SAS® Contextual Analysis 14.2: ユーザーガイド

SAS® ドキュメント
2017/08/14

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS® Contextual Analysis 14.2: ユーザーガイド*. Cary, NC: SAS Institute Inc.

SAS® Contextual Analysis 14.2: ユーザーガイド

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

April 2017

14.2-P2:utaqsug

目次

本書の利用について

利用者

本書は、SAS Contextual Analysis のユーザー用です。ここでは、SAS Contextual Analysis で使用される用語を記載し、タスクについて説明します。

SAS Contextual Analysis では、現在、英語を含む 14 言語を処理できます。このドキュメントは特定の言語向けに作成されたものではありませんが、例のテキストでは英語が使用されています。ほとんどのサポート言語で事前定義済みコンセプトのサブセットが提供されています。完全な言語リストについては、“サポート言語” (3 ページ)を参照してください。

新機能

SAS Contextual Analysis 14.2 の新機能

概要

SAS Contextual Analysis 14.2 には、次の新機能と拡張機能が含まれています。

- 追加の言語サポート。
- 圧縮データセットが作成されるようになりました。
- コンセプトスコアコードに基準形(完全形)が含まれるようになりました。
- カテゴリスコアコードにサブカテゴリを削除するオプションが含まれるようになりました。
- 事前義済みコンセプトの優先順位値をリストにするドキュメント拡張。

詳細

言語サポートとドキュメント拡張

SAS Contextual Analysis では、現在、スウェーデン語がサポートされています。すべてのサポート言語のリストについては、“サポート言語” (3 ページ)を参照してください。

現在、このドキュメントに、すべてのサポート言語の事前定義済みコンセプトの値が含まれています。詳細については、“コンセプト” (30 ページ) および 付録 2, “事前定義済みコンセプト(英語以外の言語用)” (113 ページ)を参照してください。

スコアコードの拡張

CLASSIFIER コンセプトのルールの種類に、使用可能な場合、一致文字列の基準(完全)形を返すオプションが含まれるようになりました。詳細については、表 3.2 (58 ページ)を参照してください。

カテゴリスコアコードの新しいオプションを使用すると、自動的に生成されたルールから作成されたサブカテゴリを削除できます。詳細については、“コードの表示とダウンロード” (22 ページ)を参照してください。

データファイルの圧縮

SAS Contextual Analysis プロジェクトライブラリに作成されるデータテーブルは、圧縮形式で書き込まれるようになりました。

アクセシビリティ

この製品のアクセシビリティの詳細については、support.sas.com の Accessibility Features of SAS Contextual Analysis 14.2 を参照してください。

x 新機能

1

SAS Contextual Analysis の概要

SAS Contextual Analysis について

SAS Contextual Analysis は、コンテキスト分析によって、キーテキストデータの識別とカテゴリ分けという課題に対する総合的なソリューションを提供する、Web ベースのテキスト分析アプリケーションです。このアプリケーションを使用すると、(学習ドキュメントに基づいて)ドキュメントセットを自動的に分析してカテゴリ分けするモデルを作成できます。その後で、テキストベースデータの価値を認識するために、モデルをカスタマイズできます。

SAS Contextual Analysis では、SAS Text Miner のマシン学習機能と、SAS Enterprise Content Categorization におけるカテゴリ分けおよび抽出のルールベースの言語学的方法が統合されています。これらの機能は、ドキュメントレベルのセンチメントスコアリングとともに、単一ユーザーインタフェイスに統合されています。

SAS Contextual Analysis を使用すると、ドキュメントコレクションのキーテキストデータの識別、そのデータのカテゴリ分け、コンセプトモデルの作成、無意味なテキストデータの削除を行えます。

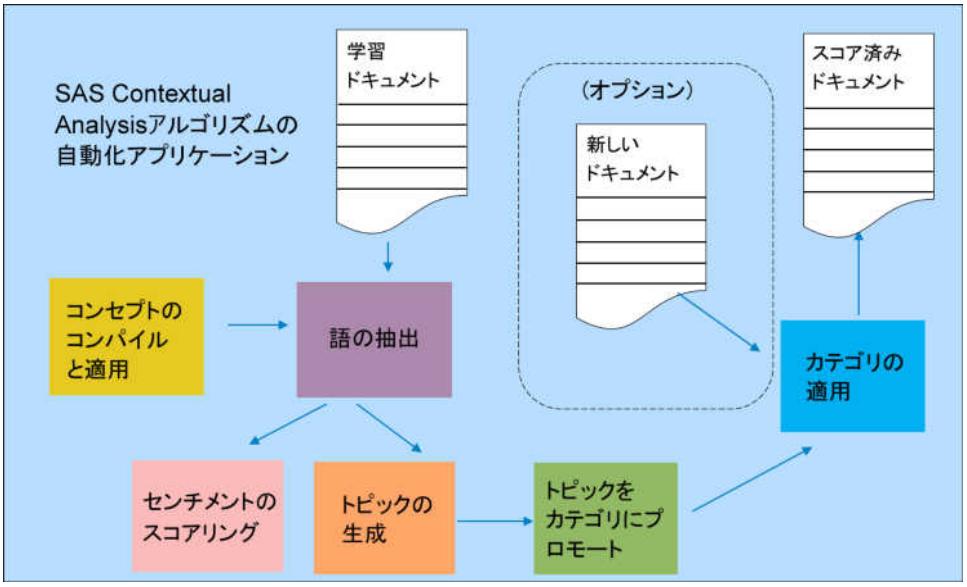
デフォルトでは、価値がほとんどないかまったくない単語は分析から除外されます。そのような単語の例としては、冠詞 *a*、*an*、*the* や、*and*、*or*、*but* などの接続詞があります。各自のドキュメントコレクションに特有で、価値がほとんどないかまったくないその他の語も、識別されて、除外されます。

SAS Contextual Analysis は、SAS プログラミングや SAS マクロ言語の経験がないユーザー用に設計されています。

SAS Contextual Analysis の動作

図 1.1 に SAS Contextual Analysis プロセスの概要を述べます。

図 1.1　プロセス概要



SAS Contextual Analysis を使用すると、事前定義済みコンセプトを抽出したり、1 ドキュメントやドキュメントセットで検出する追加カスタムコンセプトを作成したりできます。コンセプトの詳細については、“コンセプト” (30 ページ)を参照してください。

SAS Contextual Analysis アルゴリズムによって、コレクション内の類似ドキュメントがトピックにグループ分けされます。各トピックのドキュメントには、しばしば、オートバイ事故、コンピュータグラフィックス、天候パターンなどの類似する主題が含まれます。自動トピック識別を使用すると、コレクション内の各ドキュメントを容易にカテゴリ分けできます。

次の方法を使用してカテゴリを作成できます。

- SAS Enterprise Content Categorization からカテゴリをインポート
- 学習ドキュメントでカテゴリ変数を指定
- 新しいカテゴリを作成
- トピックをカテゴリにプロモート

トピックをカテゴリにプロモートしたり、学習ドキュメントでカテゴリ変数を指定したりすると、予備ルールが生成されます。そのルールには、編集や改良を行えます。

語や主題の抽出に使用可能な自動化プロセスおよびルールを使用するにせよ、プロセスやルールをカスタマイズするにせよ、コンテキスト感度はモデルの必須コンポーネントです。コンテキスト感度を高めるには、予備ルールを変更します。ブール演算子、文字、およびルール一致のコンテキスト感度を高めるその他の選択を追加または変更できます。

最後に、モデルを配置して、入力ドキュメントのセットを分類するプロセスを自動化します。

サポート言語

表 1.1 に、完全なサポート言語リストが示されています。追加言語のライセンスの詳細については、SAS の営業担当者に問い合わせてください。

表 1.1 SAS Contextual Analysis 14.2 のサポート言語

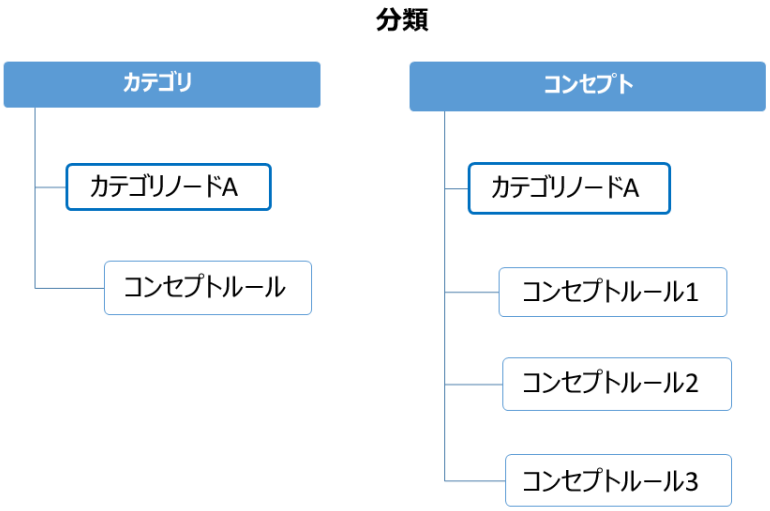
中国語	オランダ語
英語	フィンランド語

フランス語	ドイツ語
イタリア語	日本語
韓国語	ポルトガル語
ロシア語	スペイン語
スウェーデン語	トルコ語

分類の使用

SAS Contextual Analysis では、カテゴリおよびコンセプトルールを作成できます。これは分類構造で表示されます。各分類は、*ノード*のツリーで構成されます。各ノードがルールのコンテナになります。それとは対照的に、コンセプトノードの下に、複数のルールが存在することもあります。 図 1.2 (5 ページ) では、カテゴリとコンセプトの分類の違いを示します。

図 1.2 SAS Contextual Analysis の分類



インターフェイスの使用

ユーザーインターフェイスのメインコンポーネントを 図 1.3 に示します。

図 1.3　SAS Contextual Analysis インターフェイス

1　アプリケーションメニュー

2　サインアウト

3　アプリケーションツールバー

4　プロジェクトリスト

5　検索オプション

6　表示オプション

7　進捗パネル(クリックして開く)

2

SAS Contextual Analysis のプロジェクト

プロジェクトの概要

SAS Contextual Analysis では、作成したプロジェクトが、基本的にデータと分析のコンテナになります。プロジェクトには、入力データ、テキストマイニングオプション、および分析タスク(コンセプト、語、トピック、カテゴリの処理)が含まれます。SAS Contextual Analysis は、複数のプロジェクトを同時に作成して実行できるように設計されています。あるプロジェクトの分析の実行中に別のプロジェクトを開くことができるように、テキスト分析はバックグラウンドで実行されます。プロジェクトは、ユーザー間で共有し、共同で更新できます。

モデルを作成する場合、学習データセットとして使用するドキュメントコレクションを含む入力データを選択します。学習データが、このモデルの適用先となるデータの代表的なデータであるか確認することが重要です。トピックとカテゴリは、このドキュメントコレクションの語に基づいて構築されます。

次に、開始リストと停止リストのどちらを指定するかを選択します。また、類義語リストを使用するかどうかも指定できます。SAS データセットでプロジェクトを実行する前に、分析するテキストフィールドを指定する必要があります。また、分析のためにカテゴリ変数を 1 つ以上指定することもできます。

プロジェクト実行後に、語と、初期テキストマイニング中に作成された自動検出トピックを表示できます。この後、トピックを使用してカテゴリを作成します。カテゴリは、類似語を含むドキュメントのグループです。SAS Contextual Analysis では、カテゴリごとにルールのセットが作成されます。

ドキュメントコレクションの準備

SAS Contextual Analysis でプロジェクトを作成する前に、分析用にドキュメントコレクションを準備する必要があります。SAS Contextual Analysis を使用すると、SAS データセットとして保存されたか、MS Office、OpenDocument (OpenOffice)、PDF、XML、HTML などのテキストベースファイル形式で保存されたドキュメントコレクションを分析できます。SAS データセットを選択すると、分析するテキスト変数とカテゴリ変数を特定できます。または、学習データとして使用するファイルを含むディレクトリを指定できます。

入力ドキュメントコレクションを準備する場合、後でカテゴリ分けするドキュメントの代表となるドキュメントセットを選択する必要があります。入力ドキュメントコレクションに存在する語を使用して、トピックとカテゴリが作成されます。

入力ドキュメントコレクションを作成するための標準ルールはありません。ただし、次のガイドラインが、入力ドキュメントコレクションの準備に役立ちます。

- 検出するカテゴリごとに少なくとも 15 から 20 のドキュメントを含めます。
- 語の検出とルールの作成を予想するために、ドキュメントのコンテンツに精通しておきます。
- Microsoft Word や Adobe PDF のドキュメントコレクションの保存場所と同じディレクトリに SAS データセットを保存しないでください。

注: SAS データセットを使用する場合、SAS Contextual Analysis で使用可能にする前に、そのデータセットを SAS Metadata Server で登録する必要があります。SAS 管理コンソールと SAS Enterprise Guide を使用すると、データセットを登録できます。(SAS データセットではなく)フォルダ内のドキュメントのコレクションを使用する場合、SAS Contextual Analysis Workspace Server がインストールされたサーバー上にフォルダを位置付ける必要があります。詳細については、*SAS Contextual Analysis: Administrator's Guide* を参照してください。

プロジェクトのインポート

既存の SAS Contextual Analysis プロジェクトモデルのインポート

既存のカテゴリおよびコンセプトルールを再利用できるように、プロジェクトの作成中に、SAS Contextual Analysis プロジェクトをインポートできます。インポートされたプロジェクトに含まれるのはカテゴリおよびコンセプトルールのみです。トピック、語、データ、プロジェクト設定などの他のプロジェクトコンポーネントはインポートされません。インポートするファイルは JSON (JavaScript Object Notation)形式になります。

SAS Contextual Analysis プロジェクトのエクスポートの詳細については、“プロジェクトモデルのエクスポート” (23 ページ)を参照してください。

既存の SAS Enterprise Content Categorization プロジェクトのインポート

プロジェクトの作成中に、分析用に既存の SAS Enterprise Content Categorization プロジェクトをインポートできます。プロジェクトをインポートしようとする場合、次の点に注意してください。

- インポートされた SAS Enterprise Content Categorization プロジェクトに含まれる、LITI (言語解釈およびテキスト解釈)構文を使用して定義されたコンセプトは、SAS Contextual Analysis プロジェクトで使用できます。
- インポートされた SAS Enterprise Content Categorization プロジェクトに含まれる、ブールルール(MCAT 構文)を使用して定義されたカテゴリは、SAS Contextual Analysis プロジェクトで使用できます。
- SAS Enterprise Content Categorization に含まれる、言語ルールを使用して作成されたカテゴリは、サポートされません。

注: SAS Contextual Analysis で LITI コンセプトを正しく解析するためには、無効コンセプトの解析優先順位を遵守する必要があります。これを確実に行うには、SAS Enterprise Content Categorization で既存プロジェクトを開きます。無効化された子コ

ンセプトについてはいずれも、その親コンセプトを変更して、親の優先順位が子よりも高くなるようにします。プロジェクトを保存してから、SAS Contextual Analysis にインポートします。

プロジェクトのインポート時の特記事項

SAS Contextual Analysis または SAS Enterprise Content Categorization プロジェクトから新しいプロジェクトにカテゴリおよびコンセプトをインポートする場合、重複名が発生することがあるので注意してください。インポートプロセス中に重複するカテゴリまたはコンセプト名が発生した場合に考慮する事項をいくつか次に示します。

- メッセージを確認して、SAS Contextual Analysis での重複名の処理方法を調べます。メッセージを参照するには、ツールバーの  をクリックします。
- プロジェクトで事前定義済みコンセプトを使用する場合、発生し得る重複名に備えます。たとえば、インポートする LITI コンセプトが、現在のプロジェクトの事前定義済みコンセプトと同じ名前だとします。その場合、インポートされた LITI コンセプトのルールが、事前定義済みコンセプトのルールに追加されます。
- 既存プロジェクトからコンセプトをインポートする場合、名前の競合の解決時に、コンセプトのソースによって、どのコンセプト名が優先されるかが決定されます。現在のプロジェクトの事前定義済みコンセプトが最優先され、その後に SAS Contextual Analysis からインポートされたコンセプト、最後に SAS Enterprise Content Categorization からインポートされたコンセプトが続きます。
- 既存プロジェクトからカテゴリをインポートする場合、名前の競合の解決時に、カテゴリのソースによって、どのカテゴリ名が優先されるかが決定されます。SAS Contextual Analysis からインポートされたカテゴリが最優先され、その後に現在のプロジェクトの外部カテゴリ、最後に SAS Enterprise Content Categorization からインポートされたコンセプトが続きます。

プロジェクトの新規作成

プロジェクト作成ウィザードの使用

初めて SAS Contextual Analysis にログオンする際には、まず第一に、プロジェクトを作成する必要があります。プロジェクトを新規作成するには、メインウィンドウの左上隅近くの  アイコンをクリックします。**プロジェクトの新規作成**ウィザードが表示され、ここでプロジェクトの詳細をすべて入力できます。

ヒント 特定のフィールドやページの詳細については、**プロジェクトの新規作成**ウィザードのヘルプアイコン  をクリックします。

ステップ 1: プロジェクトファイル、サーバー、およびその他のプロパティの識別

1 プロジェクト名を入力し、SAS メタデータでプロジェクトフォルダにアクセス可能な場所を指定します。

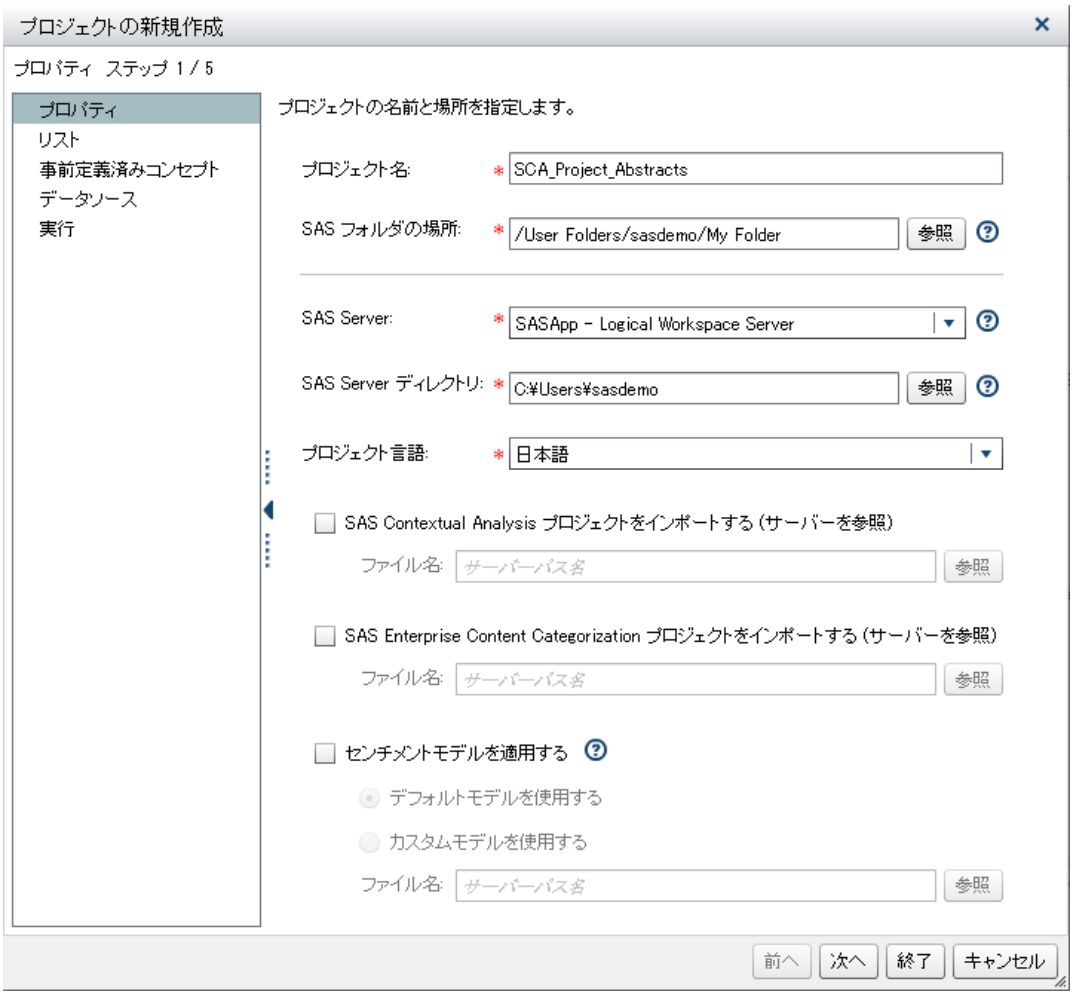
2 SAS Server と SAS Server ディレクトリを識別します。

3 プロジェクトデータの言語を選択します。これがデータファイルで処理される言語です。

4 (オプション) SAS Contextual Analysis プロジェクトをインポートします。詳細については、“既存の SAS Contextual Analysis プロジェクトモデルのインポート” (10 ページ) を参照してください。

5 (オプション) SAS Enterprise Content Categorization プロジェクトをインポートします。詳細については、“既存の SAS Enterprise Content Categorization プロジェクトのインポート” (10 ページ)を参照してください。

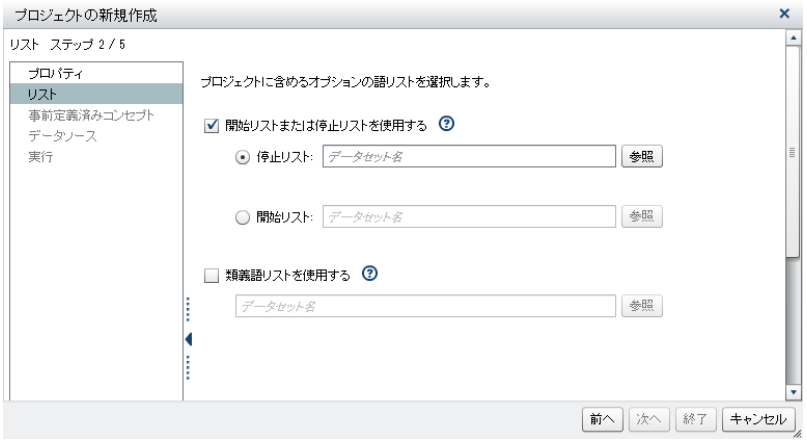
6 (オプション)センチメントモデルを適用します。詳細については、 “SAS Contextual Analysis での SAS Sentiment Analysis モデルの使用” (28 ページ)を参照してください。



ステップ 2: 語と類義語のリストの指定

1 (オプション)開始リストまたは停止リスト(どちらか一方のみ)を指定して、テキストマイニング分析中にどの語を含めたり除外したりするのかを制御します。詳細については、“開始リストと停止リスト” (34 ページ)を参照してください。デフォルトでは、デフォルト停止リストが選択されます。

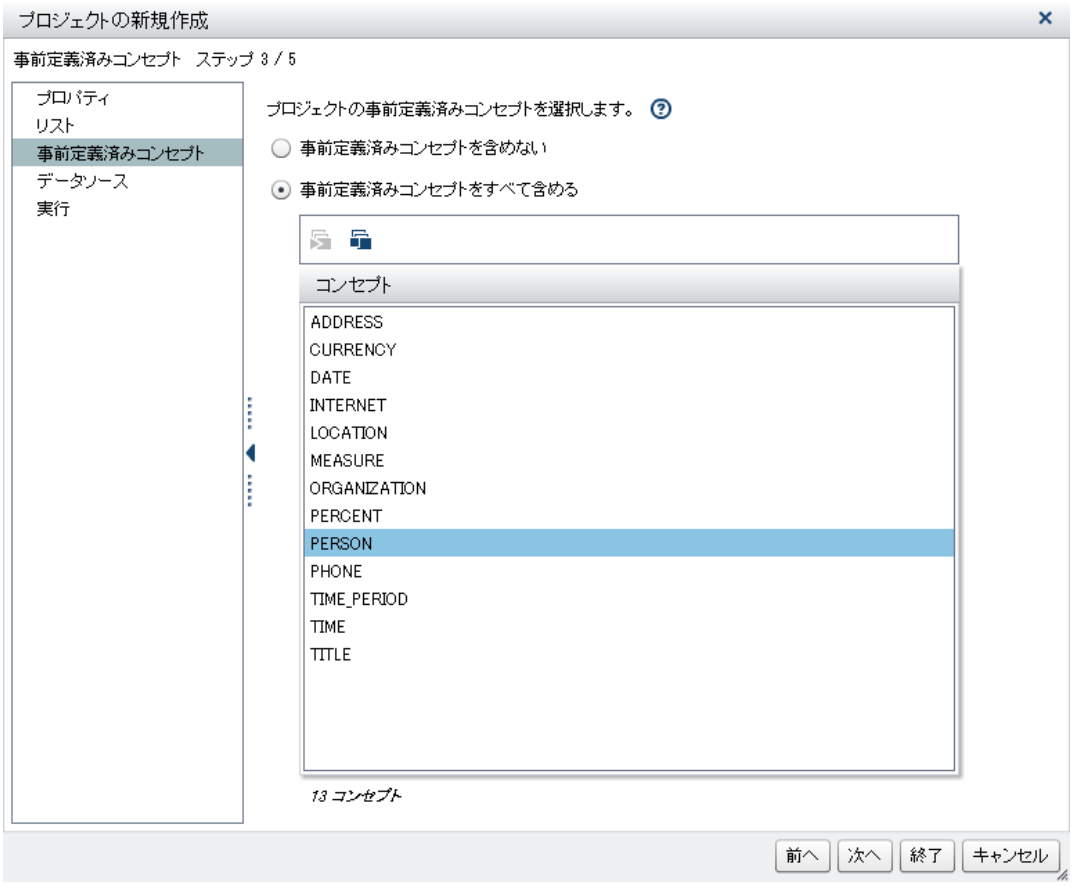
2 (オプション)類義語リストを指定して、分析のために 1 語として扱う必要がある単語のペアを識別します。詳細については、“語と類義語” (33 ページ)を参照してください。



