择保留本地版本代替更新为公用版本，或者将本地版本更新为公用版本。如果库的公用版本比本地版本新，那么可以更新本地版本以使用公用版本的内容同步其内容。更新意味着将公用版本中找到的更改合并到本地版本。

注: 如果在启动交互式工作台会话时总是更新库，或者在关闭一个，那么您很少会遇到不同步的库。请参阅主题第 157 页的『共享库』以获取更多信息。

要更新本地库

1. 从菜单，选择**资源 > 更新库**。此时"更新库"对话框将打开，缺省情况下选中需要更新的所有库。
2. 选择想要发布或重新发布的每个库左侧的复选框。
3. 单击**更新**以更新本地库。

解决冲突

本地库与公用库冲突

在启动流会话时，IBM SPSS Modeler Text Analytics 针对本地库以及"管理库"对话框中列出对执行比较。如果会话中的任何本地库与已发布版本不同步，那么"库同步警告"对话框将打开。您可以从以下选项中进行选择，从而选择想要在此处使用的库版本:

- **所有本地库文件。**此选项保留所有本地库。您总是可以稍后重新发布或更新。
- **此机器上的所有已发布的库。**此选项将显示的本地库替换为数据库中找到的版本。
- **所有更新的库。**此选项将任何旧的本地库替换为数据库中更新的公用版本。
- **其他。**此选项允许通过在表中进行选择来手动选择想要的版本。

强制术语冲突。

在添加公用库或更新本地库时，可能会在该库的术语和类型与资源中的其他库的术语和类型之间发现冲突和重复条目。如果发生此情况，将在"编辑强制的术语"对话框中要求您验证建议的冲突解决办法或者进行更改，然后才能完成操作。请参阅主题第 166 页的『强制术语』以获取更多信息。

“编辑强制的术语”对话框包含每对冲突的术语或类型。使用不同的背景色来直观区分每个冲突对。可以在“选项”对话框中更改这些颜色。请参阅主题第 69 页的『选项: 显示选项卡』以获取更多信息。 “编辑强制的术语”对话框包含两个选项卡:

- **重复。**此选项卡包含在库中找到的重复的术语。如果在术语后显示图钉图标，那么意味着已强制这一术语出现。如果显示黑色 X 图标，那么意味着在提取期间将忽略此术语出现，因为已在其他位置强制。
- **用户定义。**此选项卡包含在类型字典术语窗格中手动强制而不冲突的任何术语列表。

注: 在添加或更新库后，“编辑强制的术语”对话框将打开。如果取消此对话框，那么无法取消库更新或新增项。

要解决冲突

1. 在“编辑强制的术语”对话框中，选择想要强制的术语的“使用”列中的单选按钮。
2. 在完成时，单击**确定**以应用强制的术语并关闭对话框。如果单击**取消**，那么将取消在此对话框中执行的更改。

第 17 章 关于库字典

用于提取文本数据的资源存储为目录和库形式。库可以由三个字典组成。

- **类型字典**包含依据一个标签或类型名称分组的术语集合。在提取引擎读取您的文本数据时，其将文本中找到的字与类型字典中定义的术语进行比较。在提取期间，影响类型术语的格式，并且将依据名为概念的目标术语分组同义词。提取的概念将分配给其作为术语出现的类型字典。您可以在编辑器左上角和中心窗格中管理类型字典 - 库树和术语窗格。请参阅主题『类型字典』以获取更多信息。
- **替换字典**包含定义为同义词或可选元素的字的集合，用于依据一个名为概念的目标术语分组最终提取结果中的类似术语。您可以使用“同义词”选项卡 和“可选选项卡在编辑器的左下方窗格中管理替换字典。请参阅主题第 167 页的『替换/同义词字典』以获取更多信息。
- **排除字典**包含将从最终提取结果中除去的术语和类型的集合。您可以在编辑器的最右侧窗格中管理排除字典。请参阅主题第 170 页的『排除字典』以获取更多信息。

请参阅主题第 153 页的第 16 章，『处理库』以获取更多信息。

类型字典

类型字典由类型名称、标签和术语列表组成。类型字典在编辑器中“库资源”选项卡的左上窗格和中心窗格中进行管理。您可以使用菜单中的**视图 > 资源编辑器**来查看此视图。否则，您可以在模板编辑器中编辑特定模板的字典。

当抽取引擎读取文本数据时，它会将在文本中找到的单词与类型字典中定义的术语进行比较。术语是语言资源中的类型字典中的单词或短语。

当单词与术语匹配时，会将该单词分配到该术语的类型名称。在抽取期间读取资源时，在文本中找到的术语将经过若干处理步骤，然后成为“抽取引擎”窗格中的概念。如果属于同一类型字典的多个术语由抽取引擎确定为同义词，那么会在“抽取结果”窗格中将其分组在最频繁出现的术语下并称为**概念**。例如，术语 `question` 和 `query` 最后可能出现在概念名称 `question` 下。

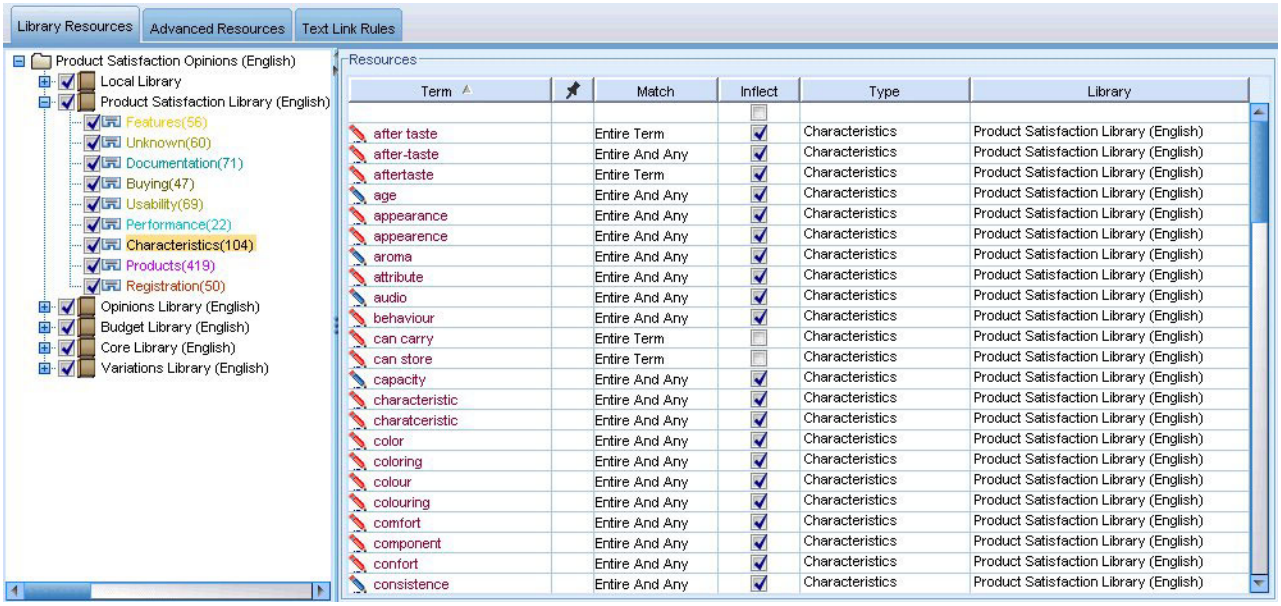


图 40. 库树和术语窗格

类型字典列表显示在左侧的库树窗格中。每个类型字典的内容显示在中心窗格中。类型字典不只是由术语列表组成。文本数据中的单词和单词短语与类型字典中定义的术语的匹配方式由所定义的匹配选项确定。**匹配选项**指定如何根据文本数据中的候选单词或短语来锚定术语。请参阅主题第 164 页的『添加术语』以获取更多信息。

注: 并非所有选项（如匹配选项和词形变化形式）都适用于日语文本。

此外，还可以通过指定是否要自动生成词形变化形式的术语并将其添加到字典来扩展类型字典中的术语。通过生成词形变化形式，可以自动向类型字典中添加单数术语的复数形式、复数术语的单数形式以及形容词。请参阅主题第 164 页的『添加术语』以获取更多信息。

注: 针对多数语言，在任何类型字典中未发现但从文本抽取的概念会自动类型化为<未知>

内置类型

IBM SPSS Modeler Text Analytics 提供一组随附库和已编译的资源形式的语言资源。随附库包含一组内置类型字典，例如，<Location>、<Organization>、<Person> 和 <Product>。

注: 对于日语文本，缺省的内置类型集不同。

提取引擎使用这些类型的字典来将类型分配给提取的概念，例如，将类型 <Location> 分配给概念 paris。虽然在内置类型字典中定义了大量术语，但是不会涵盖每一种可能性。因此，您可以添加它们或创建自己的。有关特定的随附类型字典的概念的描述，请阅读“类型属性”对话框中的注释。选择树中的类型，然后从上下文菜单中选择**编辑 > 属性**。

注: 除了随附库，已编译的资源（也供提取引擎使用）包含针对内置类型字典的大量定义补充，但是其内容在产品中不可见。但是，您可以强制已编译的字典在任何其他字典中输入术语。请参阅主题第 166 页的『强制术语』以获取更多信息。

创建类型

您可以创建类型字典以帮助分类类似术语。在提取过程中发现此字典中显示的术语时，将为它们分配此类型名称并根据概念名称进行提取。在创建库时，总是包含空类型库，以便可立即开始输入术语。

重要！: 无法针对日语资源创建新类型。

如果分析有关字体的文本并且想要分组与 vegetables 有关的术语，那么可以创建自己的 <Vegetables> 类型字典。然后，如果认为这是将在文本中显示的重要术语，那么可以将诸如 carrot、 broccoli 和 spinach 的术语。然后，提取期间，如果找到其中任何术语，将作为概念提取并分配给 <Vegetables> 类型。

不必定义每种形式的字或表达式，因为您可以选择生成术语的变化形式。通过选择此选项，提取引擎将在属于此类型的其他术语之间，自动识别单数或负数形式的术语。当您的类型包含大量名词时此选项非常有用，因为您不太可能想要动词或形容词变化形式。

“类型属性”对话框包含以下字段。

名称。为创建的类型字典指定名称。建议不要在类型名称中使用空格，尤其是如果有两个或多个类型名称以相同的字开始。

注: 对于类型名称和符号使用存在一些约束。例如，不能在名称中使用诸如“@”或“!”之类的符号。

缺省匹配。缺省匹配属性指示提取引擎如何将此术语与文本数据相匹配。在向此类型字典添加术语时，将自动为其分配匹配属性。您总是可以在术语列表中手动更改匹配选项。选项包括: **Entire Term**、**Start**、**End**、**Any**、**Start or End**、**Entire and Start**、**Entire and End**、**Entire and (Start or End)** 和 **Entire (no compounds)**.. 请参阅主题第 164 页的『添加术语』以获取更多信息。此选项不适用于日语资源。

添加到。该字段指示想要在其中创建新类型字典的字典。

缺省情况下生成变化形式。此选项告知提取引擎使用语法形态以捕获和分组添加到此字典的术语的类似形式，例如，术语的单数或复数形式。在类型大部分为名词时，此选项非常有用。在选择此选项时，添加到此类型的所有新术语将自动具有此选项，但是您可以在列表中手动更改。此选项不适用于日语资源。

字体颜色。该字段使您能够从该界面中的其他类型区分此类型的结果。如果选择**使用父代颜色**，那么将缺省类型颜色用于此类型字典，而且: 在选项对话框中设置此缺省颜色。请参阅主题第 69 页的『选项: 显示选项卡』以获取更多信息。 如果选择**定制**，那么从下拉列表中选择颜色。

注释。此字段为可选，并且可用于任何注释或描述。

要创建类型字典

1. 选择想要在其中创建新类型字典的库。
2. 从菜单，选择**工具 > 新建类型**。此时“类型属性”对话框将打开。
3. 在**名称**文本框中输入类型字典的名称，并选择想要的选项。
4. 单击**确定**以创建类型字典。在库树窗格和中心窗格中显示新类型。您可以立即开始添加术语。有关更多信息，请参阅第 164 页的『添加术语』。

注: 这些指示信息显示如何在资源编辑器视图或模板编辑器中执行更改。请记住，您还可以从“提取结果”窗格、“数据”窗格、“类别”窗格或“聚类定义”对话框（其他视图）直接执行此类微调。请参阅主题第 80 页的『优化抽取结果』以获取更多信息。

添加术语

库树窗格显示库，并且可扩展以显示它们包含的类型字典。在中心窗格中，术语列表根据树中的选择来显示选中库或类型字典中的术语。

重要！针对日语资源定义的术语不同！

在资源编辑器中，您可以直接在术语窗格中或者通过“添加新术语”对话框将术语添加到类型字典。添加的术语可以是单个字或复合词。总是在列表顶部看到一个空行，可通过其添加新术语。

注: 这些指示信息显示如何在资源编辑器视图或模板编辑器中执行更改。请记住，您还可以从“提取结果”窗格、“数据”窗格、“类别”窗格或“聚类定义”对话框（其他视图）直接执行此类微调。请参阅主题第 80 页的『优化抽取结果』以获取更多信息。

术语列

在此列中，在单元中输入单个字或复合词。术语的显示颜色取决于存储或强制术语的类型的颜色。您可以在“类型属性”对话框中更改类型颜色。请参阅主题第 163 页的『创建类型』以获取更多信息。

强制列

在此列中，通过在此单元中放置图钉图标，提取引擎知道忽略其他库中相同术语的任何其他出现。请参阅主题第 166 页的『强制术语』以获取更多信息。

匹配列

在此列中，选择匹配选项以指示提取引擎如何使此术语与文本数据相匹配。请参阅表格以了解示例。您可以通过编辑类型属性来更改缺省值。请参阅主题第 163 页的『创建类型』以获取更多信息。从菜单，选择**编辑 > 更改匹配**。以下是基本匹配选项，这些选项的组合也可行:

- **Start**。如果字典中的术语匹配从文本提取的概念中的第一个字，指定此类型。例如，如果输入 `apple`，那么 `apple tart` 将匹配。
- **End**。如果字典中的术语匹配从文本提取的概念中的最后一个字，指定此类型。例如，如果输入 `apple`，那么 `cider apple` 将匹配。
- **Any**。. 如果字典中的术语匹配从文本提取的概念的任何字，那么指定此类型。例如，如果输入 `apple`，那么 **Any** 选项将以相同的方式输入 `apple tart`、`cider apple` 和 `cider apple tart`。
- **Entire Term**。如果从文本提取的整个概念匹配字典中的精确术语，那么指定此类型。添加术语 **Entire term**，那么 **Entire and Start**、**Entire and End**、**Entire and Any** 或 **Entire (no compounds)** 将强制提取术语。

  此外，因为 `<Person>` 类型仅提取连个部分名称，例如，*edith piaf* 或 *mohandas gandhi*，如果尝试在未提及形式时提取名字，那么可能想要显式将名字添加到此类型字典。例如，如果想要捕获 *edith* 的所有实例作为名称，那么应该使用 **Entire term** 或 **Entire and Start** 将 `edith` 添加到 `<Person>`。
- **Entire (no compounds)**。如果从文本提取的整个概念匹配字典中的精确术语，那么指定此类型，并且提取停止禁止提取使术语与较长的组合词进行匹配。例如，如果输入 `apple`，那么 **Entire (no compound)** 选项将输入 `apple` 并且不会提取组合 `apple sauce`，除非某个位置强制执行。

在下表中，假定术语 `apple` 位于类型字典中。根据匹配选项，该表显示在文本中找到时应提取和输入的概念。

表 38. 匹配示例.

