THE POWER TO KNOW.

SAS® Text Miner 13.2 入門ガイド

SAS® ドキュメント

The correct bibliographic citation for this manual is as follows: SAS Institute Inc 2014. *SAS® Text Miner 13.2 入門ガイド*. Cary, NC: SAS Institute Inc.

SAS® Text Miner 13.2 入門ガイド

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

Electronic book 1, 2014 August

SAS® Publishing では、お客様が SAS ソフトウェアの可能性を最大限に利用できるようにするために、紙媒体製品および電子媒体製品の完全なセットを提供しています。e ブック、e ラーニング製品、CD、ハードコピー本に関する詳細は、SAS Publishing の Web サイト(**support.sas.com/publishing**)をご覧になるか、または 1-800-727-3228 までお電話ください。

目次

概要

利用者

本書は、SAS Text Miner を初めてお使いになるユーザーを対象としています。最初の 7 つの章では、仮想的なテキストマイニング分析の文脈において SAS Text Miner ノードの使用方法を示します。これらの章を読み終えると、プロジェクトやプロセスフローダイアグラムを作成することや、各種の SAS Text Miner ノードのプロパティを設定し、それらのノードを実行して結果を調べることができるようになります。

第 8 章から第 13 章では、次のような SAS Text Miner ノードに関するその他の例を紹介します。

- テキストインポートノード
- テキスト解析ノード
- テキストフィルタノード
- テキストクラスタノード
- テキストトピックノード
- テキストルールビルダノード
- テキストプロファイルノード

最後の 2 つの章では、テキストマイニングのヒントや、その他の機能に関する簡単な説明を紹介します。

テキストマイニング処理や SAS Text Miner ノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

1 章

テキストマイニングおよび SAS Text Miner 13.2 の概要

テキストマイニングとは何か

テキストマイニングとは、膨大なドキュメントコレクション中に潜在しているテーマやコンセプトを明らかにする作業です。テキストマイニングアプリケーションには 2 つのフェーズがあります。原文データの中身を調査するフェーズと、取り出した情報を使用して既存のプロセスを改善するフェーズです。これらのフェーズは両方とも重要であり、各フェーズはそれぞれ記述マイニングおよび予測マイニングと呼ばれることもあります。

記述マイニングとは、原文コレクション内に存在しているテーマやコンセプトを明らかにする作業です。たとえば、多くの会社では、Web、e メール、窓口センターなどのさまざまなソースから顧客のコメントを収集しています。原文コメントのマイニングには、原文コレクション内の語、フレーズ、およびその他のエンティティに関する情報を提供すること、ドキュメントを意味のあるグループへとクラスタリングすること、クラスタ内で発見されたコンセプトを報告することなどが含まれます。記述マイニングの結果を利用することで、原文コレクションをより良く理解できるようになります。

予測マイニングとは、ドキュメントをカテゴリに分類し、テキスト内に潜在している情報を利用して意思決定を行う作業です。たとえば、標準的な質問をする顧客を特定し、彼らに自動的な応答を提供したい場合などに、予測マイニングを利用できます。また、顧客が再度購入を行うかどうかや、顧客を逃がさないためにより多くの努力を行うべきかなども予測できます。

予測モデリングとは、過去のデータを調査して結果を予測する作業です。たとえば、過去の購買行動に関する情報や顧客のコメントを含んでいる顧客データセットがあるとします。これを使用して、新しい顧客のスコアリング(過去の顧客データに基づいて新規顧客を分析すること)に利用できる予測モデルを構築できます。たとえば、あなたが製薬会社の研究者であるならば、臨床研究での医師からの報告書から有害反応を手作業で符号化するのは、多くの労力のかかる誤りを起こしやすい作業であることをご存知でしょう。このような作業を行う代わりに、すべての過去の原文データを使用することで、どの医師の報告書がどの有害反応に対応しているかを示すモデルを作成できます。モデルを構築したら、原文データの処理は、入ってくる新しいレコードをスコアリングすることにより自動的に実施されます。あなたは「分類が困難な」ケースだけを調査すればよく、それ以外のケースはコンピュータに任せることができます。

テキストマイニングのこれらの両側面は、同じ要件を一部共有しています。たとえば、人間が容易に理解できる原文ドキュメントを、ソフトウェアがマイニングできるような形式で表現する必要がまずあります。生のドキュメントは、それが含んでいるパターンや関係を検出できるようにするための適切な処理を必要とします。人間は構造化されていないドキュメントに含まれている章、パラグラフ、センテンスなどを把握できますが、コンピュータは構造化された(定量的または定質的な)データを必要とします。このため、非構造化ドキュメントは、マイニングを行う前に、構造化された形式へと変換する必要があります。

SAS Text Miner 13.2 について

SAS Text Miner は、SAS Enterprise Miner 環境向けのプラグインです。SAS Enterprise Miner は、テキストマイニングの予測的な側面を促進するデータマイニングツールの豊富なセットを提供します。SAS Text Miner を SAS Enterprise Miner 内部に統合することにより、原文データを従来型のデータマイニング変数と結合できるようになります。これにより、テキストマイニングノードを SAS Enterprise Miner のプロセスフローダイアグラム内に埋め込むことが可能となります。SAS Text Miner は、ローカルデータ、SAS データセット内のオブザベーションとしてのテキスト、外部データベース、Web 上のファイルなど、原文データから構成される各種のソースをサポートしています。

SAS Text Miner 13.2 には、テキストマイニング分析で使用できる次のノードが含まれています。

- **テキストインポート**ノード
- **テキスト解析**ノード
- **テキストフィルタ**ノード
- **テキストトピック**ノード
- **テキストクラスタ**ノード
- **テキストルールビルダ**ノード
- **テキストプロファイル**ノード

SAS Text Miner の各種ノードに関する詳細は、本書の対応する章を参照するか、または SAS Text Miner のヘルプをご覧ください。

また、Text Miner ノードは、テキストマイニングの解析や調査、予測マイニングのためのデータ準備、他の SAS Enterprise Miner ノードを使用する場合のより詳細な調査などもサポートします。ユーザーは構造化されたテキスト情報を分析できるほか、Text Miner ノードの構造化された出力を、必要に応じてその他の構造化データと組み合わせることができます。Text Miner ノードは高度なカスタマイズが可能であり、ユーザーはさまざまなオプションから選択できます。たとえば、**テキスト解析**ノードを使用すると、ドキュメントを解析することで、コレクション内の語、フレーズ、およびその他のエンティティに関する詳細な情報を取得できます。**テキストクラスタ**ノードを使用すると、ドキュメントを意味のあるグループへとクラスタリングし、そのクラスタに関して検出したコンセプトを報告できます。また、語やドキュメントの並べ替え、検索、フィルタリング(サブセット化)、類似語の検出などの機能により、調査手順を強化できます。

さらに、SAS Text Miner では、%TMFILTER という名前の SAS マクロを使用できます。このマクロを使用すると、テキストの前処理ステップを実施できるほか、お使いのファイルシステムや Web ページ上に存在するドキュメントから SAS データセットを作成できます。これらのドキュメントは、多数のベンダー固有フォーマットで存在しています。

SAS Text Miner は柔軟なツールであり、さまざまな問題を解決できます。SAS Text Miner を使用して実行可能なタスクの例を次に示します。

- e メールのフィルタリング
- ドキュメントを事前定義されたカテゴリーにトピック別に分類すること
- ニュース項目のルーティング
- データベース内にある調査報告書のクラスタ分析
- 調査データのクラスタ分析
- 顧客のクレームやコメントに関するクラスタ分析
- ビジネスに関するニュース発表から株式相場を予測すること
- 顧客のコメントから顧客満足度を予測すること
- コールセンターのログに基づいてコストを予測すること

テキストマイニング処理

記述または予測(あるいはその両方)を目的として原文データの使用を意図しているかどうかにかかわらず、次の表に示すように同じ処理手順が実行されます。

アクション	結果	ツール
ファイルの前処理	お手持ちのドキュメントコレクションから単一の SAS データセットを作成します。この SAS データセットは、**テキスト解析**ノードで入力として使用されるほか、実際のテキストまたは実際のテキストへのパスを含む場合もあります。	**テキストインポート**ノード %TMFILTER マクロ — ドキュメントからテキストを抽出し、テキスト変数を含む事前定義済みの SAS データセットを作成する SAS マクロです。
テキスト解析	原文データを分解し、データマイニング用に適した定量的表現を生成します。	**テキスト解析**ノード
変換(次元縮小)	定量的表現をコンパクトで情報的な形式に変換します。	**テキストフィルタ**ノード
ドキュメント分析	ドキュメントコレクションに関して、分類、予測、コンセプトのリンク付けを実施します。データからクラスタ、トピック、ルールを作成します。	**テキストクラスタ**ノード **テキストトピック**ノード **テキストルールビルダ**ノード **テキストプロファイル**ノード SAS Enterprise Miner の予測モデリングノード

注: **Text Miner** ノードは、SAS Text Miner 13.2 の**テキストマイニング**タブからは使用できません。Text Miner ノードは、他の SAS Text Miner ノードでの機能により置き換えられています。プロセスフローダイアグラム内に **Text Miner** ノードが存在していた以前のリリースの SAS Text Miner からダイアグラムをインポートできます。た

だし、新しい **Text Miner** ノードは作成できず、インポートした **Text Miner** ノードのプロパティ値を変えることもできません。詳細については、SAS Text Miner ヘルプの「以前のバージョンに含まれていた SAS Text Miner ダイアグラムの変換」というトピックを参照してください。

最後に、クラスタリングや予測に関するルールを使用して、新しいドキュメントコレクションを任意の時点でスコアリングします。

これらの手順のすべてを分析に含める必要はありません。満足する結果を得るためには、オプションのさまざまな組み合わせを試す必要があります。

SAS Text Miner 13.2 のアクセシビリティ機能

SAS Text Miner には、お身体の不自由なユーザー向けに製品を使いやすくするアクセシビリティ機能と互換性機能が含まれています。これらの機能は、米国政府により 1973 年の米国リハビリテーション法第 508 項(Section 508)の下に修正条項として採用された電子情報テクノロジに関するアクセシビリティ標準に関連しています。SAS Text Miner は、下表に示す点を除き、Section 508 標準をサポートしています。

Section 508 アクセシビリティ基準	サポート状況	説明
ソフトウェアがキーボードの備わったシステム上で実行されるように設計されている場合、製品機能は、機能自体または機能の実行結果がテキストとして識別できるように、キーボードから実行可能であること。	例外付きでサポート。	ソフトウェアは、次に示すものを除き、すべてのユーザーアクションに対応するキーボード操作をサポートしています。 システムメニューを表示するキーボード操作は、Windows 標準の Alt + スペースバーではありません。システムメニューは、次のショートカットキーを使うことでは表示できません。(1)プライマリウィンドウ — Shift + F10 + スペースバー、または(2)セカンダリウィンドウ — Shift + F10 + ダウンキー。 データソースのポップアップメニューでのエクスプローラアクションは、キーボードからは直接呼び出せません。データソースエクスプローラを呼び出す別の方法として、**表示** —> **エクスプローラ**メニューの使用が挙げられます。
カラーコーディングは、情報の提示、アクションの表示、応答のプロンプト、視覚要素の区別を行う唯一の手段としては使用できません。	例外付きでサポート。	ノード実行の成功または失敗の表示では色を使用しています。ノードの成功や失敗を表すダイアログボックス内の対応するポッポアップメッセージも存在します。

SAS 製品のユーザー補助機能についてのご質問やご意見は、accessibility@sas.com まで電子メールでお寄せください。

2 章

事例学習:SAS Text Miner 13.2 の使用

本書のシナリオについて

最初の 7 つの章では、読者を SAS Text Miner に慣れてもらうために 1 つの大がかりな例を紹介しています。各トピックは直前のトピックに基づいて構成されているため、読者はこれらの章を順番に読む必要があります。これらの章では、SAS Text Miner のプロセスフローダイアグラムの主要なコンポーネントについて説明しています。このステップごとの例では、SAS Text Miner における基本タスク(プロジェクトの作成やプロセスフローダイアグラムの構築など)の実行方法を学ぶことができます。ユーザーが作成したダイアグラム内では、データへのアクセス、データの準備、テキスト変数を使用した複数の予測モデルの構築、モデルの比較などのタスクを実行できます。本書におけるこの大がかりな例は、SAS Text Miner ソフトウェアと組み合わせて使用するために設計されたものです。残りの章では、それぞれの SAS Text Miner ノードに焦点を当て、読者がテキストマイニング分析を行う場合に有益となる追加情報を紹介します。

ワクチン有害事象報告制度(VAERS)データは、米国保険社会福祉省(HHS)が一般に公開しているデータです。このデータは、**http://vaers.hhs.gov** から CSV 形式で誰でもダウンロードできます。同サイトには、米国がデータ収集を開始した 1990 年以来、毎年のデータが別々の CSV ファイルとして公開されています。このデータはさまざまなソースから収集されたものですが、ほとんどの報告書はワクチン製造業者や医療事業者から提供されたものです。ワクチンの提供者は、ワクチンに関する禁忌事象や重篤な合併症がある場合には、それを報告する義務があります。ワクチンの場合、禁忌事象は、そのワクチンの使用に関するリスクを高める条件または要因となります。

Getting Started Examples の zip ファイルには次のファイルが含まれています。

- ReportableEventsTable.pdf: 各ワクチンでの報告可能な事象の完全な一覧を記載しています。
- VAERS README ファイル: データ辞書および使用されている略語の一覧を記載しています。

注: **Getting Started Examples** の zip ファイルのダウンロードに関する詳細は、“このシナリオの前提条件” (9 ページ)を参照してください。

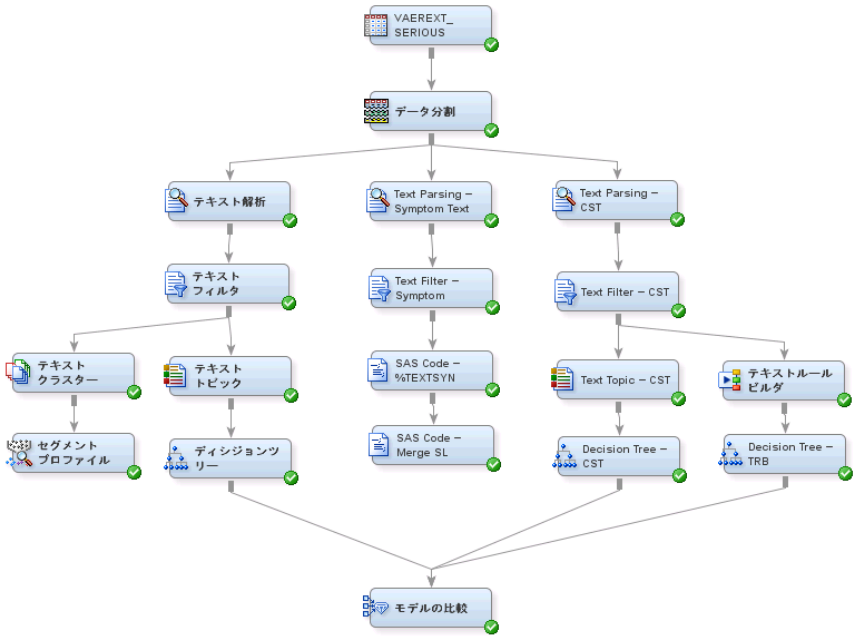
VAERS データの表の最初の 10 行の先頭から 8 列目までを次の図に示します。一意の ID、居住州、受容者の年齢などが含まれていることが分かります。それ以外の列

(次の図には示されていないもの)の中には、非構造化テキスト文字列である SYMPTOM_TEXT があり、これには報告済みの問題、特定の症状、symptom counter が含まれています。

	VAERS_ID	RECVDATE	STATE	AGE_YRS	CAGE_YR	CAGE_MO	SEX	RPT_DATE
1	179605.0	2002/01/02	FL	64.0	64.0	.	F	2001/12/26
2	179606.0	2002/01/02		29.0	29.0	.	F	2001/12/26
3	179612.0	2002/01/02	NJ	40.0	0.0	0.3	F	2001/12/23
4	179613.0	2002/01/02	NY	40.0	1.0	0.6	F	1998/02/28
5	179614.0	2002/01/02	TX	4.0	4.0	.	M	2001/12/28
6	179615.0	2002/01/02	WI	38.0	38.0	.	F	2001/12/21
7	179616.0	2002/01/02	KY	69.0	69.0	.	F	2001/12/26
8	179617.0	2002/01/02	FL	77.0	77.0	.	M	2001/12/21
9	179618.0	2002/01/02		7.0	7.0	.	M	2001/11/23
10	179619.0	2002/01/02		50.0	.	.	F	2001/12/20

この例を十分理解するためには、読者は自分がこのデータセット内にはどんな情報が含まれているかを明らかにしようとしている研究者であると仮定する必要があります。また、読者はそのような研究者として、子供や大人がこのワクチン接種から経験する有害事象についてより良く理解するためには、このデータセットをどのように使用すればよいかを知りたいと思っているとします。これらの有害事象は、1 つまたは複数のワクチン接種により引き起こされたか、または投与実験室で不適切な手順(消毒されていない針の使用など)により誘発された可能性があります。また、一部の報告は、ワクチンによる有害事象とはまったく無関係である場合もあります。たとえば、インフルエンザのワクチン接種後に風邪を引いた人がそれを報告した場合が考えられます。このため、入院を要するような、または一生引きずる障害や死亡を引き起こすような重篤な反応を調査する必要があります。

この例を完了した時点で、読者のプロセスフローダイアグラムは次のようになります。



このシナリオの前提条件

本書に紹介されているタスクを実行する前に、各サイトの管理者が SAS Text Miner 13.2 のすべての必要なコンポーネントのインストールと設定を完了している必要があります。次の操作を実行する必要があります。

1. 次の URL の SAS Text Miner 13.2 という見出しの下にあるリンクをクリックして、Getting Started with SAS Text Miner 13.2 用のサンプルデータを収めた zip ファイルをダウンロードします。

   http://support.sas.com/documentation/onlinedoc/txtminer

2. このファイルを解凍し、各自のファイルシステム内の任意のフォルダに展開します。
3. C:\ドライブ上に **`vaersdata`** という名前のフォルダを作成します。
4. 次のファイルを **C:\vaersdata** にコピーします。
   - vaerext.sas7bdat
   - vaer_abbrev.sas7bdat
   - engdict.sas7bdat

   注: 上記のファイル名は、表示環境によって大文字で表示される場合と表示されない場合があります。

SAS Text Miner 13.2 のヘルプを表示するには

SAS Text Miner のヘルプを表示するには、メインの SAS Enterprise Miner メニューバーで**ヘルプ** ⇨ **目次**を選択します。

3 章

プロジェクトの設定

実行するタスクについて

プロジェクトを設定するには、次の操作を実行します。

1. すべての成果物の格納先となる新規プロジェクトを作成します。
2. データソースの格納先となるライブラリを作成します。
3. ノードと対話する場合に使用するプロジェクト内の新規ダイアグラムを作成します。
4. VAERS データを表示して変更します。
5. SAS Enterprise Miner のデータソースである VAEREXT_SERIOUS を作成します。

プロジェクトの作成

プロジェクトを作成するには次の操作を実行します。

1. SAS Enterprise Miner を開きます。
2. SAS Enterprise Miner ウィンドウ内で**新規プロジェクト**をクリックします。

   SAS サーバーの選択ページが表示されます。
3. **次へ**をクリックします。

   プロジェクト名とサーバーディレクトリの指定ページが表示されます。
4. **プロジェクト名**フィールドにプロジェクト名(*Vaccine Adverse Events* など)を入力します。

5. 自分のプロジェクト用のデータの保存先とするサーバー上の場所へのパスを、**SAS サーバーディレクトリ**フィールドに入力します。または、参照ボタンを使用して自分のプロジェクトで使用するフォルダを指定するか、あるいは表示されているデフォルトのディレクトリパスを受け入れます。

   注: プロジェクトパスは、SAS Enterprise Miner をローカルマシン上で完全なクライアントとして実行しているか、それともクライアント/サーバーアプリケーションとして実行しているかによって異なります。SAS Enterprise Miner をローカルマシン上で完全なクライアントとして実行している場合、お使いのローカルマシンはそれ自身がサーバーとして動作します。ユーザーの SAS Enterprise Miner プロジェクトは、各自が指定した自分のローカルマシン上の場所(**C:\EM_Projects** など)に保存されます。SAS Enterprise Miner をクライアント/サーバーアプリケーションとして実行している場合、すべてのプロジェクトは SAS Enterprise Miner サーバー上に保存されます。SAS サーバーディレクトリボックスにデフォルトパスが表示されている場合、そのデフォルトのプロジェクトパスを受け入れることもできれば、独自のパスを指定することもできます。

6. **次へ**をクリックします。

   プロジェクトの登録ページが表示されます。

7. **次へ**をクリックします。

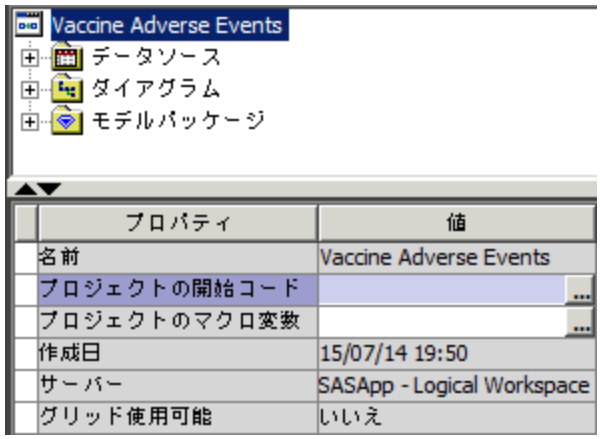
   新規プロジェクト情報ページが表示されます。

8. **完了**をクリックしてプロジェクトを作成します。

ライブラリの作成

ライブラリを作成するには次の操作を実行します。

1. プロジェクト名 **Vaccine Adverse Events** を選択し、同プロジェクトのプロパティパネルを表示します。



2. **プロジェクト開始コード**プロパティの ... をクリックします。

   プロジェクト開始コードダイアログボックスが表示されます。

3. **コード**タブ内に次のプログラムを入力します。

```
libname mylib "c:\vaersdata";
```

   注: 上記で指定する場所は、お使いのシステムでこのチュートリアルのデータを保存した場所により異なります。

4. **実行**をクリックします。
5. **OK** をクリックして、プロジェクト開始コードダイアログボックスを閉じます。

注: ライブラリウィザードを使用してライブラリを作成することもできます。ライブラリウィザードを使用するには、メインメニューから**ファイル** ⇨ **新規** ⇨ **ライブラリ**を選択します。

ダイアグラムの作成

ダイアグラムを作成するには、次の手順を実行します。

1. プロジェクトパネル内のダイアグラムフォルダを右クリックし、**ダイアグラムの作成**を選択します。

ダイアグラムの新規作成ダイアログボックスが表示されます。

2. **ダイアグラム名**フィールドに *VAERS Example* と入力します。
3. **OK** をクリックします。

空の VAERS Example ダイアグラムが、ダイアグラムワークスペース内に開かれます。

VAERS データの表示と変更

ライブラリを作成した後、SAS Enterprise Miner で使用するデータソースを作成する前に、データを表示できます。たとえば、複数の変数の値を要約するために別の変数を作成したい場合を考えます。

この場合、次の手順を実行することで、利用可能な SAS データファイルを表示し、データソースとして使用する新しい SAS データファイルを作成できます。

1. メインメニューから**表示** ⇨ **エクスプローラ**を選択します。

   エクスプローラウィンドウが表示されます。

2. SAS ライブラリツリー内の **Mylib** を選択します。

   Mylib ライブラリの内容が表示されます。このライブラリには、`Engdict`、`Vaerext`、および `Vaer_abbrev` という 3 つのファイルが含まれています。

3. `Vaerext` をダブルクリックします。

   `Vaerext` ファイルの内容が新規ウィンドウに表示されます。

4. 右スクロールして、列見出しとして表示される利用可能な変数名を表示させます。

   次の変数に注意してください。

- **DISABLE** — 二値変数であり、障害が存在した場合には値‘Y’を持ちます。
- **DIED** — 二値変数であり、死亡が存在した場合には値‘Y’を持ちます。
- **ER_VISIT** — 二値変数であり、救急外来受診が存在した場合には値‘Y’を持ちます。
- **HOSPITAL** — 二値変数であり、入院が存在した場合には値‘Y’を持ちます。

新しいデータセット **vaerext_serious** を作成するとします。このデータセットは、障害、死亡、救急外来受診、入院のいずれかが存在した場合に値‘Y’を持つ二値変数 **serious** が含むものとします。

5. **MYLIB.VAEREXT** ウィンドウとエクスプローラウィンドウを閉じます。
6. **ユーティリティ**タブを選択し、**SAS コード**ノードをダイアグラムワークスペースへとドラッグします。
7. **SAS コード**ノードを選択します。
8. **コードエディタ**プロパティのをクリックします。

   ウィンドウが表示されます。
9. **学習コード**ペイン内に次のプログラムを入力します。

```
data mylib.vaerext_serious;
   set mylib.vaerext;
   if DISABLE='Y' or DIED='Y' or ER_VISIT='Y' or HOSPITAL='Y' then serious='Y';
   else serious='N';
run;
```

   このプログラムは、**mylib** ライブラリに含まれている **vaerext** ファイルから新しい SAS ファイル **vaerext_serious** を作成し、変数 **serious** を追加した後、**DISABLE**、**DIED**、**ER_VISIT**、**HOSPITAL** の各変数の値に応じて、変数 serious に値 **Y** または **N** のいずれかを割り当てます。
10.  をクリックした後、コードエディタウィンドウを閉じます。
11. ダイアグラムワークスペース内にある **SAS コード**ノードを右クリックし、**実行**を選択します。
12. 確認ダイアログボックスで**はい**を選択します。
13. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で **OK** をクリックします。
14. メインメニューから**表示** ⇨ **エクスプローラ**を選択します。

    エクスプローラウィンドウが表示されます。
15. SAS ライブラリツリー内の **Mylib** を選択します。

    これで、Mylib ライブラリには **Vaerext_serious** ファイルの新しいエントリが含まれるようになりました。

    注: **Vaerext_serious** ウィンドウを表示するためには、Explorer ウィンドウのビューの更新が必要となる場合があります。
16. **Vaerext_serious** をダブルクリックします。

    Vaerext_serious ファイルの内容が新規ウィンドウに表示されます。
17. 右端までスクロールし、新しい列 **serious** が存在することを確認します。
18. **MYLIB.VAEREXT_SERIOUS** ウィンドウを閉じます。
19. エクスプローラウィンドウを閉じます。

データソースの作成

データソースを作成するには次の操作を実行します。

1. プロジェクトパネル内のデータソースフォルダを右クリックし、**データソースの作成**を選択します。

   データソースウィザードが表示されます。

2. ソースドロップダウンメニューで **SAS テーブル**を選択します。
3. **次へ**をクリックします。

   SAS テーブルの選択ウィンドウが表示されます。

4. **表示**をクリックします。
5. ライブラリツリー内にある **Mylib** という名前の SAS ライブラリをクリックします。

   Mylib ライブラリフォルダの内容が、SAS テーブルの選択ダイアログボックスに表示されます。

   注: Mylib フォルダ内の SAS データファイルが表示されない場合、**更新**をクリックします。

6. **Vaerext_serious** テーブルを選択します。
7. **OK** をクリックします。

   2 レベル名 **`MYLIB.VAEREXT_SERIOUS`** が**テーブル**フィールドに表示されます。

8. **次へ**をクリックします。

   テーブル情報ページが表示されます。

9. **次へ**をクリックします。

   Metadata Advisor オプションページが表示されます。

10. **詳細**をクリックします。
11. **次へ**をクリックします。

    列メタデータページが表示されます。

12. 各変数値用の役割値をクリックし、ドロップダウンリストから示された値を選択することで、次の変数役割を選択します。
    - **V_ADMINBY** の役割を **Input** に設定します。
    - **V_FUNDBY** の役割を **Input** に設定します。
    - **serious** の役割を **Target** に設定します。
13. **次へ**をクリックします。

    意思決定の構成ページが表示されます。

14. **次へ**をクリックします。

サンプルの作成ページが表示されます。

15. **次へ**をクリックします。

    データソースの属性ページが表示されます。

16. **次へ**をクリックします。

    要約ページが表示されます。

17. **完了**をクリックします。

    VAEREXT_SERIOUS テーブルが、プロジェクトパネル内のデータソースフォルダに追加されます。

4 章

SYMPTOM_TEXT 変数の分析

実行するタスクについて

SYMPTOM_TEXT 変数には、報告済みの有害事象に関するテキストが含まれています。この章では、次のタスクを実行することにより、SYMPTOM_TEXT 変数を分析する方法について説明します。

1. **入力データ**ノードを使用して、VAERS_SERIOUS データソースを特定します。
2. **データ分割**ノードを使用して、入力データを分割します。
3. **テキスト解析**ノードを使用して、ドキュメント群を解析します。
4. **テキストフィルタ**ノードを使用して、解析済みの語の総数を減らします。
5. **テキストクラスタ**ノードを使用して、ドキュメントをクラスタリングします。
6. 結果を表示します。
7. **セグメントプロファイル**ノードを使用して、データセグメントを確認します。

入力データの指定

入力データを指定するには、次の操作を実行します。

1. プロジェクトパネルのデータソースフォルダ内にある **VAEREXT_SERIOUS** データソースを選択します。
2. **VAEREXT_SERIOUS** をダイアグラムワークスペースにドラッグし、**入力データ**ノードを作成します。

入力データの分割

データ分割ノードを使用すると、入力データを次のデータセットのいずれかに分割できます。

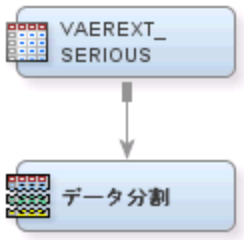
- **学習(Training)** — 事前のモデルの当てはめに使用されます。分析は、このデータセットを使用して最適なモデルの重みを見つけようとします。
- **検証(Validation)** — モデル比較ノードでのモデルの適合性評価に使用されます。検証データセットは、**デシジョンツリー**モデルノードで最適なサブツリーを作成するためのモデルの微調整にも使用されます。
- **テスト(Test)** — モデルの生成エラーに関する最終的な偏りのない評価を取得するために使用されます。

データ分割ノードに関する詳細は、SAS Enterprise Miner のヘルプを参照してください。

データ分割ノードを分析に追加するには、次の操作を実行します。

1. ノードツールバー上で**サンプル**タブを選択し、**データ分割**ノードをダイアグラムワークスペースへとドラッグします。
2. **VAEREXT_SERIOUS** 入力データノードを**データ分割**ノードに接続します。

   注: デフォルトの水平ビューで、あるノードを別のノードに接続するには、マウスポインタをノードの右端に置きます。鉛筆アイコンが表示されます。左マウスボタンを押したまま、接続したいノードの左端にまで行をドラッグした後、左マウスボタンを離します。接続されたノードのビューを垂直ビューに変更するには、ダイアグラムワークスペースで右クリックし、表示されたメニューから**レイアウト** ⇨ **垂直**を選択します。



3. **データ分割**ノードを選択し、そのプロパティを表示します。

   当該ノードに関する詳細情報がプロパティパネルに表示されます。
4. データセット割り当てプロパティを次のように設定します。
   - **Training** プロパティを **60.0** に設定します。
   - **Validation** プロパティを **20.0** に設定します。
   - **Test** プロパティを **20.0** に設定します。