ステップ 3: 事前定義済みコンセプトの選択

SAS Contextual Analysis では、*事前定義済みコンセプト*が提供されます。これは、ルールをすでに作成済みのコンセプトです。事前定義済みコンセプトでは、**COMPANY** や **TITLE** など、よく使用するコンセプトとその定義を提供することによって、時間を節約します。これは含めるか含めないかを選択できます。事前定義済みコンセプトを含めない場合、後から追加することはできません。事前定義済みコンセプトを含める場合、1 つ以上の事前定義済みコンセプトを選択して をクリックすると、1 つ以上の事前定義済みコンセプトを無効化できます。無効化されたコンセプトは、データ処理中は無視されます。コンセプトを選択して をクリックすると、任意の事前定義済みコンセプトを再有効化できます。

事前定義済みコンセプトの詳細については、“コンセプト” (30 ページ)を参照してください。



ステップ 4: データソースの識別

データソース選択のオプションを次に示します。

- すぐにでも後からでもデータソースを選択できます。後でデータソースを選択する場合でも、**プロジェクトの編集**ウィザードでプロジェクトの詳細情報を入力できます。
- SAS データセット内から分析変数を選択できます。このオプションを選択した場合は、データセットライブラリと名前を指定して、分析するテキスト変数を指定する必要があります。SAS 変数にドキュメントのフルテキストを含めるかわりに、ファイルの場所を識別するファイル参照を入力できます。ファイル参照の使用は、長さが 32,767 文字を超えるドキュメントを分析する唯一の方法です。

注: 参照ファイルはプレーンテキスト形式(TXT)である必要があります。

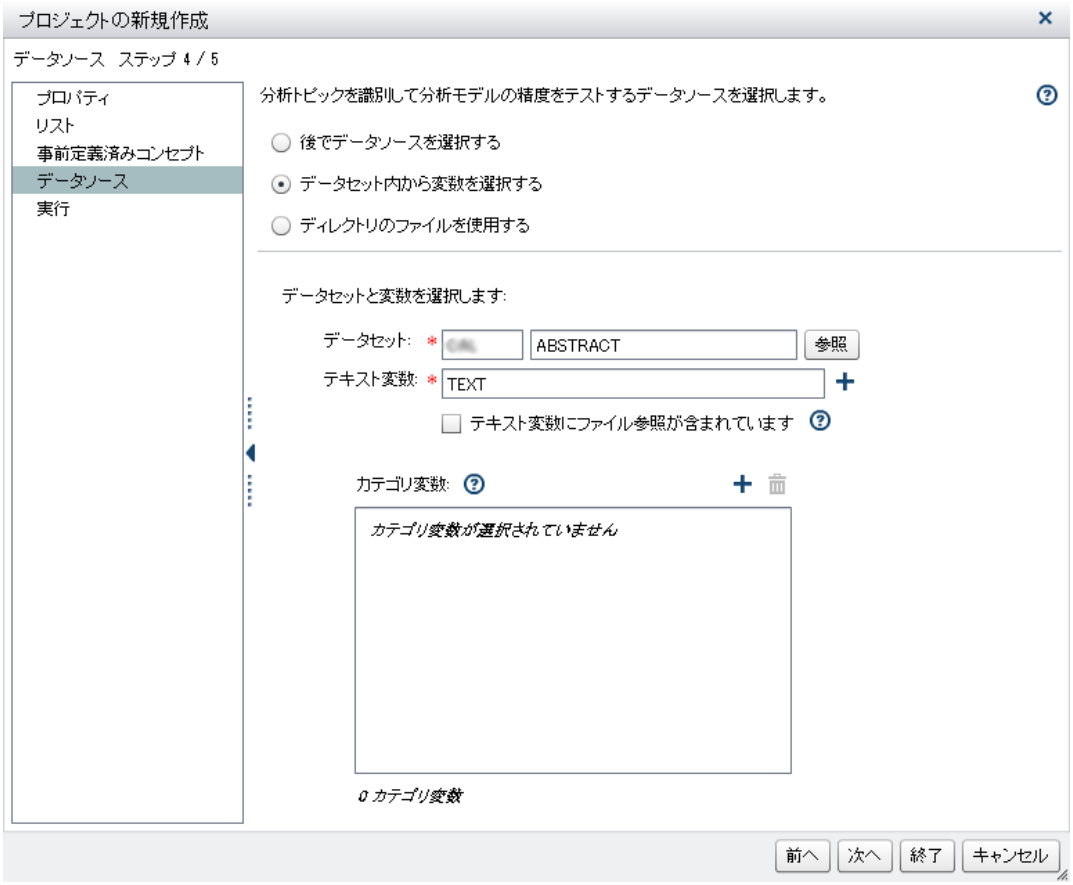
さらに、1 つ以上のカテゴリ変数を指定して、ドキュメントのグループ分けの方法を示すこともできます。たとえば、ホテルの宿泊についての顧客コメントを分析するとします。データ列 Hotels には、顧客が宿泊したホテルの名前が含まれます。カテゴリ変数として Hotels を指定した場合、カテゴリルールが自動的に生成されます。データに頻繁に出現する各ホテルに対しては、サブカテゴリルールも生成されます。

注: SAS Contextual Analysis では、アンダースコア(_)で始まる変数名が予約されています。したがって、アンダースコア(_)で始まる変数名を含むデータセットを選択した場合、エラーが発生する可能性があります。エラーが発生した場合は、データセットの変数名を変更してから再試行してください。

- MS Office、OpenDocument (OpenOffice)、PDF、XML、HTML などのテキストベースファイル形式で保存されたドキュメントコレクションを指定できます。ファイルは、フォルダ内に存在する必要があります。カテゴリは後から定義できます。

  注: コンテンツのないファイルはインポートされません。無視されます。

次の例では、テキスト変数 TEXT がデータセット ABSTRACT から選択されます。カテゴリ変数はありません。



ステップ 5: プロジェクトの実行

すぐにでも後からでもプロジェクト全体の実行を選択できます。後でプロジェクトを実行するには、**実行**ページで**なし**を選択します。プロジェクトの実行時期の詳細については、ヘルプを参照してください。

プロジェクト全体の実行を選択した場合、提供されたデータソースに対して次のイベントが発生します。

- 解析が行われます。
- トピックが生成されます。
- ルールが生成され、指定した任意のカテゴリ変数に対して実行されます。

- センチメントが適用されます(センチメントモデルを指定した場合)。
- コンセプトが適用されます(プロジェクトに事前定義済みコンセプトを含めた場合)。

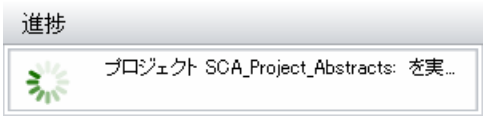
SAS Enterprise Content Categorization プロジェクトをインポートしてからプロジェクトを実行した場合、次のイベントが発生します。

- インポートされたコンセプトがコンパイルされ、提供されたデータソースに適用されます。
- インポートされたカテゴリがコンパイルされ、提供されたデータソースに適用されます。

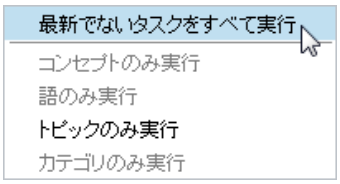
  注: サブカテゴリは、その親カテゴリが正常にインポートされた場合のみ、インポートされます。

プロパティページの**実行ステータス**フィールドは、プロジェクトが実行されるかどうかを示します。

注: メインウィンドウの左下隅にある**進捗**パネルを確認することもできます。どのプロジェクトが実行中で、どのプロジェクトが実行終了済みかを確認するには、**進捗**という単語をクリックします。



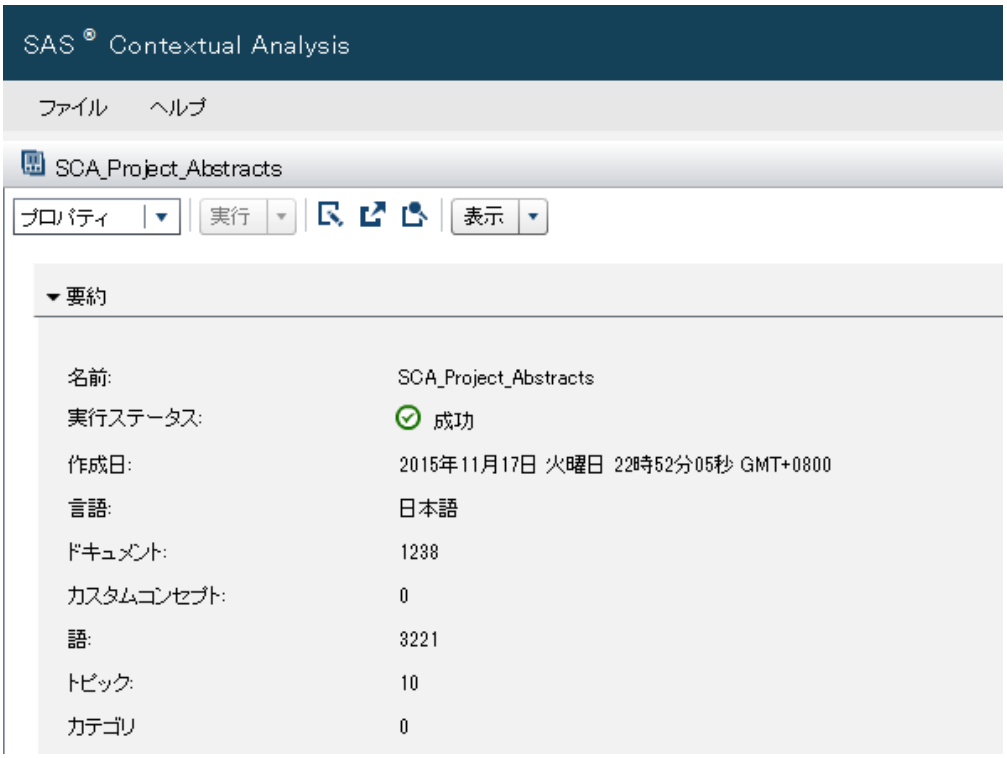
実行メニューを使用すると、分析タスクを個別に実行したり、最新ではないタスクのみ(もしあれば、その従属タスクも)実行したりできます。



プロパティページの使用

プロジェクトステータスの確認

プロパティページでは、プロジェクトが正常に実行されたかどうかが示され、分析されたデータについて基本情報が提供されます。



実行された各分析タスクの結果ステータスは、**ステータス**セクションの次のフィールドに表示されます。

タスク
分析タスクの名前

最新の状態
最後のタスク実行以後にタスクの情報が変更されたかどうかを示します。情報が変更されなかった場合、値 **Yes** となり、これ以上のアクションは必要ありません。最後の実行以後にタスクの情報が変更された場合、値は **No** となり、タスクの再実行が必要です。

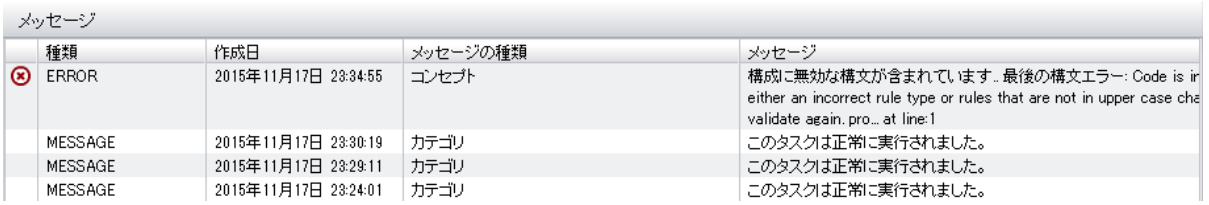
最終実行日と最終実行時刻
最後にタスクが実行された日付と時刻

最終実行期間
タスクの最終実行の期間

最終実行ステータス

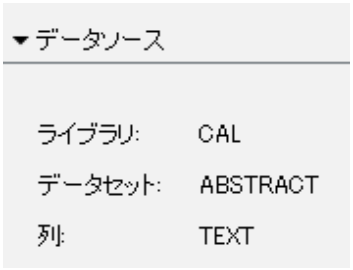
タスクの実行成功、実行失敗、タスク未実行、タスクの実行不可(情報が欠落しているか不十分なため)、警告発生のいずれであるかを示します。詳細については、ツールバーの**メッセージの表示**をクリックしてください。タスクが失敗したため従属タスクが実行されない場合、**未実行のステータス**が表示されます。次の例では、**CONCEPTS** タスクが失敗したため **TERMS** および **TOPICS** タスクが実行されません。正常に実行されるまで、エラーの修正とプロジェクトの再実行を続ける必要があります。

各タスクのステータスについてメッセージを参照するには、ツールバーのをクリックします。エラーがあるタスクの**メッセージ**ウィンドウの例を次に示します。**メッセージの種類**列は、各メッセージに対応する分析タスクを示します。


メッセージ

	種類	作成日	メッセージの種類	メッセージ
⊗	ERROR	2015年11月17日 23:34:55	コンセプト	構成に無効な構文が含まれています。最後の構文エラー: Code is ir either an incorrect rule type or rules that are not in upper case cha validate again. pro... at line:1
	MESSAGE	2015年11月17日 23:30:19	カテゴリ	このタスクは正常に実行されました。
	MESSAGE	2015年11月17日 23:29:11	カテゴリ	このタスクは正常に実行されました。
	MESSAGE	2015年11月17日 23:24:01	カテゴリ	このタスクは正常に実行されました。

データソース情報は**プロパティ**ページの下部に表示されます。


▾ データソース

ライブラリ: CAL

データセット: ABSTRACT

列: TEXT

プロジェクト情報の編集

プロジェクトの基本情報(プロジェクト名など)を編集するには をクリックします。**プロジェクトの編集**ウィザードが表示されます。編集できないアイテムはグレーで表示されます。

注: 変更の結果を確認するには、プロジェクトを再実行する必要があります。

コードの表示とダウンロード

作成された SAS スコアコードの表示とダウンロードができます。スコアコードを使用すると、プロジェクトのテキスト分析モデル(コンセプト、カテゴリ、およびセンチメント)を他のデータに適用できます。

プロパティページで、**表示**をクリックします。**コンセプトコード**、**センチメントコード**、または**カテゴリコード**を選択します。 をクリックして、他のプログラムで使用するコードをコピーします。

ヒント コードに埋め込まれたコメントがあり、生成されたコードの詳細が示されます。コードとデフォルト設定について含められたコメントを読むことをお勧めします。

スコアコードを処理する場合、次の点に注意してください。

- カテゴリがトピックプロモーションかカテゴリ変数のいずれかから作成された場合、そこに自動的に生成されたサブカテゴリのリストが含まれます。サブカテゴリのルールでは(一緒に処理する場合)カテゴリが定義されます。自動的に生成されたこれらのサブカテゴリを出力から削除(また、それによって親カテゴリの結果のみを参照)するには、次のフラグを **Y** に設定します。

  %let drop_auto_generated

- コードの生成後に SAS Contextual Analysis でいずれかの事前定義済みコンセプトが有効化または無効化された場合、スコアコードをコピーして別のプログラムに貼り付ける(またはコードをダウンロードする)前に再生成する必要があります。
- 事前定義済みコンセプトの有効化中にコンセプトを実行した場合、生成されたスコア済みデータには、完全パスを含まない一致が表示されます。

コードをファイルにダウンロードする場合は、  をクリックし、プロンプトに従ってファイルの場所を指定します。

プロジェクトモデルのエクスポート

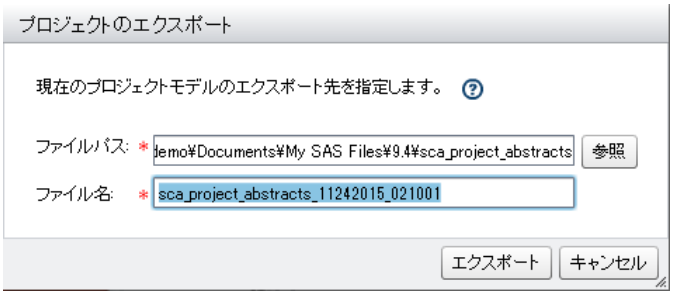
新しいプロジェクトでそのルールを再利用できるように、SAS Contextual Analysis プロジェクトモデルをエクスポートできます。プロジェクトをエクスポートすると、カテゴリおよびコンセプトルールのみエクスポートされます。選択した事前定義済みコンセプトはいずれも保持されます。トピック、語、データ、プロジェクト設定など、その他のプロジェクトコンポーネントは、エクスポートされません。エクスポートするファイルは JSON (JavaScript Object Notation)形式になります。

プロジェクトをエクスポートするには、次の操作を実行します。

1 エクスポートするプロジェクトを開きます。ツールバーの  をクリックします。

2 **ファイルパス**フィールドで、エクスポートするファイルのサーバーの場所を選択します。ファイルパスではプロジェクト名がすべて小文字で表示され、アンダースコア(_)が挿入される場合もあるので注意してください。

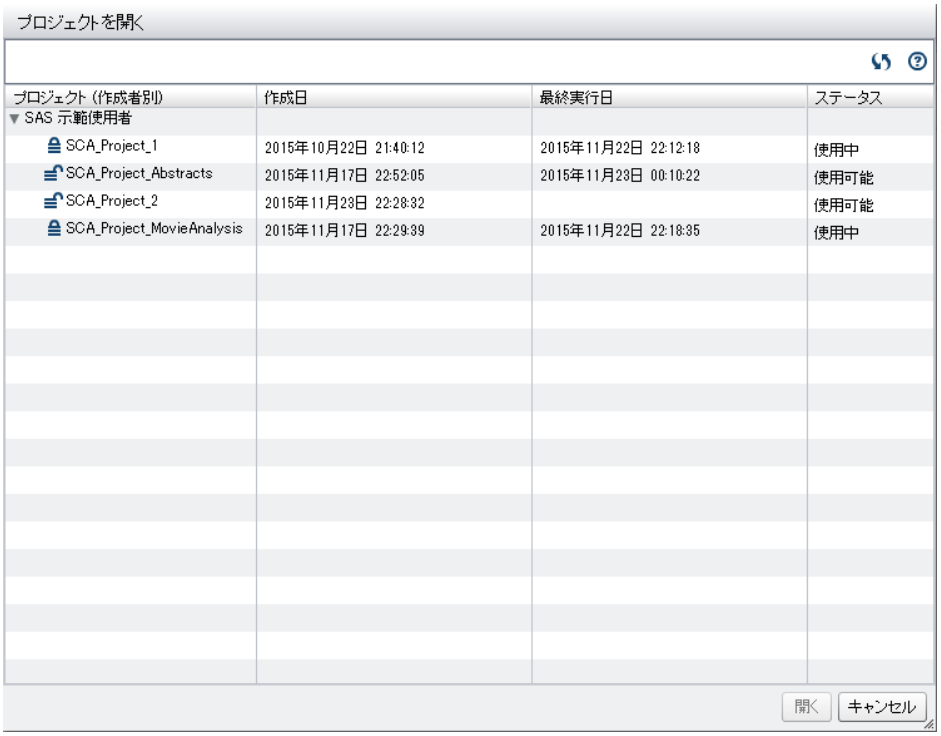
3 **ファイル名**フィールドで、エクスポートされたプロジェクトに対して自動的に生成された名前を確認し、必要があれば名前を編集します。生成されるファイル名は、プロジェクト名に加えて現在の日付(MMDDYY 形式)と時刻(MMHHSS 形式)から構成されます。



プロジェクトの共有

別のユーザーによって作成されたプロジェクトを開いて編集できます。メインウィンドウから、をクリックします。**プロジェクトを開く**ウィンドウに、すべての共有プロジェクトとその所有者が表示されます。使用中のプロジェクトはロックされ、開くことができません。ロック済みアイコンが、開くことのできないプロジェクトの横に表示さ

れます。ロック解除アイコンが、編集可能なプロジェクトの横に表示されます。



注: 他のユーザーのために共有プロジェクトをロック解除したままにしておくには、(Web ブラウザを閉じるのではなく)使用していない開いたプロジェクトを閉じて、SAS Contextual Analysis セッションをサインアウトします。

プロジェクトを開いた後で、類義語リストなどのプロジェクト情報を編集できます。プロジェクト情報編集の詳細については、“プロジェクト情報の編集” (22 ページ) を参照してください。コンセプトまたはカテゴリルール、語、あるいはその他の分析タスクの編集の詳細については、3 章, “分析タスクの実行” (29 ページ)を参照してください。

外部データセットのスコアリング

SAS Contextual Analysis プロジェクトに組み込んだモデルを使用すると、外部データセットをスコアリングできます。外部データセットをスコアリングする場合、カテゴリモデルが外部データセット(ターゲットデータセットと呼ばれる)に適用されます。する

と、ドキュメントコレクションのカテゴリ分け情報が、スコア済みデータセットに出力されます。

注: データセットは、フォルダ外のファイルに保存する必要があります。プロジェクトで、フォルダをデータセットとして使用する場合、同じフォルダ内のデータセットはスコアリングできません。

外部データセットをスコアリングするには、次の操作を実行します。

1 アプリケーションのメインメニューから**ファイル -> スコア外部データセット**を選択します。

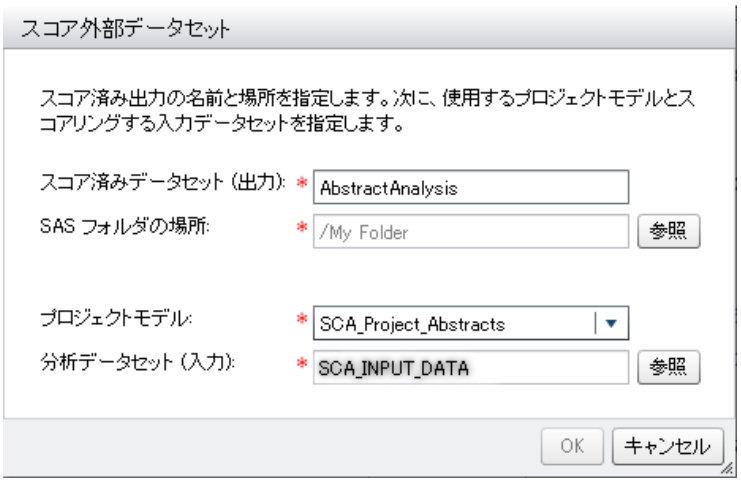
2 **スコア済みデータセット (出力)**フィールドで、生成されるスコア済みデータセットの名前を入力します。

3 データ保存用の SAS フォルダの場所を入力します。

4 **プロジェクトモデル**フィールドで、使用している分析モデルを含むプロジェクトの名前を選択します。

5 **分析データセット (入力)**フィールドで、スコアリングするデータセットを指定します。分析データセットは、選択したプロジェクトモデルのデータソースと同じテキスト変数を有する必要があります。

注: スコアリングの対象となるには、プロジェクトがカテゴリバイナリファイルをコンパイルしている必要があります。カテゴリを含むプロジェクトを実行すると、カテゴリバイナリファイルが生成されます。



スコアリングの開始後に、プロジェクトの実行ステータスが**実行中**に変更されます。スコアリングが完了すると、プロジェクトモデルとして使用したプロジェクトが保存されているライブラリフォルダにスコア済みデータセットが置かれます。

センチメント分析について

ドキュメントスコアリングの概要

センチメント分析は、ドキュメントに示された作成者のトーンまたは態度(ポジティブ、ネガティブ、またはニュートラル)を識別するプロセスです。SAS Contextual Analysis では、センチメントを示す語、フレーズ、および文字列の識別と分析を行う専用ルールセットが使用されます。次に、その分析に基づいて、センチメントスコアが割り当てられます。これらのルールを使用すると、ソフトウェアで、繰り返し可能な高品質の結果が得られます。

ドキュメントに対するセンチメントの割り当ては、ドキュメント全体に関連付けられた態度に基づいています。たとえば、次のドキュメントはポジティブセンチメントになります。**Had an awesome time yesterday. Glad I brought my tent from Store XYZ.**

ドキュメントはセンチメントを示す複数の単語または語に関連付けられる可能性があるので、SAS Contextual Analysis では、スコアリングシステムを使用して最終センチメントスコアを割り当てます。次のリストでは、センチメントの機能について基本情報が提供されます。(この情報は、キーコンセプトを説明するために単純化されてします。)

- 各ポジティブ語またはフレーズが、1(ポジティブ)ポイントに相当します。
- 各ネガティブ語またはフレーズが、ネガティブポイントに相当します。ポジティブ語またはフレーズの方がネガティブよりも多く存在する場合、最終センチメントスコアはポジティブになります。
- ネガティブ語またはフレーズの方が多く存在する場合、最終センチメントスコアはネガティブになります。
- ポジティブ語またはフレーズとネガティブ語または語句が同数存在する場合、センチメントスコアはニュートラルになります。

SAS Contextual Analysis での SAS Sentiment Analysis モデルの使用

SAS Sentiment Analysis を使用して生成されたルールは、.sam バイナリファイルに保存されます。SAS Contextual Analysis でプロジェクトを作成する場合、各自の指定に合わせて作成した.sam バイナリファイルを使用するか、プロジェクトの言語で使用可能なデフォルトファイルを使用することができます。