匹配选项 术语:  **apple**	提取的概念			
	apple	**apple** *tart*	*ripe* **apple**	*homemade* **apple** *tart*
Entire Term	✔			
Start		✔		
End			✔	
Start or End		✔	✔	
Entire and Start	✔	✔		
Entire and End	✔		✔	
Entire and (Start or End)	✔	✔	✔	
Any		✔	✔	✔
Entire and Any	✔	✔	✔	✔
Entire (no compounds)	✔	*从不抽取*	*从不抽取*	*从不抽取*

变化列

在此列中，选择提取引擎在提取期间是否应生成此术语的变化形式以便全都分组在一起。在“类型属性”中定义此列的缺省值，但是您可以直接在列中逐个更改此选项。从菜单，选择**编辑 > 更改变化**。

类型列

在此列中，从下拉列表中选择类型字典。根据字典树窗格中的选择过滤类型列表。列表中的第一个类型总是库树窗格中选中的缺省类型。从菜单，选择**编辑 > 更改类型**。

库列

在此列中，显示存储术语的库。您可以将术语拖放到库树窗格中的其他类型以更改其库。

要将单个术语添加到类型字典

1. 在库树窗格中，选择想要为其添加术语的类型字典。
2. 在中心窗格的术语列表中，在第一个可用的空单元中输入术语，并针对此术语设置想要的任何选项。

要将多个术语添加到类型字典

1. 在库树窗格中，选择想要为其添加术语的类型字典。
2. 从菜单，选择**工具 > 新术语**。此时“添加新术语”对话框将打开。
3. 通过输入术语或者复制粘贴一组术语，输入想要添加到所选类型字典的的术语。如果输入多个术语，那么必须使用“选项”对话框中定义的定界符进行分隔，或者一个新行添加每个术语。请参阅主题第 68 页的『设置选项』以获取更多信息。
4. 单击**确定**以将术语添加到字典。匹配选项将自动设置为此类型库的缺省选项。对话框关闭，并且在字典中显示新术语。

强制术语

如果想要将一个术语分配给特定类型，那么可以将其添加到相应的类型字典。但是，如果存在具有相同名称的多个术语，那么提取引擎必须知道应使用哪个类型。因此，将提示选择应使用的类型。这被称为**强制**术语为某个类型。在覆盖来自已编译（内部、不可编辑）字典的类型分配时，此选项最有用。通常，建议避免重复术语。

强制不会除去此术语的其他出现；而是，让提取引擎忽略。您可以稍后通过强制或取消强制术语来更改应使用的出现。在添加公用库或者更新公用库时，您可能还需要强制术语进入类型字典。

您可以在“强制”列（术语窗格的第二列）中查看强制或忽略的术语。如果显示图钉图标，那么意味着已强制这一术语出现。如果显示黑色 X 图标，那么意味着在提取期间将忽略此术语出现，因为已在其他位置强制。此外，在强制术语时，术语将以强制类型的颜色显示。这意味着，如果强制 `Type 1` 和 `Type 2` 中术语为 `Type 1`，那么在窗口中看到此术语时，将使用针对 `Type 1` 定义的字体颜色显示。

您可以双击图标以更改状态。如果术语在其他位置显示，“解决冲突”对话框将打开，以允许选择应使用的出现。

重命名类型

您可以通过编辑类型属性，重命名类型字典或者更改其他字典设置。

重要！建议不要在类型名称中使用空格，尤其是如果有两个或多个类型名称以相同的字开始。另外建议不要重命名“核心”或“意见”库中的类型，或者更改其缺省匹配属性。

要重命名类型

1. 在库树窗格中，选择想要重命名的类型字典。
2. 右键单击鼠标，然后从上下文菜单中选择**类型属性**。此时“类型属性”对话框将打开。
3. 在“名称”文本框中输入类型字典的新名称。
4. 单击**确定**以接受新名称。将在库树窗格中显示新的类型名称。

移动类型

您可以将类型字典拖至库中的其他位置或者树中的其他库。

要重新排序库中的类型

1. 在库树窗格中，选择想要移动的类型字典。
2. 从菜单，选择**编辑 > 上移**以将类型字典在库树窗格中上移一个位置，或者选择**编辑 > 下移**以向下移动一个位置。

要将类型移动到其他库

1. 在库树窗格中，选择想要移动的类型字典。

2. 右键单击鼠标，然后从上下文菜单中选择**类型属性**。此时“类型属性”对话框将打开。（您还可以将类型拖放至其他库）。
3. 在“添加到”列表框中，选择想要将类型字典移至的库。
4. 单击**确定**。对话框将关闭，并且类型现在位于选中的库中。

禁用和删除类型

如果想要临时除去类型字典，可以通过取消选中库树窗格中字典名称左侧的复选框来禁用。这指示您想要在库中保存字典，但是想要冲突检查和提取过程中忽略内容。

您还可以永久从库删除类型字典。

要禁用类型字典

1. 在库树窗格中，选择想要禁用的类型字典。
2. 单击空格键。此时将清除类型名称左边的复选框。

要删除类型字典

1. 在库树窗格中，选择想要删除的类型字典。
2. 从菜单中，选择**编辑 > 删除**以删除类型字典。

替换/同义词字典

替换字典是帮助将类似术语分组在一个目标术语下的术语的集合。替换字典在“库资源”选项卡的底部窗格中进行管理。您可以使用菜单中的**视图 > 资源编辑器**来查看此视图。否则，您可以在模板编辑器中编辑特定模板的字典。

可以在此字典中定义两种形式的替换：**同义词**和**可选元素**。可以单击此窗格中的选项卡以在其之间切换。

对文本数据运行抽取后，您可能会发现若干属于其他概念的同义词或词形变化形式的概念。通过标识可选元素和同义词，可以强制抽取引擎将其映射到一个目标术语。

使用同义词和可选元素进行替换会通过将其一起组合成具有更高频率文档计数的更有意义的代表性概念来减少“抽取结果”窗格中的概念数。

*注：*对于日语资源，可选元素不适用且不可用。此外，对于日语文本，同义词的处理略有不同。

同义词

同义词将具有相同含义的两个或多个单词进行关联。您也可以使用同义词将术语与其缩写进行分组，或者将经常拼写错误的单词与正确的拼写进行分组。可以在“同义词”选项卡上定义这些同义词。

同义词定义由两个部分组成。第一个是**目标**术语，它是您希望抽取引擎将所有同义词术语都在其之下分组的术语。除非此目标术语用作另一个目标术语的同义词或者将其排除，否则它可能会成为“抽取结果”窗格中显示的概念。第二个是将在目标术语下分组的同义词的列表。

例如，如果希望将 `automobile` 替换为 `vehicle`，那么 `automobile` 即是同义词，`vehicle` 即是目标术语。

可以将任何单词输入到**同义词**列中，但如果在抽取期间找不到该单词，并且术语具有包含 `Entire` 的匹配选项，那么无法发生任何匹配。不过，无需抽取目标术语即可将同义词分组在此术语下。

可选元素

可选元素标识复合词中的可选单词，在抽取期间可将其忽略，从而将类似术语保持在一起，即使其在文本中看似略有不同也如此。可选元素是单个单词，如果从复合词中将其移除，那么可能会创建与其他术语的匹配。这些单个单词可出现在复合词中的任意位置 - 开头、中间或结尾。可以在“可选”选项卡上定义可选元素。

例如，要将术语 ibm 和 ibm corp 分组在一起，应声明在本例中将 corp 作为可选元素处理。在另一个示例中，如果将术语 access 指定为可选元素且在抽取期间同时找到 internet access speed 和 internet speed，那么会在最频繁出现的术语下将其分组在一起。

注: 对于日语文本资源，由于可选元素不适用，因此没有“可选元素”选项卡。

定义同义词

在“同义词”选项卡上，您可以在表顶部的空行中输入同义词定义。首先定义目标术语及其同义词。您还可以选择想要在其中存储此定义的库。在提取期间，将根据最终提取中的目标术语分组同义词的所有出现。请参阅主题第 164 页的『添加术语』以获取更多信息。

例如，如果文本数据包含大量电信信息，那么您可以具有以下术语: cellular phone、wireless phone 和 mobile phone。在此示例中，您可能想要将 cellular 和 mobile 定义为 wireless 的同义词。如果定义这些同义词，那么 cellular phone 和 mobile phone 的每个提取出现都将视为与 wireless phone 相同的术语，并且在术语列表中一起显示。

在构建类型字典时，您可以输入术语，然后认定此术语的三个或四个同义词。在此请款下，您可以在替换字典中输入所有术语以及目标术语，然后拖动同义词。

注: 日语文本中的同义词处理略有不同。

同义词替换也适用于同义词的变化形式（例如，复数形式）。根据上下文，您可能想要对于替换术语的方式施加约束。可使用特定字符来对同义词处理的程度施加限制:

- **感叹号 (!)**。同义词直接前置感叹号，例如，!synonym，指示目标术语不替换同义词的变化形式。但是，目标术语直接前置感叹号时，即 !target-term，意味着您不不希望复合目标术语的任何部分或变体接收任何进一步替换。
- **星号 (*)**。在同义词后直接放置星号，例如，synonym*，意味着您想要此字替换为目标术语。例如，如果将 manage* 定义为同义词，并将 management 定义为目标，那么 associate managers 将替换为目标术语 associate management。您还可以在字 (synonym *) 之后添加空格和星号，例如，internet *。如果将目标定义为 internet，并将同义词定义为 internet * * 和 web *，那么 internet access card 和 web portal 将替换为 internet。在此字典中，字或字符串不能以星号开始。
- **插入标记 (^)**。同义词前置插入标记和空格，例如，^ synonym，意味着仅当术语以同此开始时，同义词分组才适用。例如，如果将 ^ wage 定义为同义词，并将 income 定义为目标，并且同时提取两个术语，那么它们一起依据术语 income 进行分组。但是，如果提取 minimum wage 和 income，那么它们不会分组在一起，因为 minimum wage 不是以 wage 开始。必须在此符号和同义词之间添加一个空格。
- **美元符号 ($)**。空格和美元符号跟在同义词后面，例如，synonym $，意味着仅在术语以同义词结束时同义词分组才适用。例如，如果将 cash $ 定义为同义词并将 money 定义为目标并且提取这两个术语，那么它们将一起依据术语 money 进行分组。但是，如果提取 cash cow 和 money，那么不会分组在一起，因为 cash cow 不是以 cash 结束。必须在此符号和同义词之间添加一个空格。
- **插入标记 (^) 和美元符号 ($)**。如果插入标记和美元符号一起使用，例如，^ synonym $，那么仅在完全匹配时术语才匹配同义词。这意味着在提取的术语中的同义词之前或之后不能出现任何字，才能发生同义词分组。例如，您可能想要将 ^ van $ 定义为同义词并将 truck 定义为目标，从而仅在 marie van guerin 保持不变时，才使用 truck 分组 van。此外，在使用插入标记和美元符号定义同义词时，并且该字在源文本中的任何位置中出现时，将自动提取同义词。

注: 日语文本不支持这些特殊字符和通配符。

要添加同义词条目

1. 使用显示的替换窗格，单击左下角中的**同义词**选项卡。
2. 在表顶部的空行中，在“目标”列中输入目标术语。输入的目标术语将以颜色显示。此颜色表示术语显示或强制（如果是此情况）的类型。如果术语以黑色显示，那么这意味着术语不会出现在任何类型字典中。
3. 单击目标右侧的第二个单元，并输入同义词集。使用“选项”对话框中定义的全局定界符来分隔每个条目。请参阅主题第 68 页的『设置选项』以获取更多信息。输入的术语将以颜色显示。此颜色表示术语显示的类型。如果术语以黑色显示，那么这意味着术语不会出现在任何类型字典中。
4. 单击最后一个单元以选择想要在其中存储此同义词定义的库。

注: 这些指示信息显示如何在资源编辑器视图或模板编辑器中执行更改。请记住，您还可以从“提取结果”窗格、“数据”窗格、“类别”窗格或“聚类定义”对话框（其他视图）直接执行此类微调。请参阅主题第 80 页的『优化抽取结果』以获取更多信息。

定义可选元素

在“可选”选项卡上，您可以针对想要的任何库定义可选元素。将针对每个库一起分组这些条目。一旦将库添加到库树窗格，就会向“可选”选项卡添加一个空的可选元素行。

所有条目将自动缓缓为小写字。提取引擎将匹配文本中所有小写或大写的字。

注: 对于日语资源，可选元素不适用并且不可用。

注: 使用“选项”对话框中定义的定界符来分隔术语。请参阅主题第 68 页的『设置选项』以获取更多信息。如果输入的可选元素包含与术语一部分相同的定界符，那么必须前置反斜杠。