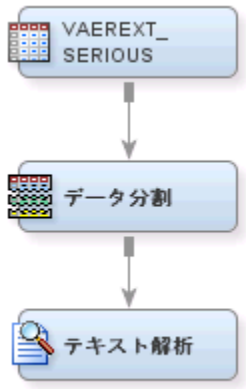
これらのデータ分割設定を行うことで、VAEREXT_SERIOUS データを使用して予測モデルを構築する場合に適切なデータを確保できます。

データの解析

テキスト解析ノードを使用すると、ドキュメント群を解析し、そこに含まれている語に関する情報を定量化できます。**テキスト解析**ノードは、e メールメッセージ、ニュース記事、Web ページ、研究報告書、調査報告書などの膨大な原文データに対して使用できます。**テキスト解析**ノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

テキスト解析ノードを分析に追加するには、次の操作を実行します。

1. ノードツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。
2. **データ分割**ノードを**テキスト解析**ノードに接続します。



3. **テキスト解析**ノードを選択します。

   テキスト解析ノードのプロパティがプロパティパネルに表示されます。
4. **品詞を区別する**プロパティの値を **No** に設定します。

   VAERS データの場合、この設定を行うと、よりコンパクトなサイズの語の集合が提供されます。
5. **類義語**プロパティの をクリックします。

   ダイアログボックスが表示されます。
6. **テーブルの交換**をクリックします。

   SAS テーブルの選択ダイアログボックスが表示されます。
7. **データセットを指定しない**を選択します。
8. **OK** をクリックして、SAS テーブルの選択ダイアログボックスを終了します。
9. 確認ダイアログボックスで**はい**を選択します。
10. **OK** をクリックして、類義語ダイアログボックスを終了します。
11. **品詞を無視する**プロパティの をクリックします。

    品詞を無視するダイアログボックスが表示されます。

12. 品詞を表す次の項目を選択します。
    - Aux
    - Conj
    - Det
    - Interj
    - Part
    - Prep
    - Pron
    - Num

    注: 複数の項目を選択する場合、CTRL キーを押しながら選択します。

    品詞を無視するダイアログボックスで選択された品詞を含む語は、解析時に無視されます。ここに示す選択では、分析において前置詞や限定子のような「低含有量」の語が確実に無視されます。
13. **OK** をクリックします。

データのフィルタリング

テキストフィルタノードを使用すると、解析済みの語や分析対象となるドキュメントの総数を減らすことができます。これにより、無関係な情報を取り除き、最も価値の高い関連性のある情報だけを検討対象とすることができます。たとえば、**テキストフィルタ**ノードを使用することで、不要な語を削除し、特定の問題について記述しているドキュメントだけを保持することができます。このような縮小されたデータセットは、数十万のドキュメントや数十万の語を含んでいるオリジナルの集合を表すデータセットよりも桁違いにサイズが小さくなります。**テキストフィルタ**ノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

データをフィルタリングするには、次の操作を実行します。

1. ノードツールバー上で**テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。
2. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。

3. **テキストフィルタ**ノードを選択します。
4. **語の重み**プロパティの値を**相互統計量**に設定します。

   これにより、語が重篤な反応に対応する場合に、その語に対する重み付けが変化するようになります。

データのクラスタリング

テキストクラスタノードは、ドキュメントをクラスタリングすることで、特定の記述語に関するドキュメントやレポートの互いに疎な集合を作成します。次の 2 つのアルゴリズムが利用できます。期待値最大化アルゴリズムは、フラット表示を使用してドキュメントをクラスタリングします。一方、階層クラスタリングアルゴリズムは、クラスタをツリー階層へとグループ化します。両アプローチとも特異値分解(SVD)を使用して、元の重み付きの語/ドキュメントの頻度マトリックスを、高密度ではあるが低次元の表現へと変換します。**テキストクラスタ**ノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

データをクラスタリングするには、次の操作を実行します。

1. ノードツールバー上で**テキストマイニング**タブを選択し、**テキストクラスタ**ノードをダイアグラムワークスペースへとドラッグします。
2. **テキストフィルタ**ノードを**テキストクラスタ**ノードに接続します。

3. **テキストクラスタ**ノードを選択します。
4. **記述語**を 12 に設定し、クラスタのラベリングを許可します。
5. ダイアグラムワークスペース内にある**テキストクラスタ**ノードを右クリックし、**実行**を選択します。
6. パスを実行するかどうかを尋ねられたら、確認ダイアログボックスで**はい**をクリックします。
7. **テキストクラスタ**ノードの実行完了後に表示される実行ステータスダイアログボックス内で **OK** をクリックします。

結果の表示

プロセスフローダイアグラムの実行が完了した後、各ノードから取得した結果を表示できます。

1. **テキスト解析**ノードを選択します。

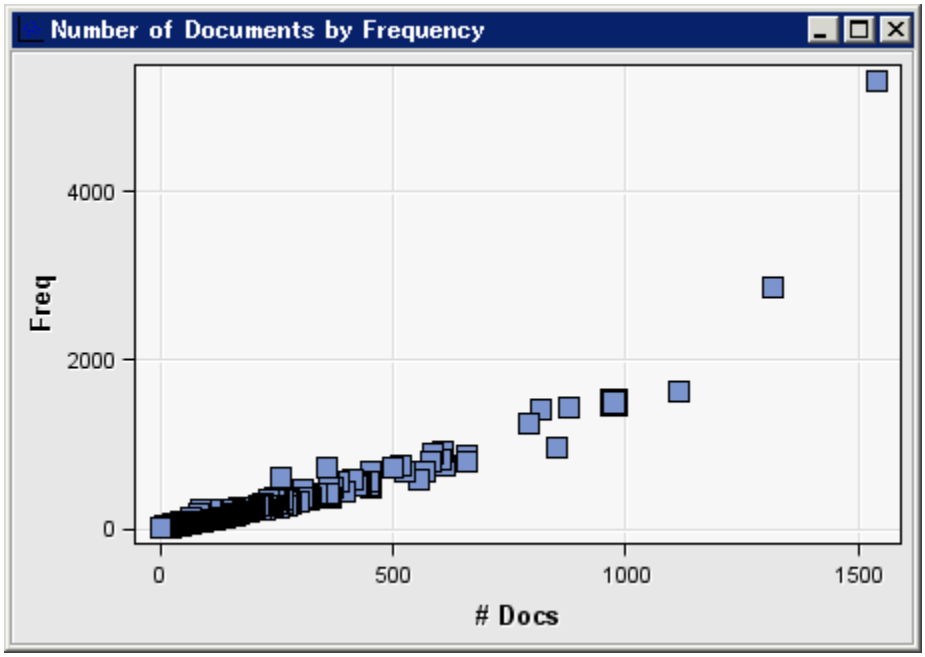
   テキスト解析ノードのプロパティがプロパティパネルに表示されます。

   テキスト解析ノードの**解析変数**プロパティには、SYMPTOM_TEXT 変数により値が割り当てられていることに注意してください。これは、SYMPTOM_TEXT 変数が、**VAEREXT_SERIOUS** 入力データソース内で**テキスト**役割を持つ最長の変数であったためです。
2. **テキスト解析**ノードを右クリックし、**結果**を選択します。

   テキスト解析ノードの結果ウィンドウが表示されます。
3. 語ウィンドウを選択します。
4. **Freq** 列見出しをクリックして、語を頻度順に並べ替えます。

語のリストをスクロールします。それぞれの語に関して、語ウィンドウには、その語が出現するドキュメントの数、その語の頻度、その語が保持されているかどうかが示されます。

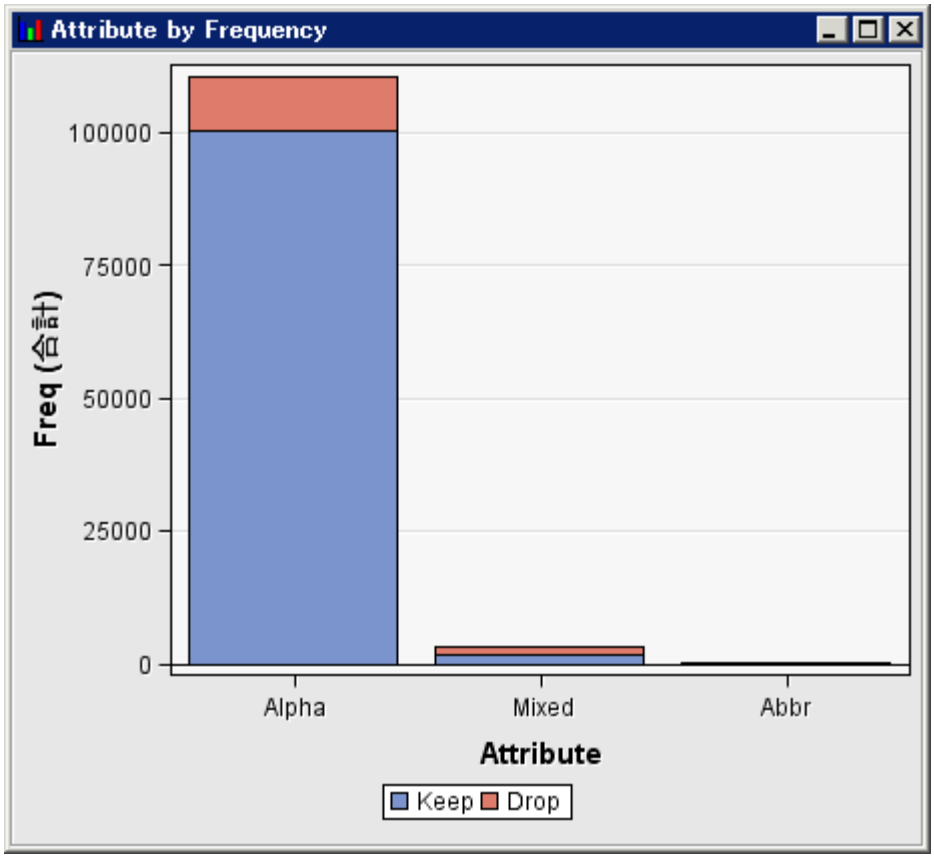
5. 語を選択します。この語に対応する点が、ZIPF プロットおよびドキュメント数と頻度プロットで選択されることに注意してください。



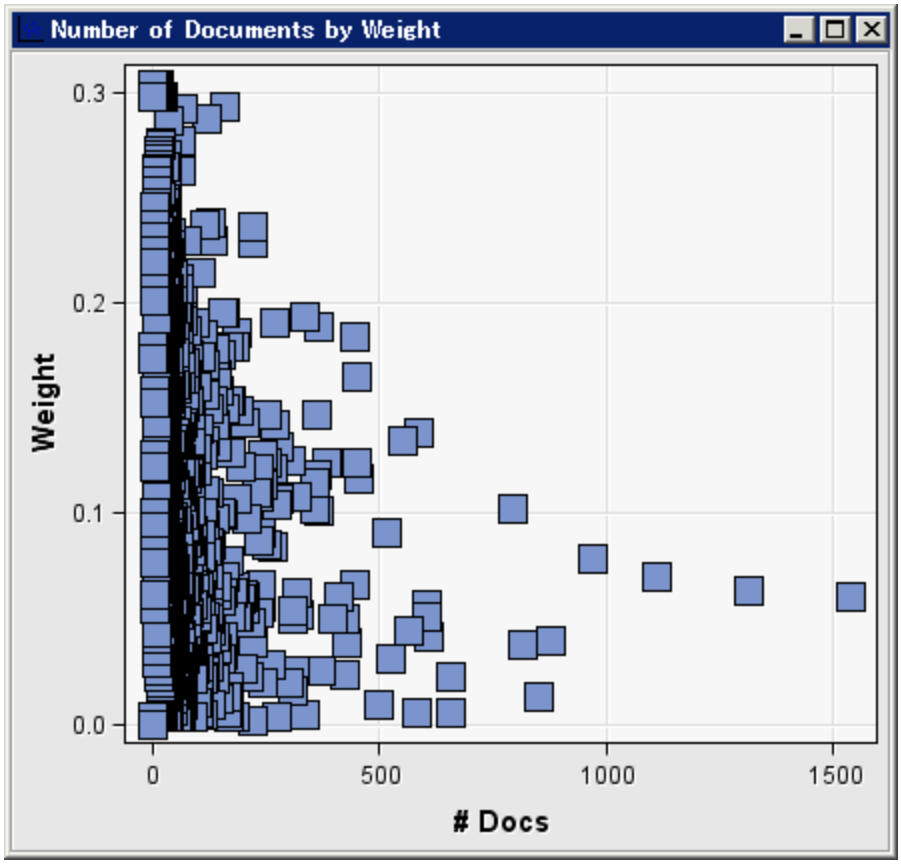
6. 結果ウィンドウを閉じます。
7. **テキストフィルタ**ノードを選択します。
8. **テキストフィルタ**ノードを右クリックし、**結果**を選択します。

   テキストフィルタノードの結果ウィンドウが表示されます。

   属性と頻度ウィンドウおよび役割と頻度ウィンドウには、各カテゴリ内で破棄された語の数または保持されている語の数が表示されることに注意してください。



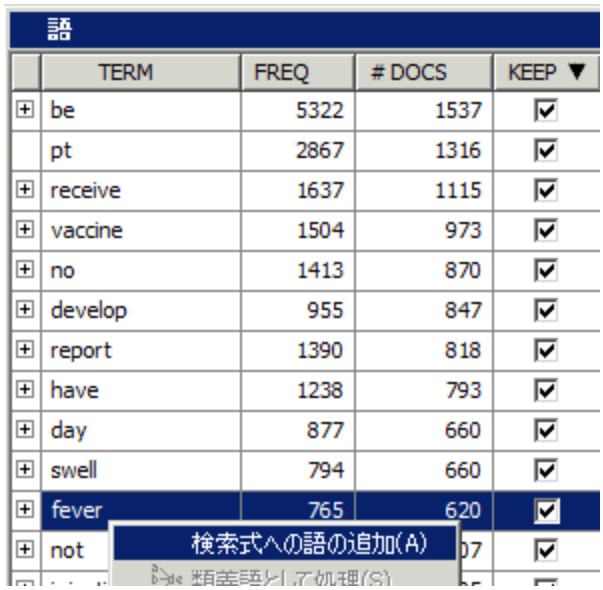
ドキュメント数と重みプロットには、それぞれの語の出現を各語に割り当てられている重みに基づいてカウントしたドキュメント数が表示されます。



9. 結果ウィンドウを閉じます。
10. **フィルタビューア**プロパティの ... をクリックします。

    対話型のフィルタビューアウィンドウが表示されます。
11. 語ウィンドウ内の語を確認するには次のようにします。語はまず保持ステータスに基づいて並べ替えられた後、それらが出現するドキュメント数に基づいて並べ替えられます。

    注: 並べ替えの順番を変更するには、列見出しをクリックします。
12. ドキュメントウィンドウ内にあるドキュメントを確認します。
13. SYMPTOM_TEXT に含まれている全文を表示するには、**SYMPTOM_TEXT** 内にあるセルを右クリックした後、**全文表示の切り替え**を選択します。
14. より詳細に調査したい有害反応に関連する語を選択します。たとえば、語ウィンドウの TERM 列の下にある **fever** を選択します。その語を右クリックし、**検索式への語の追加**を選択します。



15. **適用**をクリックします。

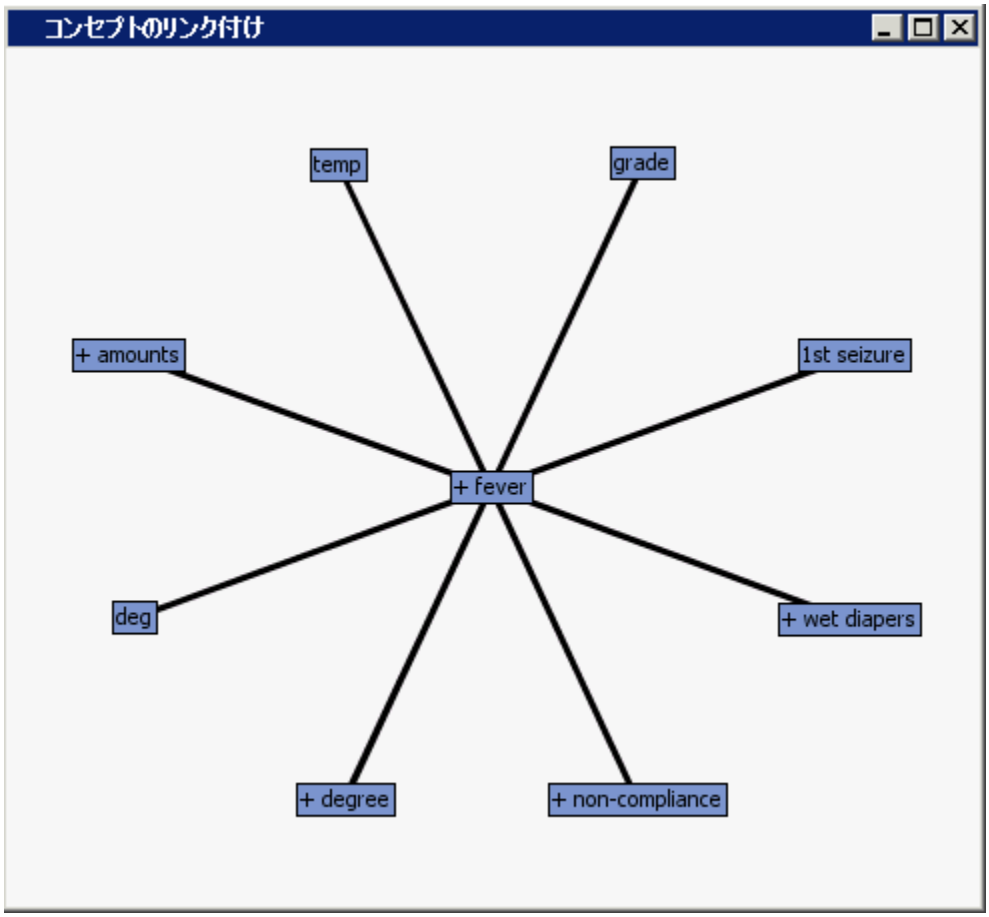
    ドキュメントウィンドウが更新され、語 **fever** を含んでいるエントリのみが同ウィンドウ内に表示されます。

16. **クリア**をクリックした後、**適用**をクリックします。

    語ウィンドウ内の語がリセットされます。

17. 語ウィンドウ内の語 **fever** を選択した後、それを右クリックし、**コンセプトリンクの表示**を選択します。

    コンセプトのリンク付けウィンドウが表示されます。コンセプトのリンク付けとは、語テーブル内の選択された語に概念的に関連付けられている語を検索し表示する方法の一つです。選択された項目は、その項目と最も強い相関を持つ語で囲まれています。コンセプトのリンク付けウィンドウには、ツリー構造の中央に語 **fever** を持つハイパーボリックツリーのグラフが表示されます。このグラフには、語 **fever** に対して強い相関を持つその他の語が表示されます。

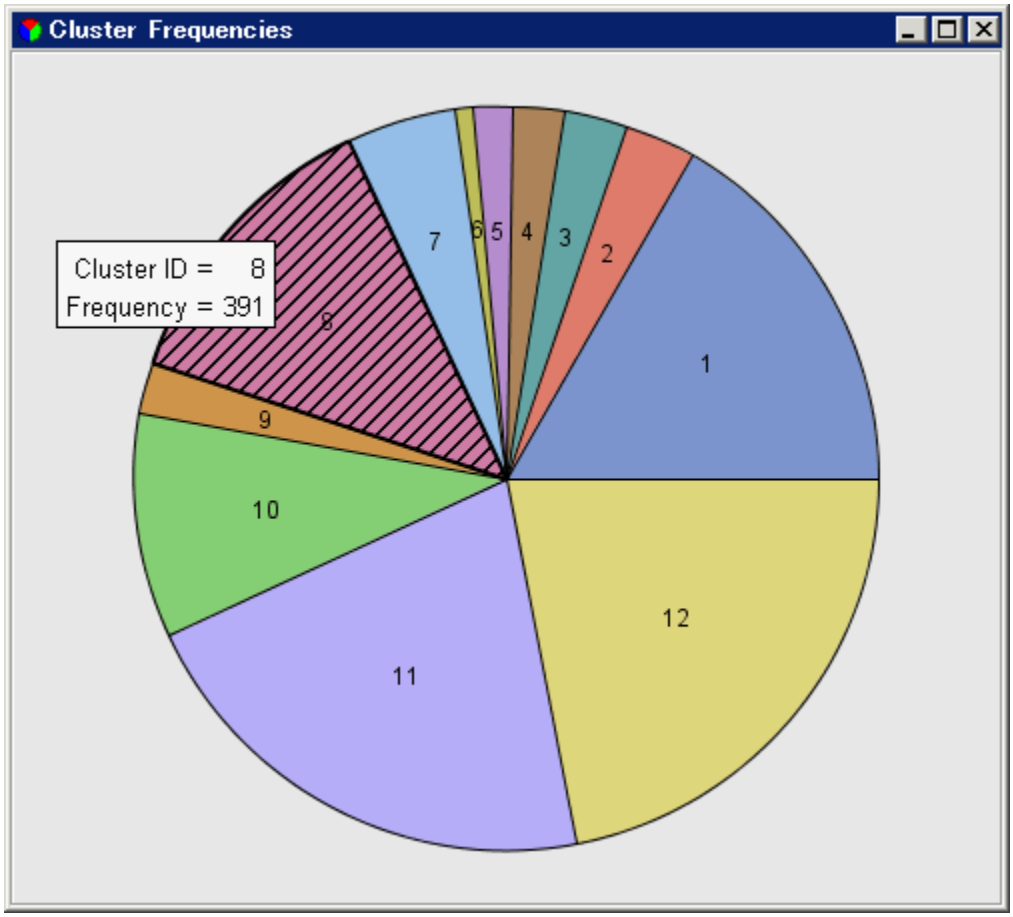


コンセプトのリンク付けのビューを展開するには、グラフの中央にない語を選択し、それを右クリックした後、**リンクの展開**を選択します。

18. 結果ウィンドウを閉じます。
19. **テキストクラスタ**ノードを選択します。
20. **テキストクラスタ**ノードを右クリックし、**結果**を選択します。

    結果ウィンドウが表示されます。
21. クラスタウィンドウ内にあるクラスタを確認します。クラスタを選択します。

    対応するクラスタが、クラスタ頻度チャート、クラスタ頻度と RMS プロット、クラスタ間の距離プロット内でどのように選択されているかを確認します。



22. 結果ウィンドウを閉じます。

データセグメントの確認

本セクションでは、**セグメントプロファイル**ノードを使用して、セグメント化されたデータやクラスタリングされたデータを確認します。セグメントとは、SAS Text Miner のクラスタリング手法を使用して分析的に導出されるクラスタ番号のことです。**セグメントプロファイル**ノードを使用すると、何が各セグメントを一意にしているか、または何が各セグメントを少なくとも母集団とは異なるようにしているかを、より良く理解できるようになります。また、同ノードを使用すると、セグメントや母集団内でのそれらの因子の分布の調査や比較に役立つ各種のレポートを生成できます。**セグメントプロファイル**ノードに関する詳細は、SAS Enterprise Miner のヘルプを参照してください。

セグメントを確認するには、次の手順を実行します。

1. ノードツールバー上で**アクセス**タブを選択し、**セグメントプロファイル**ノードをダイアグラムワークスペースへとドラッグします。
2. **テキストクラスタ**ノードを**セグメントプロファイル**ノードに接続します。

3. **セグメントプロファイル**ノードを選択します。
4. **変数**プロパティの...をクリックします。

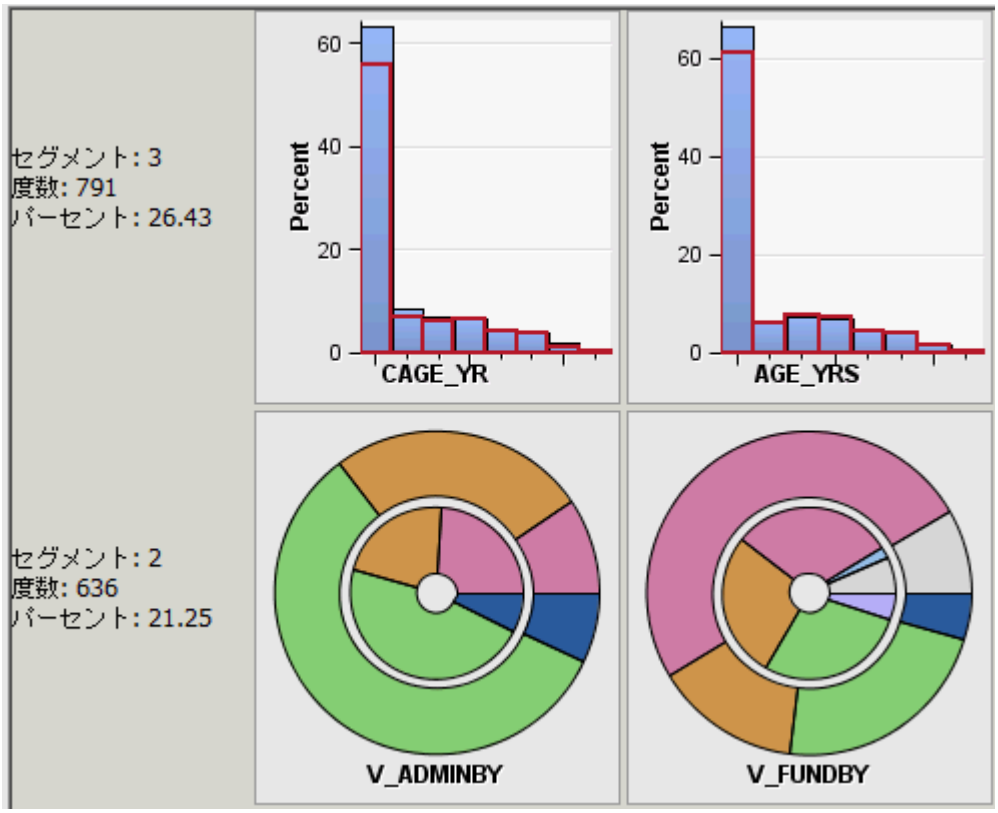
   変数ウィンドウが表示されます。
5. _prob 変数を選択し、それらの Use 値を **No** に設定します。

   注: Shift キーを押しながら、先頭の_prob 変数をクリックしてポインタをドラッグすることにより、すべての_prob 変数を選択します。すべての_prob 変数を選択した後、それらの_prob 変数の 1 つの Use 値を変更することで、個々の_prob 変数の Use 値を変更できます。これにより、それ以外の_prob 変数の Use 値が選択した値へと変更されます。

名前	使用	レポート	役割	水準
TextCluster_prob1	いいえ	いいえ	リジェクト	間隔
TextCluster_prob2	いいえ	いいえ	リジェクト	間隔
TextCluster_prob3	いいえ	いいえ	リジェクト	間隔
TextCluster_prob4	いいえ	いいえ	リジェクト	間隔
TextCluster_prob5	いいえ	いいえ	リジェクト	間隔
TextCluster_prob6	いいえ	いいえ	リジェクト	間隔
TextCluster_prob7	いいえ	いいえ	リジェクト	間隔
TextCluster_prob8	いいえ	いいえ	リジェクト	間隔
VAERS_ID	デフォルト	いいえ	ID	間隔

6. _SVD 変数をすべて選択し、それらの **Use** 値を **No** に設定します。
7. **OK** をクリックします。
8. ダイアグラムワークスペース内で**セグメントプロファイル**ノードを選択します。
9. **Minimum Worth** プロパティの値として *0.0010* を入力します。
10. **セグメントプロファイル**ノードを右クリックし、**実行**を選択します。
11. パスを実行するかどうかを尋ねられたら、確認ダイアログボックスで**はい**をクリックします。
12. 同ノードの実行が完了したら、実行状態ダイアログボックス内の**結果**をクリックします。
13. プロファイルウィンドウを最大化します。

    このウィンドウの一部を次の図に示します。



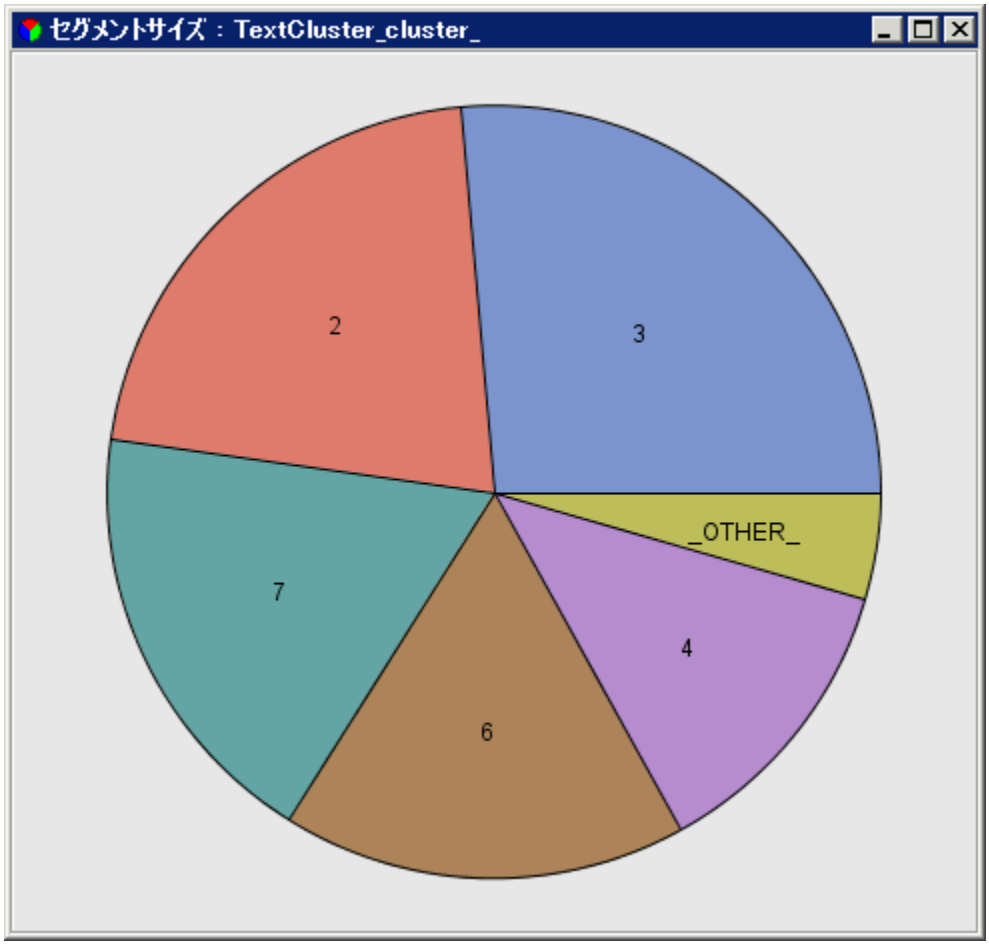
プロファイルウィンドウにはプロットの格子(グリッド)が表示されるため、これによりセグメントと母集団の両方に関して指定された変数やレポート変数の分布を比較できます。このウィンドウ内に表示されるグラフには、セグメントをその母集団から区別する因子として指定された変数が示されます。各行は単一のセグメントを表します。左端の余白には、セグメント、そのカウント、総母集団に対するパーセンテージが示されます。