注: すべての言語に、使用可能なデフォルトセンチメントモデルがあるわけではありません。

センチメント分析およびスコアリングの詳細については、*SAS Sentiment Analysis 12.2: User's Guide* を参照してください。

3

分析タスクの実行

分析タスクの概要

はじめに

プロジェクトを実行すると、次の分析タスクが実行されます(データが存在する場合)。

- コンセプト抽出
- 語の識別(類義語を含む)
- トピック検出
- カテゴリ分析

次のセクションでは、各タスクについて説明します。

コンセプト

*コンセプト*は、書名、姓、市区町村、性別などのプロパティです。コンセプトは、コンテキストの情報分析に役立ちます。自分にとって重要なコンセプトを識別するためにルールを作成して、それによりカスタムコンセプトを作成できます。たとえば、テキストで *refrigerator*、*sink*、および *countertop* という語が検出された場合に、コンセプト *kitchen* が識別されるように指定できます。

SAS Contextual Analysis では、*事前定義済みコンセプト*が提供されます。これは、ルールをすでに作成済みのコンセプトです。事前定義済みコンセプトでは、**COMPANY** や

TITLE など、よく使用するコンセプトとその定義を提供することによって、時間を節約します。事前定義済みコンセプトの名前変更はできません。また、その基本定義の表示や編集もできません。**編集**タブで処理の追加ルールを指定することはできます。

注: インポートされたコンセプトが事前定義済みコンセプトと同名の場合、インポートされたコンセプトのルールが事前定義済みコンセプトのルールに追加されます。

カスタムコンセプトには、一致の重複が発生した時に返される一致の優先順位を付けられます(たとえば、New York に一致するコンセプトノードと、New York City に一致する別のコンセプトノードなど)。そのためには優先順位値を設定します。優先順位値を設定する場合、カスタムコンセプトの優先順位をより高い値に設定できるように、事前定義済みコンセプトの事前設定値を把握しておくと便利です。優先順位設定の詳細については、“コンセプトページ” (35 ページ)を参照してください。

表 3.1 (31 ページ) では、SAS Contextual Analysis に含まれている英語の事前定義済みコンセプトのリストと一緒に、その優先順位値が示されます。英語以外のサポート言語の事前定義済みコンセプトとその優先順位値の完全なリストについては、付録 2, “事前定義済みコンセプト(英語以外の言語用)” (113 ページ) を参照してください。

注: 一部の言語では、ここにリスト表示された事前定義済みコンセプトのサブセットが使用されます。

プロジェクト作成中に(またはコンセプトタスクウィンドウで)、いずれのコンセプトでも無効化や有効化を行えます。

表 3.1 英語の事前定義済みコンセプトと優先順位

事前定義済みコンセプト	説明	優先順位値
ADDRESS	郵送先住所や番地	20
COMPANY	会社名 **注:** SAS Contextual Analysis では、このコンセプトを識別するために、用意された会社名および組織名の辞書が使用されます。このコンセプトはしばしば親と関連付けられます。たとえば、**IBM** がテキストに出現した場合、事前定義済みの親 **International Business Machines** と一緒に返されます。通常、最も長く最も正確なバージョンの名前が、親形式として使用されます。	25

CURRENCY	通貨または通貨表現。例: $300、3 億ドル	18
DATE	日付、曜日、月、年などの日付表現	18
INTERNET	URL、パス、ファイル名、メールアドレスなどを含むデジタル場所	18
LOCATION	市区町村、国、州、地理的な場所や領域、政治的な場所や領域	30
MEASURE	測定または測定表現。例: 500kg、2300 sq f	20
NOUN_GROUP	1 語として機能する複数の単語(たとえば、“for sale”や“clinical trial”など)	15
ORGANIZATION	政府機関、法機関、またはサービス機関。(COMPANY に関連付けられた注を参照)	25
PERCENT	パーセンテージまたはパーセンテージ表現。例: 97%、12 パーセントポイント	18
PERSON	氏名	40*
PHONE	電話番号	18
PROP_MISC	製品、書籍、人物などの別カテゴリに入る可能性がある不確かな(その他の)分類の固有名詞。例: Fargo	5
SSN	社会保障番号	18
TIME	時間または時間表現。例: 0800、1:15、6pm	18
TIME_PERIOD	時間の範囲。例: 9am–6pm	18
TITLE	役職または職位	18
VEHICLE	色、年、メーカー、型などを含む自動車	20

* 英語の最高優先順位値

*カスタムコンセプト*は、ルールの作成が必要なコンセプトです。

注: SAS Enterprise Content Categorization プロジェクトをインポートした場合、LITI ルールを使用して作成されたコンセプトがカスタムコンセプトとしてプロジェクトに表示されます。ルールエディタを使用すると、これをさらに編集できます。

コンセプトルールの作成の詳細については、“コンセプトルールの作成: 基本 LITI 構文” (56 ページ)を参照してください。 ブールルールの詳細については、“カテゴリルールの作成: ブールルール” (81 ページ)を参照してください。

語と類義語

*語*は、基になるルールまたはアルゴリズムによって定義されているように、1 つのコンセプト(アイデア)を表す文字列またはパターンのグループに対するラベルとして定義されます。SAS Contextual Analysis では、語は、トピック、語マップ、およびカテゴリルールの基本ビルディングブロックです。各語には、空白かまたは語の品詞を識別する役割が関連付けられています。語は、表層形を 1 つ以上反映します。*表層形*は、一致したテキストサブセットにある語のバリアントです。表層形には、活用形、類義語、スペルミス、およびその他の語の参照方法を含められます。SAS Contextual Analysis では、類似性と頻度に基づいて語のスペルミスの識別と分類を行えます。スペルミスは実際には別の語を示すため、分析中は類義語として扱われます。

*類義語リスト*は、分析のために 1 語として扱う必要がある単語のペアを識別する SAS データセットです。類義語は、親レベルで適用されます。類義語リストは、**プロジェクトの新規作成**ウィザードと**プロジェクトの編集**ウィザードで指定できます。類義語リストは、データセットに保存され、必須形式があります。次の変数を含める必要があります。

- TERM: PARENT の類義語として扱う語が含まれます。
- PARENT: TERM の割り当て先となる代表語が含まれます。

また、次の変数も含める必要があります。

- TERMROLE: この変数に指定した役割で TERM が発生した場合のみ類義語が割り当てられるように指定できます。*語の役割*は、特定のコンテキストで語によって実行される機能です。語の役割には、品詞役割、エンティティ役割、およびユーザー定義役割が含まれます。
- PARENTROLE: PARENT の役割を指定できます。

> ヒント 類義語リストのいずれかの語を使用してコンセプトを抽出する場合、PARENT 語と一致させる PARENTROLE のカスタムコンセプトの作成(またはコンセプトの存在確認)が必要になります。コンセプトが設定された後で、語を再実行します。たとえば、親の語 **Luke Skywalker** によって親の役割 **JEDI_MASTER** が指定されるとします。この場合、**Luke Skywalker** と一致するルールを含む、**JEDI_MASTER** というカスタムコンセプトを作成して、語を再実行する必要があります。

役割の詳細については、*SAS Text Miner: Reference Help* のセクション“Term Roles and Attributes” を参照してください。

注: 類義語リストに、同じ語を異なる親に割り当てる複数のエントリが含まれている場合、結果の解析には最初のエントリのみ反映されます。

開始リストと停止リスト

開始リストと停止リストを使用して、テキストマイニング分析において、どの語を使用し、どの語を使用しないかを制御します。*開始リスト*は、解析結果に含める語のリストを含むデータセットです。開始リストを使用すると、そのリストに含まれる語のみが解析結果に表示されます。*停止リスト*は、解析結果から除外する語のリストを含むデータセットです。停止リストを使用すると、情報量が少ない語や、テキストマイニングタスクと無関係の語を除外できます。デフォルト停止リストは英語用に提供されます(Sashelp.EngStop)。

開始リストと停止リストに必要な形式は同じです。変数 TERM を含めて、そこに含める語(開始)か除外する語(停止)を入れる必要があります。また、変数 ROLE を含めて、関連役割を入れることもできます。ROLE 変数を指定した場合、語は、役割が ROLE 変数で指定したものである場合に限り、保持されるか(開始リスト)、または削除されます(停止リスト)。

トピック

*トピック*は、ドキュメントに出現する重要語の自然なグループ分けから導き出されます。SAS Contextual Analysis では、トピックが自動的に生成され、ドキュメントに割り当てられます。1 つのドキュメントに 2 つ以上のトピックを含められます。

トピックページには、SAS Contextual Analysis で識別されたトピックがすべて表示されます。トピックのデフォルト名は、トピックに頻繁に出現する上位 5 つの語です。これらの語は、その重みに基づいて降順に並べ替えられます。

カテゴリ

*カテゴリ*では、共通特性を共有するドキュメントのグループが識別されます。

たとえば、カテゴリを使用すると、次を識別できます。

- ホテルの宿泊について苦情の領域
- 発行された記事の要約のテーマ
- 保証コールセンターで繰り返し発生する問題

カテゴリを作成するには、トピックをカテゴリにプロモートするか、**プロジェクトの新規作成**ウィザードでカテゴリ変数を指定するか、または新しいカテゴリを作成します。また、SAS Enterprise Content Categorization からカテゴリをインポートすることもできます。カテゴリ変数や、カテゴリにプロモートされたトピックに対して自動的に生成されたルールを編集できます。

注: カテゴリルールは、LITI 形式ではなく、SAS Enterprise Content Categorization が使用する形式(MCAT)です。カテゴリ内から LITI コンセプトを参照できます。

コンセプトルールの作成の詳細については、“コンセプトルールの作成: 基本 LITI 構文” (56 ページ)を参照してください。 ブールルールの詳細については、“カテゴリルールの作成: ブールルール” (81 ページ)を参照してください。

分析タスクページの使用

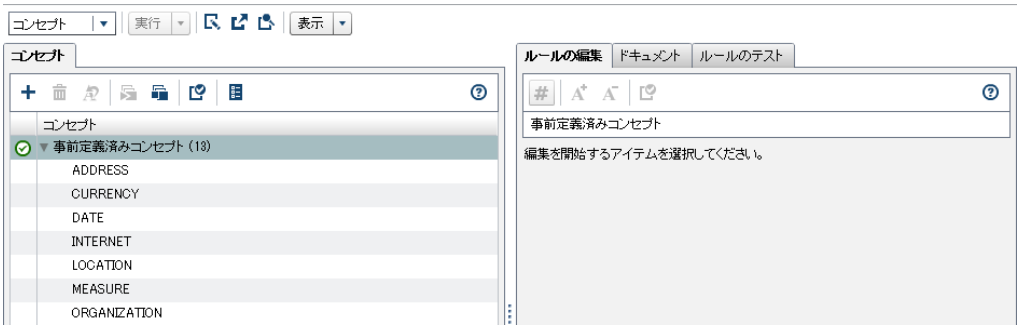
コンセプトページ

コンセプトページを使用すると、事前定義済みコンセプトとインポート済みコンセプトの表示、カスタムコンセプトの追加、コンセプトルールのテスト、コンセプトプロパティの編集、一致を含むドキュメントの表示を行えます。

ヒント　コンセプトノードに関連付けられたアイテムの表示をカスタマイズするには、**ルールの編集**、**ドキュメント**、および**ルールのテスト**タブをあるペインから別のペインにドラッグします。

分析に含める対象を参照するには、事前定義済みおよびカスタムコンセプトノードのリストを展開します。

注: プロジェクト作成中に事前定義済みコンセプトの除外を選択した場合、コンセプトページでは事前定義済みコンセプトにアクセスできません。



コンセプトノードを無効化 または有効化 するには、ツールボタンをクリックします。

注: 無効化コンセプトに関連付けれらた語はいずれも語リストから削除され、解析中は無視されます。

コンセプトページで行える、その他の重要アクションを次に示します。

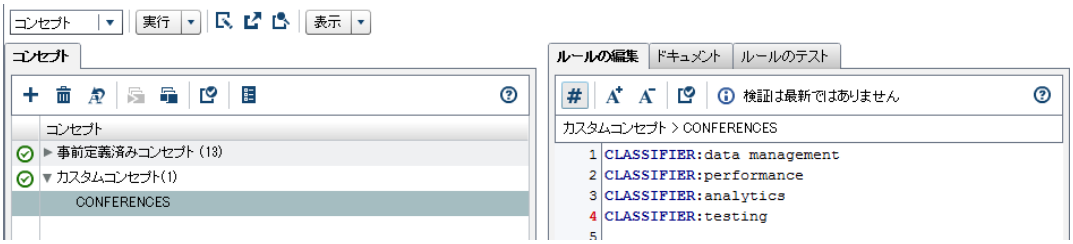
- **カスタムコンセプトの追加**

  独自のルールを作成するためのカスタムコンセプトノードを追加するには、+をクリックします。

  ヒント　カスタムコンセプトノードを作成する場合、その名前には大文字を使用します。コンセプトを後で参照する場合は、ルール内でコンセプトを認識する方が容易です。コンセプト名は大文字と小文字が区別されます。

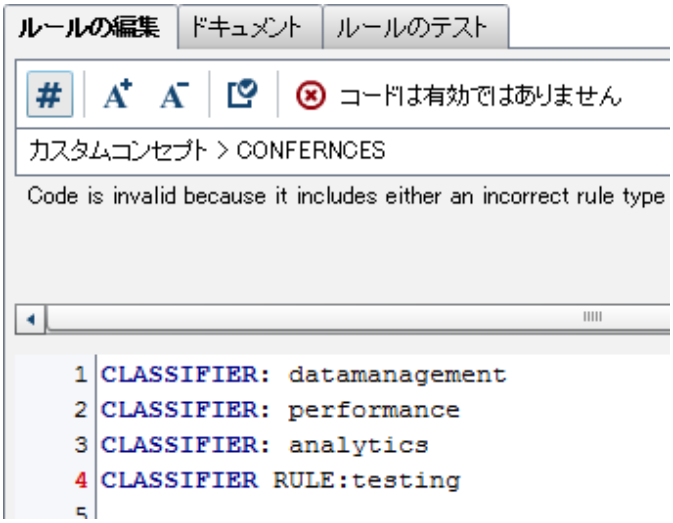
ヒント テキストで一致する可能性もある単語(たとえば、PROBLEM や MECHANICS など)をカスタムコンセプトノードの名前として使用しないでください。かわりに、MYNEWCONCEPT などの解釈不可能な名前を使用してください。

ルールの編集タブで、コンセプトノードに対して LITI ルールを入力します。分析でコンセプトノードを使用する前に、ルールを検証する必要があります。



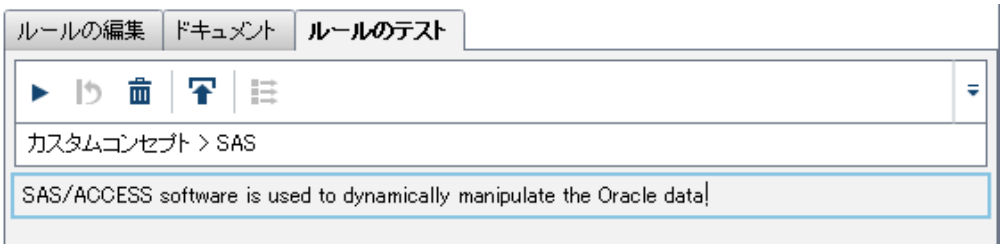
ヒント 変更を加えると、検証は最新ではありません メッセージによって、ルールを検証するように通知されます。

各ルールを個別に検証するには**ルールの編集**タブで をクリックし、すべてのルールを検証するには**コンセプト**で をクリックします。エラーおよびその他のメッセージは**ルールの編集**タブに表示されます。

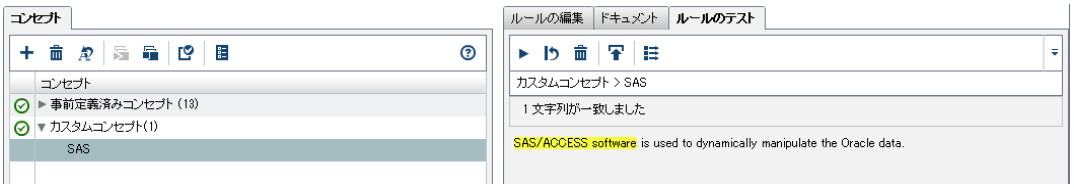


- **コンセプトノードルールのテスト**

コンセプトタブでコンセプトノードを選択し、**ルールのテスト**タブをクリックします。 をクリックしてテストするファイルをアップロードするか、単純に、選択したルールに対するテストテキストを入力します。 をクリックすると、ルールがテストされます。



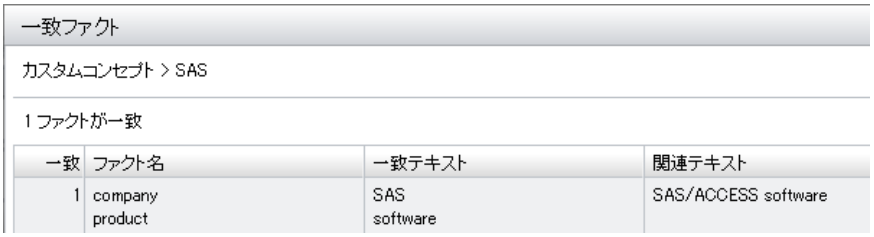
次のサンプル画面では、コンセプトノード **SAS** で一致文字列が強調表示されています。



注: 一致文字列(強調表示テキスト)には、コンセプトルールテストでの複数の一致を含められます。

テストテキストの一致文字列を検証できるように、別のウィンドウで *ファクト*(一緒に検索して一致させるテキストの関連情報)の一致を表示できます。 をクリックすると、ウィンドウが開き、一致ファクトが表示されます。次のサンプル画面は、テキスト **SAS/ACCESS software is used to dynamically manipulate the Oracle data.**と次のルールとのテストでの一致を示しています。

```
PREDICATE_RULE:(company, product):(AND, "_company{SAS}", "_product{software}")
```



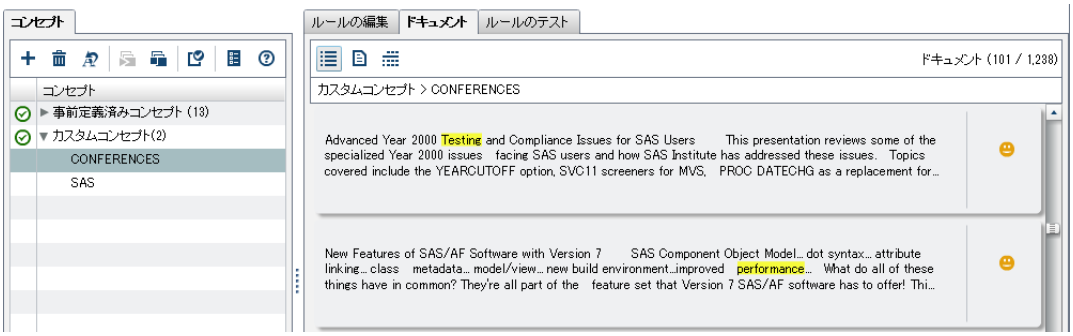
ファクトの詳細については、“コンセプトとファクト” (57 ページ)を参照してください。

- **一致ドキュメントの表示と探索**

  一致を含む学習ドキュメントを表示するには、**ドキュメント**タブをクリックします。アイコン のいずれかをクリックすると、ドキュメントビュー間の切り替えが行われます。コンセプトノード **CONFERENCES** を作成し、そこに次のルールが含まれているとします。

  CLASSIFIER:testing
  CLASSIFIER:performance

  次のサンプル画面に示されるように、ドキュメント内の一致が強調表示されます。



  注: プロジェクトの作成時にセンチメントモデルを適用した場合のみ、センチメント値が表示されます。

- **カスタムコンセプトノードプロパティを編集してコンセプトの一致を絞り込む**特定のプロパティを編集すると、カスタムコンセプトルールからの一致の絞込みに役立ちます。プロパティを表示するには、 クリックします。

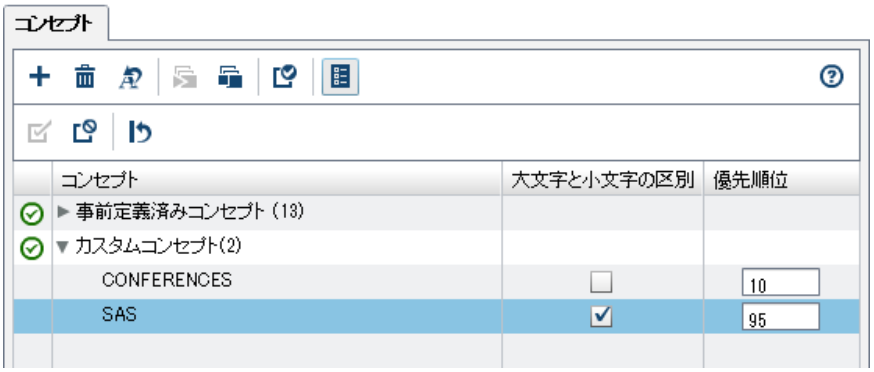
  ルールに指定された大文字/小文字に対して一致が発生することを確認するには、**大文字と小文字の区別**チェックボックスを選択します。

  一致の重複が発生した時に返される一致の優先順位を付けられます(たとえば、**New York** に一致するコンセプトノードと、**New York City** に一致する別のコンセプトノードなど)。一致が重複する場合、**優先順位**列に一番大きな数が入力されたコンセプトノードが返されます。値は正数にする必要があります(1 は優先順位が最も低い)。最も高い優先順位値に制限はありません。デフォルト値は 10 です。

  注: 自分のコンセプト一致が事前定義済みコンセプト値と重複しないようにするには、自分が含めた事前定義済みコンセプトよりも高い優先順位値を使用します。詳細

については、“コンセプト” (30 ページ) または 付録 2, “事前定義済みコンセプト(英語以外の言語用)” (113 ページ)を参照してください。

をクリックすると、すべての優先順位値がデフォルト値に戻ります。



語ページ

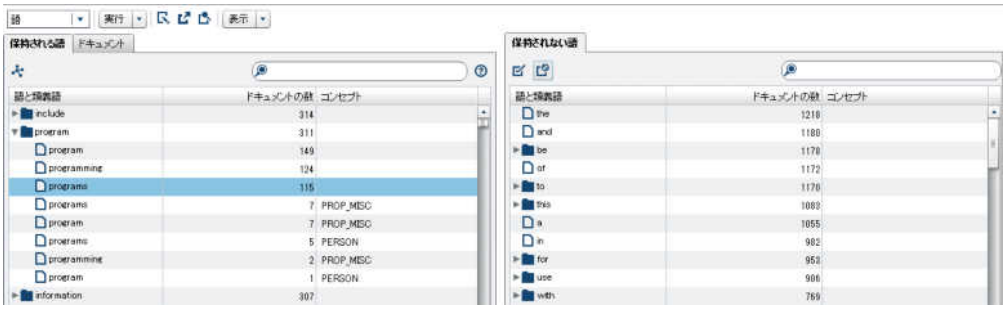
プロジェクトが正常に実行された後で、**語**ページを開いて、ドキュメントコレクションで検出された語を表示します。デフォルトビューでは、左に**保持される語**、右に**保持されない語**が示されます。タブ内でビューを切り替えるには、アイコン および を使用します。

ヒント ビューをカスタマイズするには、**ドキュメント**、**保持される語**、または**保持されない語**タブをあるペインから別のペインにドラッグします。

語ページで完了させられる、その他の重要タスクを次に示します。

- **語の表示**

  保持される語には、ドキュメントコレクションで保持されたすべての語が表示されます。**ドキュメントの数**列には、選択した語を含む学習ドキュメントの数が表示されます。**コンセプト**列には、決定可能であれば、各語の役割が表示されます。語に割り当てられた表層形を表示するには、その語の横に表示される三角形をクリックします。



- **あるタブから別のタブに語をドラッグ**

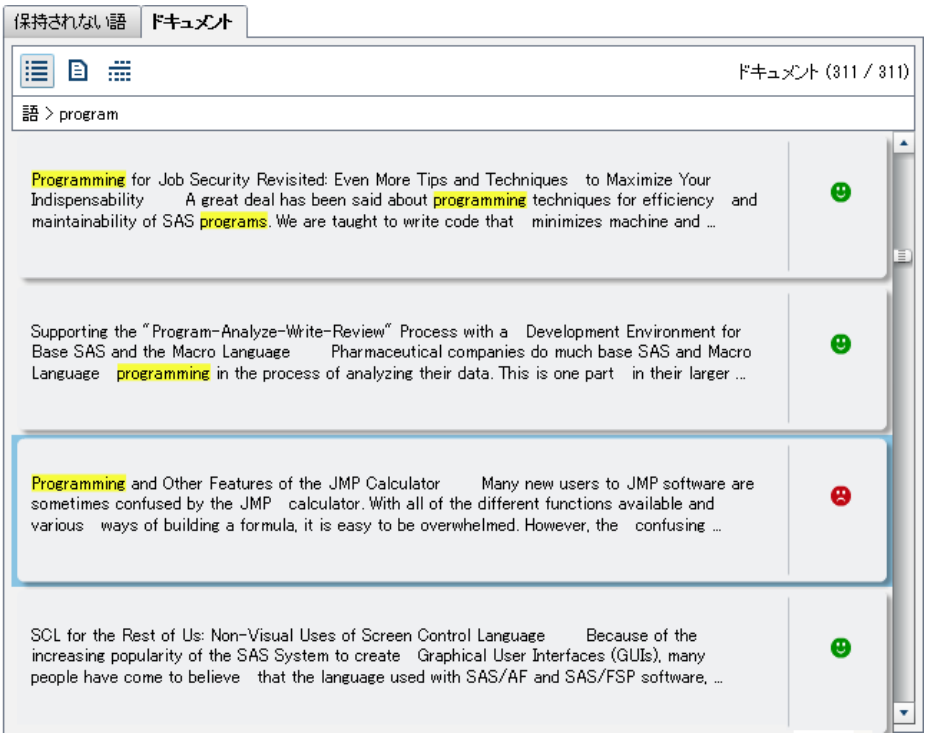
  デフォルトでは、語のリストは、各語が出現するドキュメントの数の降順に並べ替えられます。親の語を**保持される**タブから**保持されない**タブにドラッグしたり元に戻したりできます。