要添加条目

1. 使用显示的替换窗格，单击编辑器左下角中的“可选”选项卡。
2. 单击想要向其添加此条目的库的“可选元素”列中的单元。
3. 输入可选元素。使用“选项”对话框中定义的全局定界符来分隔每个条目。请参阅主题第 68 页的『设置选项』以获取更多信息。

禁用和删除替换

您可以通过在字典中禁用条目以临时方式除去此条目。通过禁用条目，提取期间将忽略此条目。

您还可以删除替换字典中任何废弃的条目。

要禁用条目

1. 在字典中，选择想要禁用的条目。
2. 单击空格键。此时将清除条目左边的复选框。

注: 您还以取消选择条目左侧的复选框以禁用条目。

要删除同义词条目

1. 在字典中，选择想要删除的条目。
2. 从菜单中，选择**编辑 > 删除**，或者按键盘上的 **Delete** 键。条目将不再位于字典中。

要删除可选元素条目

1. 在字典中，双击想要删除的条目。
2. 手动删除条目。
3. 按 Enter 键以应用更改。

排除字典

排除字典是一个字、短语或部分字符串的列表。提取时将忽略或排除匹配或包含排除字典中的条目的任何术语。在编辑器的右侧窗格中管理排除字典。通常，添加到此列表的术语是为保持连续性在文本中使用的填充字或短语，但实际上不会向文本添加任何重要内容，并且可能会干扰提取结果。通过将这些术语添加到排除字典，可确保从不会提取它们。

在编辑器"库资源"选项卡的右上方窗格中管理排除字典。您可以使用菜单中的**视图 > 资源编辑器**来查看此视图。否则，您可以在模板编辑器中编辑特定模板的字典。

在排除字典中，您可以在表顶部的空行中输入字、短语或部分字符串。您可以向排除字典添加字符串，可以是一个或多个字，甚至是使用星号作为通配符的部分字。排除字典中声明的条目将用于拦截概念参与提取。如果在界面中的某个位置也声明了条目，例如，在类型字典中，那么在其他字典中显示为带删除线，指示其当前被排除。该字符串不必出现在文本数据中，或者作为应用的任何类型字典的一部分进行声明。

注: 如果将一个概念添加到排除字典，此概念还充当同义词条目中的目标，那么也将排除目标及其所有同义词。请参阅主题第 168 页的『定义同义词』以获取更多信息。

使用通配符 (*)

对于日语之外的所有文本语言，可以使用星号通配符来表示您想要排除作为部分字符串的条目。最终提取中将排除提取引擎找到的包含以排除字典中输入的字符串开头或结尾的字的任何术语。但是，在以下两种情况下，不允许使用通配符:

- 星号通配符在连字符 (-) 前，例如，*-
- 星号通配符在撇号 (') 前，例如，*'*s*

表 39. 排除条目的示例.

条目	示例	结果
word	*next*	不会提取包含字 *next* 的概念（或其术语）。
phrase	*for example*	不会提取包含短语 *for example* 的概念（或其术语）。
partial	*copyright**	将排除匹配或包含字 *copyright* 的变体的任何概念（或其术语），例如，`copyrighted`、`copyrighting`、`copyrights` 或 `copyright 2010`。
partial	**ware*	将排除匹配或包含字 *ware* 的变体的任何概念（或其术语），例如，`freeware`、`shareware`、`software`、`hardware`、`beware` 或 `silverware`。

要添加条目

1. 在表顶部的空行中，输入术语。输入的术语将以颜色显示。此颜色表示术语显示的类型。如果术语以黑色显示，那么这意味着术语不会出现在任何类型字典中。

要禁用条目

您可以通过禁用排除列表中的条目来临时除去条目。通过禁用条目，提取期间将忽略此条目。

1. 在排除字典中，选择想要禁用的条目。
2. 单击空格键。此时将清除条目左边的复选框。

注: 您还以取消选择条目左侧的复选框以禁用条目。

要删除条目

您可以删除排除字典中任何不需要的条目。

1. 在排除字典中，选择想要删除的条目。
2. 从菜单，选择**编辑 > 删除**。条目将不再位于字典中。

第 18 章 关于高级资源

除类型字典、排除字典和替换字典以外，您还可以处理各种高级资源设置，如模糊分组设置或非语言类型定义。您可以在模板编辑器或资源编辑器视图中的“高级资源”选项卡上处理这些资源。

注意！此选项卡不适用于为日语文本调整的资源。

当转至“高级资源”选项卡时，可以编辑以下信息:

- **资源的目标语言。**用于选择将为其创建和调整资源的语言。请参阅主题第 175 页的『资源的目标语言』以获取更多信息。
- **模糊分组（异常）。**用于从模糊分组（拼写错误纠正）算法中排除单词对。请参阅主题第 175 页的『模糊分组』以获取更多信息。
- **非语言实体。**用于启用和禁用可以抽取的非语言实体，以及在抽取期间应用的正则表达式和规范化规则。请参阅主题第 176 页的『非语言实体』以获取更多信息。
- **语言处理。**用于声明构造语句（抽取模式和强制性定义）和使用所选语言的缩写的特殊方式。请参阅主题第 180 页的『语言处理』以获取更多信息。
- **语言标识。**用于配置在语言设置为**所有**时调用的自动语言标识。请参阅主题第 182 页的『语言标识』以获取更多信息。

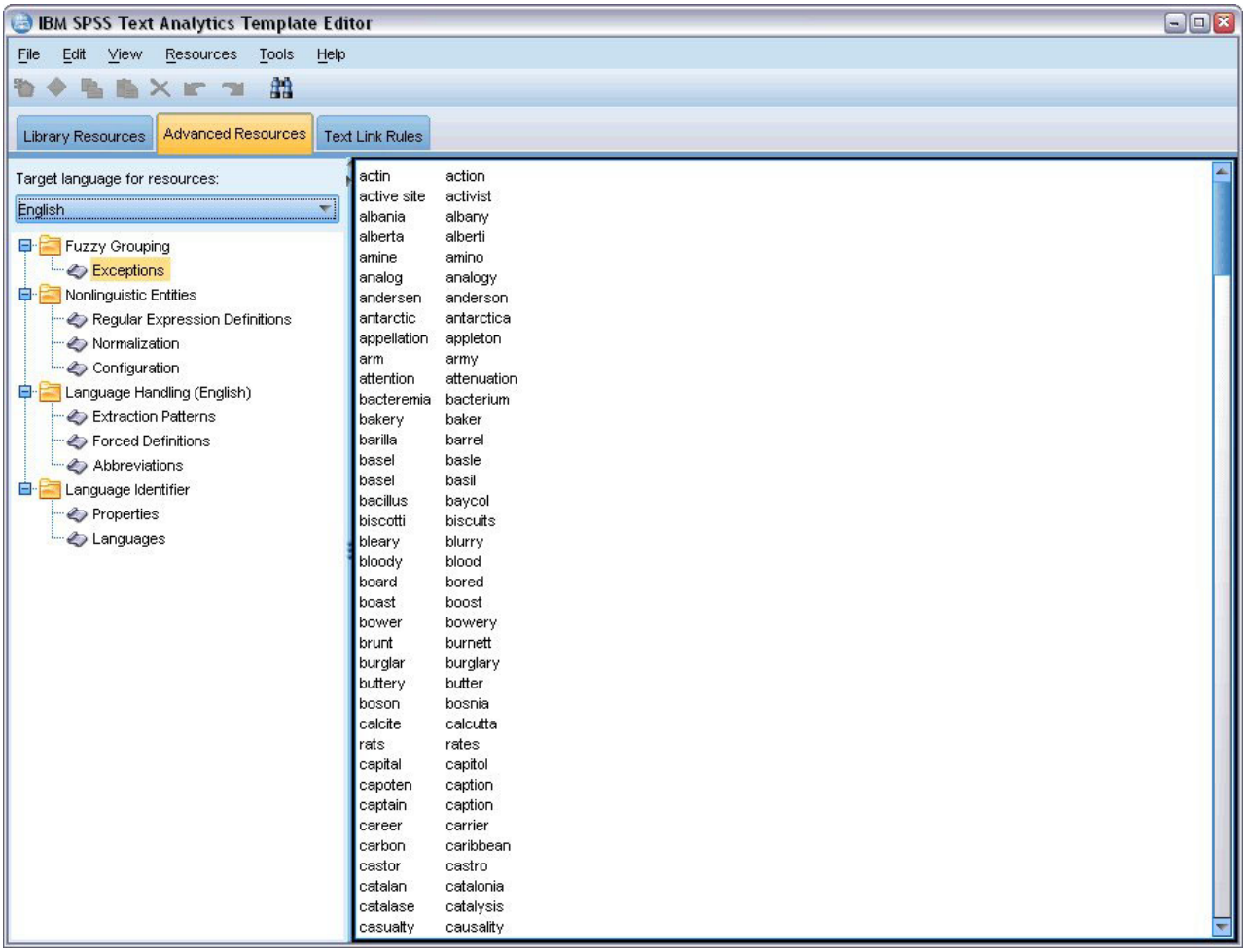


图 41. 文本挖掘模板编辑器 -"高级资源"选项卡

注: 您可以使用"查找/替换"工具栏快速查找信息或对某个部分进行统一更改。请参阅主题第 175 页的『替换』以获取更多信息。

编辑高级资源

1. 找到并选择要编辑的资源部分。内容显示在右窗格中。
2. 如有必要，使用菜单或工具栏剪切、复制或粘贴内容。
3. 使用此部分中的格式化规则编辑要更改的文件。进行更改后，随即会保存更改。使用工具栏上的撤销或重做箭头还原为先前更改。

查找

在某些情况下，您可能需要在特定部分中快速定位信息。例如，如果执行文本链接分析，那么可能会有数百个宏和模式定义。通过使用"查找"功能部件，可以快速找到特定规则。要在某个部分中搜索信息，可以使用"查找"工具栏。

使用"查找"功能部件

1. 找到并选择要搜索的资源部分。内容显示在编辑器的右窗格中。
2. 从菜单中选择**编辑 > 查找**。"查找"工具栏显示在"编辑高级资源"对话框的右上方。
3. 在文本框中输入要搜索的单词字符串。可以使用工具栏按钮来控制大小写区分、部分匹配和搜索方向。

4. 单击**查找**以开始搜索。如果找到匹配项，那么文本会在窗口中突出显示。
5. 再次单击**查找**以查找下一个匹配项。

注: 在“文本链接规则”选项卡中工作时，仅当查看源代码时，“查找”选项才可用。

替换

在某些情况下，您可能需要对高级资源进行更广泛的更新。“替换”功能部件可帮助对内容进行统一更新。

使用“更新”功能部件

1. 找到并选择要在其中进行搜索和替换的资源部分。内容显示在编辑器的右窗格中。
2. 从菜单中选择**编辑 > 替换**。这会打开“替换”对话框。
3. 在**查找内容**文本框中，输入要搜索的单词字符串。
4. 在**替换为**文本框中，输入要用于替换所找到的文本的字符串。
5. 如果要仅查找或替换完整单词，请选择**仅全字匹配**。
6. 如果要仅查找或替换大小写完全匹配的单词，请选择**匹配大小写**。
7. 单击**查找下一个**以查找匹配项。如果找到匹配项，那么文本会在窗口中突出显示。如果不希望替换此匹配项，请再次单击**查找下一个**，直至找到要替换的匹配项为止。
8. 单击**替换**以替换所选匹配项。
9. 单击**替换**以替换该部分中的所有匹配项。这会打开一条消息，其中包含进行的替换数。
10. 进行替换完成后，单击**关闭**。这会关闭对话框。

注: 如果发生替换错误，那么无法通过关闭对话框并从菜单中选择**编辑 > 撤销**来撤销替换。必须针对要撤销的每个更改执行此操作。

资源的目标语言

针对特定文本语言会创建资源。在“高级资源”选项卡中定义了会为其调整这些资源的语言。如有必要，可以切换到其他语言，方法是在**资源的目标语言**组合框中选择该语言。此外，此处所列的语言将显示为使用这些资源创建的任何文本分析包的语言。

注意！您将在极少数情况下需要更改资源中的语言。当资源与抽取语言不再匹配时，进行更改可能会导致问题。尽管很少采用，但如果计划在抽取期间使用 ALL 语言选项，那么可能会更改语言，因为预计有多种语言形式的文本。通过更改语言，可以更改例如您感兴趣的辅助语言的抽取模式、缩写和强制定义的语言处理资源。但请记住，在发布或保存已进行的资源更改或运行其他抽取之前，请将语言重置为您在抽取中感兴趣的主语言。

模糊分组

在文本挖掘节点和抽取设置中，如果选择**适应最小根字符拼写限制**，那么即已启用模糊分组算法。

模糊分组通过暂时删除所有元音（第一个元音除外）并将已抽取的单词中的辅音增加两倍或三倍，然后将其进行比较以查看其是否相同，从而帮助将经常拼写错误或拼写接近的单词分组在一起。在抽取过程中，模糊分组功能会应用于已抽取的术语，并会比较结果以确定是否找到任何匹配项。如果找到，那么会在最终抽取列表中将原始术语分组在一起。它们在数据中最频繁出现的术语下进行分组。

注: 如果进行比较的两个术语分配到不同类型（<Unknown> 类型除外），那么不会将模糊分组方法应用于此对。换句话说，术语必须属于同一类型或 <Unknown> 类型才能应用该方法。

如果启用了此功能并发现具有类似拼写的两个单词错误地分组在一起，那么可能要在模糊分组中将其排除。通过在“高级资源”选项卡中的“异常”部分中输入错误匹配对来执行此操作。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。

以下示例演示如何执行模糊分组。如果启用了模糊分组，那么这些单词看似相同并通过以下方式进行匹配:

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

在前一示例中，您可能最希望排除将 mountain 和 montana 分组在一起。因此，可以通过以下方式在“异常”部分中将其输入:

```
mountain      montana
```

注意！在某些情况下，模糊分组异常不会停止将两个单词配对，因为应用的是特定同义词规则。在该情况下，可能要尝试使用惊叹号通配符 (!) 输入同义词，以禁止单词在输出中变为同义。请参阅主题第 168 页的『定义同义词』以获取更多信息。