列は、セグメントを母集団から区別する能力に応じて、左から右へと配置されます。レポート変数(指定された場合)は、右側の部分に、選択された入力の後にアルファベット順に表示されます。格子グラフは次の機能を持ちます。

- クラス変数 — 2 つの同心円を含む、2 つの入れ子状になった円グラフとして表示されます。内側の円は、総母集団の分布を表します。外側の円は、指定のセグメントの分布を表します。

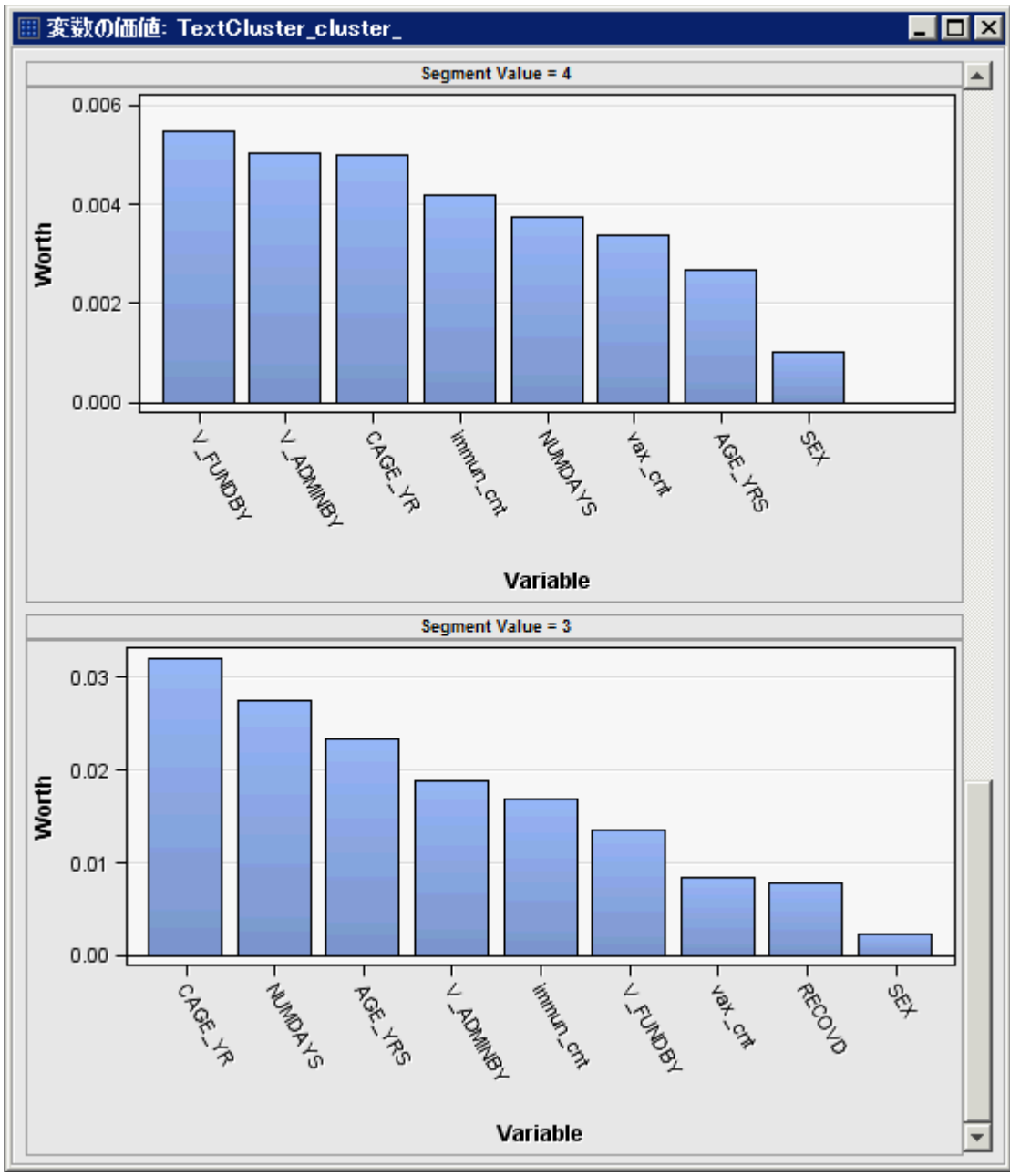
- 間隔変数 — ヒストグラムとして表示されます。青い網掛け領域は、セグメント内の分布を表します。赤い輪郭で示された領域は、母集団の分布を表します。ヒストグラムのバーの高さは、カウントまたはセグメントの母集団に対するパーセンテージを表しています。パーセンテージを使用する場合、グラフにはセグメントと母集団の間の相対的差異が表示されます。カウントを使用する場合、グラフにはセグメントと母集団の間の絶対的差異が表示されます。

14. セグメントサイズチャートを最大化します。



15. Variable Worth ウィンドウを最大化します。

    このウィンドウの一部を次の図に示します。



16. **結果**ウィンドウを閉じます。

5 章

テキストのクリーンアップ

実行するタスクについて

前の章で示したように、SAS Text Miner はデータ内で明確となっているテーマを見つける場合に良い仕事をします。ただし、データがクリーニングを必要としている場合、SAS Text Miner による有益なテーマの検出は有効性が低下します。本章では、多くのスペルミスや略語を含んでいる手動で編集を行ったデータに遭遇した場合に、そのデータをクリーニングすることで、より良い結果が得られるようにする方法を紹介します。

本書の zip ファイル内に含まれている README.TXT ファイルには、有害事象レポートで一般的に使用される略語のリストが含まれています。SAS Text Miner を使用することで、類義語リストを指定できます。本書の zip ファイル内には、VAER_ABBREV 類義語リストが含まれています。このような類義語リストを作成するために、README.TXT 内の略語リストを Microsoft Excel ファイルにコピーしたとします。この Microsoft Excel ファイル形式のリストが手動で編集された後、SAS データセットへインポートされたとします。ここで、たとえば、CT という語が"computerized axial tomography"の略語としてマークされたとします。

データを SAS データセットへとインポートする方法についての詳細は、次のドキュメントリソースを参照してください。

http://support.sas.com/documentation/

テキストをクリーニングし、その結果を確認するには、次のタスクを実行します。

1. 本書の zip ファイル内に含まれている類義語データセットを使用します。
2. **SAS コード**ノードと%TEXTSYN マクロを使用して、新しい類義語データセットを作成します。%TEXTSYN マクロはすべての語を評価することにより、スペルが誤っている語を自動的に特定し、正しいスペルの語を誤ったスペルの語に対応付ける類義語リストを作成します。
3. マージされた類義語データセットを使用して結果を確認します。

類義語データセットの使用

新しい類義語データセットを使用するには、次の操作を実行します。

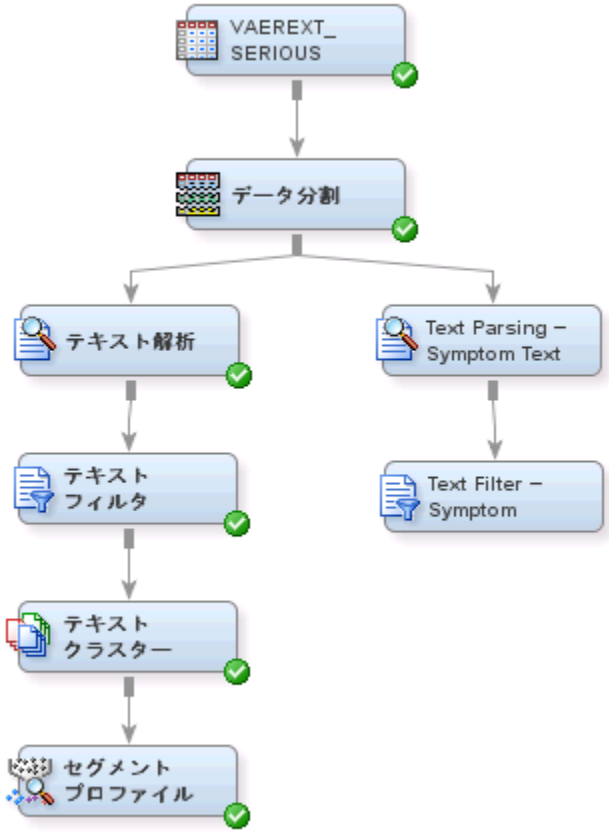
1. **テキスト解析**ノードを右クリックし、**コピー**を選択します。

   この例では、新しい**テキスト解析**ノードを作成する代わりに、既存のノードをコピーします。これは、以前に**テキスト解析**ノードのプロパティパネルで指定した設定が使われるようにするためです。

2. 空のダイアグラムワークスペースを右クリックし、**ペースト**を選択します。
3. この新しくペーストされた**テキスト解析**ノードをコピー元のノードから区別するために、新しいノードを右クリックし、**名前の変更**を選択します。
4. **ノード名**フィールドに *Text Parsing — Symptom Text* と入力した後、**OK** をクリックします。
5. **テキストフィルタ**ノードを右クリックし、**コピー**を選択します。

   この例では、新しい**テキストフィルタ**ノードを作成する代わりに、既存のノードをコピーします。これは、以前に**テキストフィルタ**ノードのプロパティパネルで指定した設定が使われるようにするためです。

6. 空のダイアグラムワークスペースを右クリックし、**ペースト**を選択します。
7. この新しくペーストされた**テキストフィルタ**ノードをコピー元のノードから区別するために、新しいノードを右クリックし、**名前の変更**を選択します。
8. **ノード名**フィールドに *Text Filter — Symptom Text* と入力した後、**OK** をクリックします。
9. **データ分割**ノードを **Text Parsing — Symptom Text** ノードに接続します。
10. **Text Parsing — Symptom Text** ノードを **Text Filter — Symptom Text** ノードに接続します。



11. **Text Parsing — Symptom Text** ノードを選択します。
12. **類義語**プロパティの ... を選択します。

    ダイアログボックスが表示されます。
13. **テーブルの交換**をクリックします。

    SAS テーブルの選択ダイアログボックスが表示されます。
14. フォルダツリーから **Mylib** ライブラリを選択します。

    Mylib ライブラリの内容が表示されます。
15. **Vaer_abbrev** を選択して **OK** をクリックします。

    Vaer_abbrev データソースの内容がダイアログボックスに表示されます。
16. 確認ウィンドウで**はい**をクリックします。
17. **OK** をクリックします。

    その他のすべての設定は、元の**テキスト解析**ノードの設定と同じままにします。
18. ダイアグラムワークスペース内にある **Text Filter — Symptom** ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
19. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で **OK** をクリックします。
20. **Text Filter — Symptom Text** ノードの**フィルタビューア**プロパティの ... をクリックします。

    対話型のフィルタ ビューアウィンドウが表示されます。

21. TERM 列見出しをクリックして、語テーブルを頻度順に並べ替えます。
22. たとえば、語ウィンドウの TERM 列の下にある **abdomen** を選択します。

    語を表示するために下スクロールが必要となる場合があります。語ウィンドウで、**abdomen** の隣にプラス記号(+)が表示されます。このプラス記号をクリックすると、この語を展開できます。これにより、この語に対応付けられているすべての類義語や語幹が表示されます。語幹とは、語の原形です。展開された語として **abd** が含まれています。**abdomen** と **abd** は両方とも同じものとして扱われます。

	TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE
	abdome	1	1	☐	0.0	
⊟	abdomen	40	39	☑	0.014	
	abdomen	38	37			
	abd	2	2			
	abdominal area	2	2	☐	0.0	Noun Group

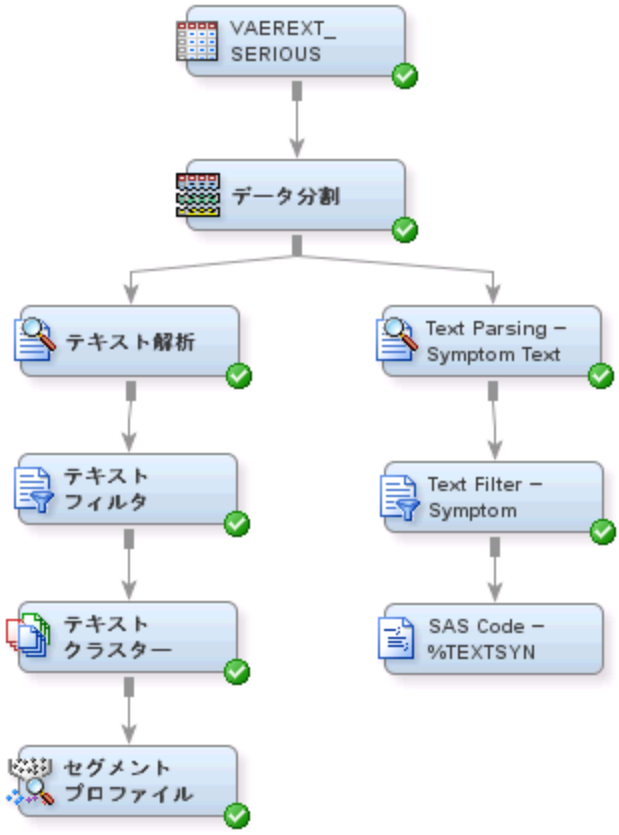
23. 対話型のフィルタビューアウィンドウを閉じます。

新しい類義語データセットの作成

SAS Text Miner の%TEXTSYN マクロを使用すると、新しい類義語データセットを作成できます。%TEXTSYN マクロはすべての語を評価することにより、スペルが間違っている語を自動的に特定し、正しいスペルの語を誤ったスペルの語に対応付ける類義語リストを作成します。

新しい類義語データセットを作成するには、次の操作を実行します。

1. ノードツールバー上で**ユーティリティ**タブを選択し、**SAS コード**ノードをダイアグラムワークスペースへとドラッグします。
2. **SAS コード**ノードを右クリックし、**名前の変更**を選択します。
3. **ノード名**フィールドに *SAS Code — %TEXTSYN* と入力した後、**OK** をクリックします。
4. **Text Filter — Symptom Text** ノードを **SAS Code — %TEXTSYN** ノードに接続します。



5. **SAS Code — %TEXTSYN** ノードを選択した後、プロパティパネル内の**コードエディタ**プロパティの ... をクリックします。

   コードエディタウィンドウが表示されます。

6. コードエディタ内に次のプログラムを入力します。

```
%textsyn( termds=<libref>.<nodeID>_terms
        , docds=&em_import_data
        , outds=&em_import_transaction
        , textvar=symptom_text
        , mnpardoc=8
        , mxchddoc=10
        , synds=mylib.vaerextsyns
        , dict=mylib.engdict
        , maxsped=15
        ) ;
```

注: 上記のプログラム内の最初の行に含まれている**<libref>**と**<nodeID>**は、それぞれ各自が使用する正しいライブラリ名とノード ID で置き換える必要があります。これらの値が何であるかを決定するには、コードエディタウィンドウを閉じた後、**Text Filter — Symptom Text** ノードを **SAS Code — %TEXTSYN** ノードに接続している矢印を選択します。<libref>の値は、プロパティパネルに表示されているテーブル名の先頭部分になります(**emws** や **emws2** など)。ノード ID は、<libref>値の後に表示されているものであり、**TextFilter** や **TextFilter2** などになります。<libref>と<nodeID>の値を正しい値で置き換えた場合、先頭行は **termds=emws2.textfilter2_terms** のようになります。実際のライブラリ参照名とノード ID の値は、どれだけの数の**テキストフィル**

タノードとダイアグラムが各自のワークスペース内に作成されているかによって異なります。

%TEXTSYN マクロに関する詳細は、SAS Text Miner のヘルプを参照してください。

7. %TEXTSYN マクロコードをコードエディタウィンドウに追加した後、`<libref>`と`<nodeID>`の値を変更したら、  をクリックして変更を保存します。
8.  をクリックして、**SAS Code — %TEXTSYN** ノードを実行します。
9. 確認ダイアログボックスで**はい**を選選択します。
10. 同ノードの実行完了後に表示されるダイアログボックス内で **OK** をクリックします。
11. **コードエディタ**ウィンドウを閉じます。
12. メインメニューから**表示** ⇨ **エクスプローラ**を選択します。

    エクスプローラウィンドウが表示されます。
13. SAS ライブラリツリー内にある **Mylib** をクリックした後、**Vaerextsyns** を選択します。

    注: **Mylib** ライブラリがすでに選択済みであり、Vaerextsyns データセットが表示されない場合、**プロジェクトデータの表示**をクリックするか、またはエクスプローラウィンドウを更新して、**Vaerextsyns** データセットを表示する必要があります。
14. **Vaerextsyns** をダブルクリックして、その内容を表示します。

	example1	example2	Term	parent
116	… 4 days following. Heat, !!anti-inflammatories!!. 1/18/2002, Muscle relaxant …	… over the counter non steroidal !!anti-inflam …	anti-inflammatories	anti-inflammatory
117	… TO TOUCH. WAS GIVEN !!ANTIBOTICS!! IF SWELLING INCREASES.	… where they gave her !!antibotics!/steroids.	antibotics	antibiotics
118	… erythematous papules over B/L !!anticubital!! spaces, elbows RLQABD, Axilla,…	… hives on face, neck, ! !anticubital!! and right foot and …	anticubital	antecubital

Vaerextsyns 列の内容を次に示します。

- **Term** はスペルが誤っている語です。
- **parent** は、その語の正しいスペルであると推測された値です。
- **example1** と **example2** は、ドキュメント内にある語を表す 2 つの例です。
- **childndocs** は、スペルが間違っている語を含んでいたドキュメントの数です。
- **numdocs** は、parent を含んでいたドキュメントの数です。
- **minsped** は、これらの語がどれだけ近似しているかを表す指標です。
- **dict** は、その語が正しい英語の単語であるかどうかを示します。正しい単語であってもスペルが間違っていると判断される場合もありますが、そのようなケースが稀であるならば、正しい単語は頻繁にターゲットとなる語のスペルに非常に近くなります。

たとえば、オブザベーション 117 では、**antibotics** は **antibiotics** のスペルミスであると示されています。これは、**antibotics** は 4 つのドキュメントにしか含まれていないのに、その parent である antibiotics は 745 件のドキュメントに含まれているためです。スペルが間違っている語は、2 個の連続する感嘆符(!!)で囲まれていることに注意してください。

15. Vaerextsyns テーブルを調べて、行われている選択に反対するかどうかを確認します。この例では、%TEXTSYN マクロがスペルミスの検出に関して十分に良い仕事をしたと仮定します。

    注: Vaerextsyns テーブルを編集するには SAS テーブルエディタを使用します。このテーブルは、SAS Enterprise Miner GUI では編集できません。スペルミスの parent が正しく表示されていない場合、その語を変更するか、または Term 列に有効な語が含まれている場合、列を削除できます。

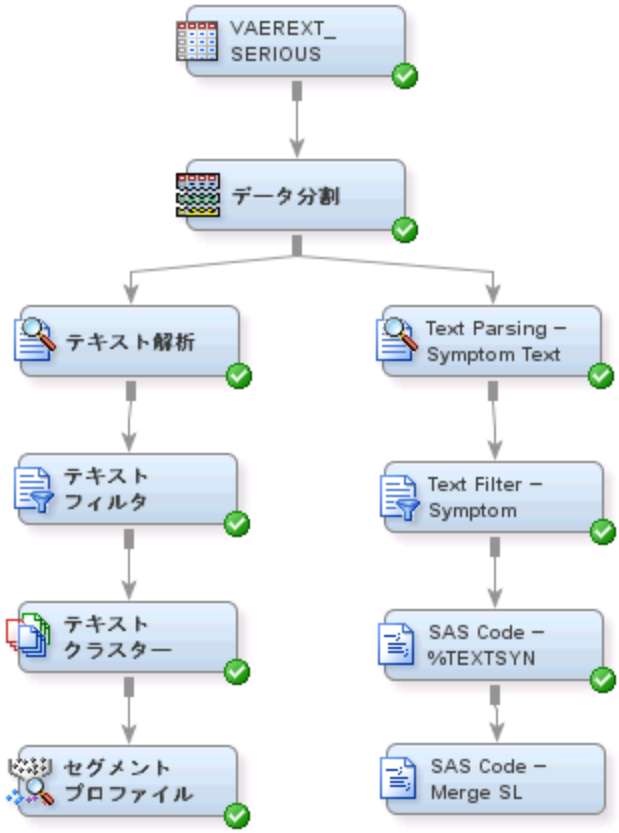
16. Mylib.Vaerextsyns テーブルとエクスプローラウィンドウを閉じます。

マージ済み類義語データセットの使用

この一連のタスクでは、Mylib.Vaerextsyns データセットと Mylib.Vaer_abbrev データセットの両方に含まれているすべてのオブザベーションを含む新しデータセットを作成します。その後、このマージされた類義語データセットを使用して結果を調べます。次の手順を実行します。

1. ノードツールバー上で**ユーティリティ**タブを選択し、**SAS コード**ノードをダイアグラムワークスペースへとドラッグします。
2. **SAS コード**ノードを右クリックし、**名前の変更**選択します。
3. **ノード名**フィールドに *SAS Code — Merge SL* と入力した後、**OK** をクリックします。

   SL は *Synonym Lists* を意味します。
4. **SAS Code — %TEXTSYN** ノードを **SAS Code — Merge SL** ノードに接続します。



5. **SAS Code — Merge SL** ノードを選択します。
6. **コードエディタ**プロパティの ... をクリックします。

   コードエディタが表示されます。
7. コードエディタ内に次のプログラムを入力します。

```
data mylib.vaerextsyns_new;
     set mylib.vaerextsyns mylib.vaer_abbrev;
run;
```

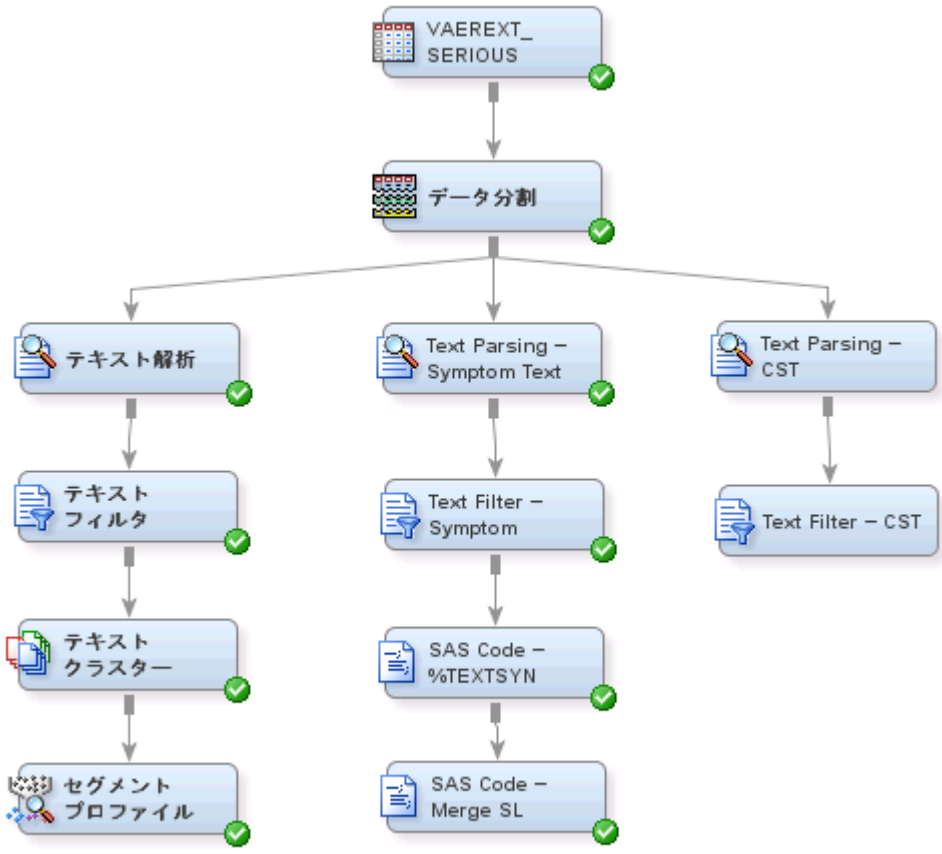
   このプログラムは、最初の **SAS Code — %TEXTSYN** ノードで作成された類義語データセットを、略語データセットとマージします。
8. をクリックします。
9. コードエディタウィンドウを閉じます。
10. **SAS Code — Merge SL** ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
11. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。
12. 結果ウィンドウで**表示** ⇨ **SAS 結果** ⇨ **ログ**を選択し、新しいデータセットを作成する SAS コードを確認します。

    結果ウィンドウを閉じます。
13. **Text Parsing — Symptom Text** ノードを右クリックし、**コピー**を選択します。

注: 新しい Text Miner ノードを作成するのではなく、**Text Parsing — Symptom Text** ノードをコピーする必要があります。これは、以前に **Text Parsing — Symptom Text** ノードで設定したのと同じプロパティ設定を保持するためです。

14. ダイアグラムワークスペース内の空のスペースを右クリックし、**ペースト**を選択します。
15. **テキスト解析**ノードを右クリックし、**名前の変更**を選択します。
16. **ノード名**フィールドに *Text Parsing — CST* と入力します。

    CST は *Cleaned Symptom Text* を意味します。
17. **OK** をクリックします。
18. **Text Filter — Symptom Text** ノードを右クリックし、**コピー**を選択します。
19. ダイアグラムワークスペース内の空のスペースを右クリックし、**ペースト**を選択します。
20. **テキストフィルタ**ノードを右クリックし、**名前の変更**を選択します。
21. **ノード名**フィールドに *Text Filter — CST* と入力します。
22. **OK** をクリックします。
23. **データ分割**ノードを **Text Parsing — CST** ノードに接続します。
24. **Text Parsing — CST** ノードを **Text Filter — CST** ノードに接続します。



25. **Text Parsing — CST** ノードを選択します。

26. **類義語**プロパティの...をクリックします。
27. **テーブルの交換**をクリックします。
28. SAS ライブラリツリー内の **Mylib** をクリックします。

    Mylib ライブラリの内容が表示されます。
29. **Mylib.Vaerextsyns_new** を選択します。
30. **OK** をクリックします。
31. 確認ダイアログボックスで**はい**を選択します。

    データセットの内容がダイアログボックスに表示されます。
32. **OK** をクリックします。
33. **Text Filter — CST** ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
34. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で **OK** をクリックします。
35. **Text Filter — CST** ノードを選択します。
36. **フィルタビューア**プロパティの...をクリックします。

    対話型のフィルタビューアウィンドウが表示されます。
37. 下スクロールし、語テーブル内の **patient** の隣にあるプラス記号(+)を選択します。**patien**、**patietn**、**paitent** が、スペルの間違っている語として表示されていることに注意してください。

	TERM ▲	FREQ	# DOCS
⊟	patient	597	306
	patient	503	227
	patients	43	32
	patien	47	43
	patient	2	2
	patietn	2	2

38. 対話型のフィルタビューアを閉じます。

6 章

トピックとルールの作成

実行するタスクについて

本章では、**テキストトピック**ノードと**テキストルールビルダ**ノードを使用してトピックとルールを作成する方法を示します。

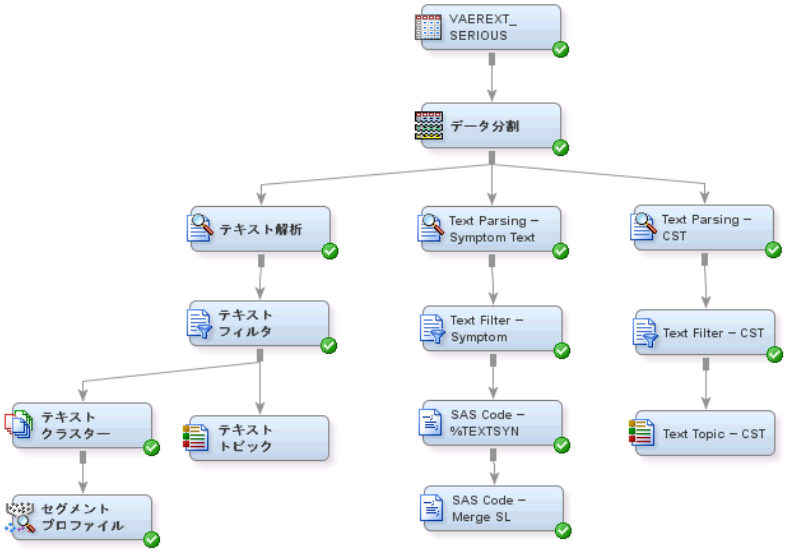
テキストトピックノードを使用すると、検出されたトピックやユーザー定義のトピックの両方に従って語とドキュメントを自動的に関連付けることにより、ドキュメントコレクションを調査できます。トピックとは、主要なテーマやアイデアを記述し特徴付ける語のコレクションです。トピックのリストを作成する目的は、分析で興味のある語の組み合わせを確立することにあります。個々の語をトピックへと結合することにより、テキストマイニング分析を改善できます。結合を通じて、分析対象となるテキストの量を、自分が興味のある語のグループ数にまで削減できます。**テキストトピック**ノードの詳細については、SAS Text Miner のヘルプを参照してください。

テキストルールビルダノードは、ターゲット変数の記述や予測に役立つルールの順序集合を、語の小規模なサブセットから生成します。この集合内の各ルールは、1 つの語または語の小規模なサブセットが存在するかどうかを示す論理積(“term1” AND “term2” AND (NOT “term3”)など)から構成される特定のターゲットカテゴリと関連付けられます。あるドキュメントが少なくとも 1 つの term1 と term2 のオカレンスを含むが term3 のオカレンスは含まない場合にのみ、そのドキュメントはこのルールにマッチします。この派生ルールの集合は、記述的かつ予測的である 1 つのモデルを生成します。新規ドキュメントを分類する場合、その作業は順序集合を通じて進められ、そのドキュメントにマッチした最初のルールと関連付けられているターゲットが選択されます。このルールは、SAS Content Categorization Studio 内部で使用可能でそこに配置可能な構文で提供されます。**テキストルールビルダ**ノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

トピックの作成

テキストをフィルタリングした後、**テキストトピック**ノードを使用してトピックを作成できます。**テキストトピック**ノードを分析で使用するには、次の操作を実行します。

1. ノードツールバー上で**テキストマイニング**タブを選択し、**テキストトピック**ノードをダイアグラムワークスペースへとドラッグします。
2. **テキストフィルタ**ノードを**テキストトピック**ノードに接続します。
3. ノードツールバー上で**テキストマイニング**タブを選択し、**テキストトピック**ノードをダイアグラムワークスペースへとドラッグします。
4. **テキストトピック**ノードを右クリックし、**名前の変更**を選択します。
5. **ノード名**フィールドに *Text Topic — CST* と入力した後、**OK** をクリックします。
6. **Text Topic — CST** ノードを選択します。
7. **複数語トピックの数**プロパティの値として 50 を入力します。
8. **Text Filter — CST** ノードを **Text Topic — CST** ノードに接続します。



9. ダイアグラムワークスペース内にある**テキストトピック**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**をクリックします。
10. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。
11. トピックテーブル内のトピックを調べて、どのような語が各トピックを形成しているかを確認します。