  注: 語に変更を加えて、その変更の結果を確認する場合は、**実行**をクリックしてプロジェクトを再実行する必要があります。

  注意! タスク(すべての最新でないタスクまたはトピックのみ)の再実行時にコンセプトルールが最新ではない場合、語に加えた変更はいずれも元の語リストで上書きされます。

- **一致ドキュメントの表示と探索**

  一致を含む学習ドキュメントを表示するには、**ドキュメント**タブをクリックします。アイコン  のいずれかをクリックすると、ドキュメントビュー間の切り替えが行われます。次のサンプル画面に示されるように、一致が強調表示されます。

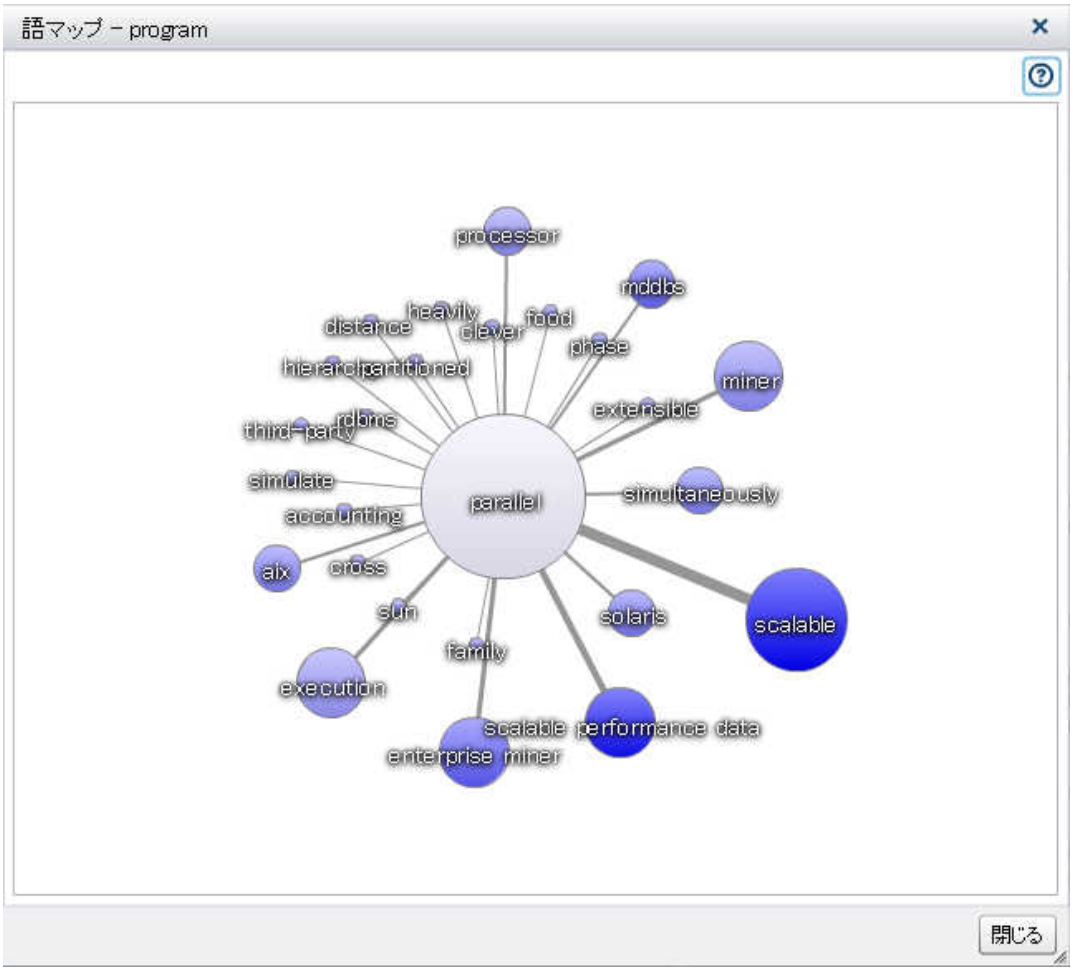


注: プロジェクトの作成時にセンチメントモデルを適用した場合のみ、センチメント値が表示されます。

- **語マップの表示**

  語の**語マップ**を表示するには、**保持される語**でその語を選択し、 をクリックします。

  ヒント コーパスに存在するドキュメントが少なすぎて語との重要な関係が見いだせないため、語マップが生成されないことがあります。その場合は、別の語を選択するか、必要に応じて、ドキュメントコレクションのサイズを増やします。



語マップウィンドウに、選択した語の語マップが表示されます。前出のサンプル画面では、選択した語は *program* で、マップでは最も大きな円で表されます。マップ判読の詳細については、語マップの上の ? をクリックしてください。

トピックページ

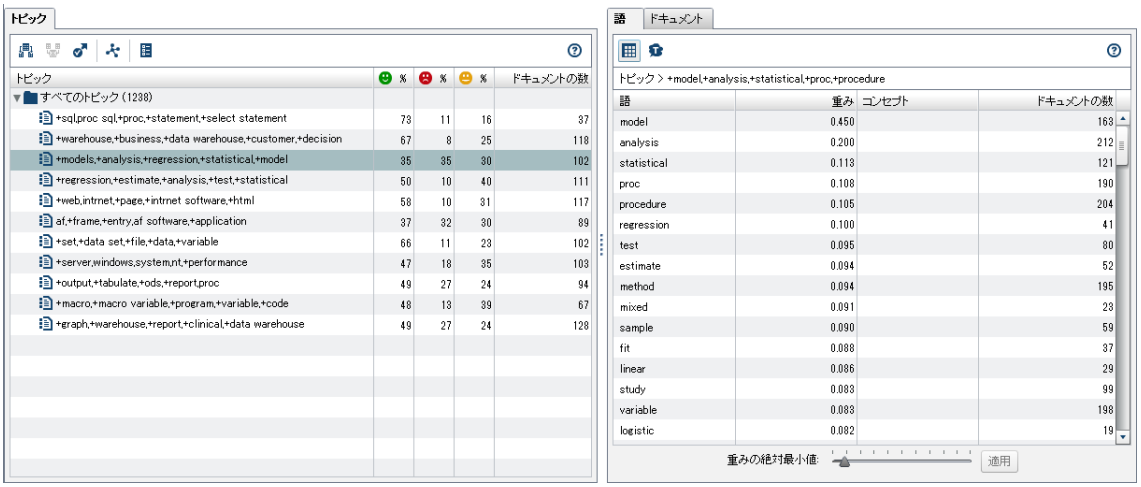
トピックを分析するには、**トピック**タブでそのトピックを選択します。選択したトピックは、その最も重要な 5 つの語で識別されます。トピックページで実行できるタスクを次に示します。

- **トピックを構成する語の表示**

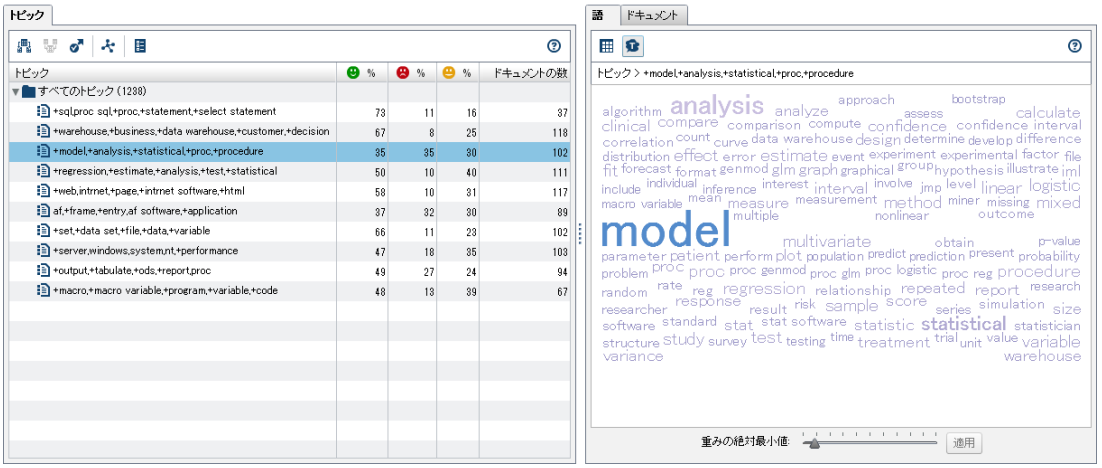
次のサンプル画面では、トピックは語 **model**、**analysis**、**regression**、**statistical**、**estimate** で識別されます。

注: プロジェクトを作成したときにセンチメントモデルを含めていた場合に限り、ポジティブ、ネガティブ、およびニュートラルのセンチメントスコアを持つ、トピック内のドキュメントのパーセンテージが、各トピックと一緒に表示されます。

トピックを選択すると、右のビューが更新されます。ビューを切り替えるには、右のツールバーのアイコンを使用します。各ビューの詳細については、右の をクリックします。語リストをテーブルとして表示するには、右の をクリックします。語テーブルには、トピック内のすべての語、計算されたその重み、そこに割り当てられた役割(コンセプト)、その語を含むドキュメントの数が表示されます。



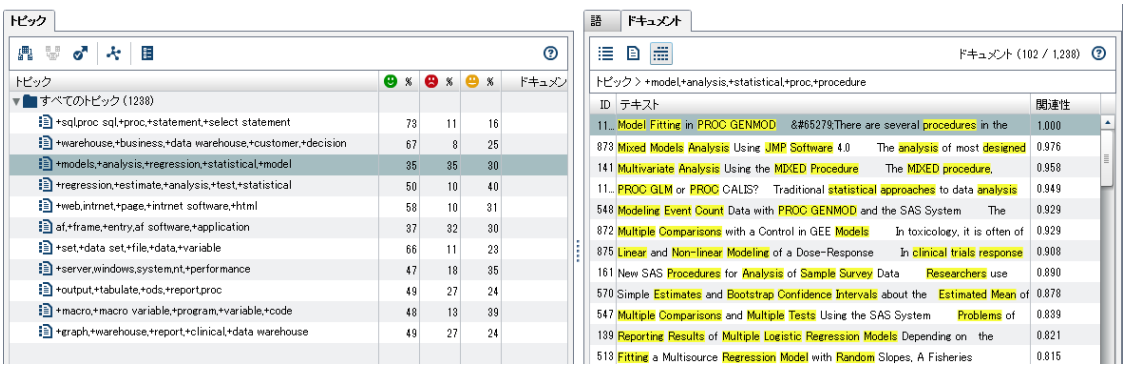
次のサンプル画面では、ワードクラウドビューアイコン が選択されました。ワードクラウドペイン下部のスライダによって、語をこのトピックに含めるために必要な重みの絶対最小値を調整できます。スライダを右に移動すると、ワードクラウドが更新されます。**適用**をクリックして、行った変更を確定します。



- **トピックに関連付けられたドキュメントの表示**

  トピックに関連付けられた学習ドキュメントを表示するには、**ドキュメント**タブをクリックします。

  選択したトピックの語を参照するには、ドキュメントビューアイコン  のいずれかを選択します。次のサンプル画面では、このトピックの一部であるとしてドキュメントをマークする語がすべて強調表示されます。



- **トピックの分割とマージ**

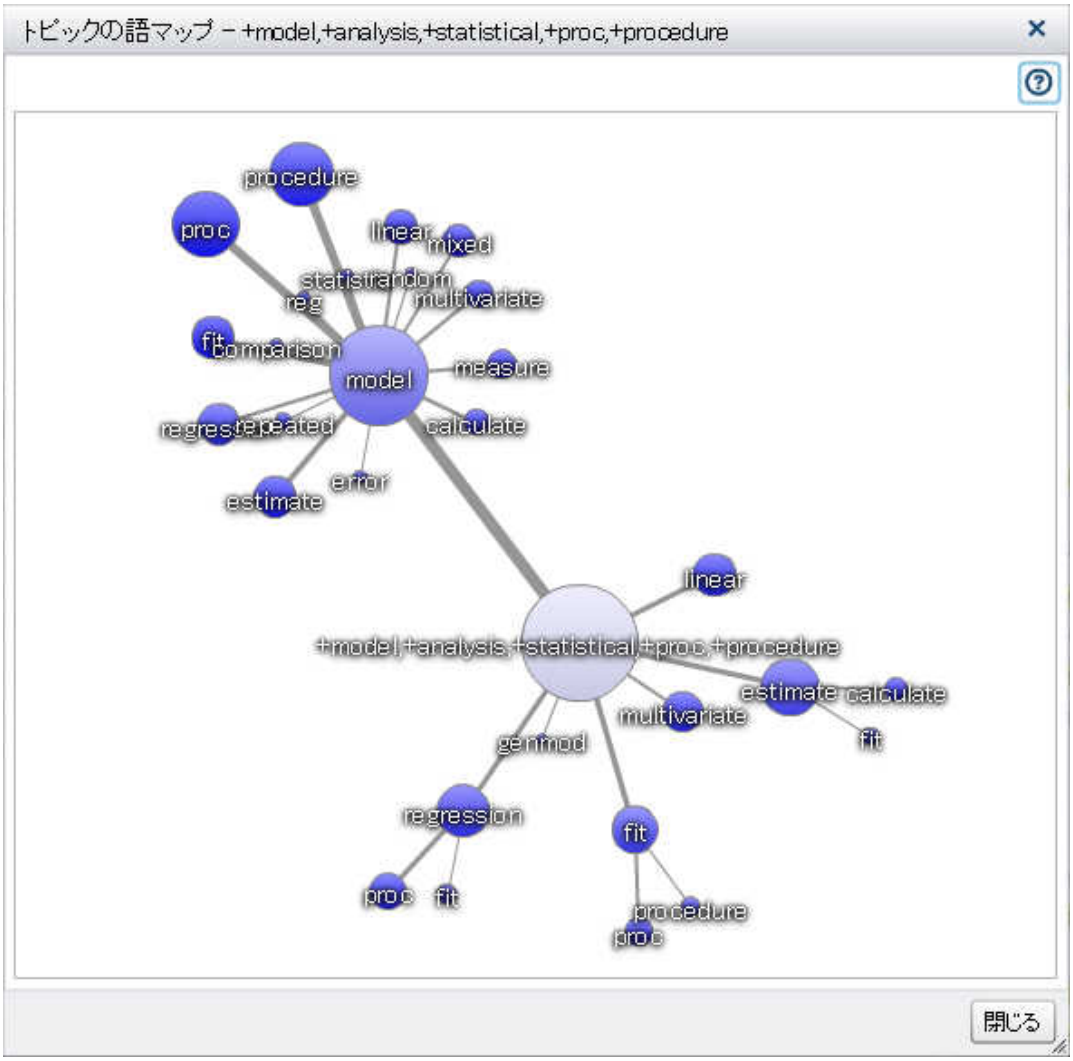
  使用目的に対してトピックが広範すぎると考えられる場合、**トピック**で対象を選択して  をクリックすると、分割できます。このアクションによって、選択したトピックが 2 つの新しいトピックに分割されます。

関連のありそうな 2 つのトピックがある場合、対象を選択して  をクリックすると、マージできます。このアクションによって、選択したトピックがすべて同じトピックに結合されます。

注: マージした後でマージ解除トピックに戻す場合は、トピックを再実行すると戻せます。その時点までの語とトピックへの変更は失われます。

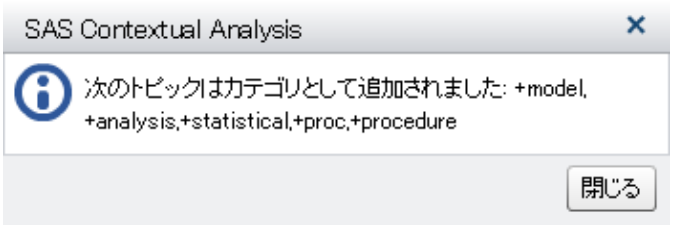
- **トピック語マップの表示**

  トピックペインからトピック語マップを表示するには、トピックを選択して  をクリックします。トピック語マップでは、語頭のチルダは NOT 演算子として扱われます。マップ判読の詳細については、トピック語マップの上の  をクリックしてください。



- **トピックをカテゴリにプロモート**

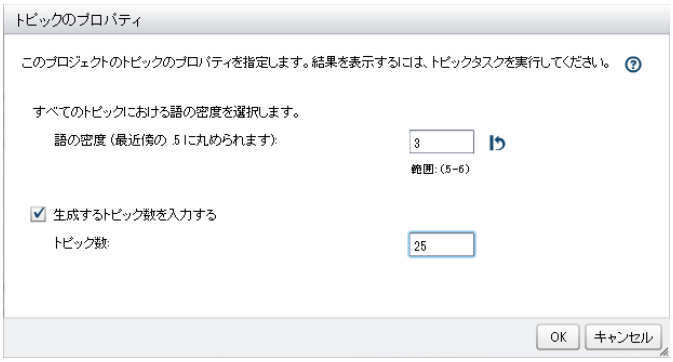
  分析における重要なステップは、どのトピックをカテゴリにプロモートするか識別することです。トピックをカテゴリにプロモートするには、そのトピックを**トピック**ペインで選択して をクリックします。このアイコンをクリックすると、SAS Contextual Analysis では、選択したトピックが**カテゴリ**ページに追加されます。一度に複数のトピックをカテゴリにプロモートできます。



- **トピックプロパティの編集**

  すべてのトピックに影響するプロパティを編集できます。語の密度とは、トピックに語がどの程度存在するかを示し、0.5 から 6 までの数字で定義されます(デフォルト値は 2)。語の密度が 0.5 に近づくほど、トピックでの語の密度が高くなります。語の密度が 6 に近づくほど、トピックでの語の密度が低くなります。この値は、トピックに属するドキュメントの数に影響します(たとえば、トピック内の語が減少すると、取り込まれるドキュメントも減少します)。入力した値は、最近傍の整数または半整数に丸められます。

  プロジェクトに対して生成されるトピックの最大数も指定できます。設定を空白にしておくと、デフォルト方法を使用して、重要語からトピックが生成されます。



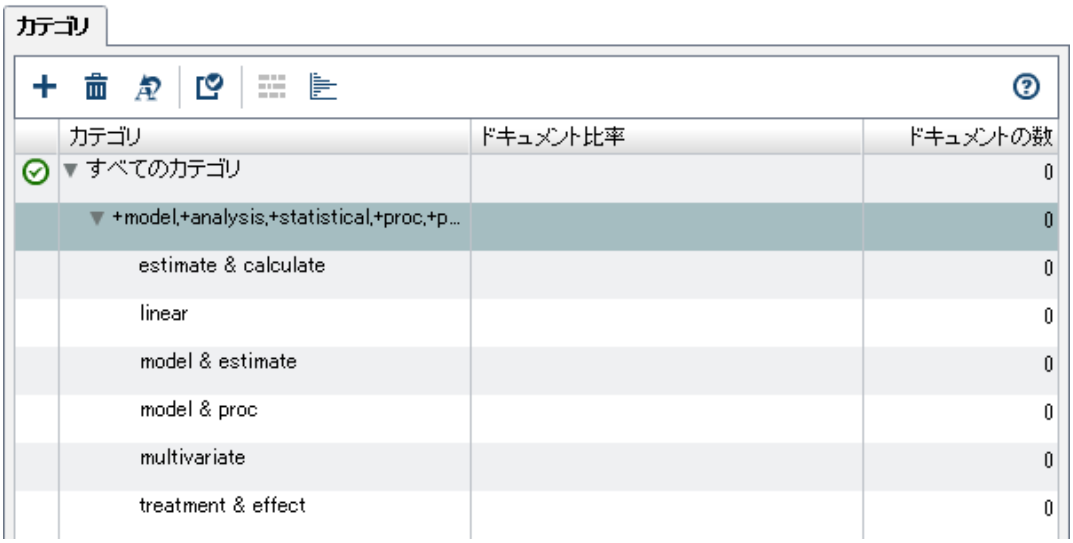
  注: 変更の結果を参照するには、トピックを実行する必要があります。

カテゴリページ

トピックページでトピックからカテゴリを作成すると、**カテゴリ**ページにカテゴリが表示されます。**ルールの編集**タブには、そのカテゴリのために生成されたルールが表示されます。

注: 外部カテゴリの変数に対して 3 つ以上の値が存在する場合、サブカテゴリが作成されます。たとえば、変数 **blueberry**、**cherry**、および **peach** を含むカテゴリ **Flavors** では、生成されたルールを含む 3 つのサブカテゴリが作成されます。サブカテゴリは、二値変数値が存在する外部カテゴリに対しては作成されません。たとえば、**Yes** と **No** の値があるカテゴリ **In stock** では、サブカテゴリは作成されません。かわりに、生成されたルールがカテゴリ **In stock** 内にリストされます。

カテゴリを実行するまでは、**ドキュメント**タブにデータは読み込まれません。



ヒント カテゴリに関連付けられたアイテムの表示をカスタマイズするには、**ルールの編集**、**ドキュメント**、および**ルールのテスト**タブを一方から他方にドラッグします。

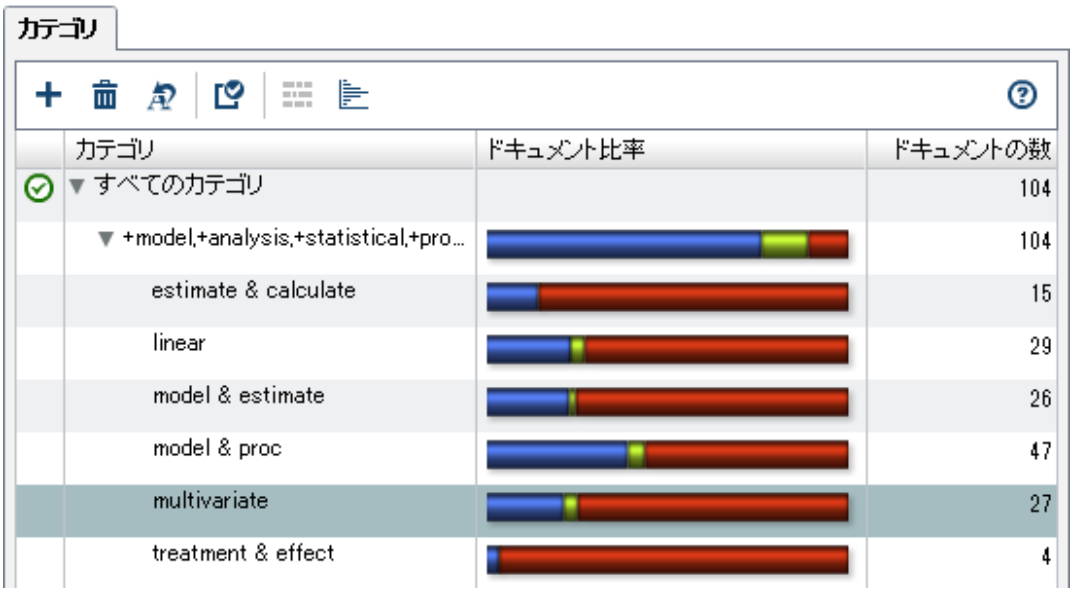
カテゴリページで実行できる重要タスクを次に示します。

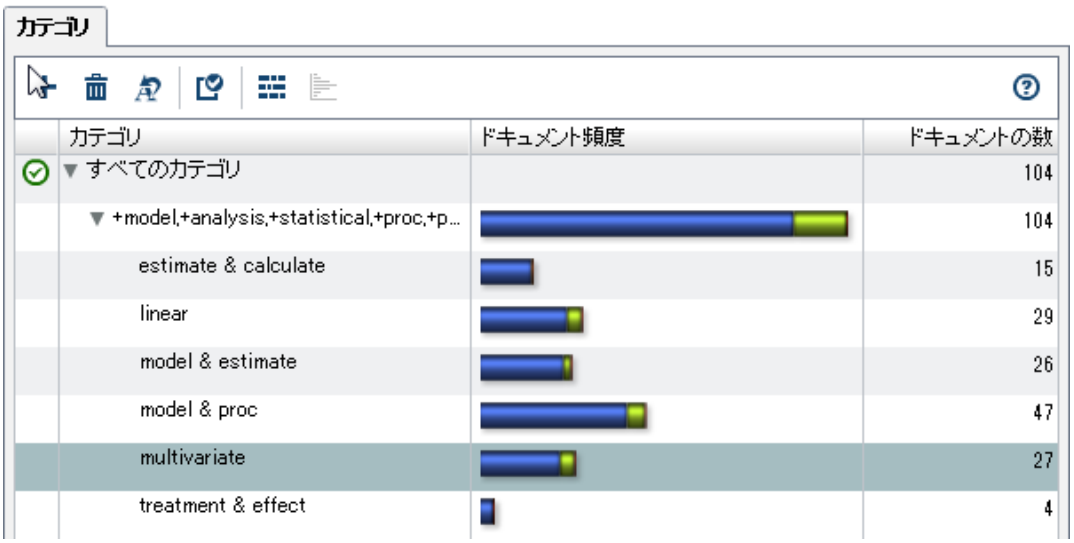
- **カテゴリの実行**

  トピックをカテゴリにまとめるには、**実行**メニューを使用します。各カテゴリのルールは、入力データセットに対して実行されます。

  注: 二値変数値が存在するカテゴリでは最小出現値が分析のターゲットとして使用されます。たとえば、カテゴリ **In stock** で、値 **Yes** が 1200 回、値 **No** が 940 回出現するとします。この場合、**No** がターゲットとして使用されます。