模糊分组异常的格式化规则

- 每行仅定义一个异常对。
- 使用简单单词或复合词。
- 仅使用单词的小写字符。将忽略大写单词。
- 使用 TAB 字符分隔对中的每个单词。

非语言实体

在处理特定种类的数据时，您可能对于抽取日期、社会安全号、百分比或其他非语言实体非常感兴趣。这些实体在可以启用或禁用实体的配置文件中进行了显式声明。请参阅主题第 179 页的『配置』以获取更多信息。为优化来自抽取引擎的输出，会根据预定义格式将来自非语言处理的输入规范化，以对类似实体进行分组。请参阅主题第 179 页的『规范化』以获取更多信息。

注: 您可以在抽取设置中打开和关闭非语言实体抽取。

可用非语言实体

可以抽取下表中的非语言实体。类型名称位于括号内。

表 40. 可以抽取的非语言实体

地址	(<Address>)
氨基酸	(<Aminoacid>)
货币	(<Currency>)
日期	(<Date>)
延迟	(<Delay>)
数字	(<Digit>)
电子邮件地址	(<email>)
HTTP/URL 地址	(<url>)
IP 地址	(<IP>)

表 40. 可以抽取的非语言实体 (续)

组织	(<Organization>)
百分比	(<Percent>)
产品	(<Product>)
蛋白质	(<Gene>)
电话号码	(<PhoneNumber>)
时间	(<Time>)
美国社会安全号	(<SocialSecurityNumber>)
权重和度量	(<Weights-Measures>)

清除文本以进行处理

在进行非语言实体抽取之前，会清除输入文本。在此步骤期间，会进行以下临时更改，以便可按如下方式标识和抽取非语言实体:

- 将两个或多个空格的任何序列替换为单个空格。
- 将制表符替换为空格。
- 将行结束字符或序列字符替换为空格，而将多个行结束序列标记为段落结束。行结束可由回车符 (CR) 和换行符 (LF) 甚至两者一起进行表示。
- 临时删除并忽略 HTML 和 XML 标记。

正则表达式定义

在抽取非语言实体时，您可能希望对其进行编辑或将其添加到用于标识正则表达式的正则表达式定义中。此操作在“高级资源”选项卡上的**正则表达式定义**部分中完成。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。

文件划分为多个不同部分。第一个部分称为 [macros]。除该部分以外，对于每个非语言实体还可能存在其他部分。可以将各部分添加到此文件中。在每个部分中，规则会进行编号（*regexp1*、*regexp2*，依此类推）。这些规则必须从 1 到 *n* 顺序编号。任何编号中断都将导致此文件的处理全部暂挂。

在某些情况下，实体根据语言而定。如果实体针对配置文件中的语言参数采用除 0 以外的值，那么将该实体视为根据语言而定。请参阅主题第 179 页的『配置』以获取更多信息。当实体根据语言而定时，必须使用语言作为部分名称的前缀，例如 [english/PhoneNumber]。该部分将会包含仅当针对英语指定 PhoneNumber 实体的值为 2 时才适用于英语电话号码的规则。

注意！如果在编辑器中对此文件或任何其他文件进行更改，并且抽取引擎不再按期望工作，请使用工具栏上的**重置为原始**选项将文件重置为原始交付内容。此文件要求对正则表达式有一定程度的熟悉。如果在此方面需要其他帮助，请联系 IBM Corp. 以获取帮助。

特殊字符包括 . [] {} () \ * + ? | ^ $

除在表达式中用于特定用途的以下特殊字符外，所有字符都与自身匹配: .[{()*+?|^$ 要照此使用这些字符，必须在定义中为其前置反斜杠 (\)。

例如，如果尝试抽取 Web 地址，那么句号字符对于实体非常重要，因此必须为其添加反斜杠，例如:

```
www\.[a-z]+\.[a-z]+
```

重复运算符和限定符包括 ? + * {}

要使定义更灵活，可以对正则表达式使用若干标准通配符。包括 * ? +

- *星号* * 指示有*零个或多个*前置字符串。例如，ab*c 与“*ac*”、“*abc*”、“*abbbc*”等等匹配。
- *加号* + 指示有*一个或多个*前置字符串。例如，ab+c 与“*abc*”、“*abbc*”、“*abbbc*”匹配，但与“*ac*”不匹配。
- *问号* ?指示有*零个或一个*前置字符串。例如，modell?ing 与“*modeling*”和“*modeling*”均匹配。
- *使用大括号 {} 限制重复*指示重复的边界。例如，

  [0-9]{n} 与恰好重复 *n* 次的数字匹配。例如，[0-9]{4} 将与“*1998*”匹配，但是既不与“*33*”匹配，也不与“*19983*”匹配。

  [0-9]{n,} 与重复 *n 次或多次*的数字匹配。例如，[0-9]{3,} 将与“*199*”或“*1998*”匹配，但不与“*19*”匹配。

  [0-9]{n,m} 与重复介于 *n 次和 m 次（包括 n 和 m）*之间的数字匹配。例如，[0-9]{3,5} 将与“*199*”、“*1998*”或“*19983*”匹配，但既不与“*19*”匹配，也不与“*199835*”匹配。

可选空格和连字符

在某些情况下，您希望在定义中包含可选空格。例如，如果要抽取货币（如“*uruguayan pesos*”、“*uruguayan peso*”、“*uruguay pesos*”、“*uruguay peso*”、“*pesos*”或“*peso*”），那么将需要考虑可能实际有两个以空格分隔的单词的情况。在此情况下，此定义应编写为 (uruguayan |uruguay)?pesos?。由于 *uruguayan* 或 *uruguay* 在与 *pesos*/*peso* 配合使用时后跟空格，因此必须在可选序列 (uruguayan |uruguay) 中定义可选空格。如果可选序列中不含空格（如 (uruguayan|uruguay)? pesos?），那么它将不与“*pesos*”或“*peso*”匹配，因为必需空格。

如果在列表中查找包括连字符在内的一系列内容，那么必须最后定义连字符。例如，如果查找逗号 (,) 或连字符 (-)，请使用 [,-] 而绝不使用 [-,]。

列表和宏中的字符串顺序

应始终在较短的序列之前定义最长的序列，否则将永不读取最长的序列，因为将对较短的序列进行匹配。例如，如果查找字符串“*billion*”或“*bill*”，那么必须在“*bill*”之前定义“*billion*”。因此，例如 (billion|bill) 而不是 (bill|billion)。这也适用于宏，因为宏是字符串列表。

定义部分中的规则顺序

每行定义一个规则。在每个部分中，规则会进行编号（*regexp1*、*regexp2*，依此类推）。这些规则必须从 1 到 *n* 顺序编号。任何编号中断都将导致此文件的处理全部暂挂。要禁用条目，请在用于定义正则表达式的每行的开头放置 # 符号。要启用条目，请移除该行前面的 # 字符。

在每个部分中，应在最广义的规则之前定义最具体的规则，以确保正确处理。例如，如果查找格式为“*month year*”和“*month*”的日期，那么必须在“*month*”规则之前定义“*month year*”规则。以下显示应如何进行定义:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

而不是

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

在规则中使用宏

只要在若干规则中使用了特定序列，即可使用宏。然后，如果需要更改此序列的定义，那么只需更改一次即可，无需在引用此序列的所有规则中都进行更改。例如，假设具有以下宏:

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

只要引用该宏的名称，就必须在 $() 中将其括起来，例如: regexp1=$(MONTH)

必须在 [macros] 部分中定义所有宏。

规范化

抽取非语言实体时，所遇到的实体会根据预定义格式规范化，以对类似实体进行分组。例如，货币符号及其在单词中的等效项视为相同。规范化条目存储在“高级资源”选项卡上的**规范化**部分中。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。文件划分为多个不同部分。

注意！此文件仅供高级用户使用。很可能无需更改此文件。如果在此方面需要其他帮助，请联系 IBM Corp. 以获取帮助。

规范化的格式化规则

- 每行添加一个规范化条目。
- 严格遵循此文件中的各个部分。不能添加新的部分。
- 要禁用条目，请在该行的开头放置 # 符号。要启用条目，请移除该行前面的 # 字符。

规范化形式的英语日期

缺省情况下，英语模板中的日期按美国样式的日期格式进行识别；即: 月，日期，年。如果需要更改为日，月，年格式，请禁用“format:US”行（通过在该行开头添加 #）并启用“format:UK”（通过从该行中移除 #）。

配置

您可以在非语言实体配置文件中启用和禁用要抽取的非语言实体类型。通过禁用不需要的实体，可以减少所需的处理时间。此操作在“高级资源”选项卡上的**配置**部分中完成。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。如果启用非语言实体，那么抽取引擎会在抽取过程中读取此配置文件，以确定应抽取哪些非语言实体类型。

此文件的语法如下:

#name<TAB>*Language*<TAB>*Code*

表 41. 配置文件的语法.

列标题	描述
#name	将在其他两个必需文件中引用非语言实体以进行非语言实体抽取时所采用的用词。此处使用的名称区分大小写。
Language	文档的语言。最好是选择特定语言；但是，存在**任意**选项。可能的选项为: 0 = 任意，只要正则表达式不是特定于某个语言并可在具有不同语言的若干模板中使用（例如 IP/URL/电子邮件地址），便会使用此项；1 = 法语；2 = 英语；4 = 德语；5 = 西班牙语；6 = 荷兰语；8 = 葡萄牙语；10 = 意大利语。

表 41. 配置文件的语法 (续).

列标题	描述
Code	部分语音代码。除少数情况以外，大多数实体采用“s”作为值。可能的值为: s = 非用词; a = 形容词; n = 名词。如果启用，那么会首先抽取非语言实体，然后应用抽取模式以在更大的上下文中标识其角色。例如，百分比的值指定为“a”。假设抽取 30% 作为非语言实体。系统会将其标识为形容词。然后，如果文本包含“30% salary increase”，那么“30%”非语言实体符合部分语音模式“ann”（形容词 名词 名词）。

实体定义顺序

此文件中的实体声明顺序非常重要，并会影响其抽取方式。实体按所列顺序进行应用。更改顺序将会更改结果。必须在更广义的非语言实体之前定义最具体的非语言实体。

例如，非语言实体“Aminoacid”定义如下:

```
regexp1=($(AA)-?$(NUM))
```

其中 $(AA) 对应于“(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”，后者是对应于特定氨基酸的三字母序列。

另一方面，非语言实体“Gene”更广义且定义如下:

```
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?
```

如果在“配置”部分中“Gene”定义在“Aminoacid”之前，那么将永不匹配“Aminoacid”，因为“Gene”中的 regexp3 将始终首先匹配。

配置的格式化规则

- 使用 TAB 字符分隔列中的每个条目。
- 不删除任何行。
- 遵循上表中显示的语法。
- 要禁用条目，请在该行的开头放置 # 符号。要启用实体，请移除该行前面的 # 字符。

语言处理

如今使用的每种语言都有特殊的表达构想、构造语句和使用缩写的方式。在“语言处理”部分中，可以编辑抽取模式，强制对这些模式进行定义，以及声明已在“语言”下拉列表中选择的语言的缩写。

- 抽取模式
- 强制性定义
- 缩写

提取模式

在从文档提取信息时，提取引擎将一组词性提取模式应用于上下文中字的“堆栈”，以识别提取的候选术语（字和短语）。您可以添加或修改提取模式。

词性包括语法元素，例如，名词、形容词、过去分词、限定词、介词、连词、名、缩写和虚词。这一系列元素构成词性提取模式。在 IBM Corp. 文本挖掘产品中，每个词性都由单个字符表示，从而更易于定义您的模式。

例如，小写字母 *a* 表示形容词。缺省情况下，支持的代码集在每个缺省提取模式部分的顶部显示，并随附一组模式和每个模式的示例，以帮助您了解使用的每个代码。

提取模式的格式化规则

- 每行一个模式。
- 在行的开头使用 # 来禁用模式。

列出的提取模式的顺序非常重要，因为提取引擎仅读取指定的字序列一次，并指定给引擎找到匹配的第一个提取模式。

强制的定义

在从文档提取信息时，提裙引擎扫描文本并针对其遇到的每个字标识词性。在某些情况下，一个字根据上下文适合不同角色。如果想要强制字采用特定词性角色或从处理中完全排除字，您可以在“高级资源”选项卡的**强制的定义**部分执行此操作。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。

要针对指定的字强制词性角色，必须使用以下语法向此部分添加一行:

term:*code*

表 42. 语法描述.