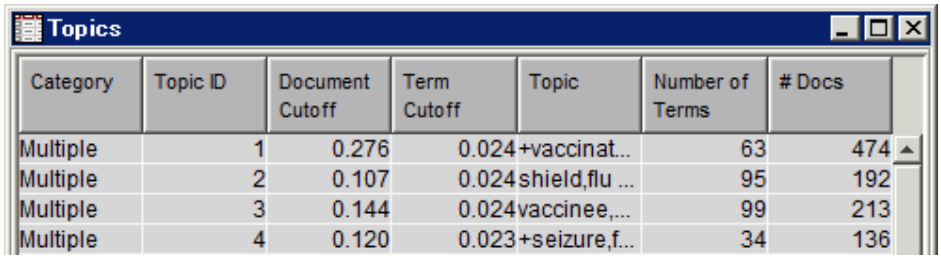
Topics

Category	Topic ID ▲	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.217	0.027	varicella,+va...	79	388
Multiple	2	0.116	0.027	shield,flu sh...	77	133
Multiple	3	0.145	0.027	vaccinee,m...	87	214
Multiple	4	0.125	0.025	+seizure,feb...	26	136

12. 結果ウィンドウを閉じます。
13. ダイアグラムワークスペース内にある **Text Topic — CST** ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**をクリックします。
14. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

15. トピックテーブル内のトピックを調べて、どのような語が各トピックを形成しているかを確認します。

    Text Topic — CST の実行により生成されたトピックは、**テキストトピック**ノードの実行により生成されたトピックとは異なることに注意してください。



Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.276	0.024	+vaccinat...	63	474
Multiple	2	0.107	0.024	shield,flu ...	95	192
Multiple	3	0.144	0.024	vaccinee,...	99	213
Multiple	4	0.120	0.023	+seizure,f...	34	136

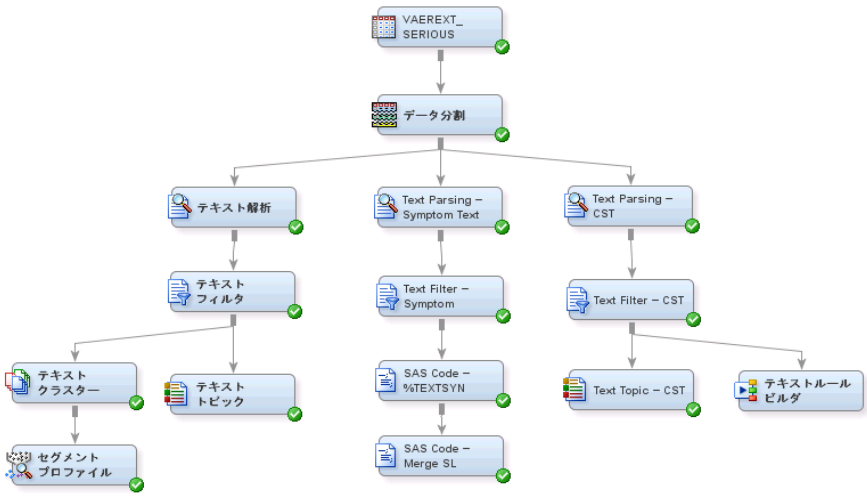
    これらの違いは、過去に実行されたテキストクリーニング活動、および複数語トピックが原因で起こります。

16. 結果ウィンドウを閉じます。

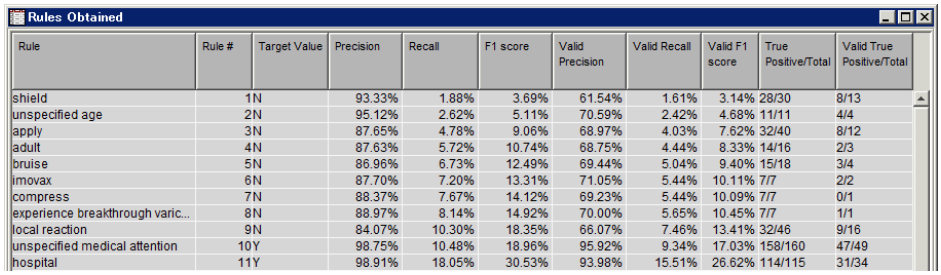
ルールの作成

テキストをフィルタリングした後、**テキストルールビルダ**ノードを使用してルールを作成できます。**テキストルールビルダ**ノードを分析で使用するには、次の操作を実行します。

1. ノードツールバー上で**テキストマイニング**タブを選択し、**テキストルールビルダ**ノードをダイアグラムワークスペースへとドラッグします。
2. **Text Filter — CST** ノードを**テキストルールビルダ**ノードに接続します。



3. ダイアグラムワークスペース内にある**テキストルールビルダ**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**をクリックします。
4. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。
5. 取得ルールウィンドウを選択します。



Rules Obtained

Rule	Rule #	Target Value	Precision	Recall	F1 score	Valid Precision	Valid Recall	Valid F1 score	True Positive/Total	Valid True Positive/Total
shield	1	N	93.33%	1.88%	3.69%	61.54%	1.61%	3.14%	28/30	8/13
unspecified age	2	N	95.12%	2.62%	5.11%	70.59%	2.42%	4.68%	11/11	4/4
apply	3	N	87.65%	4.78%	9.06%	68.97%	4.03%	7.62%	32/40	8/12
adult	4	N	87.63%	5.72%	10.74%	68.75%	4.44%	8.33%	14/16	2/3
bruise	5	N	86.96%	6.73%	12.49%	69.44%	5.04%	9.40%	15/18	3/4
imovax	6	N	87.70%	7.20%	13.31%	71.05%	5.44%	10.11%	7/7	2/2
compress	7	N	88.37%	7.67%	14.12%	69.23%	5.44%	10.09%	7/7	0/1
experience breakthrough varic...	8	N	88.97%	8.14%	14.92%	70.00%	5.65%	10.45%	7/7	1/1
local reaction	9	N	84.07%	10.30%	18.35%	66.07%	7.46%	13.41%	32/46	9/16
unspecified medical attention	10	Y	98.75%	10.48%	18.96%	95.92%	9.34%	17.03%	158/160	47/49
hospital	11	Y	98.91%	18.05%	30.53%	93.98%	15.51%	26.62%	114/115	31/34

上記の 10 番目の列で、真陽性(最初の数字)は、ルールに正しく割り当てられたドキュメントの数になります。合計(2 番目の数字)は、合計陽性になります。

上記の 11 番目の列で、有効真陽性(最初の数字)は、当該カテゴリ内の残りのドキュメントの総数になります。合計(2 番目の数字)は、残りのドキュメントの総数になります。

取得ルールウィンドウや**テキストルールビルダ**ノードに関する詳細は、13 章, “テキストルールビルダノード” (83 ページ)および SAS Text Miner のヘルプを参照してください。

6. 結果ウィンドウを閉じます。

7 章

モデルの作成と比較

実行するタスクについて

このセクションでは、**デシジョンツリー**ノードを使用してモデルを作成し、**モデル比較**ノードを使用してそれらのモデルを比較する方法を示します。

デシジョンツリーノードを使うと、名義尺度、二値、順序尺度のターゲットの値に基づいてオブザベーションを分類できます。また、間隔尺度ターゲットの結果や、複数の意思決定に関する選択肢を指定した場合の適切な決定を予測できます。エンピリカルツリーは、一連の単純なルールを適用することで作成されたデータのセグメントを表しています。各ルールにより、1 つの入力の値に基づいて、セグメントにオブザベーションが割り当てられます。

1 つのルールが順次適用され、セグメント内にセグメントの階層が作成されます。この階層はツリーと呼ばれ、各セグメントはノードと呼ばれます。最初のセグメントには、データセット全体が含まれています。このセグメントは、ツリーのルート(根)ノードと呼ばれます。1 つのノードは、そのすべての後継者と共に、それを作成したノードのブランチ(枝)を形成します。

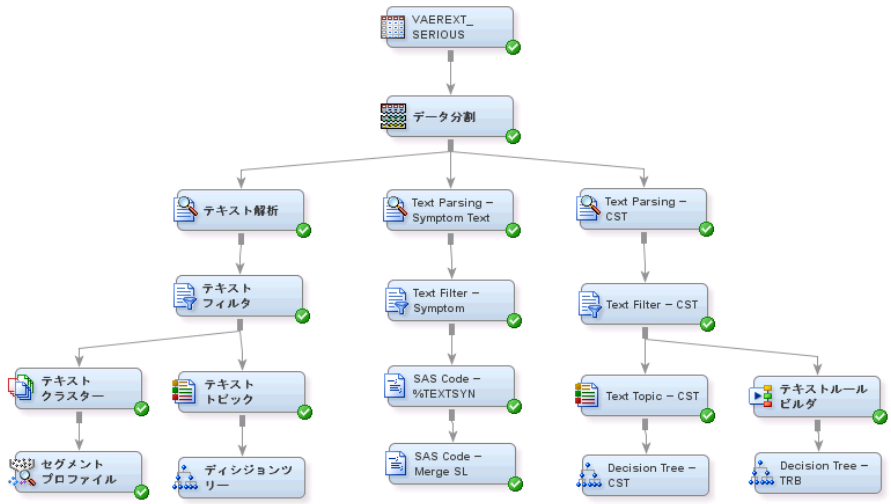
最後のノードは、リーフ(葉)と呼ばれます。リーフごとに決定が行われ、そのリーフ内のすべてのオブザベーションに適用されます。決定のタイプは文脈により異なります。予測モデリングの場合、決定は予測値になります。**デシジョンツリー**ノードに関する詳細は、SAS Enterprise Miner のヘルプを参照してください。

モデルの作成

デシジョンツリーノードを使用してモデルを作成するには次の操作を実行します。

1. ノードツールバー上で**モデル**タブを選択し、**デシジョンツリー**ノードをダイアグラムワークスペースへとドラッグします。
2. **テキストトピック**ノードを**デシジョンツリー**ノードに接続します。

3. ノードツールバー上で**モデル**タブを選択し、**デシジョンツリー**ノードをダイアグラムワークスペースへとドラッグします。
4. **デシジョンツリー**ノードを右クリックし、**名前の変更**を選択します。
5. **ノード名**フィールドに *Decision Tree — CST* と入力した後、**OK** をクリックします。
6. **Text Topic — CST** ノードを **Decision Tree — CST** ノードに接続します。
7. ノードツールバー上で**モデル**タブを選択し、**デシジョンツリー**ノードをダイアグラムワークスペースへとドラッグします。
8. **デシジョンツリー**ノードを右クリックし、**名前の変更**を選択します。
9. **ノード名**フィールドに *Decision Tree — TRB* と入力した後、**OK** をクリックします。
10. **テキストルールビルダ**ノードを **Decision Tree — TRB** ノードに接続します。



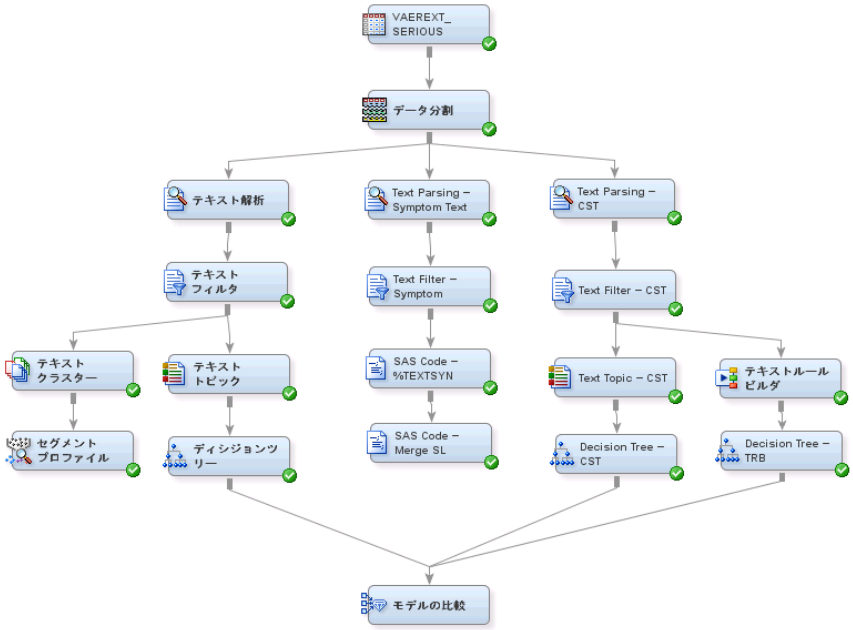
11. ダイアグラムワークスペース内にある**デシジョンツリー**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
12. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。
13. ツリーウィンドウを選択し、取得したツリーを確認します。
14. 結果ウィンドウを閉じます。
15. ダイアグラムワークスペース内にある **Decision Tree — CST** ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
16. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。
17. ツリーウィンドウを選択し、取得したツリーを確認します。このツリーは以前のツリーとどう違っているでしょうか？主な違いは、それぞれのデシジョンポイントで異なるトピックが使用されていることです。
18. 結果ウィンドウを閉じます。
19. ダイアグラムワークスペース内にある **Decision Tree — TRB** ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
20. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

21. ツリーウィンドウを選択し、取得したツリーを確認します。このツリーは以前の 2 つのツリーとどう違っているでしょうか？主な違いは、それぞれのデシジョンポイントでトピックではなく単一語または複数語のルールが使用されていることです。
22. 結果ウィンドウを閉じます。

モデルの比較

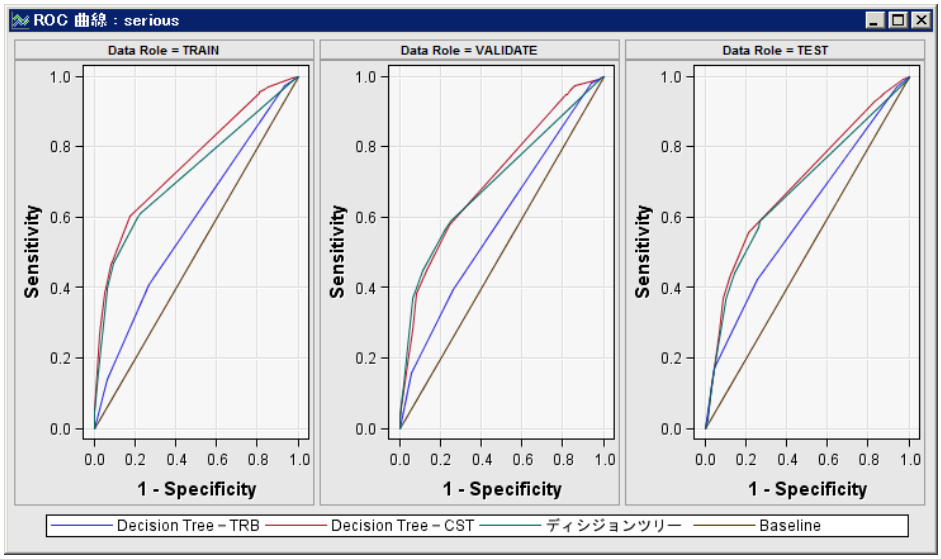
モデル比較ノードを使用してモデルを比較するには、次の操作を実行します。

1. ノードツールバー上で**アクセス**タブを選択し、**モデル比較**ノードをダイアグラムワークスペースへとドラッグします。
2. **デシジョンツリー**ノード、Decision Tree — CST ノード、Decision Tree — TRB ノードを、**モデル比較**ノードに接続します。



3. ダイアグラムワークスペース内にある**モデル比較**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。
4. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

   結果ウィンドウが表示されます。
5. ROC チャートを選択します。



曲線の下にある領域が大きいほど、モデルの性能が良いことを意味します。茶色の線は、モデルを比較する基準線を表します。青い線は、ターゲット SERIOUS の予測における **Decision Tree — TRB** モデルの性能を表します。このモデルは、**テキストルールビルダ**ノードからの入力を使用しています。緑の線は、ターゲット SERIOUS の予測における**ディシジョンツリー**モデルの性能を表します。赤い線は、ターゲット SERIOUS の予測における **Decision Tree — CST** モデルの性能を表します。

ディシジョンツリーモデルと **Decision Tree — CST** モデルは両方とも、**Decision Tree — TRB** モデルよりも優れています。**ディシジョンツリー**モデルと **Decision Tree — CST** モデルは、どちらも同じような性能です。

追加の演習として、**テキストトピック**または **Text Topic — CST** ノード内で複数または単一の語トピックの数を変更してみてください。その後、**ディシジョンツリー**ノードと **Decision Tree — CST** ノードを再実行すると、これらのモデルが改善されたことを確認できます。

8 章

テキストインポートノード

テキストインポートノードについて

テキストインポートノードは入力データノードの置き換えとして機能するものであり、これを使用することで、ディレクトリ内に含まれているファイルから、または Web 上のファイルから動的にデータセットを作成できます。**テキストインポート**ノードは、ベンダー固有フォーマット(MS Word や PDF など)のテキストファイルを含んでいる可能性のあるインポートディレクトリを入力として取得します。同ノードはこのディレクトリを調査して、ファイル内にあるテキストのフィルタリングや抽出を行い、同テキストのコピーおよび同テキストの抜粋(または全体)をプレーンなテキストファイルとして SAS データセット内に配置します。URL が指定された場合、同ノードは Web サイトをクロールし、Web からファイルを取り出し、それらをインポートディレクトリに移動した後、このフィルタリング処理を実行します。**テキストインポート**ノードの出力は、**テキスト解析**ノードにインポート可能なデータセットになります。

テキストのフィルタリングに加えて、**テキストインポート**ノードは、ドキュメントが書かれている言語を識別できるほか、ドキュメントをセッションエンコーディングへとトランスコードする処理もサポートします。エンコーディングやトランスコードに関する詳細は、SAS Text Miner のヘルプに含まれている「SAS Text Miner および SAS セッションのエンコーディング」というトピックを参照してください。

テキストインポートノードは、Windows マシン上にインストールされ実行されている SAS Document Conversion Server を利用します。このマシンは、インストール時に指定されたホスト名とポート番号を通じて SAS Enterprise Miner からアクセス可能でなければなりません。

注:

- エンコーディングが UTF-8 の SAS セッションで**テキストインポート**ノードを実行すると、同ノードは、結果データセットが UTF-8 の SAS セッションで利用できるようにするために、フィルタリングされたすべてのテキストを UTF-8 エンコーディングへとトランスコードします。その他すべての SAS セッションのエンコーディングでは、**テキストインポート**ノードはデータをトランスコードせず、入力データは当該 SAS セッションと同じエンコーディングを使用するものと仮定します。詳

細については、SAS Text Miner のヘルプに含まれている「SAS Text Miner および SAS セッションのエンコーディング」というトピックを参照してください。

- **テキストインポート**ノードは、グループ処理(開始グループノードや停止グループノード)における利用ではサポートされません。

テキストインポートノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

この章の残りの部分では、**テキストインポート**ノードの使用例を紹介します。

テキストインポートノードの使用

本セクションの内容

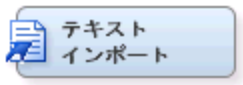
テキストインポートノードを使用すると、ディレクトリや Web からドキュメントをインポートできます。**テキストインポート**ノードの使用例については、次のページを参照してください。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

ディレクトリからのドキュメントのインポート

この例では、SAS Enterprise Miner が実行されていること、SAS Document Conversion Server が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

ディレクトリからドキュメントをインポートするには、次の手順を実行します。

1. **テキストマイニング**タブを選択し、**テキストインポート**ノードをダイアグラムワークスペースへとドラッグします。



2. **テキストインポート**ノードの**インポートファイルディレクトリ**プロパティの隣にある省略記号ボタンをクリックします。

   サーバーディレクトリの選択ダイアログボックスが開きます。

3. データセットの作成元とするドキュメントを含むフォルダへと移動し、同フォルダを選択した後、**OK** をクリックします。

   注: 選択可能なファイルの種類を確認するには、種類ドロップダウンメニューで**すべてのファイル**を選択します。

4. **言語**プロパティの隣にある省略記号ボタンをクリックします。

   言語ダイアログボックスが開きます。

5. 言語 ID を各ドキュメントの言語に割り当てる場合に必要となるライセンス言語を 1 つ以上選択した後、**OK** をクリックします。

6. (オプション)**拡張子**プロパティで、処理対象とするファイルの種類を指定します。

   たとえば、拡張子.txt および.pdf を持つファイルのみを処理対象としたい場合、**拡張子**プロパティの値として*.txt .pdf* を指定し、キーボード上の **Enter** キーを押します。

   注: 処理対象とするファイルの種類を指定しない場合、**テキストインポート**ノードは指定されたインポートファイルディレクトリ内にあるすべてのファイルタイプを処理します。

7. **テキストインポート**ノードを右クリックし、**実行**を選択します。

8. 確認ダイアログボックスで**はい**をクリックします。

9. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

10. インポートしたドキュメント内の結果を調べます。

    この時点で、**テキストインポート**ノードを、各自のテキストマイニング分析の入力データソースとして使用できるようになります。

11. **テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。

12. **テキストインポート**ノードを**テキスト解析**ノードに接続します。

13. **テキスト解析**ノードを右クリックし、**実行**を選択します。

14. 確認ダイアログボックスで**はい**を選択します。

15. これらのノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

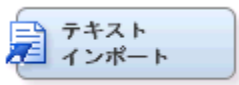
Web からのドキュメントのインポート

この例では、SAS Enterprise Miner が実行されていること、SAS Document Conversion Server が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

注: Web クロール機能は、Windows オペレーティングシステムでのみサポートされています。

Web からドキュメントをインポートするには、次の手順を実行します。

1. **テキストマイニング**タブを選択し、**テキストインポート**ノードをダイアグラムワークスペースへとドラッグします。



2. **テキストインポート**ノードの**インポートファイルディレクトリ**プロパティの隣にある省略記号ボタンをクリックします。

   サーバーディレクトリの選択ダイアログボックスが表示されます。

3. フォルダに移動し、同フォルダを選択した後、**OK** をクリックします。

   ドキュメントは、まず**インポートファイルディレクトリ**の場所に書き出されます。これらのファイルは、**インポートファイルディレクトリ**の場所で処理された後、**出力先ディレクトリ**の場所に書き出されます。

4. **テキストインポート**ノードの **URL** プロパティで、クロール対象としたい Web ページの URL を入力します。たとえば、*www.sas.com* と入力します。

5. **深さ**プロパティには、クロール対象とするレベル数として *1* を入力します。

6. **ドメイン**プロパティを**無制限**に設定します。

   注: パスワードで保護されている Web サイトをクロールしたい場合、**ドメイン**プロパティを**制限**に設定した後、**ユーザー名**プロパティにユーザー名を、**パスワード**プロパティにパスワードをそれぞれ設定します。

7. **テキストインポート**ノードを右クリックし、**実行**を選択します。

8. 確認ダイアログボックスで**はい**をクリックします。

9. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

9 章

テキスト解析ノード

テキスト解析ノードについて

テキスト解析ノードを使用すると、ドキュメント群を解析し、そこに含まれている語に関する情報を定量化できます。**テキスト解析**ノードは、e メールメッセージ、ニュース記事、Web ページ、研究報告書、調査報告書などの膨大な原文データに対して使用できます。**テキスト解析**ノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

この章の残りの部分では、**テキスト解析**ノードの使用例を紹介します。

テキスト解析ノードの使用

この例では、**テキスト解析**ノードを使用して、テキストを含んでいるデータセット内で語とそのインスタンスを特定する方法を示します。この例では、SAS Enterprise Miner が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

1. SAS データセット SAMPSIO.ABSTRACT には、さまざまな会議から収集したタイトルと概要のテキストが含まれています。ABSTRACT データソースを作成し、それをダイアグラムワークスペースに追加します。TEXT 変数および TITLE 変数のルール値を**テキスト**(`Text`)に設定します。

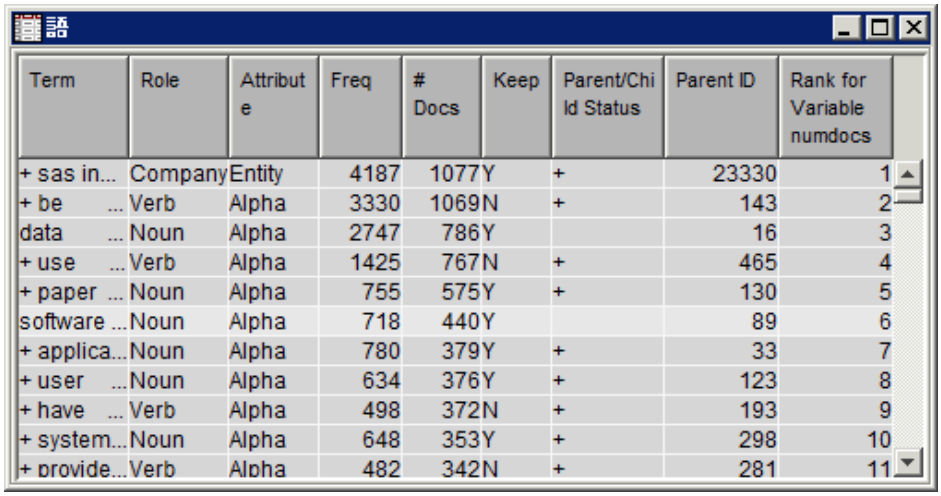
2. ツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。

3. ABSTRACT データソースを**テキスト解析**ノードに接続します。

4. ダイアグラムワークスペースで、**テキスト解析**ノードを右クリックし、**実行**を選択します。表示される確認ダイアログボックスで**はい**をクリックします。

5. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。結果ウィンドウには、ABSTRACT データソース内の語とそのインスタンスの分析に役立つ、さまざまな表形式出力やグラフィカル出力が表示されます。

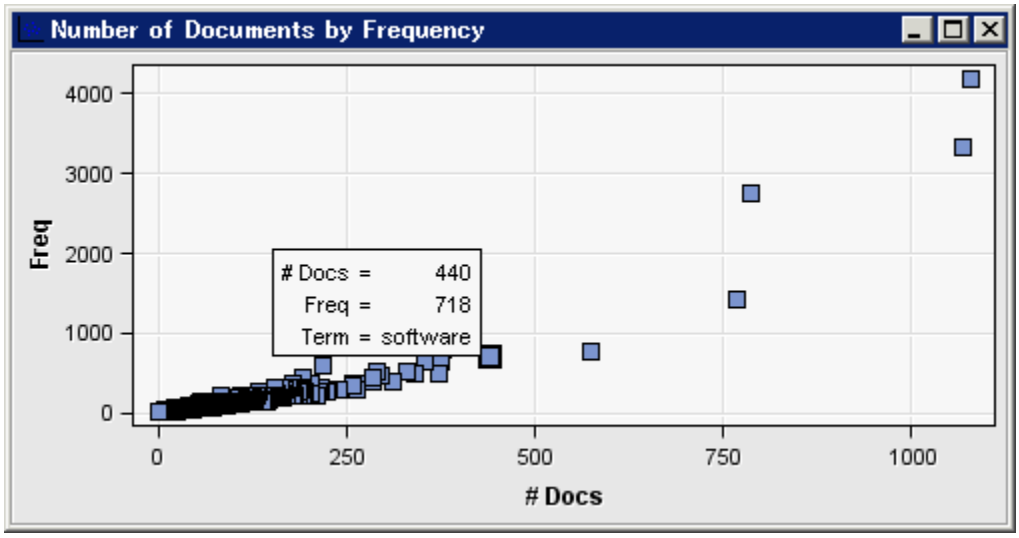
6. 語テーブル内の語を頻度順に並べ替えた後、語“software”を選択します。語テーブルに示されているように、語“software”は ABSTRACT データソース内で 440 個のドキュメントに出現している名詞であり、合計で 718 回出現しています。


語

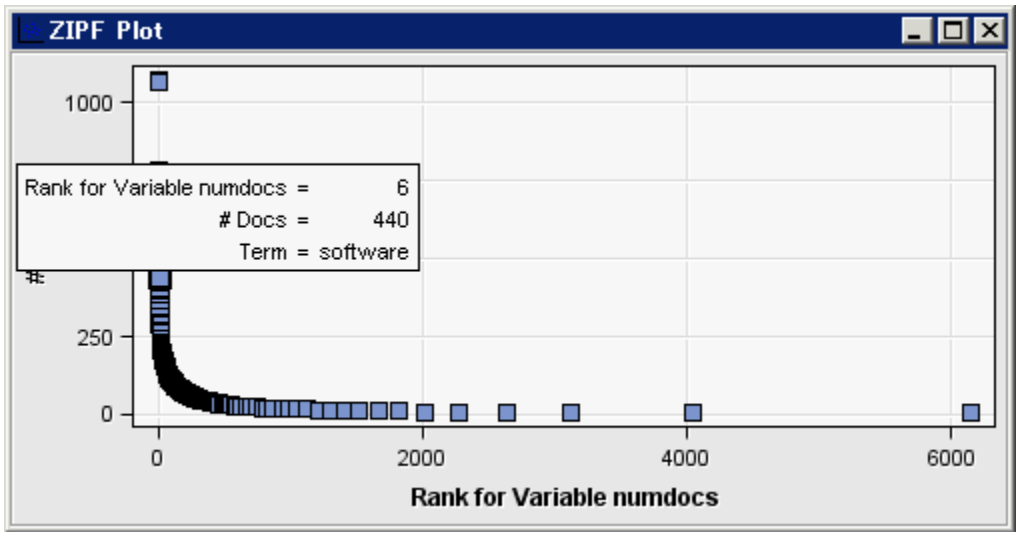
Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ sas in...	Company	Entity	4187	1077	Y	+	23330	1
+ be ...	Verb	Alpha	3330	1069	N	+	143	2
data ...	Noun	Alpha	2747	786	Y		16	3
+ use ...	Verb	Alpha	1425	767	N	+	465	4
+ paper ...	Noun	Alpha	755	575	Y	+	130	5
software ...	Noun	Alpha	718	440	Y		89	6
+ applica...	Noun	Alpha	780	379	Y	+	33	7
+ user ...	Noun	Alpha	634	376	Y	+	123	8
+ have ...	Verb	Alpha	498	372	N	+	193	9
+ system...	Noun	Alpha	648	353	Y	+	298	10
+ provide...	Verb	Alpha	482	342	N	+	281	11

語テーブルで語を選択すると、テキスト解析結果プロット内のその語に対応する点が強調表示されます。

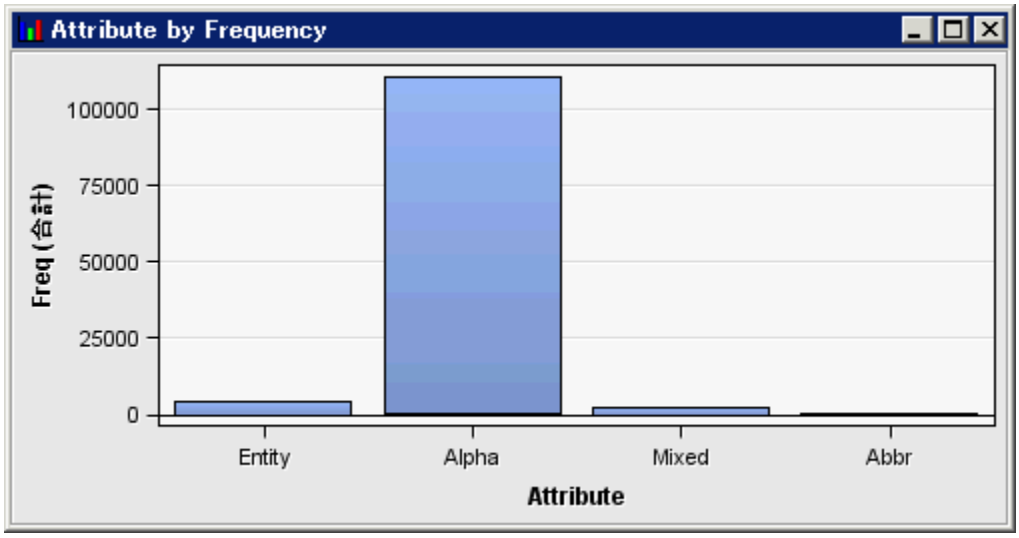
7. ドキュメント数と頻度プロットを選択し、強調表示されている点の上にカーソルを置くと、語“software”に関する情報が表示されます。