このアクションによって、**ドキュメント比率**および**ドキュメント頻度**列が更新されます。列ビュー間の切り替えには、 および を使用します。





ドキュメント比率および**ドキュメント頻度**列では、次の色が使用されます。

青 - 真陽性
カテゴリルールでは、意図していたドキュメントが取り込まれました。この数を最大にする必要があります。

緑 - 偽陽性
ドキュメントがカテゴリに当てはまらないため、カテゴリルールでは、意図していないドキュメントが取り込まれました。

赤 - 偽陰性
カテゴリルールでは、トピック内で見つかったドキュメントが見落とされました。

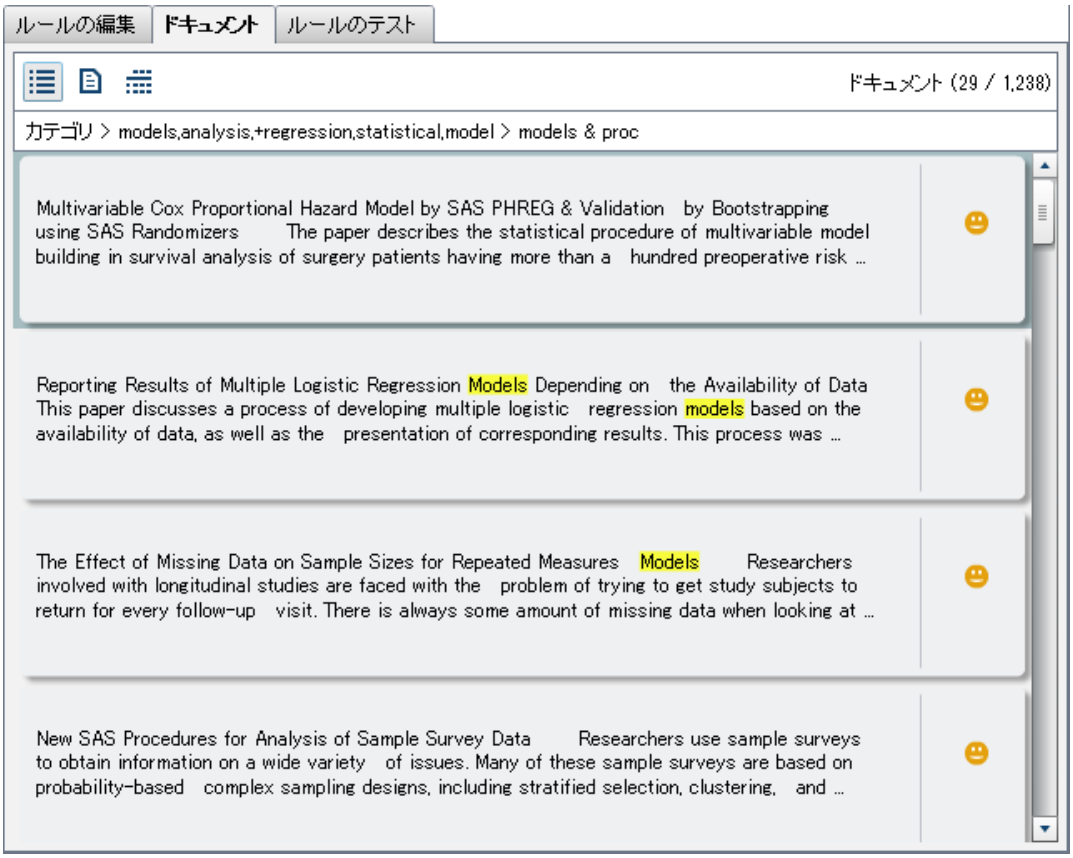
グレー - 一致
グレーのバーは、カスタムまたはインポート済みカテゴリでの一致の統計量を示します。

注: ルールもサブカテゴリルールも含まないカテゴリを実行した場合、その結果は偽陰性として示されます(プロモートされたトピックまたは外部カテゴリ変数から生成されたカテゴリの場合のみ)。

真陽性のドキュメントのみ表示するには、バーの青い部分をクリックし、**ドキュメント**タブをクリックします。偽陽性のみ表示するには、バーの緑の部分をクリックし、**ドキュメント**タブをクリックします。その他も同様です。

- **カテゴリの一致ドキュメントの表示**

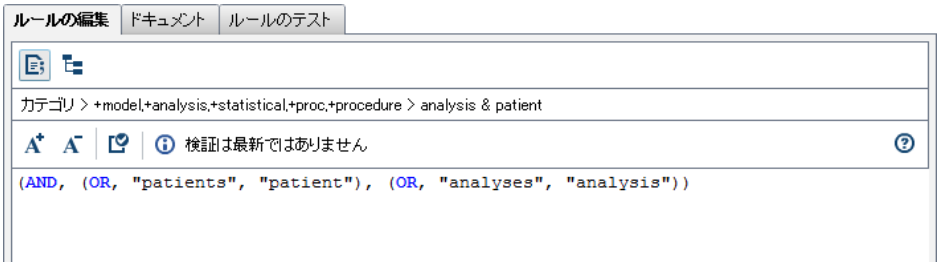
  ドキュメントタブが更新されると、選択内容を満たすドキュメントのみ表示されます。ビュー間の切り替えには、アイコン  を使用します。カテゴリ内でドキュメントのメンバシップを決定するために、強調表示された語が使用されてます。



注: プロジェクトの作成時にセンチメントモデルを指定した場合のみ、各ドキュメントのセンチメントスコアが表示されます。

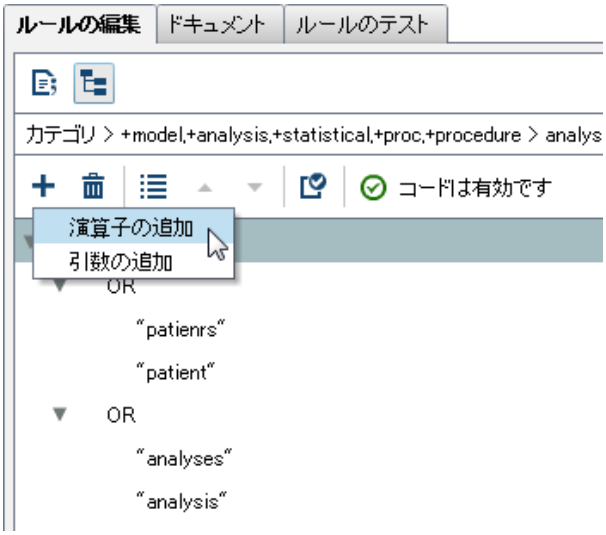
- **カテゴリルールの編集**

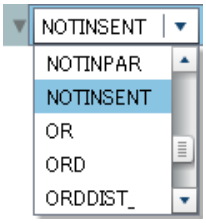
  編集を開始するには、ルールを選択して、**ルールの編集**タブをクリックします。ルールコードアイコンおよびルールツリーアイコンを使用して、編集モデルを切り替えます。次のサンプル画面は、ルールコードビューを示しています。



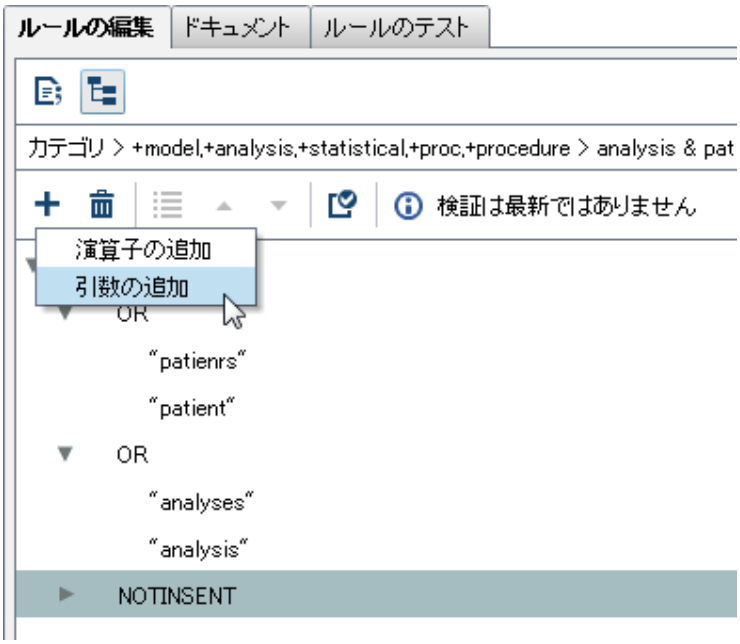
  ルールを入力し、コードを検証することによって、ルールコードを編集します。

ルールツリーを使用すると、画面上で、演算子を選択したり引数を追加したりして、ルールを作成できます。+をクリックすると、演算子または引数をルールに追加できます。次のサンプル画面は、ルールツリービューでの演算子のルールへの追加を示しています。

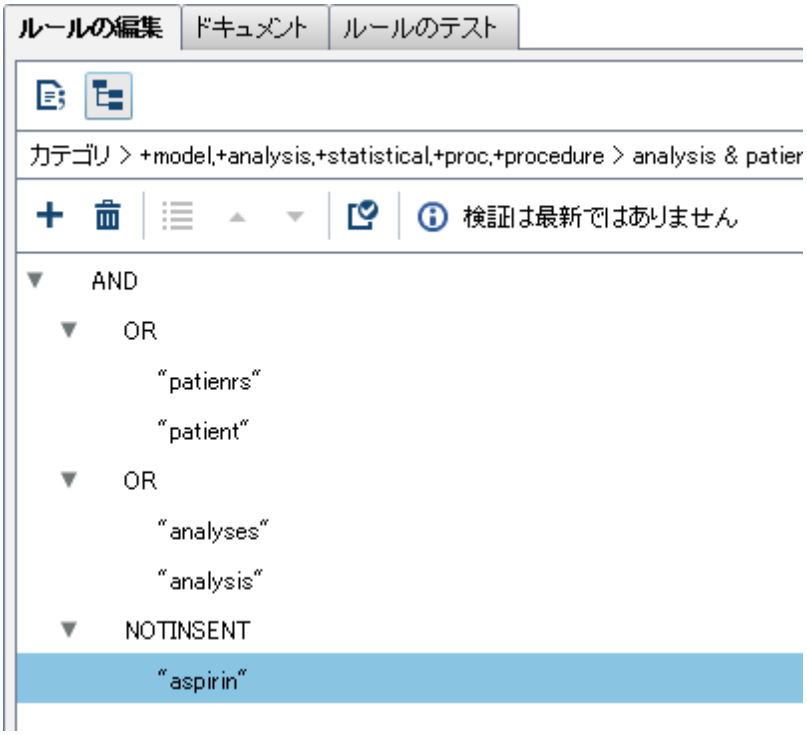




引数を追加するには、+をクリックして**引数の追加**を選択します。



与えられたスペースに引数を入力します。



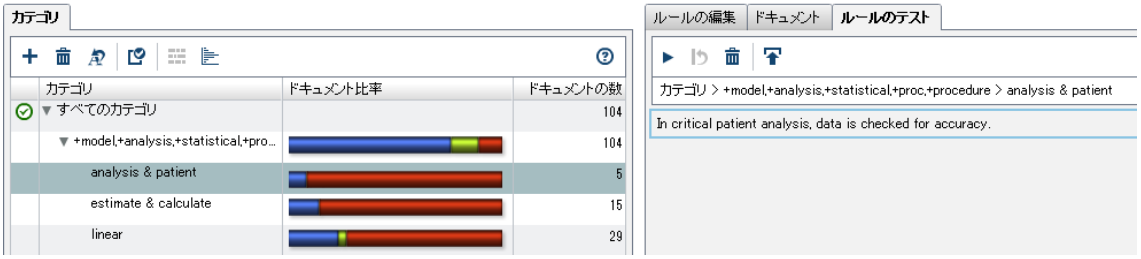
ヒント 変更を加えると、どちらのビューでも 検証は最新ではありません メッセージによって、ルールを検証するように通知されます。

各ルールを個別に検証するには**ルールの編集**タブで をクリックし、すべてのルールを検証するには**カテゴリ**タブで をクリックします。エラーおよびその他のメッセージは**ルールの編集**タブに表示されます。

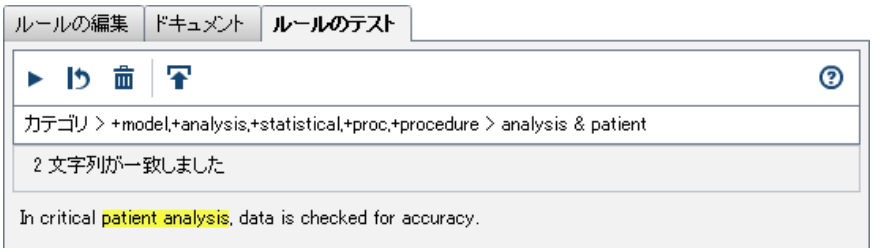
カテゴリルールの作成の詳細については、“カテゴリルールの作成: ブールルール” (81 ページ)を参照してください。

- **カテゴリルールのテスト**

  カテゴリルールをテストするには、ルールを選択して、**ルールのテスト**タブをクリックします。 をクリックしてテストするファイルをアップロードするか、選択したルールに対するテストテキストの単純入力(またはコピーと貼り付け)を行います。 ▶ をクリックすると、ルールがテストされます。



次のサンプル画面では、ルール **Analysis & patient** に対する一致文字列が強調表示されています。



をクリックすると、強調表示がクリアされます。

コンセプトルールの作成: 基本 LITI 構文

コンセプトルールの概要

コンセプトルールは、LITI (言語解釈およびテキスト解釈)構文を使用して作成されます。コンセプトルールでは、ドキュメントからルールと一致する部分のみを抽出できるように、コンテキストのアイテムを識別します。たとえば、**LaGuardia Airport Comments** という名前のカスタムコンセプトノードを作成して、ドキュメントセットで単語 **LGA** を含むすべてのドキュメントを抽出するルールを作成することができます。言い換えれば、コンセプトノード **LaGuardia Airport Comments** に対して表示されるすべてのドキュメントに **LGA** が含まれるということです。

各ドキュメントの一致は別々に評価されます。一致が複数のドキュメントに及ぶことはありません。

インターフェイスやプロパティ設定の使用によるルール編集の詳細については、“コンセプトページ” (35 ページ)を参照してください。 ルールの種類のリストについては、“使用するルールの種類について” (58 ページ)を参照してください。

次のリストでは、LITI 構文を使用してコンセプトルールを作成するための基本ガイドラインが提供されます。構文は柔軟であり、したがって、構文要素は非常に多くの方法で組み合わせられます。

- ルールは、ルールの種類(大文字で書かれる)と、それに続くコロンと引数で構成されます。たとえば、ルール **CLASSIFIER:LGA** では、**CLASSIFIER** がルールの種類、**LGA** が引数で、その 2 つがコロンで区切られます。ルール修飾子を使用すると、一致セットをさらに改良できます。ルール構文はルールの種類によって大きく変わります。基本構文は、 表 3.2 (58 ページ) および 表 3.3 (61 ページ)の各ルールの説明に記載されています。 ルール修飾子のリストについては、“ルール修飾子の追加” (62 ページ)を参照してください。
- 1 単語としては使用できない説明的なコンセプトルール名を使用します(たとえば、BASEBALLSCORE など)。ルールの種類をプレフィックスとして含めることもできます(たとえば、CONCEPT_BASEBALLSCORE など)。

- 1 つのコンセプトルールで、他のコンセプトノードを 1 つ以上参照できます。また、特定のコンテキスト内でキーワードや要素を認識するルールを作成することもできます。たとえば、文字列 **LGA** が単語 **Airport** の前に出現した場合のみ、その文字列を含むドキュメントを抽出することができます。
- 言語構造を識別するには、ルールの品詞タグを使用します。詳細については、“品詞タグとその他のタグの使用” (70 ページ) を参照してください。
- ルールの精度を高めるには、ブール演算子と近接演算子を使用します。詳細については、“ブール演算子を使用してコンセプトルールとファクトを抽出する” (64 ページ)を参照してください。
- 単語の活用形を返すには、形態的拡張演算子を使用します。
- 代名詞を解決するには、同一指示演算子を使用します。たとえば、代名詞 **he** が **Walt Disney** を指すために使用された場合、コンセプトで基準形(完全形)を指定して返すルールを作成できます。詳細については、“同一指示演算子の使用” (69 ページ)を参照してください。

コンセプトとファクト

ファクト(述語とも呼ばれる)は、一緒に検索して一致させるテキストの関連情報です。

ファクトは、カスタムコンセプト内で識別できます。たとえば、大統領にちなんで名付けられた米国の大学を識別するとします。この場合、**George Washington** を米国大統領(**US_President_Names**)として識別し、さらに **George Washington University** を大統領の名にちなんで名付けられた大学(**UNIVERSITY**)として識別するルールを作成できます。

そのため、**There are countless active student organizations at George Washington University** という文では、文字列 **George Washington** がコンセプト **US_President_Names** と一致し、**George Washington University** が **UNIVERSITY** と一致します。

次の特殊な種類のコンセプトルールを使用して、ファクトを検索できます。

- 述語ルール(PREDICATE_RULE)では、ファクトの検索のためにブール演算子と近接演算子が使用されます。たとえば、ブール演算子と近接演算子を使用して、お互いに一定の語数以内に出現する語を指定できます。次のルールでは、**flag**、**emblem**、または **crest** の 3 語以内に出現する語 **America** (**country** として示される)の出現が識別されます。

PREDICATE_RULE:(country):(DIST_3,"_country{America}",

(OR, "flag", "emblem", "crest"))

- ファクト内のアイテムの順序が重要な場合は、シーケンスルール(SEQUENCE)を使用します。シーケンスルールでは、ファクト内の各語が指定順序と一致するように、構造を検出できます。

注意! **SAS Contextual Analysis ではファクトルール(SEQUENCE および PREDICATE_RULE)の作成およびテストを行えますが、プロジェクトの実行時にはプロジェクトのドキュメントに適用されません。結果的に、ドキュメントビュー、トピック、および自動生成ルール内の一致ファクトを参照できません。** 一致ファクトルールを取得するには、次のオプションから 1 つをを選択します。(1) プロジェクトのコンセプトスコアコード機能を使用します。詳細については、“コードの表示とダウンロード” (22 ページ)を参照してください。(2) SAS Enterprise Content Categorization Server で使用するために LITI バイナリファイル(ファクトルールを含む)を配置します。詳細については、http://support.sas.com/notes/index.html で使用可能な SAS Contextual Analysis を参照してください。

使用するルールの種類について

コンセプトとファクトを抽出するために、いくつかの別個の種類のルールが存在します。各カスタムコンセプトまたはファクトに 2 つ以上のルールを指定できます。ルールの種類を理解して、最も目的に合う一致を効率的に生成する種類を選択できるようにすることが重要です。

注: 次の表に記載されたコンセプトルール構文では、< >がオプション構文要素を示します。*italics* のアイテムは、文字列やコンセプトノードなど、指定が必要な値を示します。

表 3.2 コンセプトの抽出に使用されるルールの種類をリストにしています。ルールの各種類の使用方法についての簡単な説明が、基本構文と一緒に記載されています。コンセプトルール構文の例については、“コンセプトルールの種類: 例” (78 ページ)を参照してください。

表 3.2 コンセプト抽出ルールの概要

ルールの種類	説明と基本構文

CLASSIFIER	コンテキストで一致させる 1 つの語または文字列を識別します。たとえば、コンセプト定義で、特定の空港名コードを含む CLASSIFIER ルールを作成できます。この場合、空港名コードを含むテキスト部分が、CLASSIFIER ルールと一致していると考えられます。 一致文字列に対して基準(完全)形を返すには、オプション引数 <,*information*>を使用します。引数のカンマ(,)の後に基準形を入力します。 **CLASSIFIER**:*string* <, *information*>
CONCEPT	他のコンセプトを参照することによって関連情報を識別します。たとえば、ある米国空港の名前と場所を含むドキュメントを取り込むには、ルールの種類 CONCEPT を定義で作成します。ルールの種類 CONCEPT では、ルールの種類 CLASSIFIER をその名前で参照し、それによって空港名コードのリストにアクセスできます。 CONCEPT はルールの種類です。一般的な意味の“コンセプト”と混同しないようにしてください。 **注:** ルールで参照しているコンセプトも、文字列として一致させます。たとえば、ルール **CONCEPT:SCORE** では、文字列 **SCORE** を一致させます。したがって、1 単語としては使用できないコンセプト名を使用することをお勧めします(たとえば、BASEBALLSCORE など)。 **CONCEPT**:<PRIORITY=*n*>:*argument-1*...<*argument-n*> ここでは、*argument* に、コンセプト名、ルール修飾子、または文字列を指定できます。
C_CONCEPT	指定コンテキストで発生した一致のみを返します。たとえば、大学教授の名前を含む完全一致を抽出するために、一致した名前の前に単語 **Professor** が付いている場合のみ姓を識別するコンセプト(事前定義済み)で一致を識別する C_CONCEPT ルールを作成できます。 **注:** このルールの種類には、**_c** 修飾子が必要です。 **C_CONCEPT**:<*argument*> _c{*argument*}<*argument*> ここでは、*argument* に、コンセプト名、ルール修飾子、または文字列を指定できます。

CONCEPT_RULE	ブール演算子と近接演算子を使用して、一致を決定します。ブール演算子のリストについては、“カテゴリルール用のブール演算子と近接演算子” (82 ページ)を参照してください。 **注:** このルールの種類には、**_c** 修飾子が必要です。引用符(")で、一致させる文字列を囲む必要があります。**_c{}**で囲める引数は(OR 演算子内でない限り)1 つだけです。この引数は、一致が返される際に強調表示されます。引用符内に記述されるその他の引数によって一致のコンテキストが提供されます。一致を実行するためにはこれらの引数が存在する必要があります。 **CONCEPT_RULE**:(<*Boolean-rule-1*>...<*Boolean-rule-n*> ここでは、*Boolean-rule* を *n* 回ネストすることができます。それには次のように記述します。 *Boolean-operator* “_c{*argument-1*}”,<“*argument-2*”>...<“*argument-n*”>)
NO_BREAK	文字列全体が見つかった場合のみ一致が発生するようにすることで、部分一致を防ぎます。たとえば、アイテム **National Gallery of Art** を含むテキストを取り込むとします。この場合、**National Gallery of Art** という文字列全体を一致させ、**Gallery** や **Art** を個別アイテムとして扱うことがないようにするルールを作成します。 **注:** このルールの種類には、**_c** 修飾子が必要です。 **注:** NO_BREAK は、ルールの出現場所や、ルールが有効か無効かには関係なく、分類全体にわたって適用されます。 **NO_BREAK**: _c{*argument*} ここでは、*argument* に、コンセプト名または文字列を指定できます。
REGEX	電話番号、ナンバープレートの番号と文字の組み合わせ、**merry-go-round** のような単語ペアなど、数字と文字で表現できるテキスト情報の繰り返しパターンを識別します。たとえば、表現 **32,768** と一致させる REGEX ルールを作成できます。詳細については、“正規表現(Regex)の使用” (73 ページ)を参照してください。 **REGEX**:*regular-expression*
REMOVE_ITEM	1 つの単語が 2 つ以上のコンセプトの固有識別子である場合に、正しい一致が行われるようにします。たとえば、Arizona **Cardinals** フットボールチームと St. Louis **Cardinals** 野球チームを区別するルールを作成できます。間違った一致を除去するために、各一致のコンテキストが使用されます。 **注:** このルールの種類には、**_c** 修飾子と ALIGNED 演算子が必要です。引用符(")で、一致させる文字列を囲む必要があります。 **REMOVE_ITEM**:(ALIGNED, “_c{*concept name*}”, <“*argument*”> ここでは、*argument* に、コンセプト名または文字列を指定できます。

表 3.3 ファクトの抽出に使用されるルールをリストにしています。ルールの各種類の使用方法についての簡単な説明が、基本構文と一緒に記載されています。

表 3.3 ファクト抽出ルールの概要

ルールの種類	説明と基本構文
PREDICATE_RULE	テキストで識別するファクトの定義に役立ちます。ファクトの詳細については、“コンセプトとファクト” (57 ページ)を参照してください。 **PREDICATE_RULE**:(*argument-name-1*... <*argument-name-n*>): (*Boolean-rule-1*...<*Boolean-rule-n*>) ここでは、*argument-name* はファクト一致のために指定する名前を指します。 また、ここでは、*Boolean-rule* を *n* 回ネストすることができます。それには、次のように記述します。 (*Boolean-operator*, “*_argument-name* {*argument*}”, ... “<*_argument-name*>{<*argument*>}”) ルールの種類 PREDICATE_RULE には、引数が必要です。常に順序を指定する必要はないので、ルールの種類 SEQUENCE より柔軟です。
SEQUENCE	ファクトが指定順序で出現する場合に、ドキュメントのファクトを識別します。ファクトの詳細については、“コンセプトとファクト” (57 ページ)を参照してください。 **SEQUENCE**:(*argument-name-1*... <*argument-name-n*>):*_argument-name-1*{*argument*} <*_argument-name-n argument*> ここでは、*argument-name* はファクト一致のために指定する名前を指します。 また、ここでは、*argument* は 1 つ以上のコンセプトノードの名前を指します。 **注:** この構文は、最も簡単な形式で記述されています。コンセプトルール一致のために追加の修飾子や引数を挿入できます。 ルールの種類 SEQUENCE には、引数が必要です。指定される *argument-names* の数は、*_argument-names* の数と一致させる必要があります。

句読点の使用

句読点を使用して、CLASSIFIER と CONCEPT を除くすべてのルールの種類の一致を修飾します。

Colon :
ルールの種類とタグを区切ります。コロンを使用する場合:

- コンセプトルールの種類の後(たとえば、**CLASSIFIER:**など)
- SEQUENCE または PREDICATE_RULE 定義の引数リストとルールリストの間
- 品詞タグの前(たとえば、**:Prep** など)

Comma ,
コンセプトルール定義の要素(引数など)を区切ります。カンマと次の要素の間にスペースを追加します。また、PREDICATE_RULE 定義で論理演算子を区切るためにも使用します。

Single space
ルールの種類 CONCEPT、CONCEPT_RULE、および C_CONCEPT で、文字列、コンセプトノード名、品詞タグ、ルール修飾子を区切ります。