条目	描述
term	术语名称。
code	表示词性角色的单字符代码。对于每个单元词，最多可列出 6 个不同词性角色。或者，您可以使用小写代码 s 停止将某个字提取到组合字/段元，例如，additional:s。

强制的定义的格式化规则

- 每个字一行。
- 术语不能包含冒号。
- 使用小写 s 作为词性代码以停止一起提取某个字。
- 每行最多使用 6 个词性代码。“提取模式”部分中显示支持的词性代码。请参阅主题第 180 页的『提取模式』以获取更多信息。
- 使用星号字符 (*) 作为部分匹配字符串结束处的通配符。例如，如果输入 add*:s，那么从不会作为术语或复合词术语的一部分提取诸如 add、additional、additionally、addendum 和 additive 的字。但是，如果在已编译的字典或强制的定义中明确将字匹配声明为术语，那么仍将提取。例如，如果输入 add*:s 和 addendum:n，那么在文本中找到 addendum 时，仍将进行提取。

缩写

在提取引擎处理文本时，通常将其找到的任何时间段都视为语句结束的指示。这通常正确；但是，在文本中包含缩写时，此时间段字符处理可能不适用。

如果从文本提取术语并且发现未处理某些缩写，那么应明确声明此部分中的缩写。

注: 如果缩写已在同义词定义中出现或者定义为类型字典中的术语，那么无需在此添加缩写条目。

缩写的格式化规则

- 每行定义一个缩写。

语言标识

虽然最好始终为进行分析的文本数据选择特定语言，但是也可以在文本可能采用若干不同或未知语言时指定**所有**选项。**所有**语言选项使用语言自动识别引擎，称为语言标识。语言标识会扫描文档来标识采用受支持语言的文档，并在抽取期间为每个文件自动应用最佳内部字典。**所有**选项由“属性”部分中的参数管理。

属性

语言标识使用此部分中的参数进行配置。下表描述可以在“高级资源”选项卡上的**语言标识 - 属性**部分中设置的参数。请参阅主题第 173 页的第 18 章，『关于高级资源』以获取更多信息。

表 43. 参数描述

参数	描述
NUM_CHARS	指定抽取引擎应读取的字符数，以便确定文本采用的语言。数字越小，标识语言的速度越快。数字越大，标识语言的准确性越高。如果将值设置为 0，那么将读取文档的全部文本。
USE_FIRST_SUPPORTED _LANGUAGE	指定抽取引擎是否应使用语言标识找到的第一个受支持语言。如果将值设置为 1，那么将使用第一个受支持语言。如果将值设置为 0，那么将使用回退语言。
FALLBACK_LANGUAGE	指定在标识返回的语言不受支持的情况下要使用的语言。可能的值为 english、french、german、spanish、dutch、italian 和 ignore。如果将值设置为 ignore，那么将忽略具有不受支持语言的文档。

语言

语言标识支持多种不同语言。可以在“高级资源”选项卡上的**语言标识 - 语言**部分中编辑语言列表。

您可以考虑从此列表中消除不可能使用的语言，因为存在的语言越多，假正的几率越高，并且性能速度越慢。不过，无法向此文件中添加新语言。请考虑将最可能的语言置于列表顶部，以帮助语言标识更快查找文档的匹配项。

第 19 章 关于文本链接规则

文本链接分析 (TLA) 是使用规则集来抽取在文本中找到的关系的一种模式匹配技术。为抽取启用文本链接分析后，会将文本数据与这些规则相比较。当找到匹配项时，会抽取并呈现文本链接分析模式。这些规则定义在“文本链接规则”选项卡中。

例如，抽取表示有关组织的简单构想的概念可能不足以引起您的兴趣，但通过使用 TLA，您还可以了解不同组织或与组织关联的人员之间的链接。TLA 也可用于抽取有关主题的意见（如人员对给定产品或体验的看法）。

要从 TLA 获益，您必须具有包含文本链接 (TLA) 规则的资源。当选择模板时，可以通过查看模板在 TLA 列中是否有图标来了解哪些模板具有 TLA 规则。

在抽取过程的模式匹配阶段，会在文本数据中找到文本链接分析模式。在此阶段中，规则会与文本数据进行比较，当找到匹配项时，会将此信息作为模式进行抽取。有时可能要从文本链接分析获取更多模式，或者更改匹配方式。在这些情况下，可以优化规则来使其适应特定需求。此操作在“文本链接规则”选项卡中执行。

注: 在 V13 中已停止支持变量。请改用宏。请参阅主题第 187 页的『处理宏』以获取更多信息。

在何处处理文本链接规则

您可以直接在模板编辑器或资源编辑器视图中的“文本链接规则”选项卡中编辑和创建规则。要帮助了解规则可能如何与文本匹配，可以在此选项卡中运行模拟。在模拟期间，会仅对样本模拟数据运行抽取，并会应用文本链接规则来查看是否有任何模式匹配。然后，将在模拟窗格中显示与文本匹配的任何规则。基于匹配项，可以选择编辑规则和宏来更改文本的匹配方式。

与其他高级资源不同，TLA 规则特定于库；因此，一次只能使用一个库中的 TLA 规则。从模板编辑器或资源编辑器内，转至**文本链接规则**选项卡。在此选项卡中，可以在模板中指定其中包含要使用或编辑的 TLA 规则的库。为此，除非有非常具体的原因，否则强烈建议将所有规则都存储在一个库中。

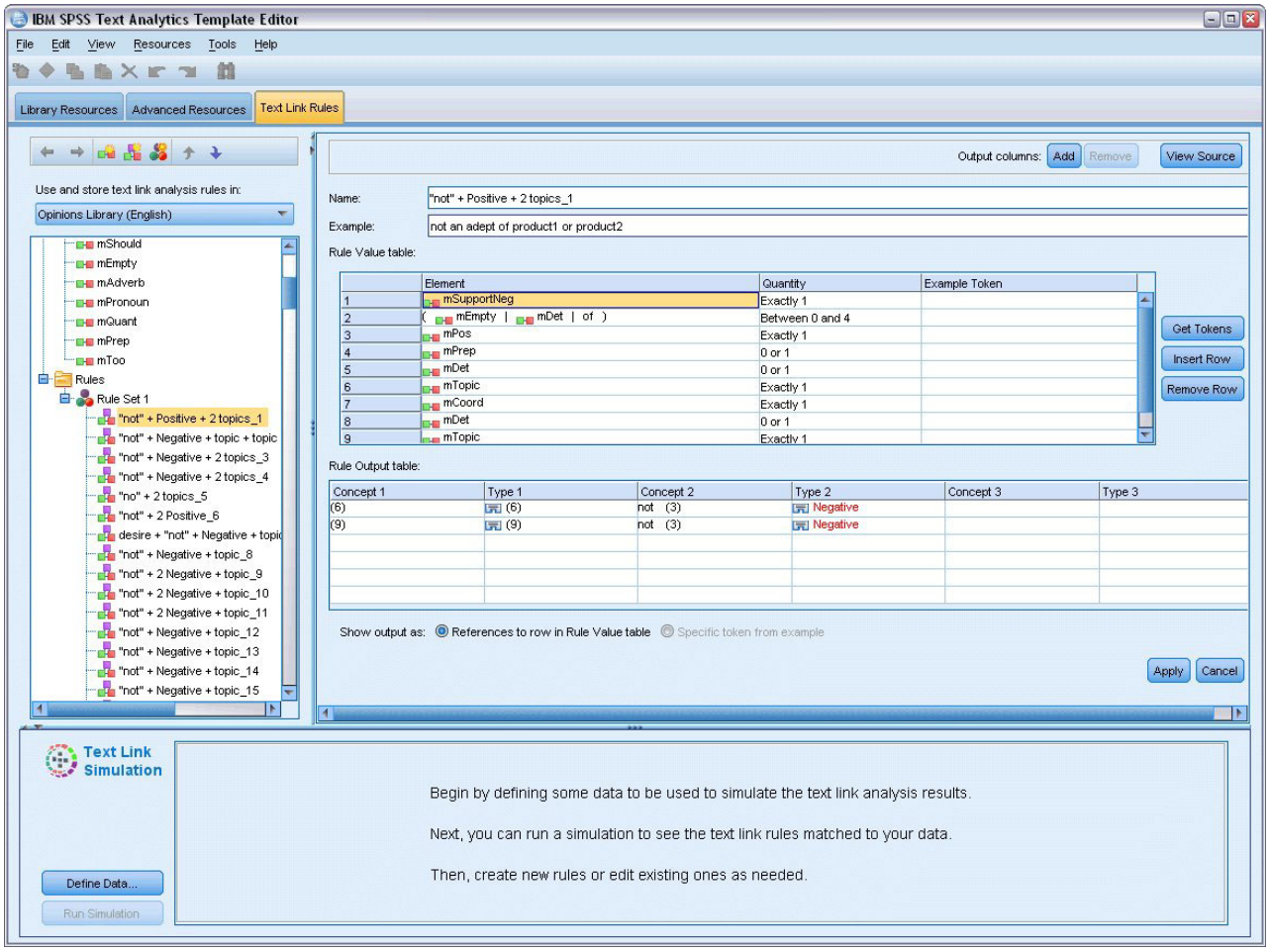


图 42. "文本链接规则"选项卡

注意！此选项卡不适用于日语语言资源。

从何处开始

在"文本链接规则"选项卡编辑器中开始工作有多种方式:

- 首先使用一些样本文本来模拟结果，并且根据当前规则集如何从模拟数据抽取模式来编辑或创建匹配规则。
- 从头开始创建新规则或编辑现有规则。
- 直接在源视图中工作。

何时编辑或创建规则

尽管随附于每个模板的文本链接分析规则通常足以用于从文本中抽取多个简单或复杂关系，但有时可能要对这些规则进行一些更改或创建自己的一些规则。例如,

- 通过创建新的规则或宏来捕获不是使用现有规则抽取的的构想或关系。
- 更改添加到资源的类型的缺省行为。这通常要求编辑诸如 `mTopic` 或 `mNonLingEntities` 之类的宏。请参阅主题第 189 页的『专用宏: mTopic、mNonLingEntities、SEP』以获取更多信息。

- 向现有文本链接分析规则和宏添加新类型。例如，如果您认为类型 <Organization> 过于广泛，那么可以在若干不同商业领域（如 <Pharmaceuticals>、<Car Manufacturing>、<Finance> 等）中为组织创建新类型。在此情况下，必须编辑文本链接分析规则和/或创建宏，以将这些新类型考虑在内并相应地对其进行处理。
- 向现有文本链接分析规则添加类型。 例如，假设您具有的规则会捕获文本 john doe called jane doe，但是您希望可捕获电话通信的此规则还捕获电子邮件交换。可以将非语言实体类型从电子邮件添加到规则，使其还会捕获如下文本：johndoe@ibm.com emailed janedoe@ibm.com。
- 略微修改现有规则，而不是创建新规则。例如，假设您具有的规则与文本 xyz is very good 匹配，但是您希望此规则还捕获 xyz is very, very good。

模拟文本链接分析结果

为帮助定义新文本链接规则或帮助了解特定语句在文本链接分析期间如何进行匹配，采用文本段样本并运行模拟往往非常有用。在模拟期间，会使用当前语言资源集和当前抽取设置仅对样本模拟数据运行抽取。目标是获取模拟结果并使用这些结果改进规则，创建新规则或更好地了解如何进行匹配。对于每个文本段（语句、单词或子句，具体视上下文而定），模拟输出会显示标记与任何已在该文本中发现模式的 TLA 规则的集合。**标记**定义为抽取过程中标识的任何单词或单词短语。

与其他高级资源不同，TLA 规则特定于库；因此，一次只能使用一个库中的 TLA 规则。从模板编辑器或资源编辑器内，转至**文本链接规则**选项卡。在此选项卡中，可以在模板中指定其中包含要使用或编辑的 TLA 规则的库。为此，除非有非常具体的原因，否则强烈建议将所有规则都存储在一个库中。

注意！强烈建议如果使用数据文件，请确保其包含的文本简短，以便尽量缩短处理时间。模拟的目标是查看如何解释文本段以及了解规则如何与此文本匹配。此信息将帮助编写和编辑规则。使用文本链接分析节点或在已启用 TLA 抽取的情况下对交互式会话运行流，以获取更完整数据集的结果。此模拟仅用于测试和规则编写目的。

为模拟定义数据

要帮助了解规则可能如何与文本匹配，可以使用样本数据运行模拟。第一步是定义数据。

定义数据

1. 单击**文本链接规则**选项卡底部的模拟窗格中的**定义数据**。或者，如果先前未定义任何数据，请从菜单中选择**工具 > 运行模拟**。这会打开“模拟数据”向导。
2. 通过选择以下选项之一指定数据类型:
   - **直接粘贴或输入文本**。文本框提供用于从剪贴板粘贴一些文本或手动输入要处理的所需文本。可以每行输入一个语句，也可以使用标点符号（如句点或逗号）进行断句。输入文本后，即可通过单击**运行模拟**来开始模拟。
   - **指定文件数据源**。此选项指示要处理其中包含文本的文件。 单击**下一步**以继续执行可在其中定义要处理的文件的向导步骤。选定文件后，即可通过单击**运行模拟**来开始模拟。支持以下文件类型：.txt 和 .text。您选择的数据文件在模拟期间“按原样”读取。整个文件的处理方式就如同已将文件列表节点连接到文本挖掘节点一样。

   要点：强烈建议如果使用数据文件，请确保其包含的文本简短，以便尽量缩短处理时间。模拟的目标是查看如何解释文本段以及了解规则如何与此文本匹配。此信息将帮助编写和编辑规则。使用文本链接分析节点或在已启用 TLA 抽取的情况下对交互式会话运行流，以获取更完整数据集的结果。此模拟仅用于测试和规则编写目的。
3. 要开始执行模拟过程，请单击**运行模拟**。这会显示进度对话框。如果您处在交互式会话中，那么模拟期间使用的抽取设置是交互式会话中当前选择的设置（请参阅“概念和类别”视图中的**工具 > 抽取设置**）。如果

您处在模板编辑器中，那么模拟期间使用的抽取设置是与文本链接分析节点的“专家”选项卡中所显示相同的缺省抽取设置。有关更多信息，请参阅『了解模拟结果』。

了解模拟结果

要帮助了解规则可能如何与文本匹配，可以使用样本数据运行模拟并复查结果。从中可以更改规则集，以更好地适用于数据。当抽取和模拟过程完成后，将会呈现模拟结果。

对于抽取期间标识的每个“语句”，将会呈现若干信息段，包括确切“语句”，在此输入文本语句中找到的标记的细分，最后是与该语句中的文本匹配的任何规则。所谓“**语句**”，是指单词、语句或子句，具体取决于抽取器如何将文本细分为可读区块。

标记定义为提取过程期间标识的任何字或短语。例如，在语句 *My uncle lives in New York* 中，提取期间可以找到以下标记: *my*、*uncle*、*lives*、*in* 和 *new york*。或者，*uncle* 可以提取为概念并且类型为 <Unknown>，并且 *new york* 可提取为概念并且类型为 <Location>。所有概念均是标记，但是并非所有标记都是概念。标记还可以是其他宏、文字串和字距。仅类型化的这些字或短语才可以是概念。

当在交互式会话或资源编辑器中工作时，即是在概念级别工作。TLA 规则更精细，并且可在规则的定义中使用语句中的个别标记，即使其从未抽取并设定类型也如此。能够使用不是概念的标记可为规则在文本中捕获复杂关系方面提供甚至更多灵活性。