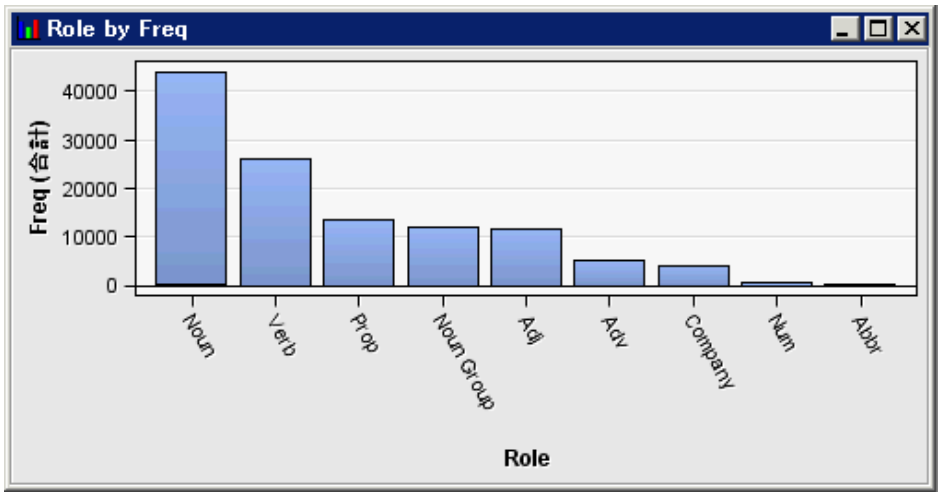
同様の情報は、ZIPF プロットでも表示されます。



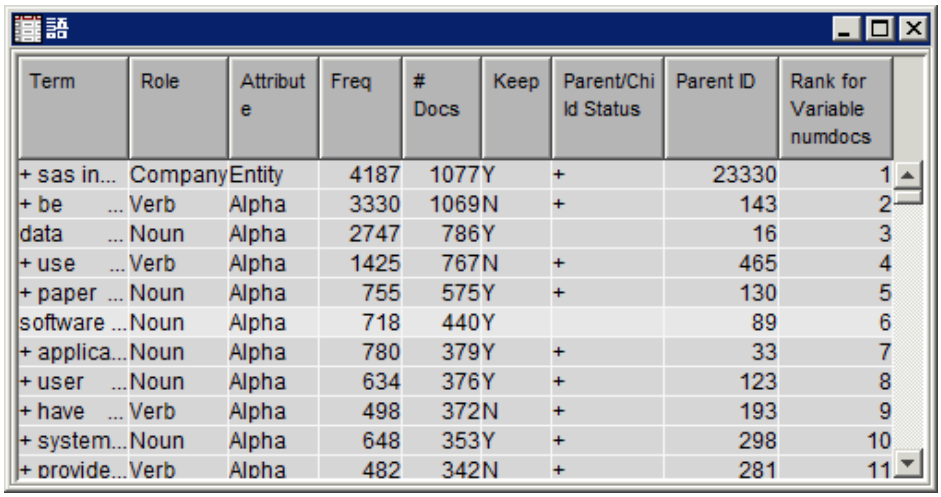
属性と頻度チャートには、**Alpha** がドキュメントコレクション内における属性の間で最高の頻度を持つことが示されます。



役割と頻度チャートには、**Noun** がドキュメントコレクション内における役割の間で最高の頻度を持つことが示されます。



8. 語テーブルに戻り、語“software”がテキスト解析分析内に保持されていることを確認します。これは、Keep 列の値が **Y** であることにより示されます。**テキスト解析**ノードをデフォルト設定で実行する場合、一部の語は保持されない場合があることに注意してください。


語

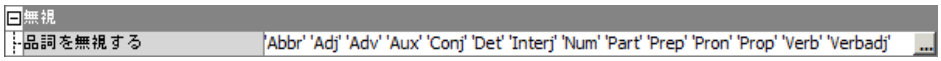
Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ sas in...	Company	Entity	4187	1077	Y	+	23330	1
+ be ...	Verb	Alpha	3330	1069	N	+	143	2
data ...	Noun	Alpha	2747	786	Y		16	3
+ use ...	Verb	Alpha	1425	767	N	+	465	4
+ paper ...	Noun	Alpha	755	575	Y	+	130	5
software ...	Noun	Alpha	718	440	Y		89	6
+ applica...	Noun	Alpha	780	379	Y	+	33	7
+ user ...	Noun	Alpha	634	376	Y	+	123	8
+ have ...	Verb	Alpha	498	372	N	+	193	9
+ system...	Noun	Alpha	648	353	Y	+	298	10
+ provide...	Verb	Alpha	482	342	N	+	281	11

テキスト解析ノードを使用すると、ドキュメントコレクション内の語に関する統計データを収集できるだけでなく、特定の品詞、エンティティの種類、属性に一致する語を破棄することにより、解析済みの語の出力セットを変更できます。語テーブル内の語リストを下スクロールし、**Noun** 以外の役割を持つ語の多くが保持されていることを確認します。ここで、テキスト解析結果を、役割が **Noun** である語に制限するとします。

9. 結果ウィンドウを閉じます。

10. **テキスト解析**ノードを選択した後、**品詞を無視する**プロパティの省略記号ボタンをクリックします。

11. 品詞を無視するダイアログボックスで、**Noun** を除くすべての品詞を選択します。これを行うには、Ctrl キーを押しながら各オプションをクリックします。**OK** をクリックします。**品詞を無視する**プロパティの値が、選択した値へと更新されたことを確認します。※ここで除外できるのは英語文法で使われる品詞のみで、例えば”Punctuation”（句読点）などは除外できません。また、テキスト解析ノードで品詞として認識できないもの（例. ”Unknown”）をあらかじめこの機能を使って除外することはできません。



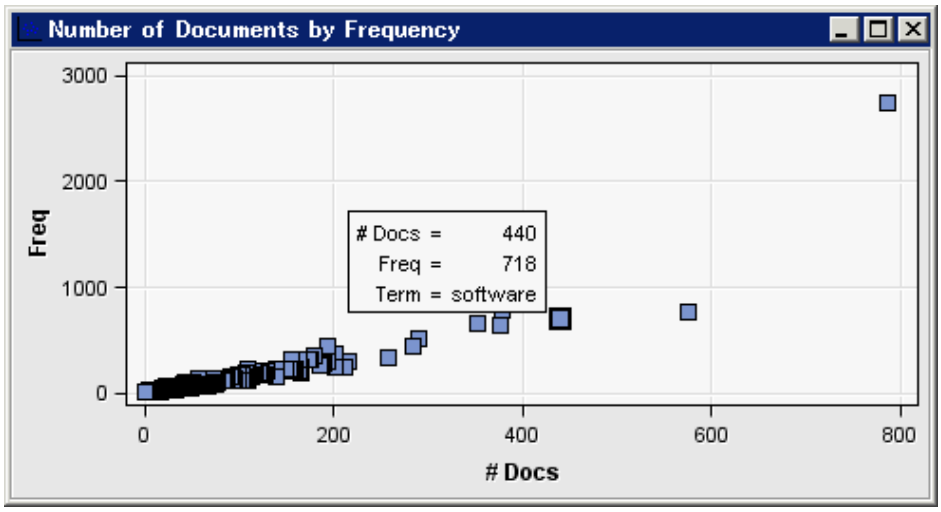
12. 続いて、名詞に加えて、名詞グループを保持するものとします。**名詞グループ**プロパティを**はい**に変更します。

13. **テキスト解析**ノードを右クリックし、**実行**を選択します。表示される確認ダイアログボックスで**はい**をクリックします。同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**を選択します。語“software”は、他の役割を含めた場合よりも、名詞または名詞グループという役割のみを含めた場合の方が、語の間でのランクがより高くなることが分かります。語テーブルを下スクロールすると、**名詞**または**名詞グループ**役割を持つ語が含まれていることを確認できます。

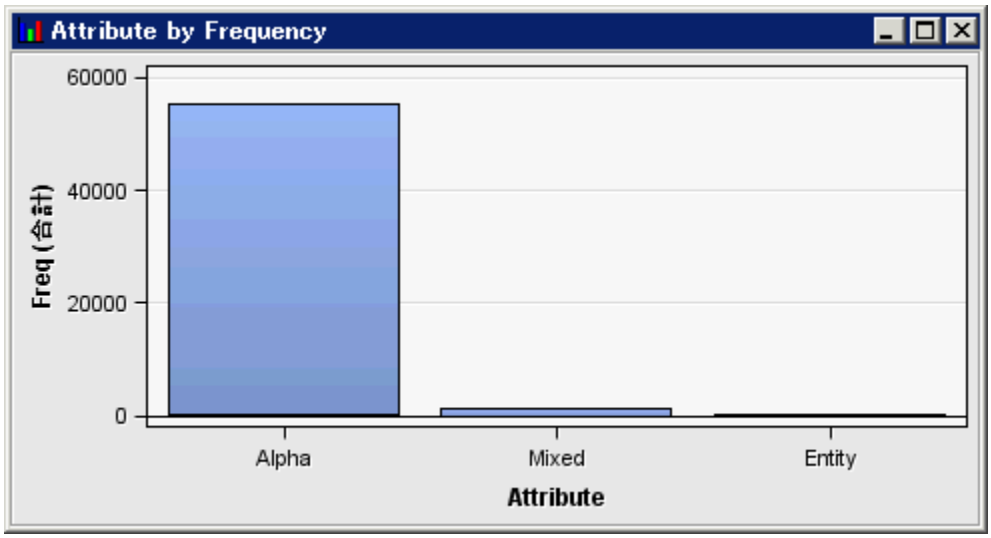
語

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
data	Noun	Alpha	2747	786	Y		14	1
+ paper	Noun	Alpha	755	575	Y	+	60	2
software	Noun	Alpha	718	440	Y		36	3
+ application	Noun	Alpha	780	379	Y	+	21	4
+ user	Noun	Alpha	634	376	Y	+	55	5

予想されるように、ドキュメント数と頻度プロット内にプロットされる語はより少なくなります。



同様に、属性と頻度チャートに示されているように、`Alpha` という属性を含む出力結果内の語の合計数も減少しています。※英語以外の場合は「複数の単語から成る語」に複数の単語から成る語を含むデータセットを指定することはできません。



10 章

テキストフィルタノード

テキストフィルタノードについて

テキストフィルタノードを使用すると、解析済みの語や分析対象となるドキュメントの総数を減らすことができます。これにより、無関係な情報を取り除き、最も価値の高い関連性のある情報だけを検討対象とすることができます。たとえば、**テキストフィルタ**ノードを使用することで、不要な語を削除し、特定の問題について記述しているドキュメントだけを保持することができます。このような縮小されたデータセットは、数十万のドキュメントや数十万の語を含んでいるオリジナルの集合を表すデータセットよりも桁違いにサイズが小さくなります。

テキストフィルタノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

この章の残りの部分では、**テキストフィルタ**ノードの使用例を紹介します。

テキストフィルタノードの使用

この例では、SAS Enterprise Miner が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトやダイアグラムの作成に関する詳細は、*Getting Started with SAS Enterprise Miner* を参照してください。

テキストフィルタノードを使用すると、テキストマイニング分析における語の総数を削減できます。たとえば、一般的な語や滅多に使われない語が分析にとって有益でない場合、それらの語をフィルタリングして取り除くことができます。この例では、**テキストフィルタ**ノードを使用して語をフィルタリングする方法を示します。この例では、ユーザーが“テキスト解析ノードの使用” (55 ページ)を実行済みであり、そこで作成されたプロセスフローダイアグラムを構築することを前提としています。

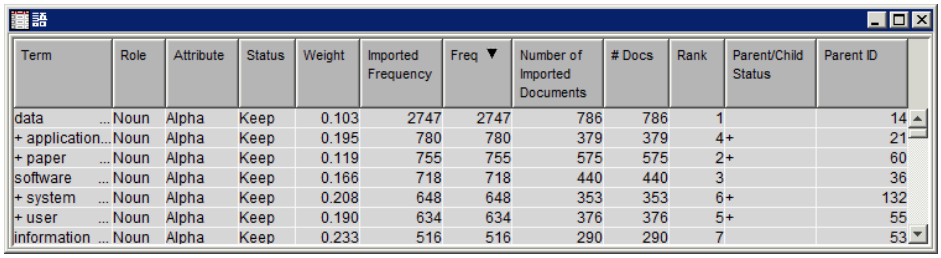
1. ツールバー上で**テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。

2. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。

3. ダイアグラムワークスペースで、**テキストフィルタ**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。

4. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

5. 語テーブルを選択します。[Freq]列見出しをクリックして、語を頻度順に並べ替えます。



語

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq ▼	Number of Imported Documents	# Docs	Rank	Parent/Child Status	Parent ID
data ...	Noun	Alpha	Keep	0.103	2747	2747	786	786	1		14
+ application...	Noun	Alpha	Keep	0.195	780	780	379	379	4+		21
+ paper ...	Noun	Alpha	Keep	0.119	755	755	575	575	2+		60
software ...	Noun	Alpha	Keep	0.166	718	718	440	440	3		36
+ system ...	Noun	Alpha	Keep	0.208	648	648	353	353	6+		132
+ user ...	Noun	Alpha	Keep	0.190	634	634	376	376	5+		55
information ...	Noun	Alpha	Keep	0.233	516	516	290	290	7		53

テキストマイニング分析を行うために、我々が分析対象とするドキュメント内では“software”および“application”という語が実際に類義語として使用されており、我々はこれらを同じ語として使用するものと仮定します。

6. 結果ウィンドウを閉じます。**テキストフィルタ**ノードを選択した後、**フィルタビューア**プロパティの省略記号ボタンをクリックします。

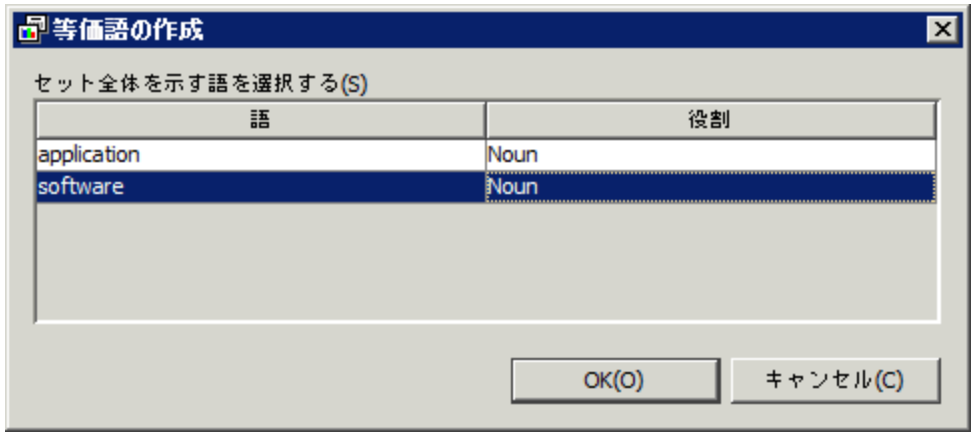
7. 対話型のフィルタビューア内で、語テーブル内の語を度数に基づいて並べ替えます。Ctrl キーを押しながら“software”と“application”を選択し、ドロップダウンメニューから**類義語として扱う**を選択します。

語

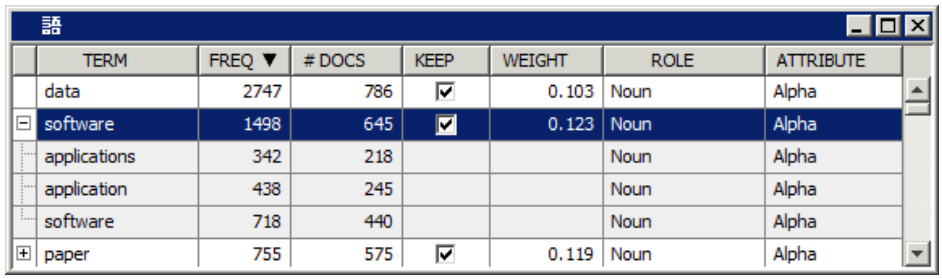
	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
	data	2747	786	☑	0.103	Noun	Alpha
⊞	application	780	379	☑	0.195	Noun	Alpha
⊞	paper	755	575	☑	0.119	Noun	Alpha
	software			☑	0.166	Noun	Alpha
⊞	system			☑	0.208	Noun	Alpha
⊞	user			☑	0.19	Noun	Alpha
	information			☑	0.233	Noun	Alpha
⊞	macro			☑	0.296	Noun	Alpha
	s			☐	0.0	Noun	Alpha
⊞	analysis			☑	0.284	Noun	Alpha
⊞	variable			☑	0.301	Noun	Alpha
⊞	use			☐	0.0	Noun	Alpha
⊞	report			☑	0.328	Noun	Alpha
⊞	program	316	176	☑	0.301	Noun	Alpha

検索式への語の追加(A)
類義語として処理(S)
類義語の削除(R)
語の保持(K)
語の削除(D)
コンセプトリンクの表示(V)
検索(N)...
次を検索(F)
選択のクリア(O)
印刷(P)...

8. 対応する語を作成ダイアログボックスで、語テーブル内にある両方の語を表す語として **software** を選択します。



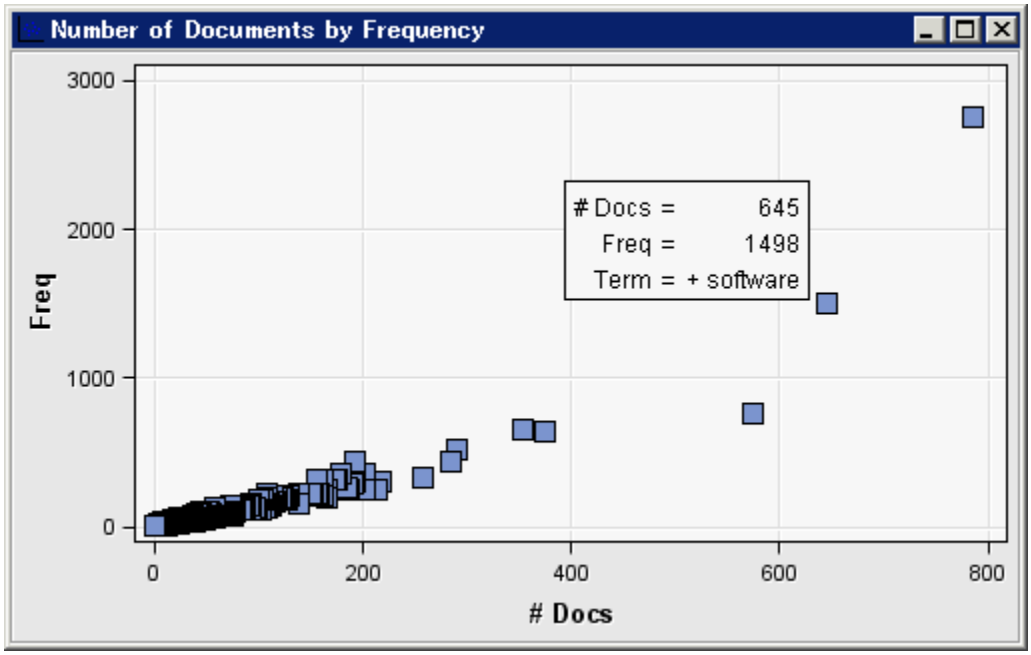
9. 対応する語を作成ダイアログボックス内で **OK** をクリックします。これで語“software”が、語テーブル内で両方の語を表すようになります。語“software”を展開します。



10. 対話型のフィルタビューアを閉じます。行った変更を保存するかどうかを尋ねるメッセージが表示されたら、**はい**を選択します。

11. **テキストフィルタ**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**を選択します。

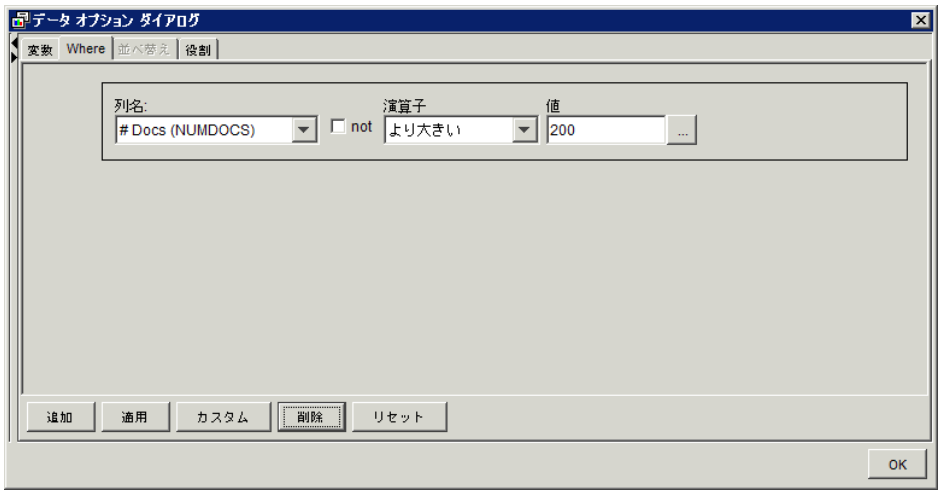
12. ドキュメント数と頻度プロットを選択し、両方の語が同じものとして扱われていることを確認しています。



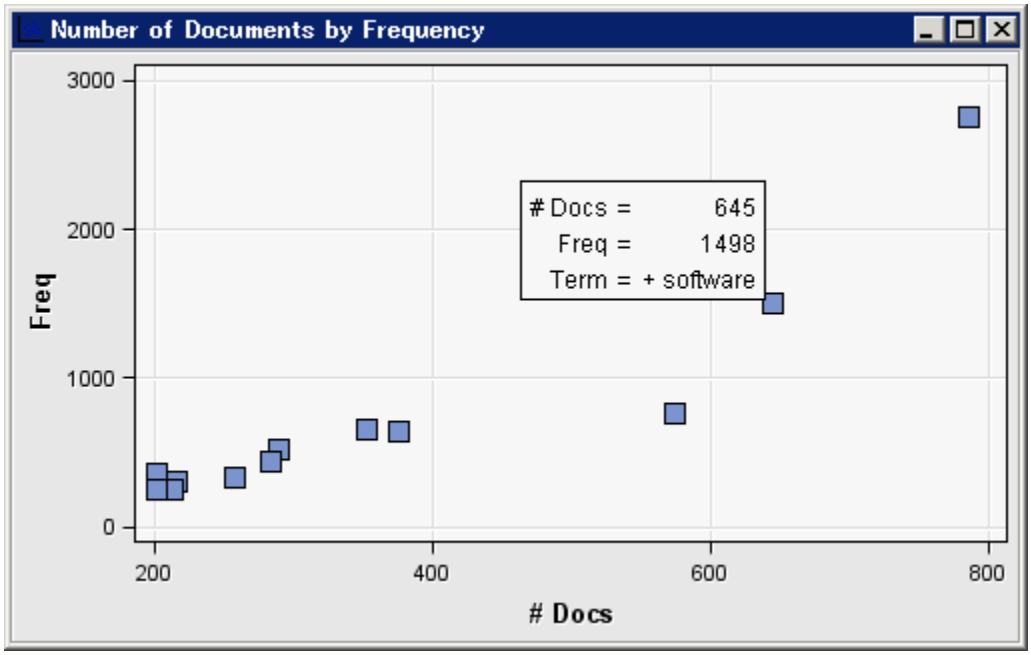
オプションを使用すると、表示を変更することや、プロットに表示する結果のサブセットを指定することもできます。たとえば、このプロットを改良し、200 個以上のドキュメントで出現した語のみを表示したいとします。

13. ドキュメント数と頻度プロットを右クリックし、**データオプション**を選択します。

14. データオプションダイアログボックスで、**Where** タブを選択します。**列名**ドロップダウンメニューから、**# Docs** を選択します。**演算子**ドロップダウンメニューから、**より大**を選択します。**値**テキストボックスに、*200* と入力します。

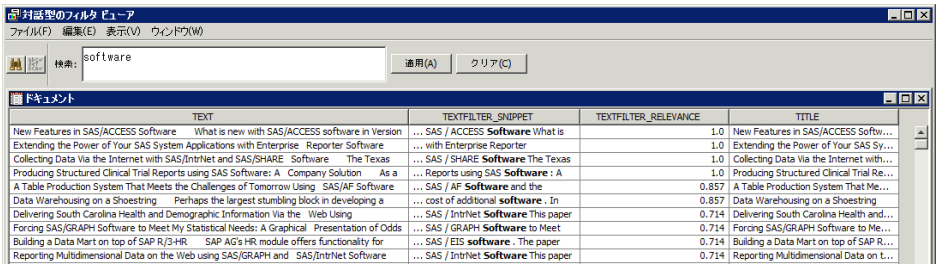


15. **適用**を選択して **OK** をクリックします。ドキュメント数と頻度プロットのサイズが変更され、200 個を超えるドキュメントで出現した語のみが同プロットに含められるようになります。



16. 結果ウィンドウを閉じます。プロットのサイズ変更やサブセット化により分析を絞り込むことに加えて、対話型のフィルタビューアを使用して語を直接検索することもできます。

17. **テキストフィルタ**ノードを選択した後、**フィルタビューア**プロパティの省略記号ボタンをクリックします。対話型のフィルタビューアで、**検索**テキストボックス内に *software* と入力し、**適用**をクリックします。



[ドキュメント]テーブルには、検索対称の語を含んでいるテキストの抜粋が表示されます。[ドキュメント]テーブル内の情報を使用すると、ドキュメントの全文およびドキュメントのタイトルに加えて、抜粋結果を調べることにより、使用されている語のコンテキストを理解できるようになります。対話型のフィルタビューアに関する詳細は、SAS Text Miner のヘルプに含まれている対話型のフィルタビューアのトピックを参照してください。

対話型のフィルタビューアで語を検索する場合、興味深い問題が発生します。先述したように、“software”の検索では大文字小文字が区別されません。ただし、見つけたい語のインスタンスが存在したが、ドキュメントコレクション内でその語のスペルが間違っていたとしたらどうなるでしょうか？語をフィルタリングする場合、辞書データセットを使用してスペルチェックを行うこともできます。

18. 対話型のフィルタビューアを閉じ、変更を保存するかどうかを尋ねられたら**いいえ**を選択します。

19. (オプション)**テキストフィルタ**ノードを選択し、**スペルチェックを行う**プロパティを**はい**に設定します。**テキストフィルタ**ノードに戻ると、語がスペルチェックされ、スペルミスが検出されるようになります。スペルチェックで使用するデータセットを指定する

には、**辞書**プロパティの隣にある省略記号ボタンをクリックし、データセットを選択します。辞書データセットの作成に関する詳細は、SAS Text Miner のヘルプに含まれている[辞書データセットの作成]というトピックを参照してください。

テキストフィルタノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択します。同ノードの実行が完了したら、実行状態ダイアログボックス内で **OK** を選択します。**スペルチェックの結果**プロパティの隣にある省略記号ボタンをクリックすると、表示されたウィンドウで、スペルチェック時に生成されたスペルの修正を含んでいるデータセットを確認できます。たとえば、語"softwae"は、語"software"のスペルミスでとして識別されます。

EMWS2.TextFilter_spellDS

	Parent # Docs	Term	# Docs	Parent	Role	Parent Role	Min Distance	Dictionary	Key	Parent ID
1	245.0	applicaion	1.0	application	Noun	Noun	5.0		7823.0	21.0
2	119.0	procedural	2.0	procedure	Noun	Noun	14.0		7850.0	33.0
3	440.0	solftware	1.0	software	Noun	Noun	6.0		879.0	36.0
4	440.0	softwae	2.0	software	Noun	Noun	7.0		5694.0	36.0
5	33.0	entr	1.0	entry	Noun	Noun	8.0		6071.0	58.0
6	5.0	dependence	1.0	dependent	Noun	Noun	14.0		11860.0	80.0
7	83.0	suport	1.0	support	Noun	Noun	4.0		5358.0	82.0
8	22.0	succe	1.0	success	Noun	Noun	14.0		4490.0	87.0
9	76.0	agility	1.0	ability	Noun	Noun	14.0		4974.0	88.0
10	13.0	enduser	1.0	end-user	Noun	Noun	7.0		10790.0	89.0

この関係は、[対話型のフィルタビューア]の[語]テーブルで確認できます。**フィルタビューア**プロパティの隣にある省略記号ボタンをクリックします。[語]テーブル内にある語"software"を展開し、その類義語を確認します。この類義語には、スペルチェック時にミススペルとして識別された語である"softwae"が含まれています。

語

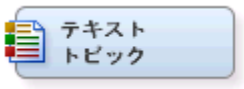
	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
	data	2747	786	☑	0.103	Noun	Alpha
⊟	software	1502	646	☑	0.123	Noun	Alpha
	solftware	1	1			Noun	Alpha
	application	438	245			Noun	Alpha
	applications	342	218			Noun	Alpha
	applicaion	1	1			Noun	Alpha
	software	718	440			Noun	Alpha
	softwae	2	2			Noun	Alpha
⊞	paper	755	575	☑	0.119	Noun	Alpha
⊞	system	648	353	☑	0.208	Noun	Alpha

この類義語には、"applicaion"（この例のステップ 7～10 で作成されたもの）が含まれているほか、"applicaion"（スペルチェック時に"application"のスペルミスとして識別された語）が含まれています。

11 章

テキストトピックノード

テキストトピックノードについて



テキストトピックノードを使用すると、検出されたトピックやユーザー定義のトピックの両方に従って語とドキュメントを自動的に関連付けることにより、ドキュメントコレクションを調査できます。トピックとは、主要なテーマやアイデアを記述し特徴付ける語のコレクションです。このアプローチはクラスタリングとは異なります。なぜなら、クラスタリングは各ドキュメントを一意のグループに割り当てますが、**テキストトピック**ノードは各ドキュメントおよび語のスコアを各トピックに割り当てるためです。ドキュメントや語が特定トピックに属していると見なすための関連付けが十分強い場合は、しきい値が使用されます。結果として、ドキュメントと語は、1 つ以上のトピックに属すか、あるいはいかなるトピックにもまったく属さないことになります。ユーザーが要求するトピックの数は、ドキュメントコレクションのサイズに対して直接的な関連があります(たとえば、大規模なコレクションでは数も大きくなります)。

最もメモリを多用するタスクは、語/ドキュメントの頻度マトリックスの特異値分解(SVD)の計算です。詳細については、SAS Text Miner のヘルプの特異値分解(SVD)に関するトピックを参照してください。インメモリリソースが制限されている場合、**テキストトピック**ノードは、完全なコレクションの代わりにドキュメントの単純なランダム標本を使用することで、同ノードを正常に実行しようと試みます。サンプリングは、サンプリングなしに SDV の計算を試みた際にノードがメモリ障害に遭遇した場合に発生します。さらに、サンプリングは通常ドキュメントコレクションが非常に大きい場合に発生するため、通常はモデリング結果に関して有害な影響はありません。サンプリングが正確にいつ発生するかは、お使いのコレクションの数、お使いのシステムが実行されているプラットフォーム、利用可能な RAM などを含む多くのパラメータに依存します。

テキストトピックノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

注: **テキストトピック**ノードは、グループ処理(開始グループノードや停止グループノード)における利用ではサポートされません。

テキストトピックノードの使用

この例では、SAS Enterprise Miner が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトやダイアグラムの作成に関する詳細は、*Getting Started with SAS Enterprise Miner* を参照してください。

テキストトピックノードを使用すると、語のリストから興味のあるトピックを作成できます。トピックのリストを作成する目的は、分析で興味のある語の組み合わせを確立することにあります。たとえば、「会社の社長(company president)」のアクティビティについて議論している記事のマイニングの興味があるとします。このタスクにアプローチする 1 つの方法は、語"company"を含んでいるすべての記事、および語"president"を含んでいるすべて記事に注目することです。**テキストトピック**ノードを使用すると、"company"および"president"という語を"company president"というトピックへと結合できます。

個々の語をトピックへと結合することにより、テキストマイニング分析を改善できます。結合を通じて、分析対象となるテキストの量を、自分が興味のある語のグループ数にまで削減できます。この例では、**テキストトピック**ノードを使用してトピックを作成する方法を示します。

1. SAS データセット SAMPSIO.ABSTRACT には、さまざまな会議から収集したタイトルと概要のテキストが含まれています。ABSTRACT データソースを作成し、それをダイアグラムワークスペースに追加します。TEXT 変数および TITLE 変数のルール値を**テキスト(Text)**に設定します。

2. ツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。

3. ABSTRACT データソースを**テキスト解析**ノードに接続します。

4. **テキスト解析**ノードを選択した後、**品詞を無視する**プロパティの省略記号ボタンをクリックします。

5. 品詞を無視するダイアログボックスで、**Noun** を除くすべての品詞を選択します。これを行うには、Ctrl キーを押しながら各オプションをクリックします。**OK** をクリックします。

6. **名詞グループ**プロパティを**はい**に変更します。

7. ツールバー上で**テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。

8. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。

9. ツールバー上で**テキストマイニング**タブを選択し、**テキストトピック**ノードをダイアグラムワークスペースへとドラッグします。

10. **テキストフィルタ**ノードを**テキストトピック**ノードに接続します。

    この時点で、プロセスフローダイアグラムは次のようになります。

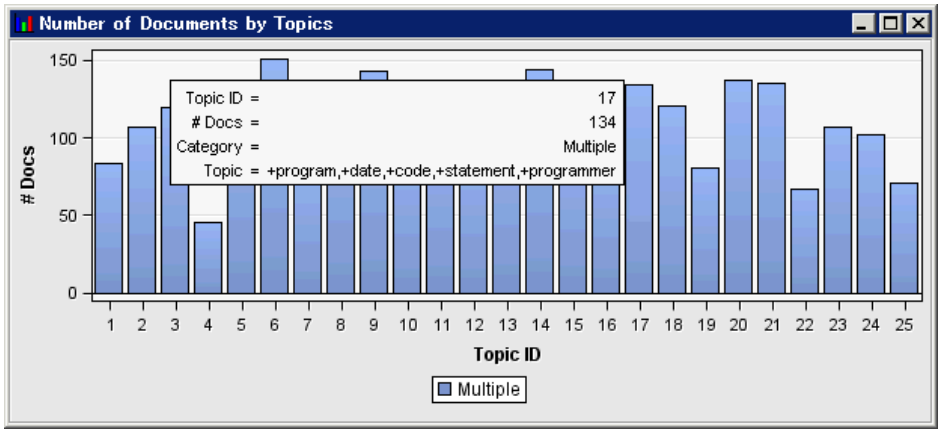
11. ダイアグラムワークスペースで、**テキストトピック**ノードを右クリックし、**実行**を選択します。表示される確認ダイアログボックスで**はい**をクリックします。同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

12. トピックテーブルを選択して、**テキストトピック**ノードのデフォルト実行により作成されたトピックを表示します。

Topics

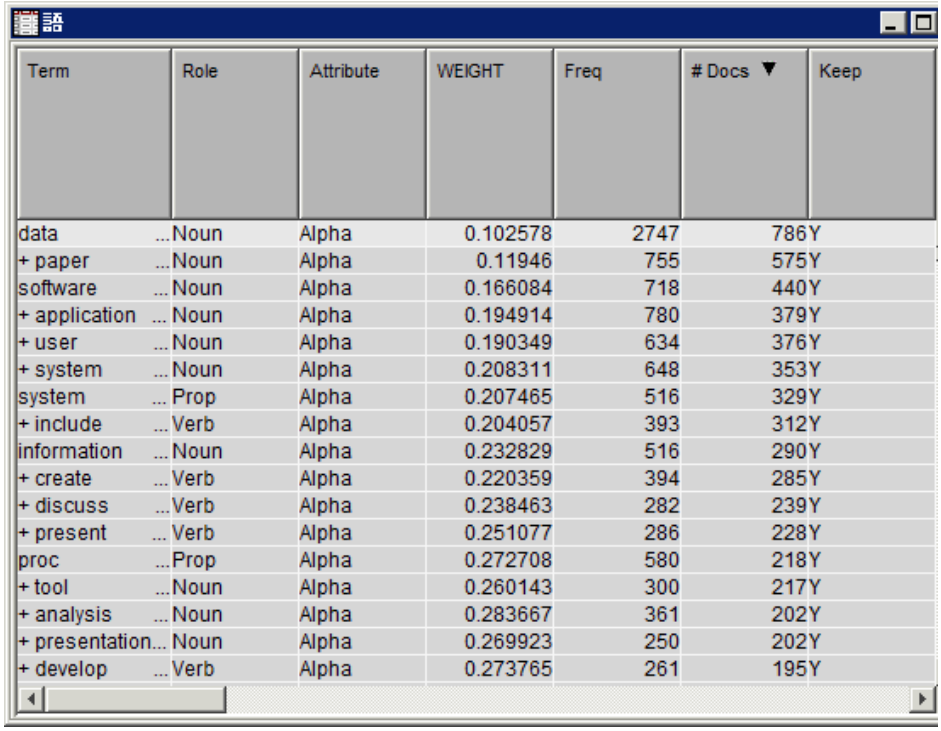
Category	Topic ID ▲	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.100	0.024	ods,+output,output,...	137	83
Multiple	2	0.085	0.026	+customer,+busin...	232	107
Multiple	3	0.094	0.026	+test,+sample,stati...	237	119
Multiple	4	0.107	0.024	sql,+select statem...	105	45
Multiple	5	0.097	0.025	+performance,+ser...	170	136
Multiple	6	0.105	0.025	+data set,+set,+set...	209	151
Multiple	7	0.105	0.024	+macro,+macro var...	94	92
Multiple	8	0.102	0.025	af,scl,+application,...	204	129
Multiple	9	0.104	0.025	web,html,intrnet,we...	154	143
Multiple	10	0.075	0.026	+treatment,clinical,...	201	105
Multiple	11	0.098	0.025	+data warehouse,+...	154	117
Multiple	12	0.089	0.025	+graph,+graph,gra...	179	105
Multiple	13	0.106	0.025	proc,+report,+tabul...	170	122
Multiple	14	0.076	0.026	+year,version,syste...	289	144
Multiple	15	0.090	0.025	java,appdev,+client...	200	113
Multiple	16	0.090	0.025	sql,dbms,access,+...	193	101
Multiple	17	0.074	0.026	+program,+date,+c...	267	134
Multiple	18	0.077	0.026	windows,excel,mic...	265	120
Multiple	19	0.080	0.025	mddb,olap,eis,md...	185	80
Multiple	20	0.090	0.026	+analysis,+progra...	234	137
Multiple	21	0.077	0.026	information,+decisi...	267	135
Multiple	22	0.080	0.025	+entry,+catalog ent...	172	67
Multiple	23	0.082	0.026	+model,+model,re...	239	107
Multiple	24	0.089	0.026	enterprise,+enterpr...	208	102
Multiple	25	0.070	0.026	tabulate,+worksho...	229	71

13. ドキュメント数とトピックチャートを選択し、それが含んでいるドキュメント数別にトピックを確認します。



注: トピック ID 値を確認するためには、場合によってはデフォルトのグラフをリサイズする必要があります。

14. **語**テーブルを選択します。同テーブル内の最初のエントリを選択します。

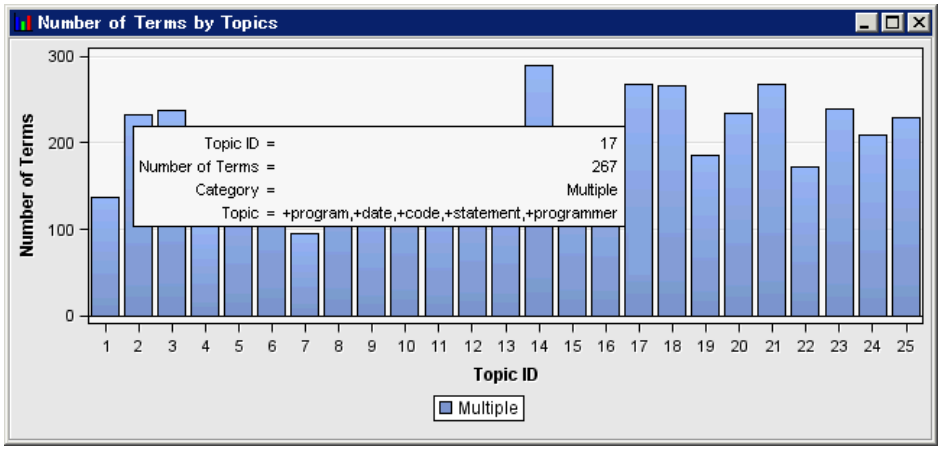


語

Term	Role	Attribute	WEIGHT	Freq	# Docs ▼	Keep
data	Noun	Alpha	0.102578	2747	786	Y
+ paper	Noun	Alpha	0.11946	755	575	Y
software	Noun	Alpha	0.166084	718	440	Y
+ application	Noun	Alpha	0.194914	780	379	Y
+ user	Noun	Alpha	0.190349	634	376	Y
+ system	Noun	Alpha	0.208311	648	353	Y
system	Prop	Alpha	0.207465	516	329	Y
+ include	Verb	Alpha	0.204057	393	312	Y
information	Noun	Alpha	0.232829	516	290	Y
+ create	Verb	Alpha	0.220359	394	285	Y
+ discuss	Verb	Alpha	0.238463	282	239	Y
+ present	Verb	Alpha	0.251077	286	228	Y
proc	Prop	Alpha	0.272708	580	218	Y
+ tool	Noun	Alpha	0.260143	300	217	Y
+ analysis	Noun	Alpha	0.283667	361	202	Y
+ presentation	Noun	Alpha	0.269923	250	202	Y
+ develop	Verb	Alpha	0.273765	261	195	Y

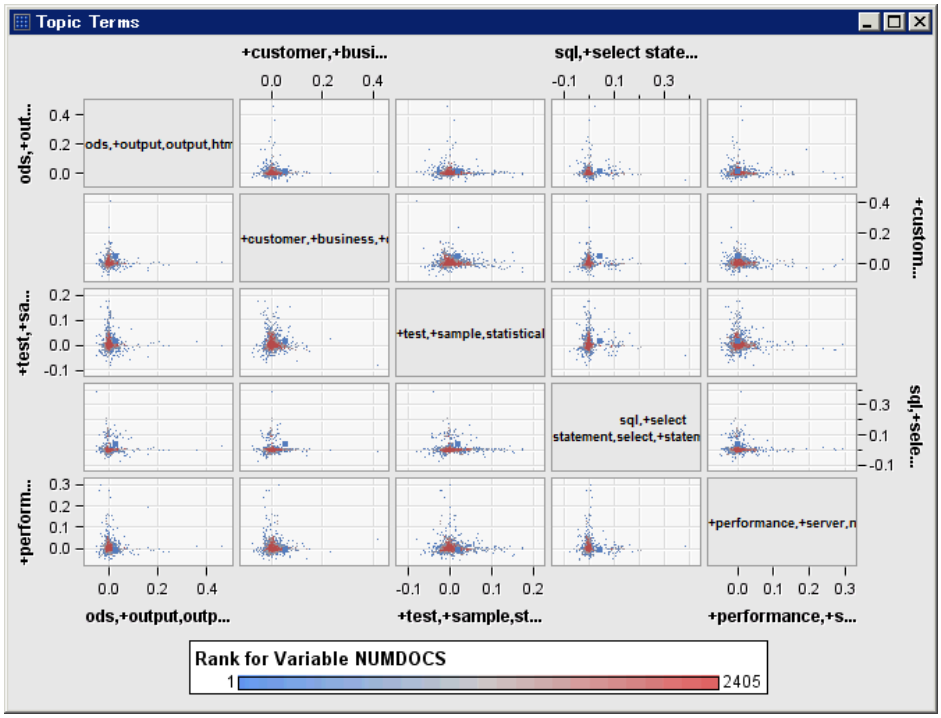
語テーブルは、各トピックに関する語とその重みを表示します。すべての保持されている語は、**名詞**または**名詞グループ**の役割を持つことに注意してください。

15. **語数とトピック**棒グラフを選択します。



マウスポインタをバーの上に置くと、ツールチップに、トピック ID、このトピックに含まれている語の数、カテゴリ、およびトピックが表示されます。

16. **トピック語**マトリックスグラフを選択します。



トピック語マトリックスグラフは、複数の語にまたがるトピック値を表示します。

注: 点をより明確に確認するためには、このマトリックスを拡大する必要があります。

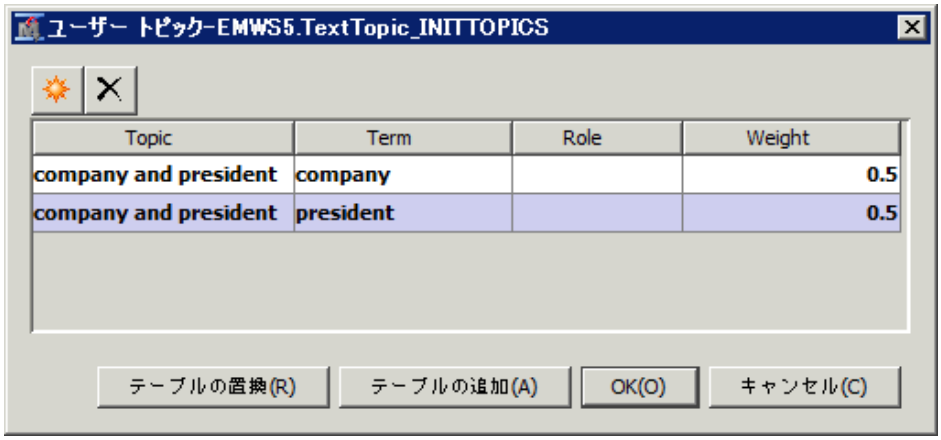
複数語トピックに加えて、**テキストトピック**ノードを使用して、単一語トピックや独自のトピックを作成できます。

17. **結果**ウィンドウを閉じ、テキストトピックノードを選択します。

18. **単一語トピックの数**プロパティを選択し、*10* を入力した後、キーボード上の **Enter** キーを押します。

19. **ユーザートピック**プロパティの隣にある省略記号ボタンをクリックします。

20. ユーザートピックダイアログボックスで、  をクリックして行を追加します。語 *company* を入力し、その後に重み *0.5* を与え、トピック *company and president* を指定します。  を再度クリックし、2 番目の行を追加します。語 *president* を入力し、その後に重み *0.5* を与え、トピック *company and president* を指定します。



21. **OK** をクリックします。

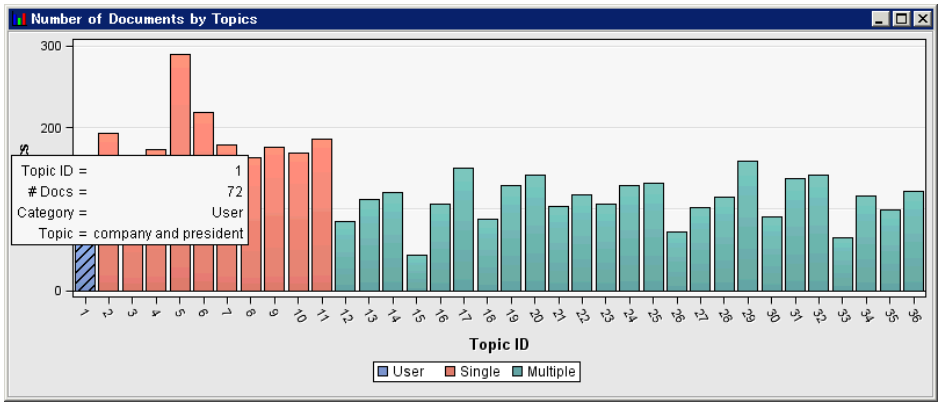
22. **テキストトピック**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**を選択した後、ノードが実行を完了した時点で、実行ステータスダイアログボックス内の**結果**を選択します。

23. トピックテーブルを選択します。10 個の新しい単一語トピックが、ユーザートピックダイアログボックスで指定したトピックと共に作成されていることに注意してください。

Topics

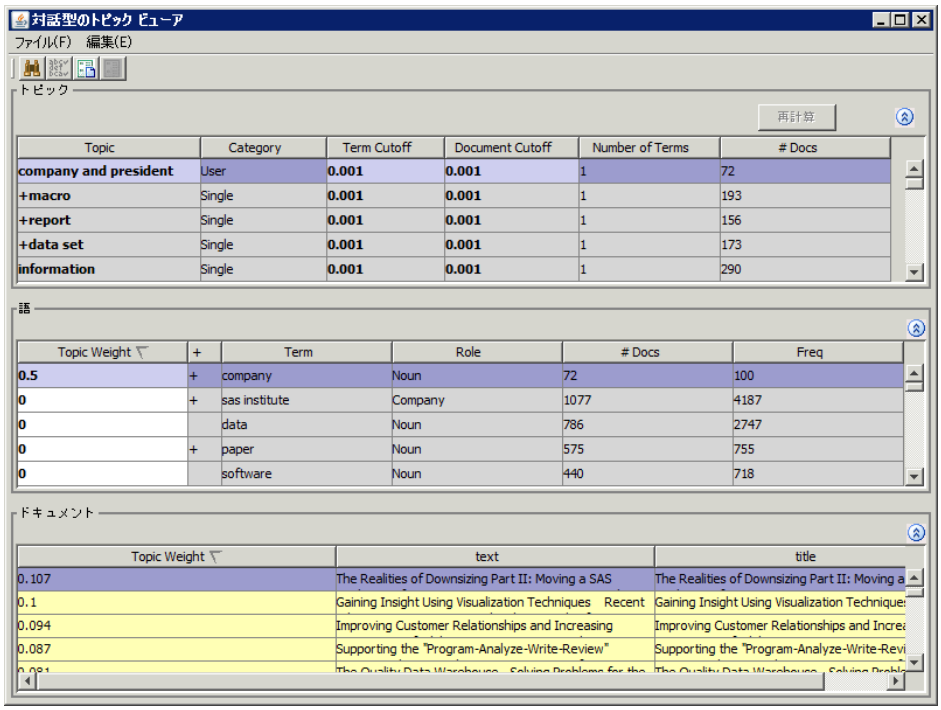
Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
User	1	0.001	0.001	company a...	1	72
Single	2	0.001	0.001	+macro	1	193
Single	3	0.001	0.001	+report	1	156
Single	4	0.001	0.001	+data set	1	173
Single	5	0.001	0.001	information	1	290
Single	6	0.001	0.001	proc	1	218
Single	7	0.001	0.001	+variable	1	179
Single	8	0.001	0.001	web	1	163
Single	9	0.001	0.001	+program	1	176
Single	10	0.001	0.001	+set	1	168
Single	11	0.001	0.001	+technique	1	186
Multiple	12	0.101	0.024	ods,+output...	138	84
Multiple	13	0.073	0.026	+date,+pro...	267	112
Multiple	14	0.090	0.026	+test,+sam...	235	120
Multiple	15	0.108	0.024	sql,+select ...	103	43
Multiple	16	0.087	0.026	+performan...	199	106
Multiple	17	0.107	0.025	+data set,+...	204	150
Multiple	18	0.105	0.024	+macro,+m...	97	87
Multiple	19	0.100	0.025	af,+applicat...	198	129
Multiple	20	0.103	0.025	web,html,in...	155	141
Multiple	21	0.076	0.026	clinical,+tre...	202	103
Multiple	22	0.096	0.025	+data ware...	150	117
Multiple	23	0.088	0.025	+graph,+gr...	175	106
Multiple	24	0.100	0.025	+report,pro...	187	129
Multiple	25	0.076	0.026	+year,versi...	284	132
Multiple	26	0.075	0.026	tabulate,+ta...	170	72
Multiple	27	0.091	0.025	sql,access,...	183	102
Multiple	28	0.069	0.026	system,odb...	305	115
Multiple	29	0.097	0.026	windows,nt,...	226	158
Multiple	30	0.075	0.026	mddb,olap,...	186	90
Multiple	31	0.088	0.026	+analysis,+...	232	137
Multiple	32	0.078	0.026	information,...	274	141
Multiple	33	0.079	0.025	+entry,+cat...	165	65
Multiple	34	0.084	0.026	+model,+m...	221	116
Multiple	35	0.090	0.026	enterprise,...	197	99
Multiple	36	0.089	0.025	java,+client,...	204	122

24. ドキュメント数とトピックウィンドウを選択し、複数語、単一語、およびユーザー作成トピックを、それらが含んでいるドキュメント数別に表示します。



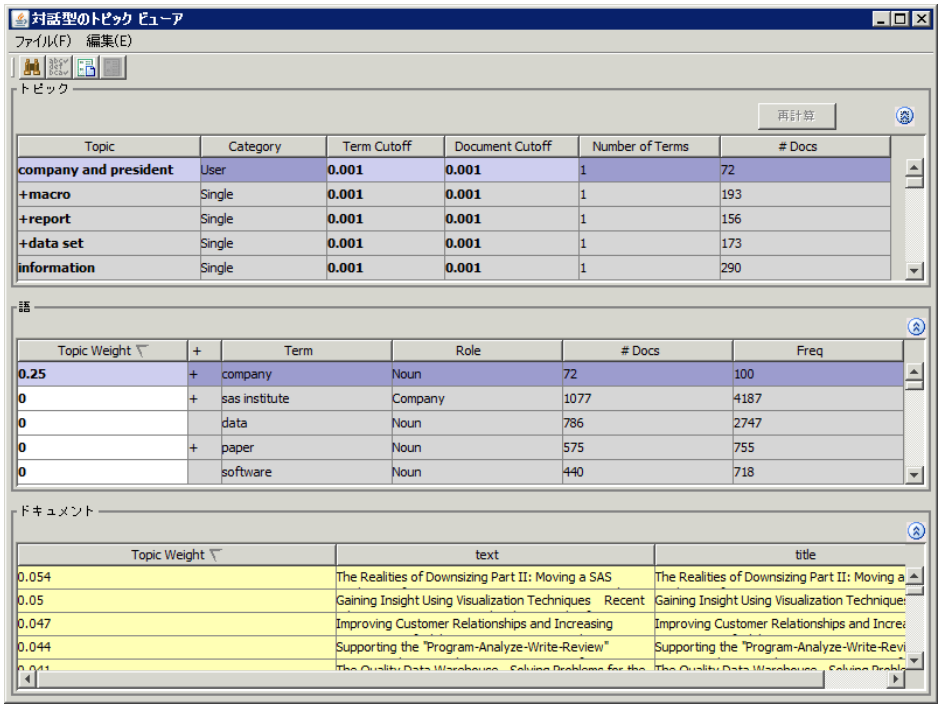
対話型のトピックビューアを使用すると、トピックのプロパティの表示や変更が行えます。

25. **結果**ウィンドウを閉じ、テキストトピックノードを選択します。**トピックビューア**プロパティの隣にある省略記号ボタンをクリックします。対話型のトピックビューアウィンドウが表示されたら、**トピック**ペインの**トピック**列に基づいて並べ替えを行います。



対話型のトピックビューア内で、トピック名、語およびドキュメントのカットオフ値、トピックの重みを変更できます。

26. トピックテーブル内でトピック値“company and president”を選択し、同トピックの名前を *company* に変更します。語テーブル内の語“company”のトピック重みを選択し、それを *0.25* に変更します。**再計算**をクリックします。



27. 対話型のトピックビューアを閉じ、変更を保存するかどうかを尋ねられたら**いいえ**を選択します。対話型のトピックビューアに関する詳細は、SAS Text Miner のヘルプに含まれている対話型のトピックビューアのトピックを参照してください。

12 章

テキストクラスタノード

テキストクラスタノードについて

テキストクラスタノードは、ドキュメントをクラスタリングすることで、特定の記述語に関するドキュメントやレポートの互いに疎な集合を作成します。次の 2 つのアルゴリズムが利用できます。期待最大アルゴリズムは、フラット表示を使用してドキュメントをクラスタリングします。一方、階層クラスタリングアルゴリズムは、クラスタをツリー階層へとグループ化します。両アプローチとも特異値分解(SVD)を使用して、元の重み付きの語/ドキュメントのマトリックスを、高密度ではあるが低次元の表現へと変換します。

テキストクラスタ処理のうちで最もメモリを多用するタスクは、重み付きのドキュメント別語の頻度マトリックスの SVD 計算です。インメモリリソースが制限されている場合、当該ノードは完全なコレクションの代わりに、ドキュメントの単純なランダム標本を使用することで、同ノードを正常に実行しようと試みます。サンプリングは、サンプリングなしに SDV の計算を試みた際に、ノードにメモリ障害が発生した場合に発生します。さらに、サンプリングは通常ドキュメントコレクションが非常に大きい場合に発生するため、通常はモデリング結果に関して有害な影響はありません。サンプリングが正確にいつ発生するかは、お使いのコレクションの数、お使いのシステムが実行されているプラットフォーム、利用可能な RAM など多くのパラメータに依存します。

テキストクラスタノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

この章の残りの部分では、**テキストクラスタ**ノードの使用例を紹介します。

テキストクラスタノードの使用

この例では、**テキストクラスタ**ノードを使用して、SAS Users Group International (SUGI) の概要をクラスタリングします。この例では、SAS Enterprise Miner が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

注: SAS Users Group International は、現在では SAS Global Forum となっています。

次の手順を実行します。

1. SAMPSIO.ABSTRACT 用のデータソースを作成します。変数 TITLE の役割を **ID** に変更します。

   注: SAMPSIO.ABSTRACT データセットには、1998～2001 年までの SUGI ミーティング(SUGI 23～26)のために用意された 1,238 件の論文に関する情報が含まれています。変数 TITLE の値は、SUGI 論文のタイトルになります。変数 TEXT は、SUGI 論文の概要を含んでいます。

2. SAMPSIO.ABSTRACT データソースをダイアグラムワークスペースに追加します。

3. ツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。

4. データ入力ノードを**テキスト解析**ノードに接続します。

5. **テキスト解析**ノードを選択した後、**停止リスト**プロパティの省略記号ボタンをクリックします。

6. **テーブルの交換**ボタンをクリックし、SAMPSIO.SUGISTOP を停止リストとして選択した後、**OK** をクリックします。確認ダイアログボックスで**はい**を選択します。**OK** をクリックして、**停止リスト**プロパティのダイアログボックスを終了します。

7. **エンティティの検索**プロパティを**標準**に変更します。

8. **エンティティの種類を無視する**プロパティの省略記号ボタンをクリックし、エンティティの種類を無視するダイアログボックスを開きます。

9. すべてのエンティティの種類を選択します。ただし、次のものは除きます。**Location**、**Organization**、**Person**、**Product**。**OK** をクリックします。

10. **テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。

11. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。

12. **テキストマイニング**タブを選択し、**テキストクラスタ**ノードをダイアグラムワークスペースへとドラッグします。

13. **テキストフィルタ**ノードを**テキストクラスタ**ノードに接続します。この時点で、プロセスフローダイアグラムは次のようになります。

14. **テキストクラスタ**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**をクリックします。

15. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

16. クラスタテーブルを選択します。

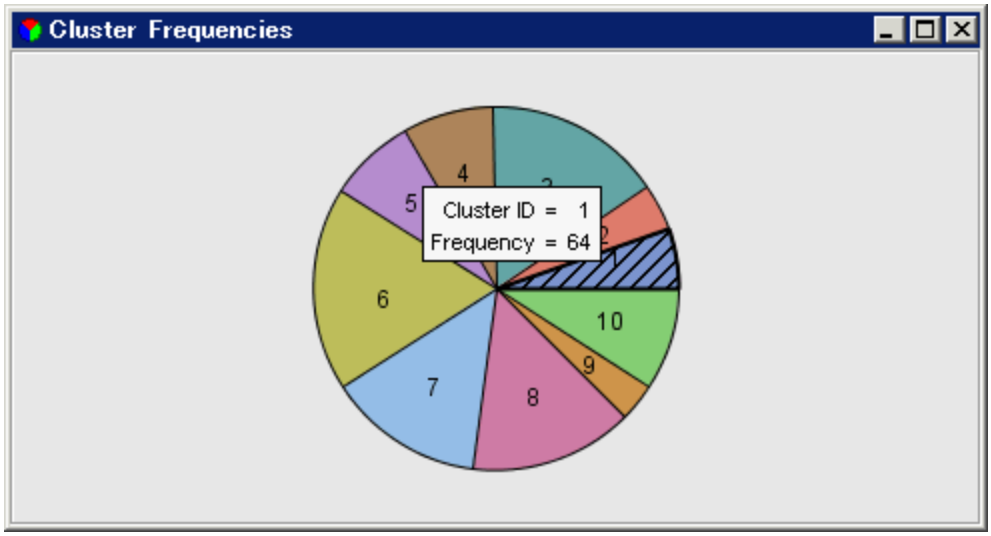
    クラスタテーブルには、各クラスタの ID、クラスタを構成している記述語、各クラスタの統計値が含まれています。

クラスター

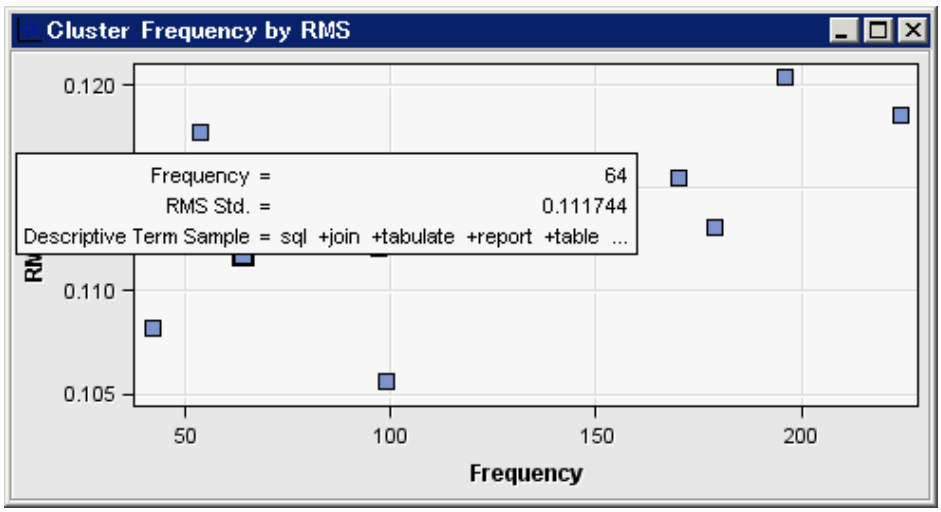
Cluster ID	Descriptive Terms	Frequency	Percentage
1	sql +join +tabulate +report +table +works...	64	5%
2	institute 'sas institute' +conference future ...	54	4%
3	+analysis +model statistical +test +study ...	196	16%
4	af +object +entry +developer +frame +scr...	97	8%
5	+program +macro macro +'macro variabl...	99	8%
6	+set +'data set' +file +format +output out...	224	18%
7	+warehouse +'data warehouse' +busines...	170	14%
8	web +graph +page intrnet graphics 'grap...	179	14%
9	+customer +market financial +business ...	42	3%
10	+server windows nt server +performance ...	113	9%