Quotation marks “ ”
ルールの種類 CONCEPT_RULE、REMOVE_ITEM、および PREDICATE_RULE で、コンセプトノード名と文字列を囲みます。

Parentheses ( )
ルールの種類 CONCEPT_RULE、REMOVE_ITEM、SEQUENCE、PREDICATE_RULE で、要素をグループ化します。

Square braces [ ]
ルールの種類 REGEX で要素をグループ化します。

Curly braces { }
一致として返される情報を区切ります。一部のルールの種類では、中かっこ**{}**を丸かっこ**()**と組み合わせて使用できます。

ルール修飾子の追加

いくつかの種類のコンセプトルール修飾子によって、ルールの一致機能を高められます。 表 3.4 および 表 3.5 では、使用可能なルール修飾子の種類をリストにして、それがどのルールの種類で使用できるかを示しています。

表 3.4 コンセプトルール修飾子と関連するルールの種類

修飾子	CLASSIFIER	CONCEPT	C_CONCEPT	CONCEPT_RULE
コメント	X	X	X	X
コンテキスト(_c)			X (必須)	X (必須)
単語(_w)		X	X	X
先頭が大文字の単語(_cap)		X	X	X
複数一致記号(>)			X	X
形態的拡張記号(@、@A、@N、@V)		X	X	X
ブール演算子と近接演算子				X
品詞タグ		X	X	X
エクスポート機能	X			
同一指示記号(_ref、_P、および_F)		X	X	X
正規表現(Regex)				
事前定義済みコンセプト		X	X	X

表 3.5 コンセプトルール修飾子と関連するルールの種類 (続き)

修飾子	REMOVE_ITEM	NO_BREAK	SEQUENCE	PREDICATE_RULE	REGEX
コメント	X	X	X	X	
コンテキスト(_c)	X (必須)	X (必須)			
単語(_w)	X	X	X	X	

先頭が大文字の単語(_cap)	X	X	X	X	
>記号					
形態的拡張記号(@、@A、@N、@V)	X	X	X	X	
ブール演算子と近接演算子				X	
品詞タグ	X	X	X	X	
エクスポート機能					
同一指示記号(_ref、_P、および_F)					
正規表現(Regex)					X (必須)
事前定義済みコンセプト	X	X	X	X	

ブール演算子を使用してコンセプトルールとファクトを抽出する

表 3.6 コンセプトルールを作成してファクトを識別する際に使用できるブール演算子をリストにしています。

表 3.6 コンセプトルールとファクトを抽出するためのブール演算子

演算子	説明

ALIGNED	2 つの引数を取ります。ドキュメントに両方の引数が存在する(並んでいる)場合に一致を返します。ルールの種類 REMOVE_ITEM、CONCEPT_RULE、および PREDICATE_RULE と一緒に使用されます。たとえば、次のルールでは、**LOC** コンセプトノードのルールの一致が、**PERSON** コンセプトノードのルールとも一致する場合、**LOC** の一致を削除するように指定します。 REMOVE_ITEM:ALIGNED, ("_c{LOC}", "PERSON")
AND	1 つ以上の引数を取ります。どのような順序でも、すべての引数がドキュメントに出現する場合に一致します。たとえば、次のルールでは、**King Louis XIV** の一致は、**France** を含むドキュメントに発生した場合に返されます。 CONCEPT_RULE:(AND, "_c{King Louis XIV}", "France")
DIST_*n*	(距離) *n* の値と 2 つ以上の引数を取ります。順序に関係なく、すべての引数同士が *n* 語以内に出現する場合に一致します。たとえば、次のルールでは、フレーズ **the picture with the best lighting** で一致が返されます。 CONCEPT_RULE:(DIST_5, "best", "_c{picture}") **注:** 計算目的では、単語間の距離に、指定語が両方とも含まれるわけではありません。たとえば、フレーズ **best in show** における単語 **best** と **show** の距離は 2 語です。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
NOT	1 つの引数を取ります。引数がドキュメントに出現しない場合に一致します。AND 演算子と一緒に使用する必要があります。たとえば、次のルールでは、**cinema**、**theater**、または **theatre** がドキュメントに出現するが、**Broadway** は出現しない場合に、一致が返されます。 CONCEPT_RULE: (AND, (OR, "_c{cinema}", "_c{theater}", "_c{theatre}"), (NOT, "Broadway")) **注:** NOT 演算子は、ドキュメント全体にわたって適用されます。AND 演算子に加えて OR 演算子を指定する場合は、OR 引数をかっこで囲む必要があります。
OR	1 つ以上の引数を取ります。少なくとも 1 つの引数がドキュメントに出現する場合に一致します。たとえば、次のルールでは、アイテム **U.S.**、**US**、または **United States** のうち 1 つ以上がドキュメントに出現する場合、一致が返されます。 CONCEPT_RULE:(OR, "_c{U.S.}", "_c{US} ", "_c{United States}") **注:** SAS Contextual Analysis で生成されるルールでは、AND 演算子内で OR 演算子がネストされます。ただし、OR 演算子は単独で使用できます。

ORD	(順序) 1 つ以上の引数を取ります。すべての引数がルールに指定した順序で出現する場合に一致します。たとえば、次のルールでは、**The warranty claim for the washing machine was denied.**という文で一致が返されます。 (ORD, “warranty”, “claim”, “denied”)
ORDDIST_*n*	(順序と距離) *n* の値と 2 つ以上の引数を取ります。すべての引数がルールでの指定と同じ順序で出現し、なおかつ、すべての引数同士が *n* 語以内である場合に一致します。たとえば、次のルールでは、フレーズ **the teacher introduced elementary statistics** で、引数が正しい順序で、かつお互いに 5 語以内に出現するため、一致が返されます。 CONCEPT_RULE:(ORDDIST_5, "elementary", "_c{statistics}") **注:** 計算目的では、単語間の距離に、指定語が両方とも含まれるわけではありません。たとえば、フレーズ **best in show** における単語 **best** と **show** の距離は 2 語です。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
PARA	(段落)どのような順序でも、すべての引数が 1 つの段落に出現する場合に一致します。たとえば、次のルールでは、段落に語 **Manhattan** が含まれ、さらに語 **apartment** も含まれる場合、一致が返されます。(**Manhattan** のみ強調表示されます。) CONCEPT_RULE:(PARA, "_c{Manhattan}", "apartment") **注:** PARA ルールは、段落区切り文字\n\n (改行)、\t\t (タブ)、または<P> (段落)を含むデータセットに適用された場合のみ、正しく機能します。PARA は、**ルールのテスト**タブでは適用できません。また、PARA は、フォルダに含まれるデータにも適用できません。

SENT

(文) 2 つ以上の引数を取ります。どのような順序でも、すべての引数が同じ文に出現する場合に一致します。たとえば、次のルールでは、**Amazon** と **river** が同じ文中に出現する場合のみ、一致が返されます。

```
CONCEPT_RULE:(SENT, "_c{Amazon}", "river")
```

区切り文字を使用して、文のトークン化が行われます。これは、文を単語、フレーズ、記号、その他の意味のある要素(トークン)のいずれかに分割する処理です。ピリオド(.)は必ずしも文末を示すとは限らないので注意してください(たとえば、**Mr. Quackenbush** や **Boston, Mass.**は文の途中に出現することがあります)。文区切り文字のリストを次に示します。

\r\n\r\n	連続する 2 つのキャリッジリターンと改行(Windows で作成されたドキュメントの場合)
\r\n \r\n	スペースで区切られた、連続する 2 つのキャリッジリターンと改行
.<SPACE>	ピリオド(.)とそれに続く ASCII スペース
.\n	ピリオド(.)とそれに続く改行
.\r	ピリオド(.)とそれに続くキャリッジリターン
!	感嘆符
!\n	感嘆符とそれに続く改行
!\r	感嘆符とそれに続くキャリッジリターン
?	疑問符
?\n	疑問符とそれに続く改行
?\r	疑問符とそれに続くキャリッジリターン
.)	ピリオドとそれに続く閉じかっこ
!)	感嘆符とそれに続く閉じかっこ
?)	疑問符とそれに続く閉じかっこ
.”	ピリオドとそれに続く二重引用符

SENT_*n*

(複数文) *n* の値と 2 つ以上の引数を取ります。*n* 文以内の一致が返されます。たとえば、次のルールでは、2 文以内のコンセプトノード **GENDER** および語 **he** の一致が返されます。**GENDER** コンセプトノードに次のルールが含まれているとします。

```
CLASSIFIER: male
```

この場合、次のルールを作成します。

```
CONCEPT_RULE:(SENT_2, "_c{GENDER}", "he")
```

詳細については、SENT 演算子を参照してください。

SENTEND_*n*	(文末) *n* の値と 1 つ以上の引数を取ります。文末の *n* 語以内の一致が返されます。たとえば、**GENDER** コンセプトノードに次のルールが含まれているとします。 CLASSIFIER: female この場合、次のルールでは、文末から 5 語以内のコンセプトノード **GENDER** および語 **she** の一致が返されます。 CONCEPT_RULE:(SENTEND_5, "_c{GENDER}", "she") 詳細については、SENT 演算子を参照してください。 **注:** *n* の値を指定する場合、文末が **0** であると考えます。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
SENTSTART_*n*	(文頭) *n* の値と 1 つ以上の引数を取ります。文頭の *n* 語以内の一致が返されます。たとえば、次のルールでは、文 **The patient experienced breathing difficulty.**で一致が見つかります。: CONCEPT_RULE:(SENTSTART_5, "_c{patient}" "breathing", "difficulty") 詳細については、SENT 演算子を参照してください。 **注:** *n* の値を指定する場合、文頭が **0** であると考えます。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
UNLESS	2 つの引数を取ります。両方の引数が同じドキュメントで一致した場合、指定したパラメータ内のある特定の一致を制限します。ルールの種類 PREDICATE_RULE および CONCEPT_RULE でのみ使用されます。UNLESS 演算子と一緒に指定するのは、ブール演算子 AND、SENT、DIST、ORD、ORDDIST のみです。ブール演算子は、例に示すように、第 2 引数と一緒に記述する必要があります。 たとえば、次のルールでは、単語 **river** は一致に含めず、川の **Mississippi** ではなく州の **Mississippi** に対する一致が返されます。 CONCEPT_RULE:(UNLESS, "river", (SENT, "_c{Mississippi}", "United States")) このルールにより、一致において **river** が **Mississippi** と **United States** の間に出現することはなくなります。 **注:** UNLESS 演算子を使用するルールでコンセプトノードを指定する場合、CLASSIFIER または REGEX ルールのみを含むコンセプトノードを指定します。

同一指示演算子の使用

代名詞やその他の単語を、参照先の語の基準形(完全形)とリンクさせる場合、同一指示演算子(_ref)を使用します。

次のルールを含むコンセプトノード **LEADERS** があるとします。

```
CLASSIFIER:Congressional leaders
```

この場合、**they** がその基準形 **Congressional leaders** を参照できるようにするコンセプトノード **THEY_SAID** を作成できます。両方の形式がドキュメントで一致します。

```
C_CONCEPT:_c{LEADERS} said _ref{they}
```

次の記号を同一指示演算子(_ref)と一緒に使用できます。_ref{*concept*}演算子の後に記号を置きます。

> (複数一致)
同一指示演算子(_ref)で指定された一致の複数インスタンスを検索します。たとえば、ニックネーム **Geri** が発生するたびに名前 **Ms. Geraldine Jones** の基準形を返すとします。>記号によって、名前の基準形が初めて検索された後でこの一致を発生させられます。

```
C_CONCEPT:_c{Ms. Geraldine Jones} _ref{Geri}>
```

_F (後)
同一指示ルールの一致後に発生した一致のみが返されます。サンプル構文:

```
C_CONCEPT:_c{PERSON} as _ref{TITLE}_F
```

_P (前)
同一指示ルールの一致前に発生した一致のみが返されます。サンプル構文:

```
C_CONCEPT:_c{MILITARY BRANCH} as _ref{HONOR}_P
```

エクスポート機能の使用

エクスポート機能を使用すると、CLASSIFIER ルールで出現する語やフレーズの一致を見つけて、1 つ以上のコンセプトにエクスポートできます。この機能は、語やフレーズの条件付き一致に役立ちます。複数のコンセプトから 1 つのコンセプトや 2 つ以上のコンセプトに一致をエクスポートできます。

注: エクスポート機能は、CLASSIFIER ルールと一緒にのみ使用できます。

たとえば、名前 **Sokolov** と一緒に出現する語 **accounts receivable** をすべて見つけて、その一致をコンセプト **AR** にエクスポートするとします。この場合、**ACCOUNT_HOLDER** という名前のコンセプトノードで次のルールを作成できます。

```
CLASSIFIER:[export=AR:accounts receivable]:Sokolov
```

このルールでは、まず語 **Sokolov** を一致させます。その一致が見つかった場合は、ルールによって、ドキュメントに出現するすべての語 **accounts receivable** がチェックされ、すべての一致がコンセプト **AR** に割り当てられます。**ACCOUNT_HOLDER** の一致リストでは、語 **Sokolov** が強調表示されます。**AR** の一致リストでは、語 **accounts receivable** が強調表示されます。ルールを機能させるためには、**accounts receivable** がコンセプトノード **AR** の一致として返される前に、主要語(例では、**Sokolov**)がドキュメントのどこかに存在する必要があるので注意してください。

エクスポート先のコンセプト(例にある **AR** など)は、コンセプトのリストに存在する必要があり、追加ルールを含める(または空にする)ことができます。

品詞タグとその他のタグの使用

品詞タグを使用すると、特定の語を検索するのではなく、検索アイテムが属する品詞によって一致を検索できます。これらのタグは、構文はわかっているけれども、求めるアイテムの正確な表現がわからない場合、役に立ちます。品詞とは見なされないその他のタグ(句読点など)も含まれます。

品詞は出現箇所のコンテキストに影響されやすいため、周辺のテキストによっては、同じ単語でもタグ付けが異なる場合があります。たとえば、単語 **will** は、法助動詞(she will be a big star someday)または名詞(a last will and testament)としてタグ付けできます。

品詞タグの前にはコロン(:)が付きます。タグは大文字と小文字が区別されます。たとえば、ニュース記事における引用の属性を一致させるとします。一致の構文が **Senator from** *state* または **Senator of** *state* として出現することはわかっていますが、上院議員(senator)の名前はわかりません。この場合、次のルールを使用できます。

```
C_CONCEPT:SENATE_TITLE _c{_cap _cap} :Prep STATE
```

このルールでは、**majority leader**、**senator**、**senators** などの単語を含むコンセプト **SENATE_TITLE** と、州名を含むコンセプト **STATE** があると仮定しています。:Prep タグは、

前置詞(たとえば、**from** や **of** など)を示します。C_CONCEPT ルールの一致は、テキスト **Senator Phineas Craymoor from North Carolina took the floor** で発生します。ただし、テキスト **Senators Phineas Craymoor and Garrett Garcia from North Carolina pushed the bill through** では、単語 **and** が前置詞ではないため、一致は発生しません。

表 3.7 英語の品詞タグをリストにしています。他の言語のタグについては、 付録 1, “品詞タグ(英語以外の言語用)” (91 ページ)を参照してください。

表 3.7 品詞タグ(英語用)

品詞タグ	定義	例
:ABBREV	省略形	etc.、Ms、cm
:Acomp	比較級形容詞	cooler、luckier、worse
:Adv	副詞	lyrically、physically
:Asup	最上級形容詞	mellowest、merriest、best
:C	接続詞	when、yet、after、except
:date	日付	2000-02-21、04/03/2012
:digit	数列	2345、234.22、21/234
:Det	限定詞	the、an、every
:F	外来語	facto、klieg、modus
:inc	未知語	slaster、lijer
:Int	間投詞	hah、hello、tallyho
:Md	法助動詞	can、should、will
:N	名詞	cake、love、shoe
:Npl	複数名詞	peas、sheep、shoes
:Num	数	one、twenty、hundred

:PN	固有名詞	SAS、Cary、Goodnight
:PossDet	所有限定詞	our、his、my
:PossPro	所有代名詞	mine、yours、hers
:PreDet	前限定辞	quite、such、all
:Prefix	接頭辞	cross、ex、multi
:Prep	前置詞	on、under、across
:Pro	代名詞 関係代名詞	he、one、somebody、me myself、oneself、themselves
:Ptl	不変化詞	away、forward、in
:sep	区切り記号と句読点	; , /
:time	時間	7AM、10:00 pm
:url	ファイル名、パス名、URL	A:/mydir/file.txt, www.sas.com
:V	格変化のない *be*、*do*、または *have* 助動詞 格変化のない動詞 一人称単数動詞	be、do、have go、see、love am
:V3sg	三人称単数の *be*、*do*、または *have* 助動詞 三人称単数動詞	is、does、has goes、sees、loves
:Ving	現在分詞形の *be*、*do*、または *have* 助動詞 現在分詞	being、doing、having bucketing、climbing
:Vpp	過去分詞形の *be*、*do*、または *have* 助動詞 過去分詞	been、done、had dashed、factored、gone

:Vpt	過去時制の *be*、*do*、または *have* 助動詞 過去時制動詞	was、were、did、have dashed、factored、went
:WAdv	副詞の *wh*	how、when、whereby
:Wdet	指示限定詞の *wh*	which、what、whatever
:WPossPro	所有限定詞の *wh*	whose
:WPro	名詞の *wh*	whose、what、whoever

正規表現(Regex)の使用

数字と文字を含むテキストで定期的に発生するパターンを特定するには、正規表現(Regex 構文)を使用します。正規表現を使用すると、ナンバープレートの番号(例: ABX-0444)、製造部品の番号(例: TMS1T3B1M5R-23)、ハイフン付きの単語(例: fifty-nine)などのパターンに一致できます。

Regex 構文には、次のガイドラインが適用されます。

- 大かっこ(**[ ]**)内に 1 つ以上の文字を含めて、その指定文字を一致させます。これにより、文字一致に柔軟性が与えられます。たとえば、次のルールでは、**c**、**r**、**a** **s**、または **h** を一致させます。

  REGEX:[crash]

  次のようにプラス記号(**+**)を追加した場合、ルールでは、**rash**、**cash**、 **ash**、**crass** など、指定文字のいずれの組み合わせでも一致します(ただし、**crashpad** や **crashdummy** は一致しません)。

  REGEX:[crash]+

- 文字列内の文字は、大かっこ(**[ ]**)なしで表された場合、順々に一致検索されます。たとえば、次のルールでは、単語 **any** のみが一致します(**anyone** や **anything** は一致しません)。

  REGEX:any

any を含む単語を一致させるには、ルールを変更してアスタリスク(*)を使用すると、**any** に他の文字が伴う(または伴わない)出現語を一致させられます。たとえば、次のルールでは、**any**、**anyone**、**anything**、**Many** が一致します。

REGEX:[A-Za-z]*any[A-Za-z]*

- 一致させる文字の範囲を指定できます。たとえば、次のルールでは、**a** から **f** までの小文字が一致します。

  REGEX:[a-f]

  大文字を追加するには、次のルールを使用します。

  REGEX:[A-Fa-f]

- 文字セットの前にキャレット(^)を挿入すると、一致させない文字(否定文字)を指定できます。たとえば、次のルールでは、**a**、**e**、**i**、**o**、**u** を除くすべての文字、数字、および記号が一致します。

  REGEX:[^aeiou]

- 特別な目的のために予約された文字(メタ文字)は、バックスラッシュ(\)を使用してエスケープさせ、正規表現で文字どおりに一致させる必要があります。メタ文字は、**[**、**]**、**(**、**)**、**?**、*****、**+**、**.**、**-**、****、**|**です。

  たとえば、**[\?]**は、テキストの疑問符**?**と一致します。

- 文字列内の数字は、大かっこ(**[ ]**)なしで表された場合、そのままで一致検索されます。たとえば、次のルールでは、**0125-**で始まって文字で終わる部品番号が一致します。

  REGEX:0125\-[A-Za-z]

- 番号は、大かっこ(**[ ]**)で囲むと、範囲を指定して一致検索されます。たとえば、次のルールでは、**0** から **9** までの間の数字について一致を返します。

  REGEX:[0-9]

Regex 構文で一致検索のために使用される特殊文字は、組み合わせて使用できます。詳細については、 表 3.8 (74 ページ)を参照してください。

表 3.8 正規表現で使用される特殊文字(メタ文字)

文字または式	説明

|	(代替)正規表現 *a* または *b* のどちらかが一致した場合に一致が発生することを示します。例: *a* | *b*
( )	グループ化メカニズム(非記憶)。式で理解しやすくするために使用されます。例: (*ababab*) |*b*)
.	(ワイルドカード)任意の文字に一致。
%	*%*または *percent* に一致
?	0 回または 1 回の出現に一致
*	0 回以上の出現に一致
+	1 回以上の出現に一致
{ }	繰り返しを示す: {*n*}は、正確に *n* 回の出現に一致　{*n,m*}は、*n* 回以上 *m* 回以下の出現に一致 {*n,*}は、*n* 回以上の出現に一致
\a	アラーム(ビープ)
\n	改行
\r	キャリッジリターン
\t	タブ
\f	フォームフィード
\e	エスケープ
\d	数字(**[0–9]**と同じ)
\D	数字以外(**[^0–9]**と同じ)
\w	単語文字(**[a-zA-Z_0–9]**と同じ)
\W	単語文字以外(**[^a-zA-Z_0–9]**と同じ)

\s	空白文字(**[\t\n\r\f]**と同じ)
\S	空白文字以外(**[^ \t\n\r\f]**と同じ)
\xh	16 進数。ここでは、*h* は 16 進文字
\xhh	16 進数。ここでは、*h* は 16 進文字
\0o	8 進数。ここでは、*o* は 8 進数字
\0oo	8 進数。ここでは、*o* は 8 進数字

Regex 構文には、次の制限が適用されます。

- Regex 構文は Perl の正規表現と似たような機能を有していますが、この 2 つは同一ではありません。
- 大かっこ([ ])内に指定された文字、数字、または記号に対する文字一致は、単語レベルでは発生しません。たとえば、次のルールでは、孤立した文字 **x**、**y**、および **z** は一致しますが、単語 **xylitol**、**yes**、または **recognize** に対しては一致は発生しません。

  REGEX:[xyz]

  次のようにプラス記号(+)を追加して複数回の出現(または 1 回の出現)と一致させる場合、ルールは、**xzx**、**yz**、**zyzy** など、指定された文字のいずれの組み合わせにも一致します。

  REGEX:[xyz]+

  ただし、単語レベルの一致は発生しないため、単語 **xxl**、**syzygy**、**diy** には一致しません。
- Regex 表現ではコンセプトを参照できません。
- テキストの一致に対する逆方向参照はサポートされていません。
- 一致が記憶されるグループ化メカニズムとしてのかっこ**( )**はサポートされていません。かっこは一致ルールを明確化するためだけに使用されます。

形態的拡張記号の使用

CLASSIFIER と REGEX を除くすべてのルールの種類で形態的拡張を使用できます。たとえば、単語 **breathe** をすべての動詞形(**breathes** と **breathing** を含む)に拡張するには、引数に対して次の構文を使用します。**“breathe@V”**

表 3.9 コンセプトルールの形態的拡張記号

記号	説明
@	コンセプトルールを拡張して、引数の単語のすべての屈折形を一致させます。たとえば、引数**“wonder@”**では、一致 **wonder**、**wonders**、**wondered**、**wondering** などが返されます。 **注:** SAS Contextual Analysis が認識しない単語に@を適用した場合、拡張は発生しません。@の前に指定された正確な文字列のみが一致します。たとえば、**“grath”**は拡張されません。このルールでは文字列 **grath** のみ一致が返されます。
@A	コンセプトルールを拡張して、引数の単語の屈折比較級および最上級形容詞形を一致させます。たとえば、引数**“happy@A”**では、一致 **happier** および **happiest** が返されます。 **注:** 形容詞ではない単語に@A を適用した場合、拡張は発生しません。
@N	コンセプトルールを拡張して、引数の単語のすべての屈折名詞形を一致させます。たとえば、引数**“quality@N”**では、一致 **quality** および **qualities** が返されます。 **注:** 名詞ではない単語に@N を適用した場合、拡張は発生しません。
@V	コンセプトルールを拡張して、引数の単語のすべての屈折動詞形を一致させます。たとえば、引数**“transfer@V”**では、一致 **transfer**、**transfers**、**transferred**、**transferring** が返されます。 **注:** 動詞ではない単語に@V を適用した場合、拡張は発生しません。

コメントの追加

CLASSIFIER ルールなど、連続行に表示される個別ルールがあるルール定義に、コメントを挿入できます。コメントは行末まで続きます。コメントは、次のように記述します。

 # comment text

注: ポンド文字(#)はコメントを示します。ルール定義で#を一致させる場合は、エスケープ文字としてバックスラッシュ(\)を#の前に使用します。(例: 表現 **99\#**では、文字列 **99#**の一致が試みられます。)

ヒント ルールをコメントアウトするには、ルールを含む行の先頭にポンド文字(#)を挿入します。

コンセプトルールの種類: 例

例の構文について考察し、さまざまな種類のコンセプトルールの作成方法を理解します。

CLASSIFIER
: 例: 米国空港名コードを含むドキュメントを抽出するために、次の CLASSIFIER ルールを含む、**US_AIRPORTS** という名前のコンセプトノードを作成します。

```
CLASSIFIER:BUF
CLASSIFIER:BUR
CLASSIFIER:BVK
```

これにより、空港名コード **BUF**、**BUR**、または **BVK** のうち 1 つ以上が一致するドキュメントで、**US_AIRPORTS** に対する一致が返されます。