如果在模拟数据中有多个语句，那么可以通过单击**下一个**和**上一个**来迁移和后移浏览结果。

在语句不与所选库中的任何 TLA 规则匹配的情况下（请参阅此选项卡中树上方的库名），系统会将结果视为不匹配并启用按钮**下一个不匹配项**和**上一个不匹配项**，以告知存在规则找不到其匹配项的文本，并允许快速浏览至这些实例。

创建新规则，编辑规则或者更改资源或抽取设置后，可能要重新运行模拟。要重新运行模拟，请单击模拟窗格中的**运行模拟**，然后将再次使用相同的输入数据。

在模拟结果中显示了以下字段和表:

输入文本。抽取过程根据向导中定义的模拟数据标识的实际“语句”。所谓语句，是指单词、语句或子句，具体取决于抽取器如何将文本细分为可读区块。

系统视图。抽取过程已标识的标记的集合。

- **输入文本标记。**在输入文本中找到的每个标记。早先在本主题中定义了标记。
- **类型设定为。**如果将标记标识为概念并设定类型，那么在此列中会显示关联类型名称（如 <Unknown>、<Person>、<Location>）。
- **匹配宏。**如果标记与现有宏匹配，那么在此列中会显示关联宏名称。

与输入文本匹配的规则。此表显示已与输入文本匹配的任何 TLA 规则。对于每个匹配的规则，将会显示**规则输出**列中的规则名称以及该规则的关联输出值（“概念”+“类型”对）。可以在匹配的规则名称上双击，以在模拟窗格上方的编辑器窗格中打开该规则。

生成规则按钮。如果单击模拟窗格中的此按钮，那么将在模拟窗格上方的规则编辑器窗格中打开新规则。它将以输入文本作为其示例。同样，在模拟期间已设定类型或与宏匹配的任何标记都会自动插入在**规则值表**的“元素”列中。如果标记已设定类型*并*与宏匹配，那么宏值即是规则中将使用的值，从而简化规则。例如，如果使用的是基本英语资源，那么语句“*I like pizza*”在模拟期间可以将类型设定为 <Unknown> 并与宏 mTopic 匹配。在此情况下，mTopic 将用作所生成的规则中的元素。请参阅主题第 190 页的『处理文本链接规则』以获取更多信息。

在树中浏览规则和宏

在抽取期间执行文本链接分析时，将使用**文本链接规则**选项卡中选择的库中存储的文本链接规则。

与其他高级资源不同，TLA 规则特定于库；因此，一次只能使用一个库中的 TLA 规则。从模板编辑器或资源编辑器内，转至**文本链接规则**选项卡。在此选项卡中，可以在模板中指定其中包含要使用或编辑的 TLA 规则的库。为此，除非有充分或具体的原因，否则强烈建议将所有规则都存储在一个库中。

可以在“文本链接规则”选项卡中指定要在其中工作的库，方法是在此选项卡中的**文本链接分析规则的使用和存储位置**：下拉列表中选择该库。在抽取期间执行文本链接分析时，将使用**文本链接规则**选项卡中选择的库中存储的文本链接规则。因此，如果已在多个库中定义文本链接规则（TLA 规则），那么将仅使用在其中找到 TLA 规则的第一个库进行文本链接分析。为此，除非有非常具体的原因，否则强烈建议将所有规则都存储在一个库中。

在树中选择宏或规则时，其内容显示在右侧的编辑器窗格中。如果右键单击树中的任何项目，那么将打开一个上下文菜单以显示可能的其他任务，例如:

- 在树中创建新宏并在右侧的编辑器中将其打开。
- 在树中创建新规则并在右侧的编辑器中将其打开。
- 在树中创建新规则集。
- 剪切、复制并粘贴项目以简化编辑。
- 删除宏、规则和规则集以将其从资源中移除。
- 禁用宏、规则和规则集以指示在处理期间应将其忽略。
- 将规则上移或下移以影响处理顺序。

树中的警告

警告在树中显示带有黄色三角形，并且用于告知可能存在问题。将鼠标指针悬停在错误的宏或规则上方会显示弹出说明。在大多数情况下，将会显示诸如**警告：未提供示例；请输入示例**之类的内容，因此需要输入示例。

如果缺少示例，或者如果示例与规则不匹配，那么将无法使用“获取标记”功能，因此建议每行仅输入一个示例。

当规则以黄色突出显示时，意味着类型或宏对于 TLA 编辑器未知。消息将类似于：**警告：未知类型或宏**。这旨在告知在源视图中将按 `$something` 进行定义的项目（例如 `$myType`）既不是库中的旧类型，也不是宏。

要更新语法检查程序，需要切换到其他规则或宏；无需重新编译任何内容。因此，例如，如果规则 `A` 显示警告（因为缺少示例），那么需要添加示例，单击上面或下面的规则，然后返回到规则 `A` 以检查其现在是否正确。

处理宏

宏通过允许使用 `OR` 运算符 (|) 将类型、其他宏和文字（单词）串分组在一起，可以简化文本链接分析规则的外观。使用宏的优点是，不仅能够在多个文本链接分析规则中复用宏来将其简化，还支持在一个宏中进行更新，而不必在所有文本链接分析规则中都进行更新。大多数交付的 TLA 规则都包含预定义宏。宏显示在“文本链接规则”选项卡的最左侧窗格中的树顶部。

在模拟结果中显示了以下字段和表:

名称。标识此宏的唯一名称。建议使用小写 m 作为宏名称前缀，以帮助在规则中快速标识宏。在规则中手动引用宏时（通过内嵌编辑或在源视图中），必须使用 $ 字符前缀，以便抽取过程知道查找此特殊名称。但是，如果拖放宏名称或通过上下文菜单添加该名称，那么产品会自动将其识别为宏且不添加 $。

宏值表。

- 表示此宏可以表示的所有可能的值的多个行。这些值区分大小写。
- 这些值可以包含一个类型、文字串、单词间隙或宏，也可以包含其组合。请参阅主题第 195 页的『规则和宏的受支持元素』以获取更多信息。
- 要在宏中输入元素的值，请双击要在其中工作的行。这会显示一个可编辑文本框，可在其中输入类型引用、宏引用、文字串或单词间隙。或者，在单元格中右键单击以显示上下文菜单，其中提供常用的宏、类型名称和非语言类型名称的列表。要引用类型或宏，必须将"$"字符前置于宏或类型名称，例如 `$mTopic` 表示宏 `mTopic`。在组合自变量时，必须使用括号 `( )` 将自变量和字符 `|` 分组以指示布尔 `OR`。
- 可以使用宏值表右侧的按钮在该表中添加或移除行。
- 在各自的行中输入每个元素。例如，如果要创建将 3 个文字串其中之一表示为 `am OR was OR is` 的宏，那么会在视图中不同的行上输入每个文字串，并且宏表会包含 3 行。

创建和编辑宏

您可以创建新宏或编辑现有宏。请遵循宏编辑器的准则和描述进行操作。请参阅主题第 187 页的『处理宏』以获取更多信息。

创建新宏

1. 从菜单中选择**工具 > 新建宏**。或者，单击树工具栏中的"新建宏"图标以在编辑器中打开新宏。
2. 输入唯一名称并定义宏值元素。
3. 完成后单击**应用**以检查错误。

编辑宏

1. 单击树中的宏名称。这会在右侧的编辑器窗格中打开该宏。
2. 进行更改。
3. 完成后单击**应用**以检查错误。

禁用和删除宏

禁用宏

如果要在处理期间忽略宏，那么可以禁用该宏。执行此操作可能会在任何仍然引用此已禁用宏的规则中导致警告或错误。请谨慎删除和禁用宏。

1. 单击树中的宏名称。这会在右侧的编辑器窗格中打开该宏。
2. 右键单击名称。
3. 从上下文菜单中，选择**禁用**。宏图标变为灰色，并且宏本身变为不可编辑。

删除宏

如果要去除宏，那么可以将其删除。执行此操作可能会在任何仍然引用此宏的规则中导致错误。请谨慎删除和禁用宏。

1. 单击树中的宏名称。这会在右侧的编辑器窗格中打开该宏。
2. 右键单击名称。
3. 从上下文菜单中，选择**删除**。宏从列表中消失。

检查错误、保存和取消

应用宏更改

如果在宏编辑器外部单击，或者如果单击**应用**，那么会自动扫描宏以查找错误。如果找到错误，那么将需要对其进行修订，然后再继续移至应用程序的其他部分。

但是，如果检测到不太严重的错误，那么仅会发出警告。例如，如果宏包含类型或其他宏的不完整或未引用的定义，那么会显示警告消息。 一旦单击**应用**，任何未更正的警告就会导致在左窗格中的“规则和宏树”中的宏名称左侧出现警告图标。

应用宏并不意味着会永久保存宏。应用将导致验证过程检查错误和警告。

在交互式工作台会话内保存资源

1. 要在交互式工作台会话期间保存对资源进行的更改，从而可在下次运行流时获取这些资源，必须执行以下操作:
   - 更新建模节点，以确保可在下次执行流时获取这些相同资源。请参阅主题第 70 页的『更新建模节点并保存』以获取更多信息。然后，保存流。要保存流，请在更新建模节点后在 IBM SPSS Modeler 主窗口中执行此操作。
2. 要在交互式工作台会话期间保存对资源进行的更改，从而可在其他流中使用这些资源，可以执行以下操作:

- 更新所使用的模板或创建新模板。请参阅主题第 141 页的『创建和更新模板』以获取更多信息。这将不会保存当前节点的更改（请参阅上一步）
- 或者，更新所使用的 TAP。请参阅主题第 118 页的『更新文本分析包』以获取更多信息。

在模板编辑器内保存资源

1. 首先，发布库。请参阅主题第 158 页的『发布库』以获取更多信息。
2. 然后，通过菜单中的**文件 > 保存资源模板**来保存模板。

取消宏更改

1. 如果希望废弃更改，请单击**取消**。

专用宏: mTopic、mNonLingEntities、SEP

“意见”模板（和类似模板）以及“基本资源”模板随附两个专用宏，称为 `mTopic` 和 `mNonLingEntities`。

mTopic

缺省情况下，宏 `mTopic` 对模板中交付的可能与意见连接的所有类型进行分组，例如 *Core* 库类型: `<Person>`、`<Organization>`、`<Location>` 等，只要类型不是意见类型（例如 `<Negative>` 或 `<Positive>`）或在“高级资源”中定义为非语言实体的类型即可。

只要在“意见”或类似模板中创建新类型，产品便会假设除非在另一个宏中或在“高级资源”选项卡的非语言实体部分中指定此类型，否则将按照处理宏 `mTopic` 中定义的其他类型的方式对其进行处理。

假设从“意见”模板在资源中创建了新类型: `<Vegetables>` 和 `<Fruit>`。在不必进行任何更改的情况下，新类型作为 `mTopic` 类型处理，因此可以自动发现关于新类型的肯定、否定、中立和上下文意见。例如，在抽取期间，语句“*I enjoy broccoli, but I hate grapefruit*”会产生以下两个输出模板:

```
broccoli <Vegetables> + like <Positive>

grapefruit <Fruit> + dislike <Negative>
```

但是，如果要以不同于 mTopic 中的其他类型的方式处理这些类型，那么可以将类型名称添加到现有宏（例如 mPos，用于对所有肯定意见类型进行分组），或者创建稍后可在一个或多个规则中引用的新宏。

注意！如果创建诸如 <Vegetables> 之类的新类型，那么此新类型将包含作为 mTopic 中的类型，但是此类型名称在宏定义中将不会显式可视。

mNonLingEntities

同样，如果在“高级资源”选项卡的**非语言实体**部分中添加新的语言实体，那么除非另外指定，否则会自动将其作为 mNonLingEntities 处理。请参阅主题第 176 页的『非语言实体』以获取更多信息。

SEP

您还可以使用预定义宏 SEP，它与本地计算机上定义的全局分隔符（通常为逗号 (,)）对应。

处理文本链接规则

文本链接分析规则是用于对语句执行匹配的布尔查询。文本链接分析规则包含以下一个或多个自变量：类型、宏、文字串或单词间隙。您必须具有至少一个文本链接分析规则，以便抽取 TLA 结果。

在规则编辑器的“文本链接规则”选项卡中显示了以下区域和字段:

名称字段。文本链接规则的唯一名称。

示例字段。（可选）可以包含将会由此规则捕获的示例语句或单词序列。建议使用示例。在此编辑器中，将能够从此示例文本生成标记，以了解其如何与规则匹配以及将如何输出。**标记**定义为提取过程期间标识的任何字或短语。例如，在语句 *My uncle lives in New York* 中，提取期间可以找到以下标记：*my*、*uncle*、*lives*、*in* 和 *new york*。或者，*uncle* 可以提取为概念并且类型为 <Unknown>，并且 *new york* 可提取为概念并且类型为 <Location>。所有概念均是标记，但是并非所有标记都是概念。标记还可以是其他宏、文字串和字距。仅类型化的这些字或短语才可以是概念。

规则值表。此表包含用于将规则与语句匹配的规则元素。可以使用表右侧的按钮在该表中添加或移除行。该表由 3 列组成:

- **元素**列。将值输入为一个类型、文字串、单词间隙 (<Any Token>) 或宏，或者输入为各项的组合。请参阅主题第 195 页的『规则和宏的受支持元素』以获取更多信息。双击元素单元格以直接输入信息。或者，在单元格中右键单击以显示上下文菜单，其中提供常用的宏、类型名称和非语言类型名称的列表。请记住，如果通过键入来将信息输入到单元格中，请将“$”字符前置于宏或类型名称，例如 $mTopic 表示宏 mTopic。创建元素行所采用的顺序对于规则将如何与文本匹配至关重要。在组合自变量时，必须使用括号 ( ) 将自变量和字符 | 分组以指示布尔 OR。请记住，这些值区分大小写。
- **数量**列。这指示为发生匹配而必须找到元素的最小和最大次数。例如，如果要在 0 到 3 个单词的其他两个元素之间定义间隙或一系列单词，那么可以从列表中选择**介于 0 和 3 之间**，或者直接在对话框中输入数量。缺省值为“**刚好 1 个**”。在某些情况下，将要使元素为可选。如果情况如此，那么其最小数量将为 0，最大数量大于 0（即，0 或 1，介于 0 和 2 之间）。请注意，规则中的第一个元素不能为可选，意味着其数量不能为 0。
- **示例标记**列。如果单击**获取标记**，那么程序会将**示例**文本细分为多个标记，并且使用这些标记将此列填充为与所定义的元素匹配的标记。您也可以选择在输出表中查看这些标记。