17. クラスタテーブル内の最初のクラスタを選択します。

18. クラスタ頻度ウィンドウを選択し、クラスタを頻度別に表した円グラフを確認します。マウスポインタをセクション上に置くと、そのクラスタの頻度がツールチップ内に表示されます。

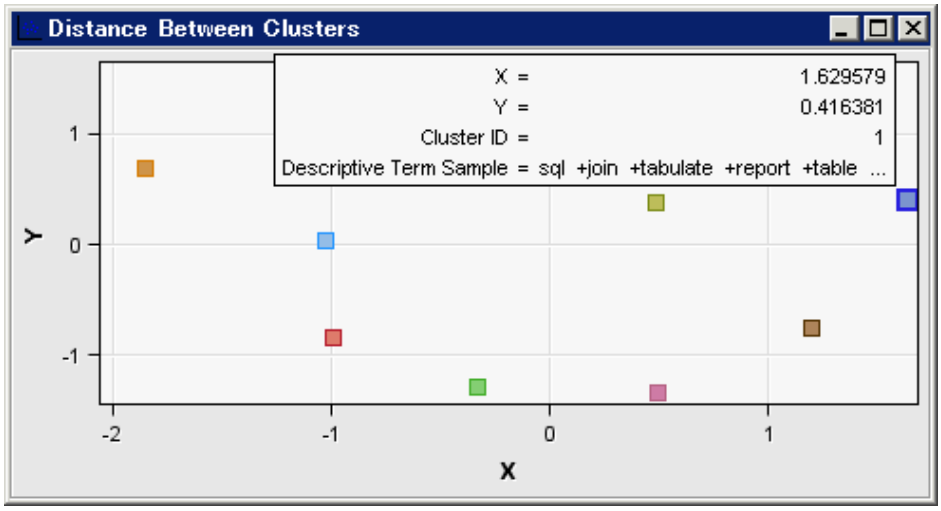


19. クラスタ頻度と RMS ウィンドウを選択した後、強調表示されているクラスタ上にマウスポインタを置きます。



最初のクラスタとそれ以外のクラスタは距離に関してどう異なるでしょうか？

20. クラスタ間の距離ウィンドウを選択した後、強調表示されているクラスタ上にマウスポインタを置き、XY 座標グリッド内で最初のクラスタの位置を確認します。



距離を比較したいその他のクラスタ上にマウスポインタを置きます。

21. 結果ウィンドウを閉じます。

    期待値最大化クラスタリングアルゴリズムにより取得したクラスタリングの結果を、階層クラスタリングアルゴリズムを使用した場合の結果と比較します。

22. **テキストクラスタ**ノードを選択します。

23. **指定値または最大数**プロパティで**指定値**を選択します。

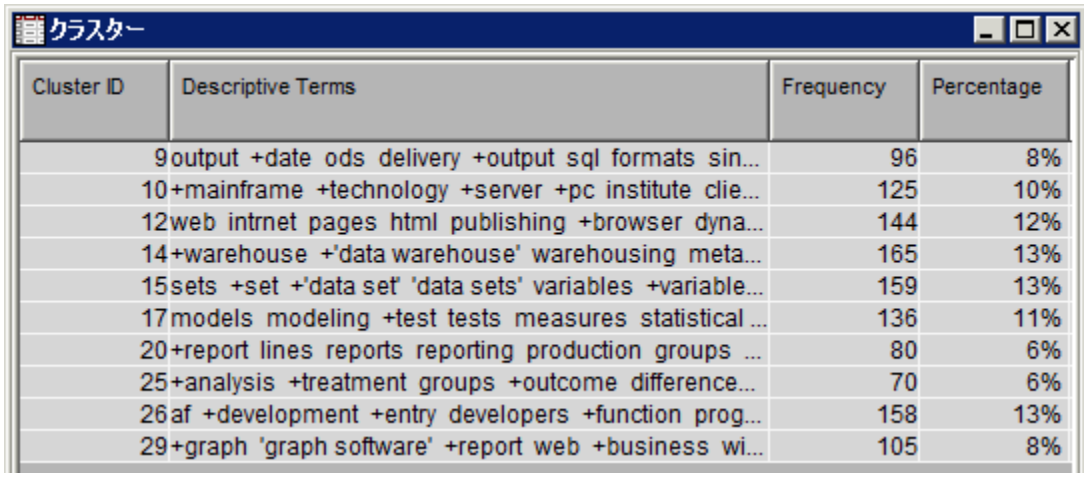
24. **クラスタ数**プロパティで *10* を指定します。

25. **クラスタアルゴリズム**プロパティで**階層**を選択します。

26. **テキストクラスタ**ノードを右クリックし、**実行**を選択します。確認ダイアログボックスで**はい**をクリックします。

27. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

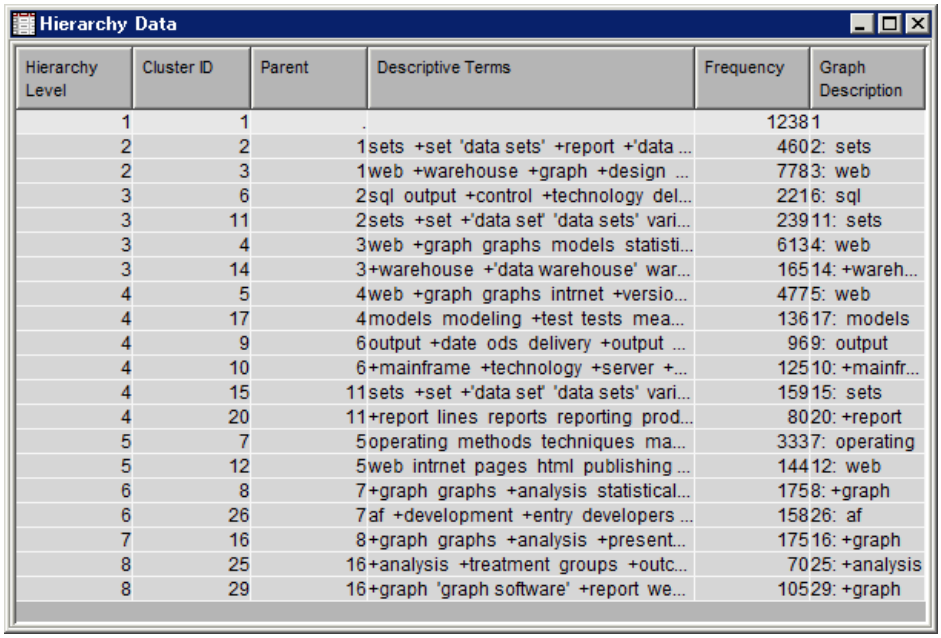
28. クラスタテーブルを選択します。



Cluster ID	Descriptive Terms	Frequency	Percentage
9	output +date ods delivery +output sql formats sin...	96	8%
10	+mainframe +technology +server +pc institute clie...	125	10%
12	web intrnet pages html publishing +browser dyna...	144	12%
14	+warehouse +'data warehouse' warehousing meta...	165	13%
15	sets +set +'data set' 'data sets' variables +variable...	159	13%
17	models modeling +test tests measures statistical ...	136	11%
20	+report lines reports reporting production groups ...	80	6%
25	+analysis +treatment groups +outcome difference...	70	6%
26	af +development +entry developers +function prog...	158	13%
29	+graph 'graph software' +report web +business wi...	105	8%

このテーブルには 10 個のクラスタがありますが、クラスタ ID の範囲は 1～10 ではないことに注意してください。

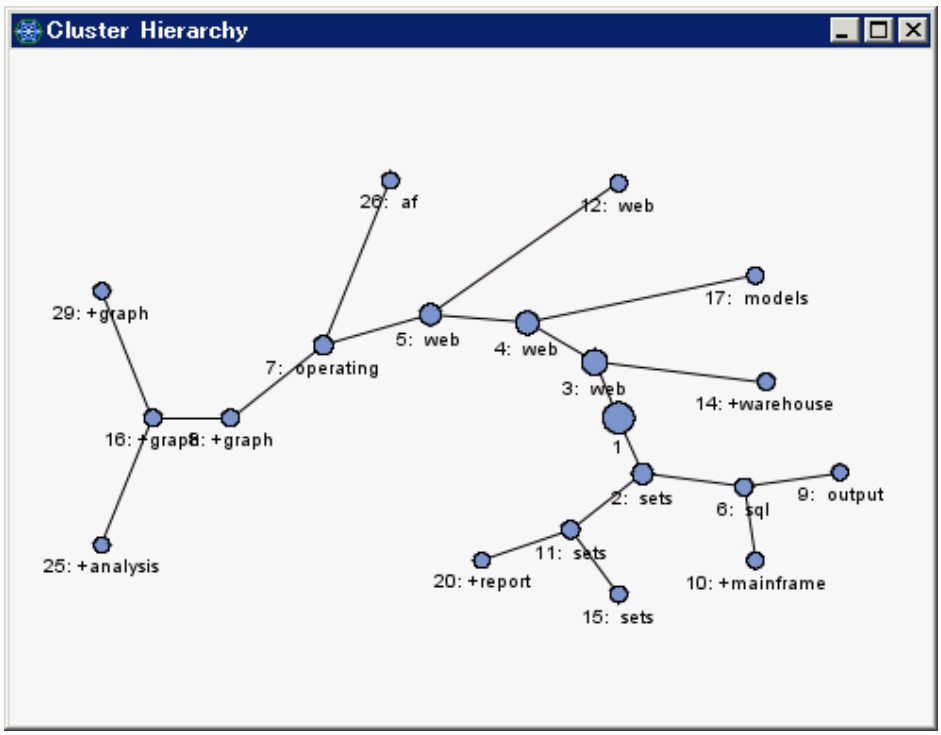
29. Hierarchy Data テーブルを選択し、クラスタテーブル内に非表示されているクラスタに関する詳細情報を確認します。



Hierarchy Data

Hierarchy Level	Cluster ID	Parent	Descriptive Terms	Frequency	Graph Description
1	1			12381	
2	2	1	sets +set 'data sets' +report +'data ...	4602	2: sets
2	3	1	web +warehouse +graph +design ...	7783	3: web
3	6	2	sql output +control +technology del...	2216	6: sql
3	11	2	sets +set +'data set' 'data sets' vari...	2391	11: sets
3	4	3	web +graph graphs models statisti...	6134	4: web
3	14	3	+warehouse +'data warehouse' war...	1651	4: +wareh...
4	5	4	web +graph graphs intrnet +versio...	4775	5: web
4	17	4	models modeling +test tests mea...	1361	7: models
4	9	6	output +date ods delivery +output ...	969	9: output
4	10	6	+mainframe +technology +server +...	1251	0: +mainfr...
4	15	11	sets +set +'data set' 'data sets' vari...	1591	5: sets
4	20	11	+report lines reports reporting prod...	802	0: +report
5	7	5	operating methods techniques ma...	333	7: operating
5	12	5	web intrnet pages html publishing ...	1441	2: web
6	8	7	+graph graphs +analysis statistical...	175	8: +graph
6	26	7	af +development +entry developers ...	1582	6: af
7	16	8	+graph graphs +analysis +present...	1751	6: +graph
8	25	16	+analysis +treatment groups +outc...	702	5: +analysis
8	29	16	+graph 'graph software' +report we...	1052	9: +graph

30. クラスタ階層グラフを選択し、クラスタの階層的なグラフィカル表現を確認します。

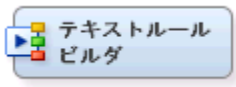


31. 結果ウィンドウを閉じます。

13 章

テキストルールビルダノード

テキストルールビルダノードについて



テキストルールビルダノードは、ターゲット変数の記述や予測に役立つルールの順序集合を生成します。この集合内の各ルールは、1 つの語または語の小規模なサブセットが存在するかどうかを示す論理積(“term1” AND “term2” AND (NOT “term3”)など)から構成される特定のターゲットカテゴリと関連付けられます。あるドキュメントが少なくとも term1 と term2 のオカレンスを含むが term3 のオカレンスは含まない場合にのみ、そのドキュメントはこのルールにマッチします。

この派生ルールの集合は、記述的かつ予測的である 1 つのモデルを生成します。新規ドキュメントを分類する場合、その作業は順序集合を通じて進められ、そのドキュメントにマッチした最初のルールと関連付けられているターゲットが選択されます。このルールは、SAS Content Categorization Studio 内部で使用可能でそこに配置可能な構文で提供されます。

テキストルールビルダノードは、標準的なレポート作成機能を備えた、標準的な SAS Enterprise Miner のモデリングツールです。このツールを使用すると、生成されたモデルに基づいて、どの予測ターゲット値が最も間違っている可能性が高いかを確認できます。オプションで、一部のオブザベーションに割り当てられたターゲットを変更し、結果を再実行できます。これにより、ユーザーがアルゴリズムと動的に対話して予測モデルを繰り返し構築できるような「アクティブな学習」を推進できます。

テキストルールビルダノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

この章の残りの部分では、**テキストルールビルダ**ノードの使用例を紹介します。

テキストルールビルダノードの使用

この例では、SAS Enterprise Miner が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

テキストルールビルダノードは、小規模な語のサブセットからブールルールを作成し、分類ターゲット変数を予測します。このノードの前には、**テキスト解析**ノードと**テキストフィルタ**ノードを配置する必要があります。

この例では、SAMPSIO.NEWS データセットを使用して、**テキストルールビルダ**ノードを使って分類ターゲット変数を予測する方法を示します。結果には、モデルが高度な解釈が可能であることや、説明や要約に役立つことも示されます。

SAMPSIO.NEWS データセットは、600 件の簡潔なニュース記事から構成されます。これらのニュース記事のほとんどは、コンピュータグラフィックス、ホッケー、医療問題のうちいずれか 1 つのカテゴリに分類されます。

SAMPSIO.NEWS データセットには 600 個のオブザベーションと次の変数が含まれています。

- **TEXT** は名義尺度変数であり、ニュース記事のテキストを含んでいます。
- **graphics** は二値変数であり、ドキュメントがコンピュータグラフィックスカテゴリに属すかどうかを示します(属す場合は 1、属さない場合は 0 となる)。
- **hockey** は二値変数であり、ドキュメントがホッケーカテゴリに属すかどうかを示します(属す場合は 1、属さない場合は 0 となる)。
- **medical** は二値変数であり、ドキュメントが医療問題カテゴリに属すかどうかを示します(属す場合は 1、属さない場合は 0 となる)。
- **newsgroup** は名義尺度変数であり、ニュース記事が当てはまるグループを含んでいます。

テキストルールビルダノードを使用して SAMPSIO.NEWS データセット内の分類ターゲット変数 **newsgroup** を予測するには、次の操作を実行します。

1. データソースウィザードを使用して、データセット SAMPSIO.NEWS 用のデータソースを定義します。
   a. 変数 **graphics**、**hockey**、**medical** の測定レベルを**二値**に設定します。
   b. 変数 **newsgroup** のモデル役割を**ターゲット**に設定し、変数 **graphics**、**hockey**、**medical** の役割を**入力**に設定します。
   c. 変数 **TEXT** が役割**テキスト**を持つように設定します。
   d. データソースウィザード — 意思決定の構成ダイアログボックスで**いいえ**を選択します。
   e. ターゲット **newsgroup** ではデフォルトのターゲットプロファイルを使用します。
2. **NEWS** データソースを作成した後、それをダイアグラムワークスペースへとドラッグします。

3. ツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。

4. **NEWS** データソースを**テキスト解析**ノードに接続します。

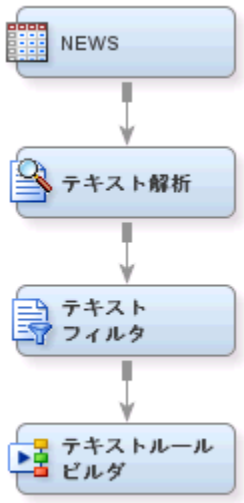
5. ツールバー上で**テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。

6. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。

7. ツールバー上で**テキストマイニング**タブを選択し、**テキストルールビルダ**ノードをダイアグラムワークスペースへとドラッグします。

8. **テキストフィルタ**ノードを**テキストルールビルダ**ノードに接続します。

   この時点で、プロセスフローダイアグラムは次のようになります。



9. プロセスフローダイアグラム内で**テキストルールビルダ**ノードを選択します。

10. **一般化誤差**プロパティの値をクリックし、**最低**を選択します。

11. **ルールの純度**プロパティの値をクリックし、**最低**を選択します。

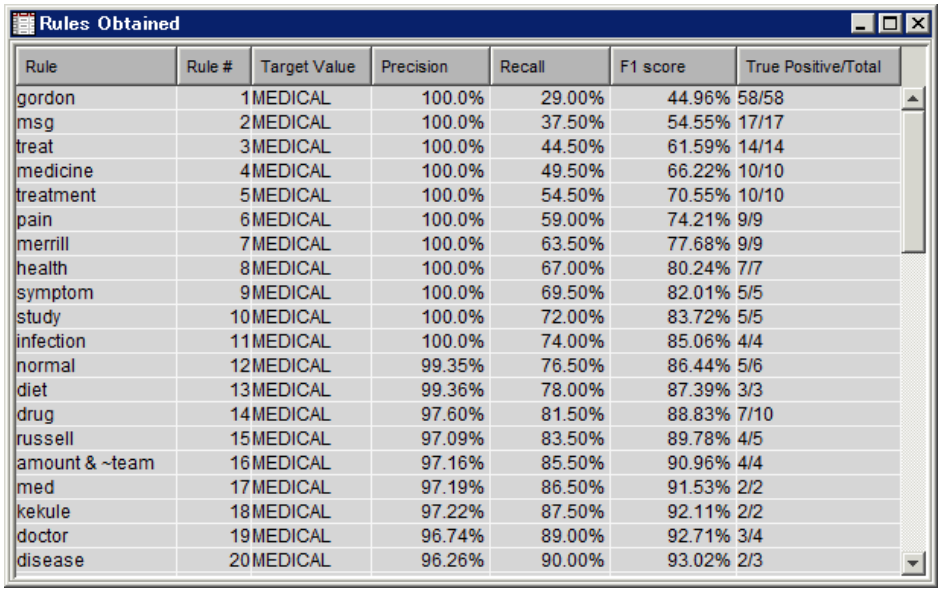
12. **全数**プロパティの値をクリックし、**最低**を選択します。

13. ダイアグラムワークスペースで、**テキストルールビルダ**ノードを右クリックし、**実行**を選択します。表示される確認ダイアログボックスで**はい**をクリックします。

14. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

15. 取得ルールテーブルを選択し、取得済みのルールに関する情報を確認します。

    Rule 列にある語は、ターゲット **newsgroup** を示すための対応する精度を持っています。



Rules Obtained

Rule	Rule #	Target Value	Precision	Recall	F1 score	True Positive/Total
gordon	1	MEDICAL	100.0%	29.00%	44.96%	58/58
msg	2	MEDICAL	100.0%	37.50%	54.55%	17/17
treat	3	MEDICAL	100.0%	44.50%	61.59%	14/14
medicine	4	MEDICAL	100.0%	49.50%	66.22%	10/10
treatment	5	MEDICAL	100.0%	54.50%	70.55%	10/10
pain	6	MEDICAL	100.0%	59.00%	74.21%	9/9
merrill	7	MEDICAL	100.0%	63.50%	77.68%	9/9
health	8	MEDICAL	100.0%	67.00%	80.24%	7/7
symptom	9	MEDICAL	100.0%	69.50%	82.01%	5/5
study	10	MEDICAL	100.0%	72.00%	83.72%	5/5
infection	11	MEDICAL	100.0%	74.00%	85.06%	4/4
normal	12	MEDICAL	99.35%	76.50%	86.44%	5/6
diet	13	MEDICAL	99.36%	78.00%	87.39%	3/3
drug	14	MEDICAL	97.60%	81.50%	88.83%	7/10
russell	15	MEDICAL	97.09%	83.50%	89.78%	4/5
amount & ~team	16	MEDICAL	97.16%	85.50%	90.96%	4/4
med	17	MEDICAL	97.19%	86.50%	91.53%	2/2
kekule	18	MEDICAL	97.22%	87.50%	92.11%	2/2
doctor	19	MEDICAL	96.74%	89.00%	92.71%	3/4
disease	20	MEDICAL	96.26%	90.00%	93.02%	2/3

上記の 7 番目の列で、真陽性(最初の数字)は、ルールに正しく割り当てられたドキュメントの数になります。合計(2 番目の数字)は、合計陽性になります。

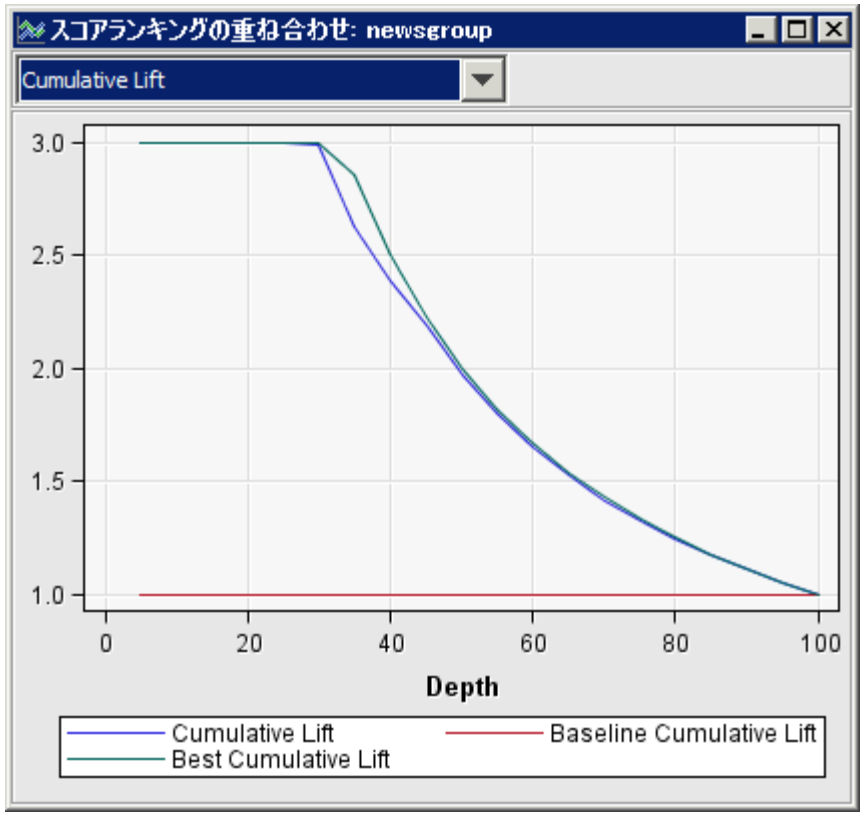
上記の例では、最初の行で、58 個のドキュメントがルール“gordon”に割り当てられている(58 個が正しく割り当てられている)ことが示されています。これは、ドキュメントが語“gordon”を含んでいる場合に、これらのドキュメントをすべて MEDICAL ニュースグループに割り当てるならば、58 個のうち 58 個が正しく割り当てられることを意味します。次の行では、17 個のドキュメントが、ルール“msg”に正しく割り当てられています。これは、ドキュメントが語“msg”を含んでいる場合に、これらのドキュメントをすべて MEDICAL ニュースグループに割り当てるならば、17 個のうち 17 個が正しく割り当てられることを意味します。

NEWS データセットのサイズが制限されているため、これらのルールのほとんどは単一語ルールです。ただし、複数語ルールが 1 つだけ含まれています。16 番目の行で、ルール“amount & ~team”は、ドキュメントが語“amount”を含んでいるが語“team”は含んでいない場合、残りのドキュメントのうち 4 個が MEDICAL ニュースグループに正しく割り当てられることを意味します。

注: ~は論理 NOT を示します。

16. スコアランキングオーバーレイグラフを選択し、ターゲット変数に関する次の種類の情報を表示します。
    - Cumulative Lift(累積リフト)
    - Lift(リフト)
    - Gain(利得)
    - % Response(応答%)
    - Cumulative % Response(累積応答%)
    - % Captured Response(捕捉済み応答%)
    - Cumulative % Captured Response(累積捕捉済み応答%)

    注: 統計量を変更するには、ドロップダウンメニューから上記の選択肢のいずれかを選択します。

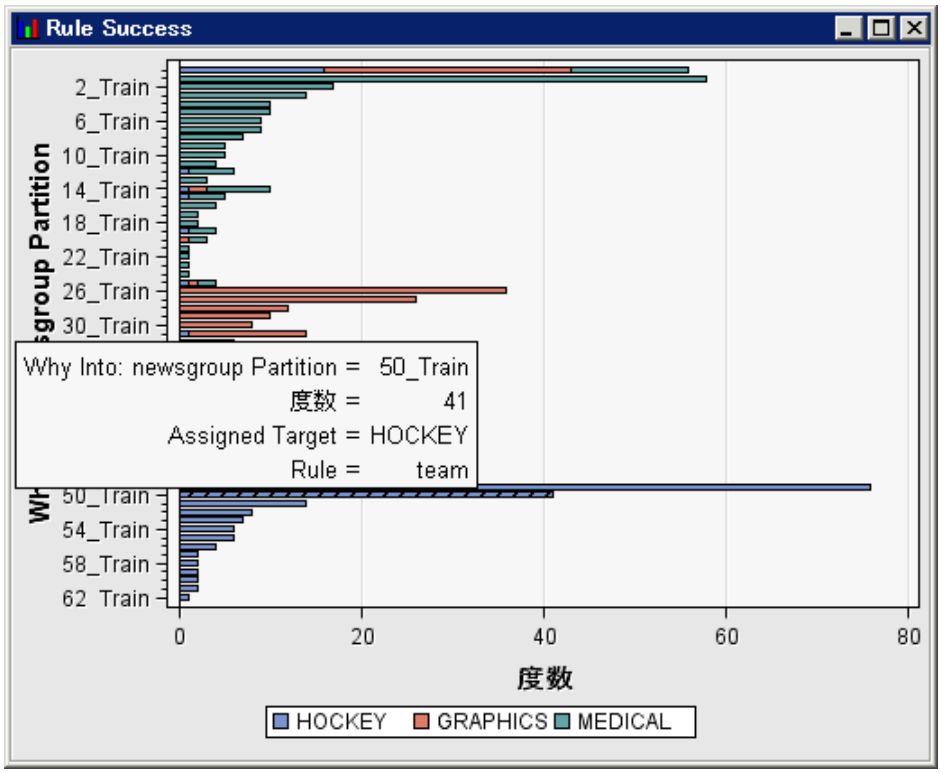


17. 当てはめ統計量ウィンドウを選択し、ターゲット変数 **newsgroup** に関する統計情報を表示します。

当てはめの統計量

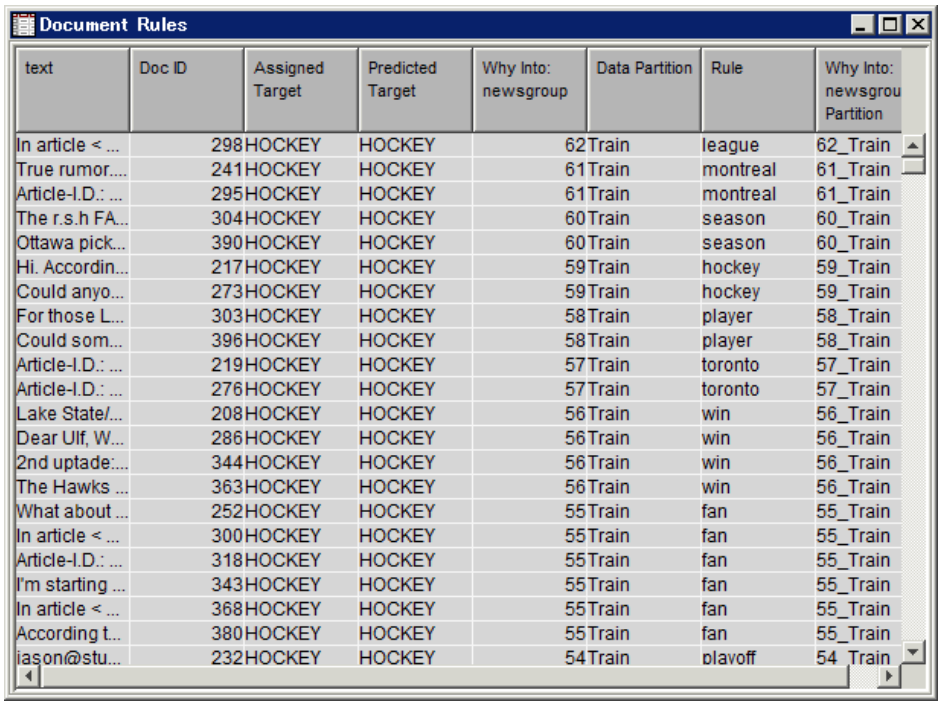
Target ▼	Target Label	Fit Statistics	Statistics Label	Train
newsgroup		_ASE_	Average Squared Error	0.037681
newsgroup		_DIV_	Divisor for ASE	1800
newsgroup		_MAX_	Maximum Absolute Error	1
newsgroup		_NOBS_	Sum of Frequencies	600
newsgroup		_RASE_	Root Average Squared Error	0.194116
newsgroup		_SSE_	Sum of Squared Errors	67.82585
newsgroup		_DISF_	Frequency of Classified Cases	600
newsgroup		_MISC_	Misclassification Rate	0.075
newsgroup		_WRONG_	Number of Wrong Classifications	45

18. ルール成功グラフを選択し、カーソルをバーの上に置くと、より詳細な情報を表示できます。



19. メニューから**表示** ⇨ **ルール** ⇨ **ドキュメントルール**を選択します。

ドキュメントルールテーブルが表示され、ルール成功グラフ内のルールに関するより詳細な情報を確認できます。



text	Doc ID	Assigned Target	Predicted Target	Why Into: newsgroup	Data Partition	Rule	Why Into: newsgrou Partition
In article < ...	298	HOCKEY	HOCKEY	62	Train	league	62_Train
True rumor....	241	HOCKEY	HOCKEY	61	Train	montreal	61_Train
Article-I.D.: ...	295	HOCKEY	HOCKEY	61	Train	montreal	61_Train
The r.s.h FA...	304	HOCKEY	HOCKEY	60	Train	season	60_Train
Ottawa pick...	390	HOCKEY	HOCKEY	60	Train	season	60_Train
Hi. Accordin...	217	HOCKEY	HOCKEY	59	Train	hockey	59_Train
Could anyo...	273	HOCKEY	HOCKEY	59	Train	hockey	59_Train
For those L...	303	HOCKEY	HOCKEY	58	Train	player	58_Train
Could som...	396	HOCKEY	HOCKEY	58	Train	player	58_Train
Article-I.D.: ...	219	HOCKEY	HOCKEY	57	Train	toronto	57_Train
Article-I.D.: ...	276	HOCKEY	HOCKEY	57	Train	toronto	57_Train
Lake State/...	208	HOCKEY	HOCKEY	56	Train	win	56_Train
Dear Ulf, W...	286	HOCKEY	HOCKEY	56	Train	win	56_Train
2nd uptade:...	344	HOCKEY	HOCKEY	56	Train	win	56_Train
The Hawks ...	363	HOCKEY	HOCKEY	56	Train	win	56_Train
What about ...	252	HOCKEY	HOCKEY	55	Train	fan	55_Train
In article < ...	300	HOCKEY	HOCKEY	55	Train	fan	55_Train
Article-I.D.: ...	318	HOCKEY	HOCKEY	55	Train	fan	55_Train
I'm starting ...	343	HOCKEY	HOCKEY	55	Train	fan	55_Train
In article < ...	368	HOCKEY	HOCKEY	55	Train	fan	55_Train
According t...	380	HOCKEY	HOCKEY	55	Train	fan	55_Train
jason@stu...	232	HOCKEY	HOCKEY	54	Train	playoff	54_Train