CONCEPT
: 例: フライト到着情報を含むドキュメントを抽出するために、コンセプトノード **ON_TIME_ARRIVALS** を作成します。**ON_TIME_ARRIVALS** のルール定義には、ルールの種類 CONCEPT が含まれます。ルールの種類 CONCEP では、コンセプトノード **US_AIRPORTS** を参照することによって、空港名コードを検出できます。コンセプトノード **ON_TIME_ARRIVALS** のルール定義は次のとおりです。**CONCEPT:at US_AIRPORTS on time** (ここでは、**US_AIRPORTS** に、米国空港名コードを識別する CLASSIFIER ルールが含まれます)。

C_CONCEPT
: 例: 大学教授の名前を含むドキュメントを抽出するために、**PROFESSORS** という名前の C_CONCEPT ルールを作成し、その定義には、**C_CONCEPT: Professor _c{FIRSTNAME LASTNAME}**が含まれます。このルールは、**FIRSTNAME** と **LASTNAME** (事前定義済み)が見つかった場合、その前に単語 **Professor** が付くときのみ、一致が返されることを

示します。修飾子**_c** を使用し、一致させる引数を中かっこ(**{}**)で囲むことで、一致のコンテキストを提供します。

ルール修飾子**_c** は、指定コンセプトノードのコンテキスト内で一致が発生することを示します。

NO_BREAK

例: **National Gallery of Art** を抽出するとします。そこで、CLASSIFIER ルール **National Gallery of Art** を含むコンセプトノード **US_ART_GALLERIES** を定義しました。また、CLASSIFIER ルール **Art** を含む、**CLASS_TYPES** というコンセプトノードも存在します。次のルールを作成すると、**CLASS_TYPES** での部分一致を防ぎ、**National Gallery of Art** の文字列全体を一致します。**NO_BREAK:_c{US_ART_GALLERIES}**

ルール修飾子**_c** は、別のコンセプトノードのコンテキスト内で一致が発生することを示します。

注: NO_BREAK は、ルールの出現場所や、ルールが有効か無効かには関係なく、分類全体にわたって適用されます。

REMOVE_ITEM

例: 野球チーム St. Louis **Cardinals** は抽出しますが、フットボールチーム Arizona **Cardinals** は抽出しないとします。ルール **CLASSIFIER:Cardinals** を含む、**FOOTBALL** という名前のコンセプトノードがあります。ルール **CLASSIFIER:Cardinals** を含む、**BASEBALL** という名前の別のコンセプトノードもあります。次のルールでは、野球チームの一致のみが返されます。

REMOVE_ITEM:(ALIGNED, "_c{FOOTBALL}", "BASEBALL")

注: ルールの種類 REMOVE_ITEM は、使用されているコンセプトノード外にも影響を及ぼす可能性があります。この場合、そのルールが FOOTBALL ルールの一致に影響を及ぼす可能性があります。これは、そのルールでアイテムの削除が指定されているためです。

REGEX

例: テキストの整数(**1**、**23**、**456** など)を抽出するには、ルール

```
REGEX:[0-9]+
```

を使用します。

このルールでは、小数点なしで、1 つ以上の連続する数字が出現する必要があります。

例: **392.55**、**45.25**、**0,987654321** など、小数点表記を使用する数を抽出するには、次のルールを使用します。

REGEX:[0-9]+[,\.][0-9]+

このルールでは、0 から 9 までの任意の数字の後に(任意の組み合わせの)任意の数の数字、カンマ、またはピリオドが続き、数字で終わる一致が返されます。

Regex ルール作成の詳細については、“正規表現(Regex)の使用” (73 ページ)を参照してください。

CONCEPT_RULE

例: Amazon 社を抽出し、Amazon 川は抽出しないとします。このルールを使用すると、**company** の 3 語内の社名が返されますが、ドキュメントに自然関連の単語がある場合は返されません。

CONCEPT_RULE:(AND, (DIST_3, "_c{COMPANY}", "company"), (NOT, "NATURE"))

SEQUENCE

例: 姓名とミドルネームのリストから姓名のみを抽出するとします。この場合、SEQUENCE ルールを使用すると、引数 **first** および **last** を定義できます。これらの引数を使用すると、コンセプトノード **FIRST_NAME**、**MIDDLE_NAME**、および **LAST_NAME** で一致が発生しますが、**FIRST_NAME** および **LAST_NAME** でのみ一致が返されます。

SEQUENCE:(first, last): _first{FIRST_NAME} MIDDLE_NAME _last{LAST_NAME}

PREDICATE_RULE

例: 会社をその製品に一致するとします。この場合、次の PREDICATE_RULE を使用できます。ここでは、コンセプトノード **COMPANY** に、社名をリストにする CLASSIFIER ルールが含まれ、コンセプトノード **PRODUCTS** に、製品をリストにする CLASSIFIER ルールが含まれていると仮定しています。アイテムは同じ文中に出現する必要があります。

PREDICATE_RULE:(company, product):(SENT, "_company{COMPANY}",

"produces", "_product{PRODUCTS}")

カテゴリルールの作成: ブールルール

カテゴリルールの概要

カテゴリルールが解決して真または偽になります。“真”の結果は一致です。ブールルールは、ブール演算子および近接演算子、引数、ならびに修飾子を使用して、カテゴリー致に必要な条件を定義します。カテゴリルールは、LITI ルールよりも作成が簡単なので、データから特定の情報を抽出する必要がない場合はお勧めです。演算子のリストについては、 表 3.10 (82 ページ)を参照してください。

ブールルールには次の構文を使用します。

(***OPERATOR***, <*argument1*>, <*argument2*>, ...)
ここでは、語、文字列、またはネストされたルールを引数に指定できます。

構文の一般ルール:

- ブール演算子と近接演算子はかっこで囲み、カンマで区切ります。引数内の文字列は引用符(“ ”)に入れます。例: (AND, “holiday”, “vacation”)
- ルールはネストできます。例: (AND, (OR, “courage”, “courageous”), (OR, “brave”, “bravery”))
- あるカテゴリを別のカテゴリから参照するには、*tmac syntax* (_tmac)という特殊構文を使用します。詳細については、 “_tmac を使用したカテゴリ参照” (89 ページ)を参照してください。
- コンセプトノード名は、カテゴリルールで参照できます。コンセプトノード名を参照する場合、すべてのコンセプト一致がカテゴリでも一致します。コンセプトノード名は、大かっこ(**[]**)と引用符(**“ ”**)で囲む必要があります。たとえば、カテゴリルールでコンセプトノード **GAME_SHOWS** を参照するには、ルール(OR, “[GAME_SHOWS]”)を作成します。

  注: カテゴリで名前が指定されたコンセプトノードでは、カテゴリ外で実行されたコンセプトよりも多くの一致が返される可能性があります。カテゴリでは、コンセプトの一致は“全一致”法に基づいています。ここでは、テキストで見つかった全一致が返されます。これとは対照的に、コンセプトでは、一致は“最適一致”法に基づいていま

す。最適一致法では、あるコンセプトに一致するテキストが別のコンセプトに一致するテキストと重複する場合に検出をします(たとえば、**New York** に一致するコンセプトと、**New York City** に一致する別のコンセプトなど)。コンセプト一致が重複していて、なおかつ最適一致法が使用されている場合、優先順位に最も大きな数が割り当てられているコンセプトのみ返されます(1 が最も小さい)。2 つ以上のコンセプトに同じ優先順位が割り当てられている場合、SAS Contextual Analysis では、“最長一致”法を使用して一致を選択します。

- カテゴリで名前が指定されたコンセプトの有効化または無効化ステータスは、カテゴリ一致中は無視されます。結果的に、コンセプトは、事前に無効化されたかどうかには関係なく、すべて有効であるかのように処理されます。
- 特殊記号を使用すると、含めるルール、ワイルドカード、大文字と小文字の区別などを変更できます。記号のリストについては、表 3.11 (88 ページ)を参照してください。

注: XPath 表現はサポートされていません。

カテゴリルール用のブール演算子と近接演算子

表 3.10 カテゴリルールの作成に使用できるブール演算子と近接演算子のリストを示します。

表 3.10 カテゴリルール用のブール演算子と近接演算子

演算子	説明
AND	1 つ以上の引数を取ります。どのような順序でも、すべての引数がドキュメントに出現する場合に一致します。たとえば、ルール(AND, “King”, “Louis”, “XIV”)では、**King**、**Louis**、および **XIV** がすべてドキュメントに出現する場合に一致が返されます。

DIST_*n*	(距離) *n* の値と 2 つ以上の引数を取ります。順序に関係なく、すべての引数同士が *n* 語以内に出現する場合に一致します。たとえば、ルール (DIST_5, “best”, “picture”)では、フレーズ **the picture with the best lighting** で一致が返されます。 **注:** 計算目的では、単語間の距離に、指定語が両方とも含まれるわけではありません。たとえば、フレーズ **best in show** における単語 **best** と **show** の距離は 2 語です。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
END_*n*	(ドキュメントの最後から) *n* の値と 1 つ以上の引数を取ります。ドキュメントの最後から *n* 語以内に引数が出現する場合に一致します。たとえば、ルール(END_35, “conclusion”)では、ドキュメントの最後の単語から 35 語以内に **conclusion** が見つかった場合に一致が返されます。 **注:** ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
MIN_*n*	(最小) *n* の値と 1 つ以上の引数を取ります。指定した引数が(どのような順序でも)*n* 以上ドキュメントに含まれる場合に一致します。たとえば、ルール(MIN_2, “Hollywood”, “tinseltown”, “movies”)では、**Hollywood** と **movies** がドキュメントに出現する場合、一致が返されます。ただし、**Hollywood** が 2 回出現し、他の引数が出現しない場合、一致は発生しません。
MINOC_*n*	(最小出現) *n* の値と 1 つ以上の引数を取ります。ドキュメントで(どのような順序や組み合わせでも)指定引数の出現回数が *n* 以上の場合に一致します。たとえば、ルール(MINOC_2, “Hollywood”, “tinseltown”, “movies”)では、**Hollywood** と **movies** がドキュメントに出現する場合、一致が返されます。また、**Hollywood** が 2 回出現し、他の引数が出現しない場合も、一致が発生します。
MAXOC_*n*	(最大出現) *n* の値と 1 つ以上の引数を取ります。ドキュメントで(どのような順序や組み合わせでも)引数の出現回数が *n* 以下の場合に一致します。スパムドキュメントのフィルタリングに役立ちます。たとえば、ルール(MAXOC_8, “savings”, “offer”, “best”)では、**savings** がドキュメントに 6 回発生した場合、一致が返されます。また、ドキュメントに **offer** が 6 回と **best** が 2 回出現した場合も一致が発生します。

MAXPAR_*n*	(最大段落) *n* の値と 1 つ以上の引数を取ります。どのような順序でも、すべての引数がドキュメントの最初から *n* 段落以内に出現する場合に一致します。たとえば、ルール(MAXPAR_4, “seasonal", “herbs”, “native”)では、4 段落目に **seasonal**、2 段落目に **herbs** と **native** が出現する場合、一致が返されます。 **注:** MAXPAR ルールは、段落区切り文字(\n\n)を含むデータセットに適用された場合のみ、正しく機能します。MAXPAR は、**ルールのテスト**タブでは適用できません。また、MAXPAR は、**カテゴリ**タブで、フォルダに含まれるデータに適用することもできません。
MAXSENT_*n*	(最大文) *n* の値と 1 つ以上の引数を取ります。どのような順序でも、すべての引数がドキュメントの最初の *n* 文以内に出現する場合に一致します。たとえば、ルール(MAXSENT_4, “weight loss”, “plan”)では、**weight loss** と **plan** がドキュメントの 3 文目に出現する場合、一致が返されます。文区切り文字のリストについては、SENT 演算子を参照してください。
NOT	1 つの引数を取ります。引数がドキュメントに出現しない場合に一致します。AND 演算子と一緒に使用する必要があります。たとえば、ルール(AND, (OR, “cinema”, “theater”, “theatre”), (NOT, “Broadway”))では、**cinema**、**theater**、または **theatre** がドキュメントに出現して **Broadway** は出現しない場合、一致が返されます。 **注:** NOT 演算子は、ドキュメント全体にわたって適用されます。AND 演算子に加えて OR 演算子を指定する場合は、OR 引数をかっこで囲む必要があります。
NOTIN	(中になし) 2 つの引数を取り、第 1 引数が第 2 引数内に出現しない場合に一致します。たとえば、ルール(NOTIN, “butter”, “peanut butter”)では、**butter** が、名詞句 **peanut butter** 内に出現しない場合に識別されます。次の文で、一致が返されます。**Early American colonists churned their own butter.**

NOTINDIST_*n*	(距離中になし) *n* の値と 2 つの引数を取ります。引数がお互いに *n* 語以内に出現しない場合、またはルールに記述された第 1 引数がドキュメントに出現して第 2 引数は出現しない場合に一致します。たとえば、ルール(NOTINDIST_3 “orange”, “green”)では、**orange** と **green** がお互いに 3 語以内に出現しない場合、または **orange** のみドキュメントに出現する場合、一致が返されます。次の文では、ルールに指定された単語が 4 語以上離れているため、一致が返されます。**How green is my valley, how orange is the sunset?** **注:** 計算目的では、単語間の距離に、指定語が両方とも含まれるわけではありません。たとえば、フレーズ **best in show** における単語 **best** と **show** の距離は 2 語です。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
NOTINPAR	(段落中になし) 2 つ以上の引数を取り、すべての引数がドキュメント内に出現しますが、別々の段落に出現する場合に一致します。たとえば、ルール(NOTINPAR, “China”, “export”)では、**China** と **export** が、別々の段落(他方の引数は存在しない)に出現する場合、一致が返されます。 **注:** NOTINPAR ルールは、段落区切り文字(\n\n)を含むデータセットに適用された場合のみ、正しく機能します。NOTINPAR は、**ルールのテスト**タブでは適用できません。また、NOTINPAR は、**カテゴリ**タブで、フォルダに含まれるデータに適用することもできません。
NOTINSENT	(文中になし) 2 つ以上の引数を取り、すべての引数がドキュメント内に出現しますが、別々の文に出現する場合に一致します。たとえば、ルール(NOTINSENT, “China”, “trade”)では、**China** と **trade** が、次のように、別々の文(他方の引数は存在しない)に出現する場合、一致が返されます。**China is our biggest partner.The trade it generates is huge.**文区切り文字のリストについては、SENT 演算子を参照してください。
OR	1 つ以上の引数を取ります。少なくとも 1 つの引数がドキュメントに出現する場合に一致します。たとえば、ルール(OR, "U.S.", "US ", "United States")では、アイテム **U.S.**、**US**、または **United States** のうち 1 つ以上がドキュメントに出現する場合、一致が返されます。 **注:** SAS Contextual Analysis で生成されるルールでは、AND 演算子内で OR 演算子がネストされます。ただし、OR 演算子は単独で使用できます。
ORD	(順序) 1 つ以上の引数を取ります。すべての引数がルールに指定した順序で出現する場合に一致します。これは、SENT (または一致範囲を制限するその他の演算子)と一緒には使用できません。たとえば、ルール(ORD, “warranty”, “claim”, “denied”)では、**The warranty claim for the washing machine was denied.**という文で一致が返されます。

ORDDIST_*n*	(順序と距離) *n* の値と 2 つ以上の引数を取ります。両方の引数がルールでの指定と同じ順序で出現し、なおかつ、両方の引数同士が *n* 語以内である場合に一致します。たとえば、ルール(ORDDIST_5, “elementary”, “statistics”)では、フレーズ **the teacher introduced elementary statistics** で一致が返されます。 **注:** 計算目的では、単語間の距離に、指定語が両方とも含まれるわけではありません。たとえば、フレーズ **best in show** における単語 **best** と **show** の距離は 2 語です。ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。
PAR	(段落) 1 つ以上の引数を取ります。どのような順序でも、すべての引数が 1 つの段落に出現する場合に一致します。たとえば、ルール(PAR, “director”, “budget”)では、**director** と **budget** の両方が段落に含まれている場合、一致が返されます。 **注:** PAR ルールは、段落区切り文字(\n\n)を含むデータセットに適用された場合のみ、正しく機能します。PAR は、**ルールのテスト**タブでは適用できません。また、PAR は、**カテゴリ**タブで、フォルダに含まれるデータに適用することもできません。
PARPOS_*n*	(段落位置) *n* の値と 1 つ以上の引数を取ります。どのような順序でも、すべての引数が第 *n*段落内に出現する場合に一致します。たとえば、ルール(PARAPOS_2, “journalists”, “detained”, “overseas”)では、ドキュメントの 2 段落目内に **journalists**、**detained**、および **overseas** が発生する場合、一致が返されます。 **注:** PARPOS ルールは、段落区切り文字(\n\n)を含むデータセットに適用された場合のみ、正しく機能します。PARPOS は、**ルールのテスト**タブでは適用できません。また、PARPOS は、**カテゴリ**タブで、フォルダに含まれるデータに適用することもできません。

SENT	(文) 2 つ以上の引数を取ります。どのような順序でも、すべての引数が同じ文に出現する場合に一致します。たとえば、ルール(SENT, “growth”, “hormone”)では、**Patients who take a growth hormone might experience side effects** という文で一致が返されます。文区切り文字は次のとおりです。 \r\n\r\n 連続する 2 つのキャリッジリターンと改行(Windows で作成されたドキュメントの場合) \r\n \r\n スペースで区切られた、連続する 2 つのキャリッジリターンと改行 .<SPACE> ピリオド(.)とそれに続く ASCII スペース .\n ピリオド(.)とそれに続く改行 .\r ピリオド(.)とそれに続くキャリッジリターン ! 感嘆符 !\n 感嘆符とそれに続く改行 !\r 感嘆符とそれに続くキャリッジリターン ? 疑問符 ?\n 疑問符とそれに続く改行 ?\r 疑問符とそれに続くキャリッジリターン .) ピリオドとそれに続く閉じかっこ !) 感嘆符とそれに続く閉じかっこ ?) 疑問符とそれに続く閉じかっこ .” ピリオドとそれに続く二重引用符
START_*n*	(ドキュメントの先頭から) *n* の値と 1 つ以上の引数を取ります。ドキュメントの先頭から *n* 語以内に引数が出現する場合に一致します。たとえば、ルール(START_22, “infection”)では、ドキュメントの先頭の単語から 22 語以内に **infection** が出現する場合に一致が返されます。 **注:** ハイフンを含む単語は 1 語としてカウントされます(たとえば、**merry-go-round** は 1 語です)。

ブールルールでの記号の使用

表 3.11 のリストにある記号を使用すると、カテゴリ一致のためのブールルールを変更できます。記号は、引数の文字列に対する接尾辞として記述されます。たとえば、単語

breathe をすべての屈折動詞形(**breathes** と **breathing** を含む)に拡張するには、引数に対して次の構文を使用します。**“breathe@V”**

表 3.11 ブールルールで使用される特殊記号

記号	説明
*	(ワイルドカード一致)単語の先頭または末尾に出現する任意の文字に一致します。たとえば、引数**“travel*”**では、一致 **travels**、**traveled**、**traveler**、**traveling** などが返されます。引数**“*room”**は、**bedroom**、**cloakroom**、**ballroom**、**room** などに一致します。
^	(文頭)文頭から検索を開始して一致を探します。たとえば、引数**“^Independent”**では、次の文で一致が返されます。**Independent research was conducted.** **注:** 一致を考慮する場合は特に、トークン(単語、フレーズ、記号、その他の意味のある要素)を入力する必要があります。たとえば、****In this case** を検索する場合は、引数**“^**In this case”**を使用します。バックスラッシュ(\)をアスタリスク(*)のエスケープ文字として使用して、アスタリスクがワイルドカードとして扱われないようにしている点にも注意してください。
$	(文末)文末から検索を開始して一致を探します。たとえば、引数**“deleted. $”**では、次の文で一致が返されます。**All the files were hastily deleted.** **注:** 一致を考慮する場合は特に、トークン(単語、フレーズ、記号、その他の意味のある要素)を入力する必要があります。たとえば、引数**“deleted$”**では、末尾のピリオド(.)が指定されていないため、**All the files were hastily deleted.**という文で一致は発生しません。
@	(形態的拡張)カテゴリルールを拡張して、引数の単語のすべての屈折形を一致させます。たとえば、引数**“wonder@”**では、一致 **wonder**、**wonders**、**wondered**、**wondering** などが返されます(ただし **wonderful** では一致は返されません)。 **注:** SAS Contextual Analysis が認識しない単語に@を適用した場合、拡張は発生しません。@の前に指定された正確な文字列のみが返されます。たとえば、**“grath”**は拡張されません。このルールでは文字列 **grath** のみ一致が返されます。

@A	(形容詞の形態的拡張)カテゴリルールを拡張して、引数の単語の屈折比較級および最上級形容詞形を一致させます。たとえば、引数**"happy@A"**では、一致 **happier** および **happiest** が返されます。 **注:** 形容詞ではない単語に@A を適用した場合、拡張は発生しません。
@N	(名詞の形態的拡張)カテゴリルールを拡張して、引数の単語のすべての名詞形を一致させます。たとえば、引数**"quality@N"**では、一致 **quality** および **qualities** が返されます。 **注:** 名詞ではない単語に@N を適用した場合、拡張は発生しません。
@V	(動詞の形態的拡張)カテゴリルールを拡張して、引数の単語のすべての動詞形を一致させます。たとえば、引数**"transfer@V"**では、一致 **transfer**、**transfers**、**transferred**、**transferring** が返されます。 **注:** 動詞ではない単語に@V を適用した場合、拡張は発生しません。
_L	(リテラル一致)リテラル文字列を一致させます。記号を含む文字列を一致させる場合に役立ちます。たとえば、引数**"$USD_L"**では、一致**$USD** が返されます。 **注:** 一致を考慮する場合は特に、トークン(単語、フレーズ、記号、その他の意味のある要素)を入力する必要があります。
:C	(大文字と小文字を区別した一致)大文字と小文字を区別した一致を指定します。たとえば、引数**"Iris_C"**では、一致 **Iris** が返されますが、**iris** は返されません。

_tmac を使用したカテゴリ参照

カテゴリを参照すると、ルールを複製しなくても、既存カテゴリのルールを使用できます。カテゴリルールで既存カテゴリを参照するには、tmac 構文(_tmac)を使用します。既存ルールの定義は、その参照元のカテゴリで処理されます。

カテゴリを参照するには、そのパスを識別する必要があります。すべてのカテゴリパスは、**@Top/**から始まります。そこから、カテゴリ階層に続けてパスを指定できます。

たとえば、**すべてのカテゴリ**の下に次のカテゴリ構造があるとします。

NAME
    FIRST

LAST

カテゴリ **FIRST** は、**@Top/NAME/FIRST** として参照します。

tmac 構文は、ブール演算子と一緒に使用できます。たとえば、**FIRST_NAME** というカテゴリからカテゴリ **FIRST** を参照するとします。**FIRST_NAME** 定義で次のルールを追加できます

(OR,_tmac:"@Top/NAME/FIRST")

後に姓が続く名(FIRST LAST)を適用するには、**COMPLETE_NAME** というカテゴリで次のルールを追加します。

(ORD,_tmac:"@Top/NAME/FIRST",_tmac:"@Top/NAME/LAST")

FIRST と **LAST** に記述された定義が自動的に処理されます。

付録 1

品詞タグ(英語以外の言語用)

品詞タグとその他のタグの概要

英語以外の言語の品詞タグを次に表に示します。品詞とは見なされないその他のタグ(句読点など)も含まれます。コンセプトルールでは、すべてのタグで大文字と小文字が区別され、前にコロン(:)が付きます。英語タグも含めて、詳細については、“品詞タグとその他のタグの使用” (70 ページ)を参照してください。

品詞タグ

中国語

表 A1.1 中国語の品詞タグ

品詞タグ	説明	例
:A	形容詞	俊俏、开心、兇險、凌亂
:ASCII	ASCII 文字	sas、do、happy、day2136456
:C	接続詞	或、与、雖然
:D	副詞	非常、偏偏、稍微、永遠
:digit	数	1051、1.9
:E	間投詞	咦、呸、哦喲
:F	場所/方向	中間、下边、南側
:G	その他の形態素	馨、慚
:H	その他の接頭辞	亚、非
:K	その他の接尾辞	们、者、們
:L	慣用句(成語)	囫囵吞枣、博古通今、一廂情願
:M	数量詞	十、卅、成千上万、上萬、１０５１
:N	名詞	人、桌子、香蕉、枷鎖
:NR	固有名詞、名前	习近平、梁振英、奥巴马

:NS	固有名詞、地理	中国、美國、山東
:NT	固有名詞、組織	北京大学、上汽集團
:NZ	固有名詞、その他	潘婷、劍南春
:O	擬音語	吱呀、叽叽喳喳、劈裏啪啦
:P	前置詞	依照、对于
:Q	分類辞	个、斤、艘、加侖
:R	代名詞	我、他們、这
:S	準国の場所(一般。漢字文化圏内にのみ特有)	地上、上空、高处、內廳
:T	時制句	今天、夜间、十月、去歲
:U	不変化詞	的、了、着
:UNKNOWN	未知語	嫚、禪
:V	動詞	看、认为、彈奏、徵納
:W	句読点または記号	！、。、$、￥
:Y	間投詞的不変化詞	吧、吗、麼

オランダ語

表 A1.2 オランダ語の品詞タグ

品詞タグ	定義	例
:A	形容詞	betrouwbaar、gelukkig、mooi
:ABB	省略形	enz、kg、zgn

:ADV	副詞	eenmaal、hier、nu
:CONJ	接続詞	als、doch、hoe
:DET	限定詞	de、der、een
:digit	数	21
:DNUM	限定詞、数	acht、elf、miljard、duizend
:inc	未知語	xrxx
:N	名詞	geluk、schoonheid
:PFX	接頭辞	anti
:PN	固有名詞	Amerika、Nederland
:PREP	前置詞	met、per、te、van
:PREPDET	前置詞および限定詞縮約形	ten、ter
:PRO	代名詞	alles、beide、hetgeen
:sep	区切り記号または句読点	,
:url	URL	www.sas.com
:V	動詞	helpt、vernieuwt
:VB	不定詞	helpen、vernieuwen
:VE	現在進行形	helpende、vernieuwende
:VH	過去分詞	geholpen、vernieuwd
:XI	古形	hoofde、tijde、voordele