规则输出表。此表中的每行定义 TLA 模式输出将如何显示在结果中。规则输出可以产生最多 6 个“概念/类型”列对，每对表示一个*槽*。例如，类型模式 <Location> + <Positive> 是一个两槽模式，意味着它由两个“概念/类型”列对组成。

注: **规则值表**的**元素**列中或**规则输出表**的任何**概念**列中的术语不能以下列任何字符开头: `、#、%、^、*、_、-、:、<、>、/、 \ 或 ″。

如同语言给予我们自由以多种不同方式表达相同的基本构想一样，您可能定义有多个规则来捕获同一基本构想。例如，文本*"Paris is a place I love"*和文本*"I really, really like Paris and Florence"*表示同一基本构想（即，喜欢巴黎），但以不同方式表达，并且将会要求同时捕获两个不同的规则。但是，如果将类似构想分组在一起，那么可更轻松地处理模式结果。为此，虽然您可能具有两个不同规则来捕获这两个短语，但是可以为两个规则均定义相同的输出（例如类型模式 `<Location> + <Positive>`），以便其同时表示两个文本。通过此方式，可以看到输出并非总是模仿在原始文本中找到的单词的结构或顺序。此外，这样的类型模式可与其他短语匹配，并可产生如下概念模式: `paris + like` 和 `tokyo + like`。

要帮助快速定义错误更少的输出，可以使用上下文菜单选择要在输出中显示的元素。或者，也可以将元素从规则值表拖放到输出中。例如，如果具有的规则在规则值表的第 2 行中包含对 `mTopic` 宏的已用，并且希望在输出中显示该值，那么将 `mTopic` 的元素拖放到规则输出表中的第一个列对即可。执行此操作将自动填充已选定的对的"概念"和"类型"。或者，如果希望输出以规则值表的第三个元素（第 3 行）定义的类型开头，请将该类型从规则值表拖放到输出表中的**类型 1** 单元格。该表将更新以在括号 (3) 中显示行引用。

或者，可以通过双击要输出的每个**概念**列中的单元格并输入后跟行号的 `$` 符号（例如 `$2`）将这些引用手动输入到表中，以引用规则值表的第 2 行中定义的元素。手动输入信息时，还需要定义**类型**列，输入后跟行号的 `#` 符号（例如 `#2`），以引用规则值表的第 2 行中定义的元素。

此外，甚至可能会组合方法。假设您在规则值表的第 4 行中具有类型 `<Positive>`。您可以将其拖至类型 2 列，再双击概念 2 列中的单元格，然后在其前面手动输入单词"*not*"。输出列之后会在表中读取 `not (4)`，如果您处于编辑方式或源方式下，那么会读取 `not $4`。然后，可以在"类型 1"列中右键单击并选择例如名为 `mTopic` 的宏。然后，此输出可产生如下概念模式: `car + bad`。

大多数规则仅有一个输出行，但有时可能需要多个输出。在此情况下，请在规则输出表中每行定义一个输出。

要点: 请记住，在 TLA 模式的抽取期间会执行其他语言处理操作。因此，当输出读取 `t$3\t#3` 时，这意味着在应用所有语言处理（同义词和其他分组）后，模式终将显示第三个元素的最终概念和第三个元素的最终类型。

- **输出显示为**。缺省情况下，将会选择选项**对规则值表中的行的引用**，并且通过按照规则值表中的定义使用对行的数字引用来显示输出。如果先前单击"获取标记"并具有规则值表中"示例标记"列中的标记，那么可以通过选择该选项来查看这些特定标记的输出。

注: 如果在输出表中没有显示足够的概念/类型输出对，那么可以通过单击编辑器工具栏中的"添加"按钮来添加其他对。如果当前显示 3 对并单击添加，那么会向表中再添加 2 列（"概念 4"和"类型 4"）。这意味着现在将在所有规则的输出表中看到 4 对。您也可以移除未使用的对，只要此库中的规则集中没有其他规则使用该对即可。

示例规则

假设资源包含以下文本链接分析规则，并且您已启用 TLA 结果的抽取:

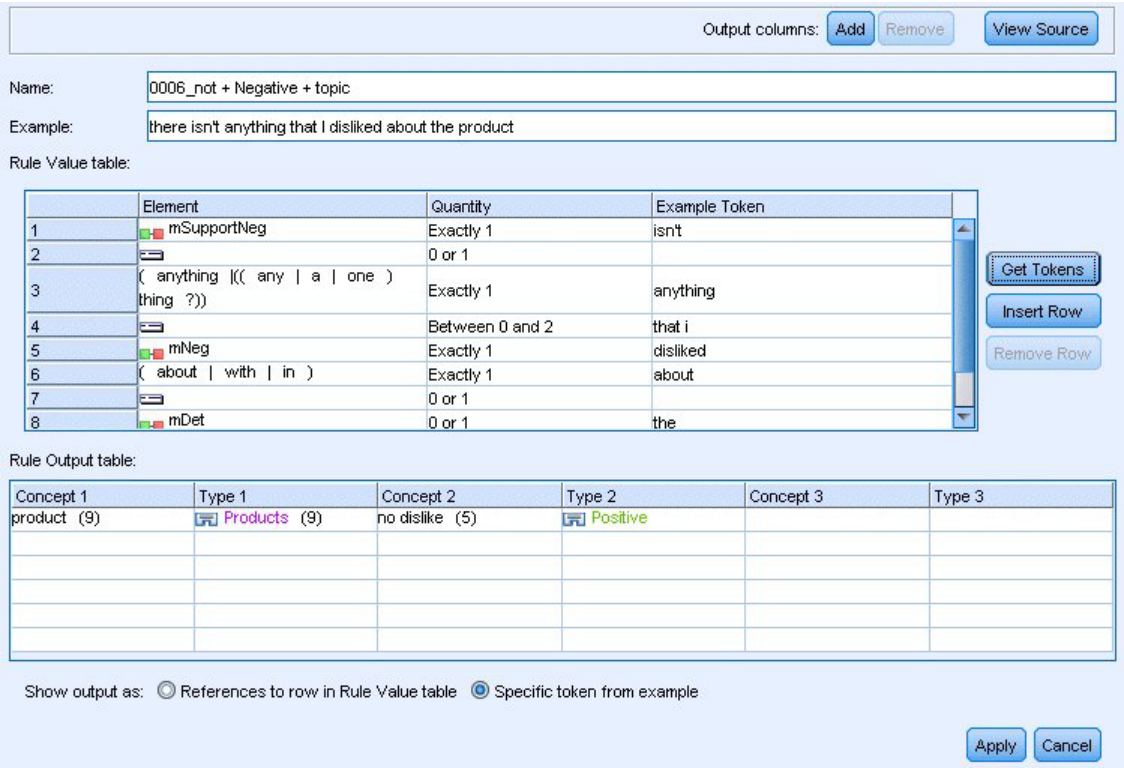


图 43. "文本链接规则"选项卡: 规则编辑器

只要进行抽取，抽取引擎就将读取每个语句，并将尝试匹配以下序列:

表 44. 抽取序列示例

元素（行）	自变量的描述
1	来自宏 mPos 或 mNeg 表示的其中一个类型或来自类型 <Uncertain> 的概念。
2	类型为宏 mTopic 表示的其中一个类型的概念。
3	宏 mBe 表示的其中一个单词。
4	可选元素（0 或 1 个单词），也称为单词间隙或 <Any Token>
5	类型为宏 mTopic 表示的其中一个类型的概念。

输出表显示此规则中所需的全部内容都是模式，其中概念或类型对应于**规则值表**的第 5 行中定义的宏 mTopic + 任何概念或类型都对应于**规则值表**的第 1 行中定义的 mPos、mNeg 或 <Uncertain>。这可以是 sausage + like 或 <Unknown> + <Positive>。

创建和编辑规则

您可以创建新规则或编辑现有规则。请遵循规则编辑器的准则和描述进行操作。请参阅主题第 190 页的『处理文本链接规则』以获取更多信息。

创建新规则

1. 从菜单中选择**工具 > 新建规则**。或者，单击树工具栏中的"新建规则"图标以在编辑器中打开新规则。
2. 输入唯一名称并定义规则值元素。
3. 完成后单击**应用**以检查错误。

编辑规则

1. 单击树中的规则名称。这会在右侧的编辑器窗格中打开该规则。
2. 进行更改。
3. 完成后单击**应用**以检查错误。

禁用和删除规则

禁用规则

如果要在处理期间忽略规则，那么可以禁用该规则。请谨慎删除和禁用规则。

1. 单击树中的规则名称。这会在右侧的编辑器窗格中打开该规则。
2. 右键单击名称。
3. 从上下文菜单中，选择**禁用**。规则图标变为灰色，并且规则本身变为不可编辑。

删除规则

如果要去除规则，那么可以将其删除。请谨慎删除和禁用规则。

1. 单击树中的规则名称。这会在右侧的编辑器窗格中打开该规则。
2. 右键单击名称。
3. 从上下文菜单中，选择**删除**。规则从列表中消失。

检查错误、保存和取消

应用规则更改

如果在规则编辑器外部单击，或者如果单击**应用**，那么会自动扫描规则以查找错误。如果找到错误，那么将需要对其进行修订，然后再继续移至应用程序的其他部分。

但是，如果检测到不太严重的错误，那么仅会发出警告。例如，如果规则包含类型或宏的不完整或未引用的定义，那么会显示警告消息。一旦单击**应用**，任何未更正的警告就会导致在左窗格中的规则名称左侧出现警告图标。

应用规则并不意味着会永久保存规则。应用将导致验证过程检查错误和警告。

在交互式工作台会话内保存资源

1. 要在交互式工作台会话期间保存对资源进行的更改，从而可在下次运行流时获取这些资源，必须执行以下操作:
   - 更新建模节点，以确保可在下次执行流时获取这些相同资源。请参阅主题第 70 页的『更新建模节点并保存』以获取更多信息。然后，保存流。要保存流，请在更新建模节点后在 IBM SPSS Modeler 主窗口中执行此操作。
2. 要在交互式工作台会话期间保存对资源进行的更改，从而可在其他流中使用这些资源，可以执行以下操作:
   - 更新所使用的模板或创建新模板。请参阅主题第 141 页的『创建和更新模板』以获取更多信息。这将不会保存当前节点的更改（请参阅上一步）
   - 或者，更新所使用的 TAP。请参阅主题第 118 页的『更新文本分析包』以获取更多信息。

在模板编辑器内保存资源

1. 首先，发布库。请参阅主题第 158 页的『发布库』以获取更多信息。

2. 然后，通过菜单中的**文件 > 保存资源模板**来保存模板。

取消规则更改

1. 如果希望废弃更改，请单击编辑器窗格中的**取消**。

规则的处理顺序

在抽取期间执行文本链接分析时，“语句”（子句、单词、短语）将顺序与每个规则进行匹配，直至找到匹配项或已用完所有规则为止。树中的顺序规定尝试处理规则的顺序。最佳实践表明，应从最具体到最通用对规则进行排序。 最具体的规则应位于树的顶部。要更改特定规则或规则集的顺序，请从“规则和宏树”上下文菜单中选择**下移**或**上移**，或者选择工具栏中的向上和向下箭头。

如果您*在源视图中*，那么无法通过在编辑器中将规则四处移动来更改其顺序。规则在源视图中的显示位置越高，便会越快对其进行处理。强烈建议仅在树中对规则重新排序，以避免复制/粘贴问题。

注意！在先前版本的 IBM SPSS Modeler Text Analytics 中，要求具有唯一的数字规则标识。从 V18 开始，只能通过在树中将规则上移或下移或者按照规则在源视图中的位置来指示处理顺序。

例如，假设文本包含以下两个语句:

I love anchovies

I love anchovies and green peppers

此外，假设存在两个具有以下值的文本链接分析规则:

A

	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4			
5			
6			
7			

B

	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4	(SEP | and | or)	1 or 2	
5	mDet	0 or 1	
6	mTopic	Exactly 1	
7			

图 44. 两个示例规则

在源视图中，规则值可能如下所示:

A: `value = $Positive $mDet? $mTopic`

B: `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

如果规则 **A** 比规则 **B** 在树中的位置更靠上（更接近顶部），那么将首先处理规则 **A**，语句 *I love anchovies and green peppers* 将首先按 `$Positive $mDet? $mTopic` 进行匹配，并将产生不完整的模式输出 (`anchovies + like`)，因为它是按不是查找两个 `$mTopic` 匹配项的规则进行匹配。

因此，要捕获文本的真正实质，必须将最具体的规则（在本例中为 **B**）放在树中比更通用的规则（在本例中为规则 **A**）更高的位置。

处理规则集（多重通过）

通过规则集，有助于在“规则和宏树”中将相关规则集分组在一起，从而执行多重通过处理。规则集除名称外自身没有定义，并且用于将规则组织到有意义的组中。在某些上下文中，文本过于丰富且多变而无法以一重通过进行处理。例如，处理安全情报数据时，文本可能会包含通过联系人方法（*x 呼叫 y*）、通过家庭关系（*y 是 x 的姐夫/妹夫*）、通过金钱交换（*x 向 y 汇款 $100*）等发现的个人之间的链接。在此情况下，创建专用文本链接分析规则集可有所帮助，其中每个规则集专注于特定关系类型，例如一个用于发现联系人，另一个用于发现家庭成员，依此类推。

要创建规则集，请从“规则和宏树”上下文菜单或从工具栏中选择“创建规则集”。然后，可以直接在树上的“规则集”节点下创建新规则，或者将现有规则移至“规则集”。

使用资源（其中规则会组成规则集）运行抽取时，系统会强制抽取引擎多次通过文本，以便在每次通过时与不同类型的模式匹配。通过此方式，“语句”可与每个规则集中的规则匹配，而在没有规则集的情况下，它只能与单个规则匹配。