フィンランド語

表 A1.3 フィンランド語の品詞タグ

品詞タグ	定義	例
:A	形容詞	loistava、korkea
:ADV	副詞	ohitse、juuri
:CLX	接語	kinko、pas
:CONJC	等位接続詞	ja、vaan
:CONJS	従位接続詞	ellei、jotta
:date	日付	12-14、2001-12-02
:digit	数	1234、7
:inc	未知語	auttonkkan、eggs
:N	名詞	siltoineen、postiksi
:PN	固有名詞	Pertti、Fazer
:PREP	前置詞	pitkin、kanssaan
:PRO	代名詞	noihin、muussa、ketkä
:sep	区切り記号または句読点	; / +
:time	時間	12:00:00、7PM
:PROP	人称代名詞	sinun、heissä、me
:url	URL	http://www.sas.com
:VB	不定詞動詞	heilahtamassa、heilauttaen、olla

:VC	可能法現在動詞	lähennemme、luvannette
:VE	現在分詞動詞	kumarrettava、ilmaisevaa
:VH	過去分詞動詞	jaettu、ilmaistu
:VJ	直説法過去動詞	meditoitpa、matkattu
:VP	直説法現在動詞	ihastele、hörähdä
:VS	条件法現在動詞	omistautuisi、hehkuisikaan
:VY	命令法動詞	parannuttako、pakkaa

フランス語

表 A1.4 フランス語の品詞タグ

品詞タグ	定義	例
:A	形容詞	comparable、compassionnelle、intraduisibles
:ADV	副詞	plutôt、individuellement
:CONJC	等位接続詞	et、ou
:CONJS	従位接続詞	lorsque、puisque
:DET	限定詞	sa、tes
:digit	数	123、12.3、12.3.2003、12/3/2003
:inc	未知語	analytics
:INTJ	間投詞	tralala、zzz

:N	名詞	zèbre、encyclopédie
:PN	固有名詞	Eurotunnel、Égypte
:PFX	接頭辞	anglo、éco
:PREP	前置詞	après、jusque
:PREPDET	前置詞および限定詞縮約形	aux、du
:PRO	代名詞	lui、ce
:sep	区切り記号または句読点	, .!
:url	URL	http://www.sas.com
:V	動詞	vais、obligez
:Vpp	過去分詞	mangé、relaxée、travaillées
:VB	不定詞	traduire、rompre
:VE	現在分詞	ceignant、tramant
:XI	外来語	vitae、ab

ドイツ語

表 A1.5 ドイツ語の品詞タグ

品詞タグ	定義	例
:A	形容詞	schön、zuverlässig
:ADV	副詞	gern、sehr
:CONJ	接続詞	und、oder
:CPO	複合語(接頭辞のみ)	Lustigkeits

:DET	限定詞	der、eine
:digit	数	21
:DNUM	限定詞、数	fünf、zwölf
:EMP	強調用法/強意語	ganz
:inc	未知語	xrxx
:N	名詞	Schönheit、Zuverlässigkeit
:PFX	接頭辞	Irr、lob
:PN	固有名詞	Mozart
:PN.gen	固有名詞、属格	Nirvanas
:PNG.dat	固有名詞、地理、与格	Niederlanden
:PREDET	前限定辞	manch
:PREP	前置詞	kontra、ober
:PRO	代名詞	er、sie
:PXPRO	代名詞的副詞	heraus
:sep	区切り記号または句読点	,
:url	URL	www.sas.com
:V	動詞	ging、half
:VI	不定(不定詞と分詞)	gehen、helfen

イタリア語

表 A1.6 イタリア語の品詞タグ

品詞タグ	定義	例
:A	形容詞	affidabile、bellissimo、felice
:AVV	副詞	felicemente、rapidamente
:CONG	接続詞	ma、oppure、sebbene
:DET	限定詞	il、la、uno
:digit	数	21
:ESC	間投詞	ah、ahimè
:inc	未知語	Xrxx
:N	名詞	affidabilità、bellezza、felicità
:PN	固有名詞	Roma、Italia
:PRON	代名詞	io、ne、tu
:PREFIX	接頭辞	anti、ri
:PREP	前置詞	con、in、per
:sep	区切り記号または句読点	,
:SUFFIX	接尾辞	anza、issimo
:url	URL	www.sas.com
:V	動詞	andare、vedono
:VGerund	動名詞	andando、vedendo

:VH	定過去	andasse、vedessero
:Vpastpart	過去分詞	andato、visto

日本語

表 A1.7 日本語の品詞タグ

品詞タグ	説明	例
:AJ	形容詞	長い、いい、忙しい、便利だ
:AV	副詞	別に、相変わらず、年年歳歳
:AVC	形式や状態の副詞(述語の形式や状態を示す副詞)	正々堂々、淡々と、きらり
:AVD	程度の副詞	結構、とっても
:AVE	評価の副詞	たまたま、幸い、無論
:AVF	頻度の副詞	次々と、次次に
:AVO	意見の副詞	実は、即ち、すなわち
:AVQ	量の副詞	大方、いくら、半分
:AVS	陳述または宣言の副詞	何でも、多分、絶対
:AVT	時制または相の副詞	徐々に、急遽、直ぐ
:AX	助動詞	らしい、みたいだ、様だ、わけだ
:CN	接続詞	並びに、でも、但し、だけど
:CP	連結動詞	だ,、なんだ

:DA	指示詞、副詞類	こう、そう、あのように、この様に
:DM	指示詞、名詞前位修飾語句 (DN またはその他の代名詞の名詞前位形)	この、あの、そのような、そんな
:DN	指示詞、名詞類(英語の *this*、*that*、およびその他の代名詞に類似)	あれ、こちら、あそこ
:MD	名詞前位修飾語句	明くる、小さな、主たる、色んな
:IT	間投語	あれれ、あれー、あ〜、ええ、ええと
:NA	副詞的名詞	所所、前、間、後、挙句
:NC	普通名詞	風, 学校, 雑誌, 椅子
:NT	時間の名詞	永年、長年、夏、先月
:NK	内容名詞: 非主要部関係詞節の関係代名詞として機能(英語の *that* 節または *what* 節に類似)	ぐらい、の、もの、こと
:NN	数詞	千、十、〇、零、レイ、六、６
:NP	固有名詞	ＷＴＯ繊維協定、米州
:NPO	組織の固有名詞	米軍、米国、米国際貿易委員会
:NL	場所の名詞	米国、越南、奈井江町、奈央島
:NH	人間の名詞	中川秀直、中川浩明、中川勝
:NHM	名前	奈江子, 太郎, 那恵子

:NHS	姓	鈴木, 佐藤, 田中
:NV	動詞的名詞	請求, 弁解, 勉強
:VSN	品詞が与えられる動詞 **注:** 漢語動詞が単独で現れる場合は、名詞の一種としてカテゴリ分けされます。動詞活用形態素 *suru* と一緒に現れる場合、漢語動詞は VSN としてカテゴリ分けされます。	くるくる、くよくよ、伸びのび
:PC	格標識の不変化詞	を、で、の、へ、から
:PE	文末に現れて、話者の気分を表す不変化詞	っけ、な、ナ、なぁ
:PN	名詞類を結合する不変化詞(英語の *and* に類似)	ないし、ないしは、並びに
:PP	節を結合する不変化詞	ながら、なら、なり、のに、きり
:PQ	引用文の不変化詞	って、っと、て、と
:PS	英語の *only* や *too* を意味する不変化詞	ったら、って、等、など、なら
:PRJ1	形容詞の接頭辞(いで終わるもの)	か、こ、真
:PRJ2	形容詞の接頭辞(なで終わるもの)	無、不、非
:PRN	名詞類の接頭辞	高、前、全
:PRV	述語の接頭辞	相、猛、最
:SC	特殊カテゴリのカンマ	、,
:SCP	特殊カテゴリの閉じかっこ	’、”、＞、)、》
:SCP	特殊カテゴリのピリオド	。.

:SS	特殊カテゴリのスペース	
:SJN	名詞の接尾辞で、形容詞的名詞を形成する	っぽい、くさい
:SJV	動詞の接尾辞で、形容詞を形成する	ない、たい、づらい
:SNA	形容詞の接尾辞	さ
:SNC	分類辞またはカウンタの接尾辞	頁、ページ、杯、版
:SNN	名詞の接尾辞	っ子、下手、内、等、など、制、性、生、製、席、説、線、船
:SNV	名詞的述語の接尾辞	っきり、っぱなし、っ放し
:SV	動詞の接尾辞	る、ある、得る
:V1	一段動詞(*-eru* または*-iru* で終わる)	飛び始める、べんじる、便じる、直せる、流れ落ちる
:V5	五段動詞	並ぶ、せめぐ、泣付く、流し込む、往ぬ
:VK	くる動詞	やってくる、くる
:VS1	する動詞	する
:VS2	する動詞	辞する、じする、無くする
:VZ	ずる動詞	べんずる、便ずる、べんずる

韓国語

表 A1.8 韓国語の品詞タグ

品詞タグ	説明	例

:AD	副詞	매우、정말、빨리
:AJ	形容詞	예쁘다、귀엽다、차분하다
:ASCII	外来語	Korean、iPhone、SK
:DATE	日付	2015-04-28、20150428
:DEFAULT	未知語	하페즈、샤리프、쿠레쉬
:GAC	格文法接辞	가、를、로
:GAD	決定文法接辞	은、을、는
:GAH	変化文法接辞	이다、기、음
:GAJ	接続文法接辞	는데、는지、느라고
:GAP	述語文法接辞	다、습니다、더구만
:GAR	尊敬文法接辞	시、으시、옵
:GAT	時間文法接辞	겠、었、였었
:GAX	補助文法接辞	도、만、까지
:IJ	間投詞	아、네、그래
:NN	名詞	하늘、산、바다
:NNB	拘束名詞	것、수、개
:NNP	固有名詞	서울、이순신、국립국어원
:NUMBER	数詞	하나、둘、셋
:PF	接頭辞	제-、햇-、명-
:PN	名詞前位	각、첫、기초적
:PR	代名詞	이것、언제、이분

:PUNC	句読点	.? ! ( )
:SF	接尾辞	-꾼、꾸러기、-감
:TIME	時間	23:59:59
:URL	URL	www.sas.com
:VB	動詞	웃다、뛰다、날다

ポルトガル語

表 A1.9 ポルトガル語の品詞タグ

品詞タグ	定義	例
:A	形容詞	confiável、belo、feliz
:ADV	副詞	belamente、felizmente
:CONJ	接続詞	e、que
:DET	限定詞	alguns、cada、os
:digit	数	21
:DNUM	数値限定詞	bilionésimo、cinco
:inc	未知語	xrxx
:INTJ	間投詞	caramba、eh
:N	名詞	beleza、felicidade
:PFX	接頭辞	anti、circum
:PN	固有名詞	Brasil、Portugal
:PREP	前置詞	com、de、em

:PREPDET	前置詞および限定詞縮約形	dessas、dum
:PRO	代名詞	me、nós、quem
:sep	区切り記号または句読点	,
:url	URL	www.sas.com
:V	動詞	agradecem、 garanto
:VB	不定詞	agradecer、garantir
:VG	動名詞	agradecendo、garantindo
:VH	定過去	agradecido、garantido
:XL	外来語	cf、ibid、sic

ロシア語

表 A1.10 ロシア語の品詞タグ

品詞タグ	定義	例
:A	形容詞	духовитый、красивая、лучших
:ABBREV	省略形	др、км
:adv	比較級副詞	дальше
:adverbial	副詞	хорошо、сколько-нибудь
:conj	接続詞	если、и
:digit	数	123、12.3、12.3.2003、12/3/2013
:inc	未知語	геминг、analytics

:idet	疑問限定詞	который
:INT	間投詞	ах
:intadv	疑問副詞	где
:intquant	疑問数量詞	сколькие、почём
:N	名詞	велосипед、история、малолетство
:NONDECL	不変化語	мартини、маэстро
:NONDECL-ADJ	不変化形容詞	баскервиллей
:NONDECL-PN	不変化固有名詞	Шевроле、Айдахо
:NONDECL-PRO	不変化代名詞	всяко
:num	数	один, десятью
:particle	不変化詞	бы、же
:PN	固有名詞	Миа、Тузла
:PNA	固有形容詞	Роханский、Сашина
:PNN	固有名詞、名前	Свердловск、Мария、Давыдович
:prep	前置詞	до、вроде
:pron	代名詞	я、её
:sep	区切り記号または句読点	, .!
:url	URL	http://www.sas.com
:VB	不定詞	автоматизировать、менять、кончить

:V	動詞	нажимает、кладите、плавала
:VG	動名詞	адаптировав、вальсируя

スペイン語

表 A1.11 スペイン語の品詞タグ

品詞タグ	定義	例
:A	形容詞	confiable、feliz、hermoso
:ABBREV	省略形	km、pág、Sra
:Adv	副詞	ahora、felizmente
:CONJ	接続詞	ni、pero、y
:DET	限定詞	el、las、mi、nuestro
:digit	数	21
:inc	未知語	xrxx
:INTJ	間投詞	hola
:N	名詞	belleza、felicidad
:PN	固有名詞	Chile、España
:PREP	前置詞	con、de、en、por
:PREPDET	前置詞および限定詞縮約形	al、del
:PRON	代名詞	alguien、ellos、me
:sep	区切り記号または句読点	,

:url	URL	www.sas.com
:V	動詞	ayudan、pide
:VB	不定詞	ayudar、pedir
:VE	現在進行形	ayudando、pidiendo
:VH	過去分詞	ayudado、pedido

スウェーデン語

表 A1.12 スウェーデン語の品詞タグ

品詞タグ	定義	例
:A	形容詞	fört
:ABB	省略形	st.
:ADV	副詞	väl
:CONJ	接続詞	samt
:DET	限定詞	ens
:DNUM	数	två
:INT	間投詞	hej
:INV	不変詞	morse
:N	名詞	bok
:PN	固有名詞	Øsel
:PNF	固有名詞 - 名	Tove
:PNG	固有名詞 - 地理	Östmark

:PNH	固有名詞 - 姓	Viklund
:PNO	固有名詞 - 組織	Toshiba
:PREP	前置詞	till
:PRO	非人称代名詞	somlig
:PROP	人称代名詞	du
:VB	不定詞動詞	vara
:VD	能動態動名詞動詞	varit
:VE	現在分詞	varande
:VF	受動態動名詞動詞	varats
:VH	完了分詞	sedd
:VI	受動態現在形	ses
:VJ	能動態過去形	såg
:VK	受動態過去形	sågs
:VP	能動態現在動詞	varar
:VY	命令形	vara

トルコ語

表 A1.13 トルコ語の品詞タグ

品詞タグ	説明	例
:A	形容詞	iyi、zor
:ADV	副詞	yine、zaten

:CONJ	接続詞	veya、hem
:date	日付	12/30/2000、 12/30/00、 2000-30-12
:digit	数	12.302.000、5
:inc	未知語	wug
:N	名詞	kitap、insan
:NUMERAL	数詞	dokuz、onbir
:PARTICLE	不変化詞	beri、diye
:PN	固有名詞	Ayşe、Türkçe
:PRONOUN	非人称代名詞	bunlar、kendi
:PROP	人称代名詞	onlar、sen
:QUANT	数量詞	çok、her
:sep	区切り記号または句読点	! .,
:time	時間	12:30:00
:url	URL	sas.com、 www.sas.com、 http://sas.com

:V	動詞(この後に、次のいずれかを任意の数だけ任意の組み合わせで続けられます)。	
	A (習慣/普通)	bilir
	B (不定詞)	bilmek
	C (条件文)	bilse
	F (未来形)	bilecek
	I (不定/推論)	bilmiş
	J (過去形)	bildi
	N (必要形)	bilmeli
	P (古形を含む進行形)	biliyor、bilmekte
	O (仮定法形)	bile
	Y (命令形)	bil
:V_GER_STEM	付加なし動詞	該当なし

付録 2

事前定義済みコンセプト(英語以外の言語用)

事前定義済みコンセプトの優先順位値の使用

優先順位を正確に設定して自分の言語のカスタムコンセプトを一致させるためには、トピック “事前定義済みコンセプトの優先順位値” を参照してください。優先順位設定の詳細については、 “コンセプトページ” (35 ページ)を参照してください。。英語の優先順位値については、 表 3.1 (31 ページ) を参照してください。

注: ドキュメント処理中にカスタムコンセプトとの競合がないことを確認するには、言語ごとの最高優先順位値を使用します。次のセクションの表では各言語の最高優先順位値にアスタリスク(*)のマークが付いています。

事前定義済みコンセプトの優先順位値

オランダ語

表 A2.1 オランダ語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	20*
COMPANY	19
CURRENCY	18
DATE	18
INTERNET	18
LOCATION	20
MEASURE	18
NOUN_GROUP	15
ORGANIZATION	20
PERCENT	18
PERSON	20
PHONE	18
PROP_MISC	13

SSN	18
TIME	18
TIME_PERIOD	18
TITLE	18
VEHICLE	15

* この言語の最高値。

フィンランド語

表 A2.2 フィンランド語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
COMPANY	25*
LOCATION	25
NOUN_GROUP	15
PERSON	20

* この言語の最高値。

フランス語

表 A2.3 フランス語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	20*
COMPANY	19

CURRENCY	18
DATE	18
INTERNET	18
LOCATION	20
MEASURE	18
NOUN_GROUP	15
ORGANIZATION	20
PERCENT	18
PERSON	20
PHONE	18
PROP_MISC	13
SSN	18
TIME	18
TIME_PERIOD	18
TITLE	18

* この言語の最高値。

ドイツ語

表 A2.4 ドイツ語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	25

COMPANY	25
CURRENCY	25
DATE	18
INTERNET	18
LOCATION	40*
MEASURE	18
NOUN_GROUP	15
ORGANIZATION	20
PERCENT	18
PERSON	20
PHONE	18
PROP_MISC	13
SSN	18
TIME	18
TIME_PERIOD	18
TITLE	18

* この言語の最高値。

イタリア語

表 A2.5 イタリア語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値

ADDRESS	25*
COMPANY	25
CURRENCY	18
DATE	18
INTERNET	18
LOCATION	25
MEASURE	18
NOUN_GROUP	15
PERCENT	18
PERSON	20
PHONE	18
PROP_MISC	13
SSN	18
TIME	18
TIME_PERIOD	18
TITLE	18

* この言語の最高値。

日本語

表 A2.6 日本語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値

LOCATION	20*
ORGANIZATION	20
PERSON	20

* この言語の最高値。

韓国語

表 A2.7 韓国語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	20*
COMPANY	19
CURRENCY	19
DATE	20
INTERNET	18
LOCATION	20
MEASURE	19
NUMBER	18
ORDNUMBER	18
ORGANIZATION	20
PERCENT	18
PERSON	20
PHONE	18

TIME	20
TIME_PERIOD	18
TITLE	18

* この言語の最高値。

ポルトガル語

表 A2.8 ポルトガル語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	25*
COMPANY	25
CURRENCY	18
DATE	18
INTERNET	18
LOCATION	25
MEASURE	18
NOUN_GROUP	15
PERCENT	18
PERSON	20
PHONE	18
PROP_MISC	13
SSN	18

TIME	18
TIME_PERIOD	18
TITLE	18

* この言語の最高値。

ロシア語

ロシア語では、事前定義済みコンセプトに特定の優先順位値はありません。デフォルト値の 10 が使用されます。

スペイン語

表 A2.9 スペイン語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	25*
COMPANY	25
CURRENCY	18
DATE	18
INTERNET	18
LOCATION	25
MEASURE	18
NOUN_GROUP	15
ORGANIZATION	20
PERCENT	18

PERSON	20
PHONE	18
PROP_MISC	13
SSN	18
TIME	18
TIME_PERIOD	18
TITLE	18

* この言語の最高値。

スウェーデン語

表 A2.10 スウェーデン語の事前定義済みコンセプト優先順位

事前定義済みコンセプト	優先順位値
ADDRESS	20*
COMPANY	19
CURRENCY	18
DATE	18
INTERNET	18
LOCATION	20
MEASURE	18
NOUN_GROUP	15
ORGANIZATION	20

PERCENT	18
PERSON	20
PHONE	19
PROP_MISC	13
TIME	18
TIME_PERIOD	18

* この言語の最高値。

トルコ語

トルコ語では、事前定義済みコンセプトに特定の優先順位値はありません。デフォルト値の 10 が使用されます。

推奨資料

このタイトルの推奨資料リストを次に示します。

- *SAS Contextual Analysis: Administrator's Guide*
- *SAS Content Categorization Single User Servers: Administrator's Guide*
- *SAS Text Miner: Reference Help*
- *SAS Encoding: Understanding the Details*
- *SAS Enterprise Content Categorization Studio: User's Guide*
- *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*

SAS 刊行物の一覧については、sas.com/store/books から入手できます。必要な書籍についての質問は SAS 担当者までお寄せください:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
電話: 1-800-727-0025
ファクシミリ: 1-919-677-4444
メール: sasbook@sas.com
Web アドレス: sas.com/store/books

用語集

重み
アイテムに割り当てられて、度数分布や母集団におけるアイテムの相対的な重要度を示す数値インジケータ。

解析
SAS ステートメントなどのテキストを、その構成要素の単語、フレーズ、句読点、値、その他の種類の情報などに分ける目的で分析すること。その後、定義やルールセットに従って情報を分析できます。

カテゴリ
共通特性に基づくドキュメントの分類。カテゴリメンバシップは、バイナリプロパティとして示されます。どのような場合にドキュメントがカテゴリのメンバとなる可能性があるかを決定するためには、カテゴリテキスト定義を構成する 1 つ以上のブールルールが満たされる必要があります。

関連性スコア
ドキュメントがルールまたはモデルをどの程度満たしているかを示すスコア。最高の一致は 1 のスコアで、完全(100%)一致を表しています。

語
ルールまたはアルゴリズムによって定義された、1 つ以上のテキスト形式での 1 コンセプトの表現。

語幹処理
単語の原形を検索して返すプロセス。たとえば、grind、grinds、grinding、ground の原形は grind です。

語テーブル
ドキュメントコレクション内に表示される各語、その役割、およびその表層形すべてに対する代表テキスト形式を含むドキュメントコレクション内のすべての語のリスト。

語の役割
特定のコンテキストで語によって実行される機能。語は、品詞、エンティティの種類、またはユーザー定義のその他の目的として機能します。

語マップ
"対象とするオブジェクト"を中心とするノードアークグラフです。オブジェクトには、カテゴリ、コンセプト、トピック、語などがあります。グラフ内の対応するノードが、対象とするオブジェクトを予測するルールを示します。よいルールほど大きなノードとして示されます。アークは、ルールを作り上げるために使用される語の追加や除外を表します。

コンセプト
意味の抽象クラス。どのような場合にコンセプトがテキストのサブセットで参照される可能性があるかを決定するためには、コンセプトテキスト定義を構成するルールが満たされる必要があります。

スコアリング
モデルスコアリングを参照してください。

センチメント
テキストのセグメント、テキストセグメントのグループ、特定の興味のある話題など、分析対象のアイテムに示される態度。

センチメント分析
トピックやドキュメントなどの分析アイテムに対する話者や書き手の態度を特定するための、自然言語処理、計算言語学、テキスト分析の使用。センチメント分析の結果は、分析のターゲットのスコア(ポジティブ、ネガティブ、ニュートラル)になります。

停止リスト
テキストマイニング分析から除外する低品質な情報や無関係の単語の単純なコレクションです。

テキストのサブセット
コンセプトテキスト定義の一致テキスト。これは、ドキュメントに含まれる 1 つ以上の文字列で構成されます。

テキスト文字列
いずれかの種類の隣接する文字で構成されるテキストのサブセット。指定オプションに応じて、文字列は、大文字と小文字を区別するか、または区別しないように指定できます。

トークン
SAS プログラミング言語において、SAS に意味を伝達し、なおかつ、これ以上小さい機能単位に分割できない文字のコレクション。変数名などのトークンは、英単語に似ている場合もありますが、数学演算子や、さらにはセミコロンなどの個々の文字である可能性もあります。トークンには、最大 32,767 文字を含められます。

トピック
ドキュメントの内容を示すことを目的とするマシン生成カテゴリ。トピックによって、ドキュメントコレクションの重要語のグループ分けが指定されます。1 つのドキュメントに 1 つ以上のトピックを含めることも、トピックを含めないことも可能です。

トピック語の重み
トピック固有語の重みを参照してください。

トピック固有語の重み
トピック内の語を他の語と比較した相対的重要度のインジケータ。指定カットオフ値を上回る値を持つ語は、ドキュメントのトピックへの割り当てに寄与します。

トピック固有ドキュメントの重み
ドキュメントにとってのトピックの重要性のインジケータ。指定カットオフ値を上回る値は、ドキュメントにそのトピックが含まれることを示します。

トピックドキュメントの重み
トピック固有ドキュメントの重みを参照してください。

表層形
1 つ以上のドキュメントで一致したテキストサブセットに含まれる語のバリアントです。この形式には、語幹、類義語、スペルミス、同じエンティティに対する別の参照方法が含まれます。

分類
親と子のカテゴリノードの階層関係。真の分類では、カテゴリが検出された場合は必ず、すべての親も表されることが示されます。たとえば、人間として識別されたものがある場合、それは必ず霊長類、哺乳動物、動物などでもあります。

文字列
テキスト文字列を参照してください。

モデルスコアリング
出力を計算するためにモデルを新しいデータに適用するプロセス。