注: 每个规则集最多可以添加 512 个规则。

创建新规则集

1. 从菜单中选择**工具 > 新建规则集**。或者，单击树工具栏中的“新建规则集”图标。这会在规则树中显示规则集。
2. 将新规则添加到此规则集，或者将现有规则移至该集中。

禁用规则集

1. 右键单击树中的规则集名称。
2. 从上下文菜单中，选择**禁用**。规则集图标变为灰色，并且在处理期间还会禁用和忽略该规则集内包含的所有规则。

删除规则集

1. 右键单击树中的规则集名称。
2. 从上下文菜单中，选择**删除**。这会从资源中删除规则集及其包含的所有规则。

规则和宏的受支持元素

针对文本链接分析规则和宏，接受以下自变量:

宏

可以直接在文本链接分析规则中使用宏，或在另一个宏内使用宏。如果是手动或从源视图内输入宏名称（相对于从上下文菜单中选择宏名称），请确保名称以美元符号字符 ($) 为前缀，例如 `$mTopic`。宏名称区分大小写。通过上下文菜单选择宏时，可以从当前“文本链接规则”选项卡中定义的任何宏进行选择。

类型

可以直接在文本链接分析规则或宏中使用类型。如果是手动或在源视图中输入类型名称（相对于从上下文菜单中选择类型），请确保类型名称以美元符号字符 ($) 为前缀，例如 $Person。类型名称区分大小写。如果使用上下文菜单，那么可以从使用的当前资源集中的任何类型进行选择。

如果引用无法识别的类型，那么将收到警告消息，并且规则在“规则和宏树”中将具有警告图标，直至更正为止。

文字串

要包含从未抽取的信息，那么可以定义抽取引擎将搜索的文字串。所有已抽取的单词或短语都已分配到相应的类型，因此，不能在文字串中使用。如果使用已抽取的单词，那么会将其忽略，即使其类型为 <Unknown> 也如此。

文字串可以是一个或多个单词。定义文字串列表时，以下规则适用:

- 将字符串列表用括号括起来，例如 (his)。如果存在文字串选项，那么每个字符串必须以 OR 运算符分隔，例如 (a|an|the) 或 (his|hers|its)。
- 使用单个单词或复合词。
- 列表中的每个单词以 | 字符（类似于布尔 OR）分隔。
- 如果要与单数和复数形式均匹配，请同时输入两者。不会自动生成词形变化。
- 仅使用小写。
- 要复用文字串，请将其定义为宏，然后在其他宏和文本链接分析规则中使用该宏。
- 如果字符串包含句点（句号）或连字符，那么必须将其包含在内。例如，要与文本中的 a.k.a 匹配，请输入句点以及字母 a.k.a 作为文字串。

排除运算符

使用 ! 作为排除运算符，以阻止任何求反表达式占用特定槽。只能通过内嵌单元格编辑（双击规则值表或宏值表中的单元格）或在源视图中手动添加排除运算符。例如，如果将 $mTopic @{0,2} !($Positive) $Budget 添加到文本链接分析规则中，那么表明查找的文本中包含 (1) 分配给 mTopic 宏中的任何类型的术语，(2) 零个或两个单词长度的单词间隙，(3) 分配给 <Positive> 类型的术语的实例不存在以及 (4) 分配给 <Budget> 类型的术语。这可能会捕获“*cars have an inflated price tag*”，但会忽略“*store offers amazing discounts*”。

要使用此运算符，必须通过双击单元格在元素单元格中手动输入惊叹号和括号。

单词间隙 (<Any Token>)

单词间隙（也称为 <Any Token>）定义两个元素之间可能存在的标记的数字范围。在匹配由于存在附加的限定词、介词短语、形容词或其他此类单词而可能仅略有不同的非常类似的短语时，单词间隙非常有用。

表 45. 规则值表中无单词间隙的元素的示例

#	元素
1	 Unknown
2	 mBeHave

表 45. 规则值表中无单词间隙的元素的示例 (续)

3	 Positive

注: 在源视图中，该值定义为: `$Unknown $mBeHave $Positive`

该值将匹配类似于"*the hotel staff was nice*"的语句，其中 *hotel staff* 属于类型 `<Unknown>`，*was* 位于宏 `mBeHave` 下，并且 *nice* 为 `<Positive>`。但是，它将不与"*the hotel staff was very nice*"匹配。

表 46. 规则值表中具有 <Any Token> 单词间隙的元素的示例

#	元素
1	 Unknown
2	 mBeHave
3	
4	 Positive

注: 在源视图中，该值定义为: `$Unknown $mBeHave @{0,1} $Positive`

如果向规则值添加单词间隙，那么它将与"*the hotel staff was nice*"和"*the hotel staff wasvery nice*"均匹配。

在源视图中或对内嵌编辑而言，单词间隙的语法为 `@{#,#}`，其中 `@` 表示单词间隙，而 `{#,#}` 定义前置元素和尾随元素之间接受的最小和最大单词数。例如，`@{1,3}` 意味着如果存在至少一个单词但在两个已定义元素之间出现不超过三个单词，那么可在这两个元素之间进行匹配。`@{0,3}` 意味着如果存在 0、1、2 或 3 个单词但不超过三个单词，那么可在两个已定义元素之间进行匹配。

在源方式下查看和工作

对于每个规则和宏，TLA 编辑器会生成供抽取器用于匹配和产生 TLA 输出的底层源代码。如果首选处理代码本身，那么可以直接通过单击编辑器顶部的"查看源代码"按钮来查看此源代码和对其进行编辑。源视图将跳至当前选定的规则或宏并将其突出显示。但是，建议使用编辑器窗格来减少错误几率。

查看或编辑源代码完成后，单击**退出源代码**。如果为规则生成无效语法，那么将需要对其进行修订，然后再退出源视图。

要点: 如果在源视图中进行编辑，那么强烈建议逐个编辑规则和宏。编辑宏之后，请通过抽取来验证结果。如果对结果满意，那么建议先保存模板，然后再进行其他更改。如果对结果不满意或发生错误，请还原为已保存的资源。

源视图中的宏

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

表 47. 宏条目

[macro]	每个宏必须以标记为 [macro] 的行开头以表示宏开始。
name	宏定义的名称。每个名称必须唯一。
value	一个或多个类型、文字串、单词间隙或宏的组合。请参阅主题第 195 页的『规则和宏的受支持元素』以获取更多信息。在组合自变量时，必须使用括号 ( ) 将自变量和字符 | 分组以指示布尔 OR。

除有关“宏”的部分中涵盖的准则和语法以外，源视图还具有在编辑器视图中工作时不需要的少数其他准则。在源方式下工作时，宏还必须遵循以下准则:

- 每个宏必须以标记为 [macro] 的行开头以表示宏开始。
- 要禁用某个元素，请在每行前面放置注释指示符 (#)。

示例。此示例定义一个名为 mTopic 的宏。mTopic 的值表明存在与以下类型之一匹配的术语: <Product>、<Person>、<Location>、<Organization>、<Budget> 或 <Unknown>。

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

源视图中的规则

```
[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 48. 规则条目

[pattern (<ID>)]	指示文本链接分析规则开始，并提供用于确定处理顺序的唯一数字标识。
name	为此文本链接分析规则提供唯一名称。
value	提供要与文本匹配的语法和自变量。请参阅主题第 195 页的『规则和宏的受支持元素』以获取更多信息。
output	在文本中发现的所生成的匹配模式的输出格式。输出并不总是与源文本中元素的确切原始位置相似。此外，通过将每个输出放在单独的行上，针对给定的文本链接分析规则可具有多个输出行。 输出的语法: • 使用跳格代码 \t 来分隔输出，例如 $1\t#1\t$3\t#3 • $ 和数字需要所找到的术语与该位置中 value 参数中定义的自变量匹配。因此 $1 意味着该术语与为 value 定义的第一个自变量匹配。 • # 和数字需要该位置中的元素的类型名称。如果项目是文字串列表，那么将分配类型 <Unknown>。 • 值为 Null\tNull 将不创建任何输出。

除有关“规则”的部分中涵盖的准则和语法以外，源视图还具有在编辑器视图中工作时不需要的少数其他准则。在源方式下工作时，规则还必须遵循以下准则:

- 只要定义了两个或多个元素，无论其是否可选，都必须用括号将其括起来（例如，($Negative|$Positive) 或 ($mCoord|$SEP)?）。$SEP 表示逗号。

- 文本链接分析规则中的第一个元素不能是可选元素。例如，不能以 value = $mTopic? 或 value = @{0,1} 开头。
- 可以将数量（或实例计数）与标记关联。这在仅编写一个包含所有案例的规则而不是为每个案例编写单独的规则时非常有用。例如，如果尝试匹配 ,（逗号）或 and，那么可以使用文字串 ($SEP|and)。如果通过添加数量对此进行扩展，以便文字串变为 ($SEP|and){1,2}，那么现在将与以下任何实例匹配: “,”、“and”、“, and”。
- 在文本链接分析规则 value 中的宏名称与 $ 和 ? 字符之间不支持空格。
- 在文本链接分析规则 output 中不支持空格。
- 要禁用某个元素，请在每行前面放置注释指示符 (#)。

示例。假设资源包含以下 TLA 文本链接分析规则，并且您已启用 TLA 结果的抽取:

```
## 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
```

只要进行抽取，抽取引擎就将读取每个语句，并将尝试匹配以下序列:

表 49. 抽取序列示例

位置	自变量的描述
1	人员的姓名 ($Person)
2	以下一项或两项: 逗号 ($SEP)、定界符 ($mDet)、助动词 ($mSupport)、字符串“then”或“as”
3	0 或 1 个单词 (@{0,1})
4	职能 ($Function)
5	以下字符串之一: “of”、“with”、“for”、“in”、“to”或“at”
6	0 或 1 个单词 (@{0,1})
7	组织的名称 ($Organization)
8	0、1 或 2 个单词 (@{0,2})
9	位置的名称 ($Location)

此样本文本链接分析规则将与如下语句或短语匹配:

Jean Doe, the HR director of IBM in France

Jean Doe was the former HR director of IBM in France

IBM appointed Jean Doe as the HR director of IBM in France

此样本文本链接分析规则将产生以下输出:

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

其中:

- jean doe 是对应于 $1（文本链接分析规则中的第 1 个元素）的术语，<Person> 是 jean doe (#1) 的类型，
- hr director 是对应于 $4（文本链接分析规则中的第 4 个元素）的术语，<Function> 是 hr director (#4) 的类型，
- ibm 是对应于 $7（文本链接分析规则中的第 7 个元素）的术语，<Organization> 是 ibm (#7) 的类型，

- france 是对应于 $9（文本链接分析规则中的第 9 个元素）的术语，<Location> 是 france (#9) 的类型

源视图中的规则集

```
[set(<ID>)]
```

其中 [set (<ID>)] 指示文本集的开始，并提供用于确定集的处理顺序的唯一数字标识。

示例。以下语句包含有关个人、其在公司内的职能以及该公司的合并/收购活动的信息。

```
Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO of Org2
Ltd.
```

您可以编写具有一个具有若干输出的规则来处理所有可能的输出，例如:

```
## 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
```

它将产生以下两个输出模式:

- org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd<Organization>
- john doe <Person> + ceo <Function> + org2 ltd<Organization>

注意！请记住，在 TLA 模式的抽取期间会执行其他语言处理操作。在此情况下，在抽取过程的同义词分组阶段中，merger 在 merges with 下分组。并且，由于 merges with 属于 <ActiveVerb> 类型，此类型名称会显示在最终 TLA 模式输出中。因此，当输出读取 t$3\t#3 时，这意味着在应用所有语言处理（同义词和其他分组）后，模式终将显示第三个元素的最终概念和第三个元素的最终类型。

可以轻松管理和处理两个规则，而不是如同先前编写复杂规则。第一个规则专用于了解公司之间的合并/收购:

```
[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
```

它将产生 org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd <Organization>

第二个规则专用于个人/职能/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
```

它将产生 john doe <Person> + ceo <Function> + org2 ltd <Organization>

声明

本信息是为在全球供应的产品和服务而编写的。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您所在区域当前可获得的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。但是，评估和验证任何非 IBM 产品、程序或服务的操作，由用户自行负责。

IBM 可能已拥有或正在申请与本文档内容有关的各项专利。提供本文档并不意味着授予用户使用这些专利的任何许可。您可以用书面形式将许可查询寄往:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某特定用途的保证。某些管辖区域在某些交易中不允许免除明示或暗含的保证。 因此本条款可能不适用于您。

本信息可能包含技术方面不够准确的地方或印刷错误。本信息将定期更改；这些更改将编入本信息的新版本中。IBM 可以随时对本出版物中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的: (i) 使其能够在独立创建的程序和其它程序（包括本程序）之间进行信息交换，以及 (ii) 使其能够对已经交换的信息进行相互使用，请与下列地址联系:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

只要遵守适当的条件和条款，包括某些情形下的一定数量的付费， 都可获得这方面的信息。

本文档中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际程序许可协议或任何同等协议中的条款提供。

提供的性能数据和引用的客户机示例，仅供参考。根据特定配置和操作条件的不同，实际性能结果也可能不同。

涉及非 IBM 产品的信息可从这些产品的供应商、其出版说明或其他可公开获得的资料中获取。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问题应当向这些产品的供应商提出。

所有关于 IBM 未来方向或意向的声明都可随时更改或收回，而不另行通知，它们仅仅表示了目标和意愿而已。

本信息包含在日常业务操作中使用的数据和报告的示例。为了尽可能完整地说明这些示例，示例中可能会包括个人、公司、品牌和产品的名称。所有这些名字都是虚构的，若现实生活中实际人员或业务企业使用的名字与此相似，纯属巧合。

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。当前的 IBM 商标列表，可从 Web 站点 www.ibm.com/legal/copytrade.shtml 上的“版权和商标信息”获取。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、 Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国和@3B72其他国家或地区的注册商标。

Microsoft、Windows、Windows NT 以及 Windows 徽标是 Microsoft Corporation 在美国和/或其他国家或地区的商标。

UNIX 是 The Open Group 在美国和／或其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 和/或其子公司的商标或注册商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。

索引

[G]

[H]

[J]

[K]

[X]

[Y]

[Z]

A

F

H

I

M

Printed in China