20. 結果ウィンドウを閉じます。

21. **一般化誤差**プロパティの値をクリックし、**中**を選択します。

22. **ルールの純度**プロパティの値をクリックし、**中**を選択します。

23. **全数**プロパティの値をクリックし、**中**を選択します。

24. **News** データソースを選択します。

25. **変数**プロパティの隣にある省略記号ボタンをクリックします。

26. HOCKEY 変数の役割を**ターゲット**に、NEWSGROUP 変数の役割を**入力**にそれぞれ変更します。

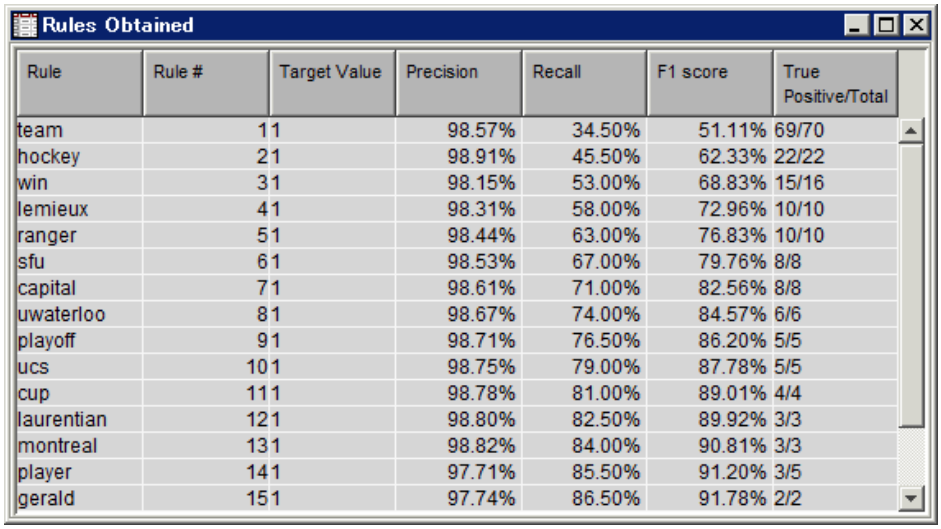
27. **OK** をクリックします。

28. ダイアグラムワークスペースで、**テキストルールビルダ**ノードを右クリックし、**実行**を選択します。表示される確認ダイアログボックスで**はい**をクリックします。

29. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

30. 取得ルールテーブルを選択し、ターゲットである HOCKEY ニュースグループを予測したルールに関する情報を確認します。

    Rule 列にある語は、ターゲット hockey を示すための対応する精度を持っています。


Rules Obtained

Rule	Rule #	Target Value	Precision	Recall	F1 score	True Positive/Total
team	11		98.57%	34.50%	51.11%	69/70
hockey	21		98.91%	45.50%	62.33%	22/22
win	31		98.15%	53.00%	68.83%	15/16
lemieux	41		98.31%	58.00%	72.96%	10/10
ranger	51		98.44%	63.00%	76.83%	10/10
sfu	61		98.53%	67.00%	79.76%	8/8
capital	71		98.61%	71.00%	82.56%	8/8
uwaterloo	81		98.67%	74.00%	84.57%	6/6
playoff	91		98.71%	76.50%	86.20%	5/5
ucs	101		98.75%	79.00%	87.78%	5/5
cup	111		98.78%	81.00%	89.01%	4/4
laurentian	121		98.80%	82.50%	89.92%	3/3
montreal	131		98.82%	84.00%	90.81%	3/3
player	141		97.71%	85.50%	91.20%	3/5
gerald	151		97.74%	86.50%	91.78%	2/2

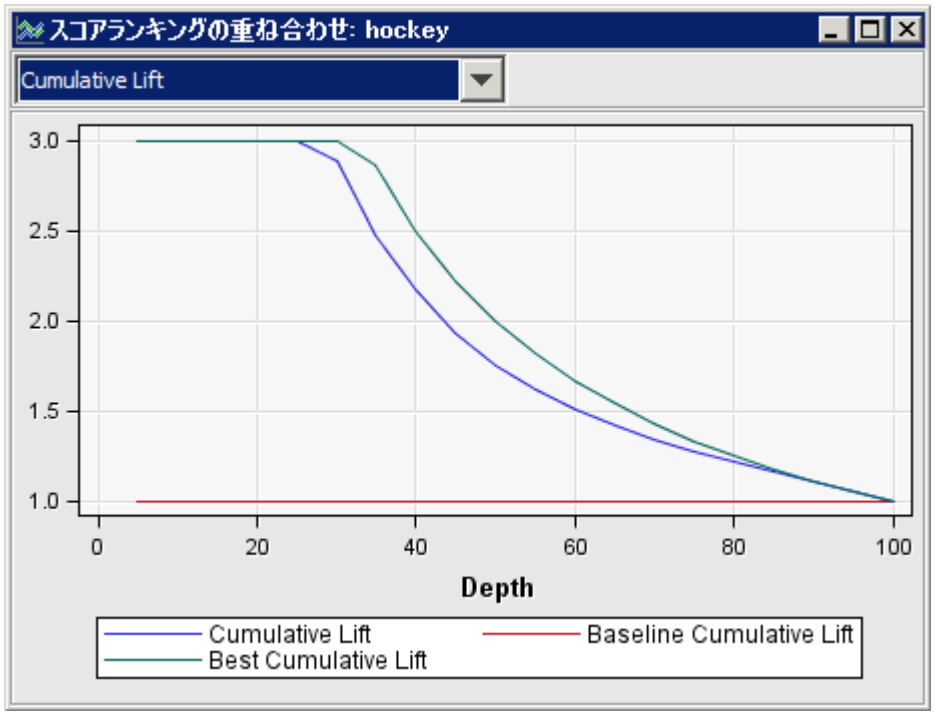
    上記の例では、ターゲット値は“HOCKEY”ではなく 1 になります。これは、**newsgroup** 変数ではなく、**hockey** 変数がターゲットになっているためです。70 個のドキュメントがルール“team”に割り当てられています(69 個が正しく割り当てられている)。これは、ドキュメントが“team”という語を含んでおり、これらのドキュメントをすべて HOCKEY ニュースグループに割り当てるならば、70 個のうち 69 個が正しく割り当てられることを意味します。次の行では、22 個のドキュメントが、ルール“hockey”に正しく割り当てられています。これは、ドキュメントが“hockey”という語を含んでおり、これらのドキュメントをすべて HOCKEY ニュースグループに割り当てるならば、22 個のうち 22 個が正しく割り当てられることを意味します。

31. スコアランキングオーバーレイグラフを選択し、ターゲット変数に関する次の種類の情報を表示します。

    - Cumulative Lift(累積リフト)
    - Lift(リフト)

- Gain(利得)
- % Response(応答%)
- Cumulative % Response(累積応答%)
- % Captured Response(捕捉済み応答%)
- Cumulative % Captured Response(累積捕捉済み応答%)

注: 統計量を変更するには、ドロップダウンメニューから上記の選択肢のいずれかを選択します。

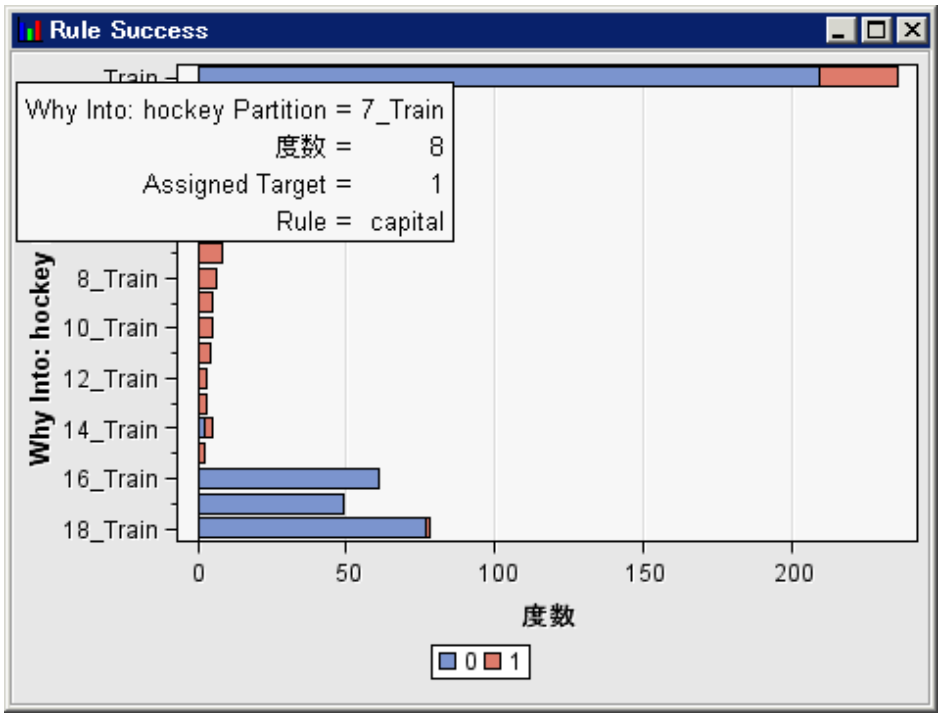


32. 当てはめ統計量テーブルを選択し、ターゲット変数 hockey に関する統計情報を表示します。

当てはめの統計量

Target	Target Label	Fit Statistics	Statistics Label	Train
hockey		_ASE_	Average Squared Error	0.009274
hockey		_DIV_	Divisor for ASE	1200
hockey		_MAX_	Maximum Absolute Error	0.957888
hockey		_NOBS_	Sum of Frequencies	600
hockey		_RASE_	Root Average Squared Error	0.096302
hockey		_SSE_	Sum of Squared Errors	11.12879
hockey		_DISF_	Frequency of Classified Cases	600
hockey		_MISC_	Misclassification Rate	0.051667
hockey		_WRONG_	Number of Wrong Classifications	31

33. ルール成功グラフを選択し、カーソルをバーの上に置くと、より詳細な情報を表示できます。



34. メニューから**表示** ⇨ **ルール** ⇨ **ドキュメントルール**を選択します。

    ドキュメントルールテーブルが表示され、ルール成功グラフ内のルールに関するより詳細な情報を確認できます。


Document Rules

text	Doc ID	Assigned Target	Predicted Target	Why Into: hockey	Data Partition	Rule	Why Into: hockey Partition
Ottawa picks f...	3901		1		5Train	ranger	5_Train
Well, I looked ...	3931		1		5Train	ranger	5_Train
Tampa Bay 1 ...	2111		1		4Train	lemieux	4_Train
NY Rangers 3...	2121		1		4Train	lemieux	4_Train
Article-I.D.: oz....	2191		1		4Train	lemieux	4_Train
Article-I.D.: alc...	2751		1		4Train	lemieux	4_Train
In article < 19...	3111		1		4Train	lemieux	4_Train
Boston 2 2 0--...	3331		1		4Train	lemieux	4_Train
In article < 1q...	3541		1		4Train	lemieux	4_Train
In article < 92...	3551		1		4Train	lemieux	4_Train
In article 6143...	3571		1		4Train	lemieux	4_Train
Pens-6 NJ De...	3671		1		4Train	lemieux	4_Train
Lake State/Ma...	2081		1		3Train	win	3_Train
First of all, the...	2131		1		3Train	win	3_Train
In article < 19...	2341		1		3Train	win	3_Train
Article-I.D.: cb...	2361		1		3Train	win	3_Train
In < rauser.73...	2491		1		3Train	win	3_Train
In article < 1p...	2611		1		3Train	win	3_Train

35. 結果ウィンドウを閉じます。

36. **コンテンツ分類コード**プロパティの隣にある省略記号ボタンをクリックします。

    コンテンツ分類コードウィンドウが表示されます。このウィンドウ内に提供されるコードは、SAS コンテンツ分類の出力となるコードであり、コンパイルの用意ができています。

37. **キャンセル**をクリックします。

38. **ターゲット値の変更**プロパティの隣にある省略記号ボタンをクリックします。

    ターゲット値の変更ウィンドウが表示されます。

    ターゲット値の変更ウィンドウを使用するとモデルを改善できます。

39. **割り当てターゲット**列内にある 1 つ以上のセルを選択し、新しいターゲット値を選択します。

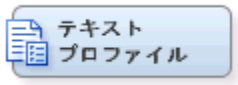
40. **OK** をクリックします。

41. **テキストルールビルダ**ノードに戻り、モデルが改善されたかどうかをチェックします。

14 章

テキストプロファイルノード

テキストプロファイルノードについて



テキストプロファイルノードを使用すると、ドキュメント内で見つかった語を使ったターゲット変数のプロファイリングが行えます。ターゲット変数のレベルごとに、同ノードは、そのレベルを特徴付け記述するコレクションに含まれている語のリストを出力します。

このアプローチは、TMBelief プロシジャを使用する階層ベイズモデルを使用して、当該レベルを記述するのに最もふさわしい語は何であるかを予測します。最も一般的な語を単に選択するのを防ぐために、事前の確率を使用して、ターゲット変数の 2 つ以上のレベルで共通する語には低い重みが付けられます。二値ターゲット変数の場合、2 ウェイ比較を使用することで、語の選択を強化しています。名義尺度変数の場合、n ウェイ比較が使用されます。順序尺度および時間変数(内部的に順序尺度へと変換される)の場合、シーケンシャルな 2 ウェイ比較が実施されます。これは、レベル n で報告された語が、レベル n-1 で報告された語と比較されることを意味します。例外は最初のレベルであり、この場合、比較するための先行レベルが存在しないため、レベル 2 と比較されます。

変数タイプのあらゆるケースにおいて、コーパスレベルのプロファイル出力も提供されます。これは、コレクション全体で最良の記述語として解釈されます。

テキストプロファイルノードに関する詳細は、SAS Text Miner のヘルプを参照してください。

この章の残りの部分では、**テキストプロファイル**ノードの使用例を紹介します。

テキストプロファイルノードの使用

この例では、SAMPSIO.NEWS データセットを使用して、**テキストプロファイル**ノードを使った語のプロファイリングを行う方法を示します。この例では、SAS Enterprise Miner

が実行されていること、およびダイアグラムワークスペースがプロジェクトで開かれていることを前提としています。プロジェクトとダイアグラムの作成に関する詳細は、3 章, “プロジェクトの設定” (11 ページ)を参照してください。

テキストプロファイルノードの前には、1 つの**テキスト解析**ノードおよび少なくとも 1 つの**テキストフィルタ**ノードを配置する必要があります。

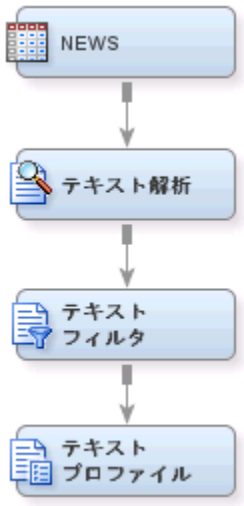
SAMPSIO.NEWS データセットは、600 件の簡潔なニュース記事から構成されます。これらのニュース記事のほとんどは、コンピュータグラフィックス、ホッケー、医療問題のうちいずれか 1 つのカテゴリに分類されます。

SAMPSIO.NEWS データセットには 600 個のオブザベーションと次の変数が含まれています。

- **TEXT** は名義尺度変数であり、ニュース記事のテキストを含んでいます。
- **graphics** は二値変数であり、ドキュメントがコンピュータグラフィックスカテゴリに属すかどうかを示します(属す場合は 1、属さない場合は 0 となる)。
- **hockey** は二値変数であり、ドキュメントがホッケーカテゴリに属すかどうかを示します(属す場合は 1、属さない場合は 0 となる)。
- **medical** は二値変数であり、ドキュメントが医療問題カテゴリに属すかどうかを示します(属す場合は 1、属さない場合は 0 となる)。
- **newsgroup** は名義尺度変数であり、ニュース記事が当てはまるグループを含んでいます。

1. データソースウィザードを使用して、データセット SAMPSIO.NEWS 用のデータソースを定義します。
   a. 変数 **graphics**、**hockey**、**medical** の測定レベルを**二値**に設定します。
   b. 変数 **hockey** のモデル役割を**ターゲット**に設定し、変数 **newsgroup**、**graphics**、**medical** の役割を**入力**に設定します。
   c. 変数 **TEXT** が役割**テキスト**を持つように設定します。
   d. データソースウィザード — 意思決定の構成ダイアログボックスで**いいえ**を選択します。
   e. ターゲット **hockey** ではデフォルトのターゲットプロファイルを使用します。
2. **NEWS** データソースを作成した後、それをダイアグラムワークスペースへとドラッグします。
3. ツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。
4. **NEWS** データソースを**テキスト解析**ノードに接続します。
5. ツールバー上で**テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。
6. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。
7. ツールバー上で**テキストマイニング**タブを選択し、**テキストプロファイル**ノードをダイアグラムワークスペースへとドラッグします。
8. **テキストフィルタ**ノードを**テキストプロファイル**ノードに接続します。

この時点で、プロセスフローダイアグラムは次のようになります。

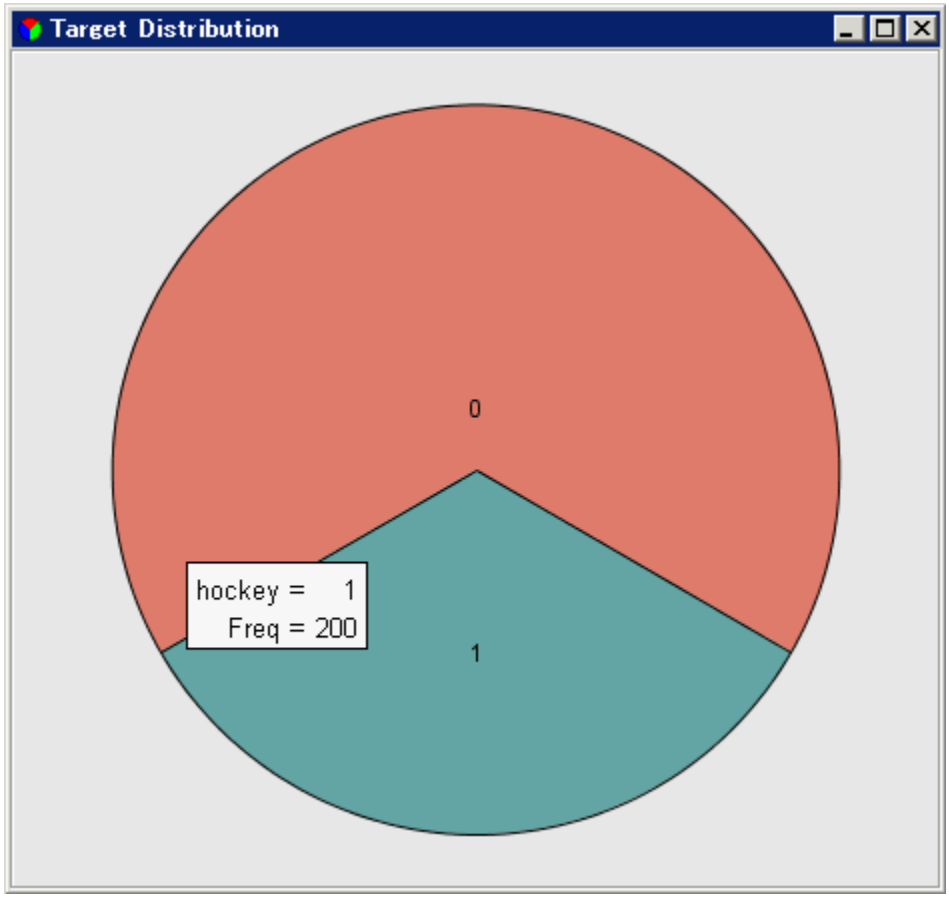


9. プロセスフローダイアグラム内で**テキストプロファイル**ノードを選択します。

10. ダイアグラムワークスペースで、**テキストプロファイル**ノードを右クリックし、**実行**を選択します。表示される確認ダイアログボックスで**はい**をクリックします。

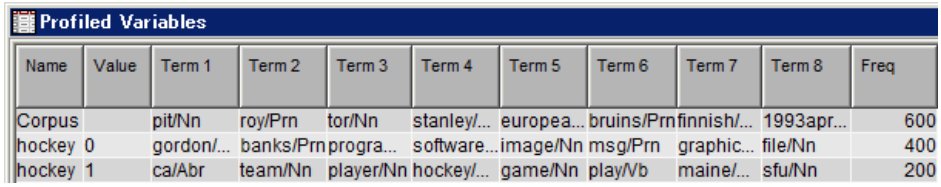
11. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

12. **ターゲットの分布**円グラフを選択します。



ターゲットの分布円グラフには、ターゲット値の頻度が表示されます。このグラフは階層レベル別にグループ化されており、プロファイル済み変数テーブルにリンクされています。

13. **プロファイル済み変数**テーブルを選択します。



Profiled Variables

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq
Corpus		pit/Nn	roy/Prn	tor/Nn	stanley/...	europea...	bruins/Prn	finnish/...	1993apr...	600
hockey	0	gordon/...	banks/Prn	progra...	software...	image/Nn	msg/Prn	graphic...	file/Nn	400
hockey	1	ca/Abr	team/Nn	player/Nn	hockey/...	game/Nn	play/Vb	maine/...	sfu/Nn	200

プロファイル済み変数テーブルには、ターゲット変数値の個々の組み合わせと、それらが関連付けられている最も高いビリーフ(確信度)を持つ語が表示されます。各オブザベーションは最大で、指定された最大数の語(各オブザベーションに関連付けられているもの)を持ちますが、それより少ない数を持つ場合もあります。**テキストプロファイル**ノードの結果ウィンドウに表示されるすべてのグラフィカルな結果は、このテーブルにリンクされています。このため、プロファイル済み変数テーブル内のオブザベーションを選択すると、それに対応するデータ点がグラフィックス内で強調表示されます。または、グラフィックス内のデータ点を選択すると、それに対応するオブザベーションがプロファイル済み変数テーブル内で強調表示されます。

14. 結果ウィンドウを閉じます。

15. **News** データソースを選択します。

16. **変数**プロパティの隣にある省略記号ボタンをクリックします。

変数ダイアログボックスが表示されます。

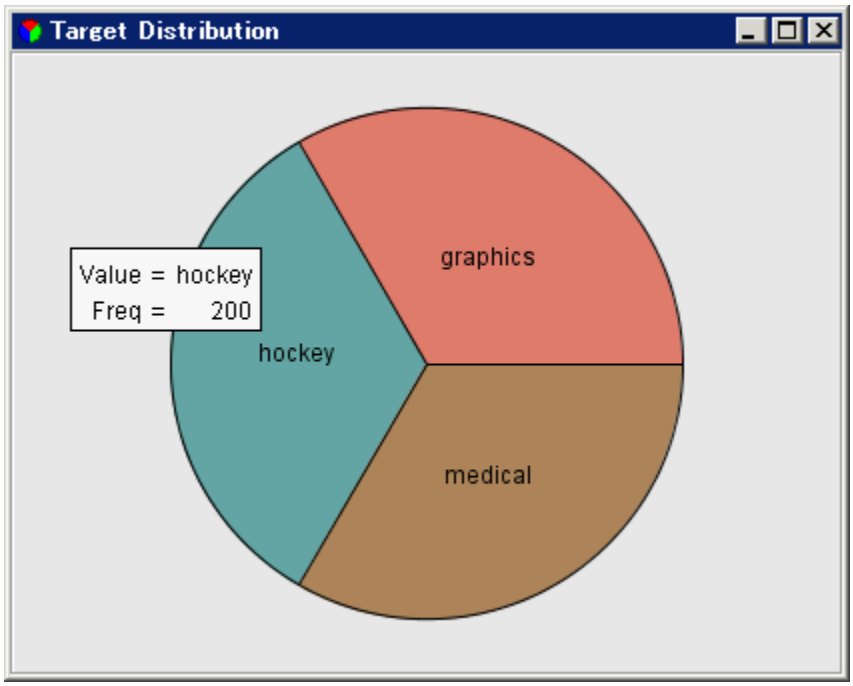
17. **newsgroup** 変数の役割を**ターゲット**に、**hockey** 変数の役割を**入力**にそれぞれ設定します。

18. **OK** をクリックします。

19. ダイアグラムワークスペースで、**テキストプロファイル**ノードを右クリックし、**実行**をクリックします。表示される確認ダイアログボックスで**はい**をクリックします。

20. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

21. **ターゲットの分布**円グラフを選択します。



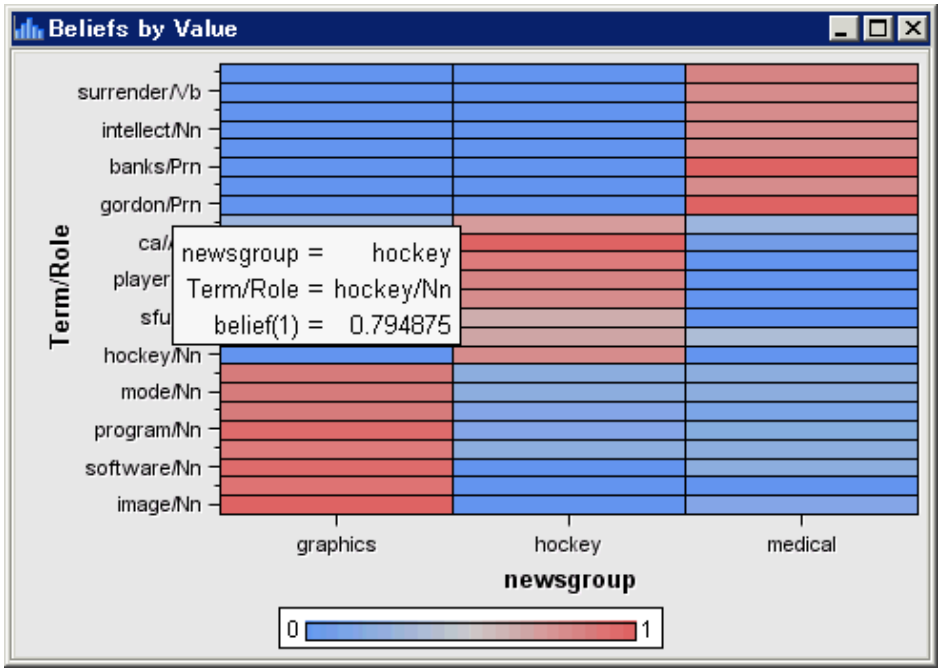
ターゲットの値を、二値の hockey 変数から名義尺度の newsgroup 変数へと変更したため、3 つの可能な newsgroup 値(hockey、medical、graphics)の分布を確認できます。

22. **ターゲットの類似点**コンスタレーションプロット

ターゲットの類似点コンスタレーションプロットは、さまざまなターゲット値間における類似性を表示します。類似性は、語のビリーフ(確信度)に対して PROC DISTANCE を使用することにより計測されます。リンクは、階層の同じレベル上のターゲット値間でのみ表示されます。コンスタレーションプロットは、プロファイル済み変数テーブルにリンクされています。

注: **ターゲットの類似点**コンスタレーションプロットは、名義尺度および順序尺度ターゲットで利用できます。

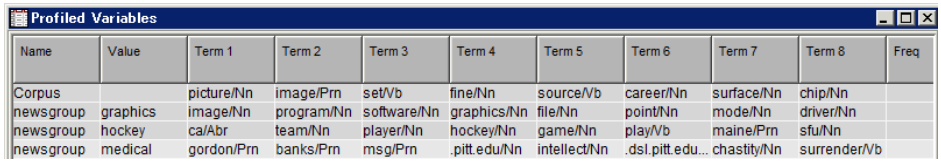
23. **ビリーフ(値)**グラフを選択します。



ビリーフ(値)グラフは、さまざまなターゲット値の語と役割のペアに関するビリーフ(確信度)値を表示します。マウスポインタをセルの上に置くと、ツールチップにターゲット値、語と役割のペア、ビリーフ値が表示されます。

注: **ビリーフ(値)**グラフは、名義尺度および順序尺度ターゲットの場合に表示されます。

24. **プロファイル済み変数**テーブルを選択します。



Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq
Corpus		picture/Nn	image/Prn	set/Vb	fine/Nn	source/Vb	career/Nn	surface/Nn	chip/Nn	
newsgroup	graphics	image/Nn	program/Nn	software/Nn	graphics/Nn	file/Nn	point/Nn	mode/Nn	driver/Nn	
newsgroup	hockey	ca/Abr	team/Nn	player/Nn	hockey/Nn	game/Nn	play/Vb	maine/Prn	sfu/Nn	
newsgroup	medical	gordon/Prn	banks/Prn	msg/Prn	.pitt.edu/Nn	intellect/Nn	.dsl.pitt.edu...	chastity/Nn	surrender/Vb	

25. 結果ウィンドウを閉じます。

26. **テキストプロファイル**ノードを選択します。

27. **語の最大数**プロパティの値をクリックし、16 を入力します。

    キーボードの **Enter** キーを押します。

28. ダイアグラムワークスペースで、**テキストプロファイル**ノードを右クリックし、**実行**をクリックします。表示される確認ダイアログボックスで**はい**をクリックします。

29. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

30. **プロファイル済み変数**テーブルを選択します。ターゲット値ごとに 16 個の語と役割のペアが表示されていることを確認します。

31. 結果ウィンドウを閉じます。

32. (オプション)**テキストプロファイル**ノードを、出力形式 DATE または DATETIME を持つターゲット変数とともに実行します。この結果、語(時系列)ラインプロットが作成されます。

    a. **Sampsio** ライブラリを使用してデータソース **VRTEXT** を作成します。**VAX_DATE** 変数の役割値を `Target` に設定します。

       このデータソースには、ワクチン接種に対する有害事象反応が含まれています。たとえば、報告された副作用には、痛み、充血、発熱などが含まれる可能性があります。

    b. **VRTEXT** ノードをダイアグラムワークスペースに追加します。

    c. ツールバー上で**テキストマイニング**タブを選択し、**テキスト解析**ノードをダイアグラムワークスペースへとドラッグします。

    d. **VRTEXT** データソースを**テキスト解析**ノードに接続します。

    e. ツールバー上で**テキストマイニング**タブを選択し、**テキストフィルタ**ノードをダイアグラムワークスペースへとドラッグします。

    f. **テキスト解析**ノードを**テキストフィルタ**ノードに接続します。

    g. ツールバー上で**テキストマイニング**タブを選択し、**テキストプロファイル**ノードをダイアグラムワークスペースへとドラッグします。

    h. **テキストフィルタ**ノードを**テキストプロファイル**ノードに接続します。

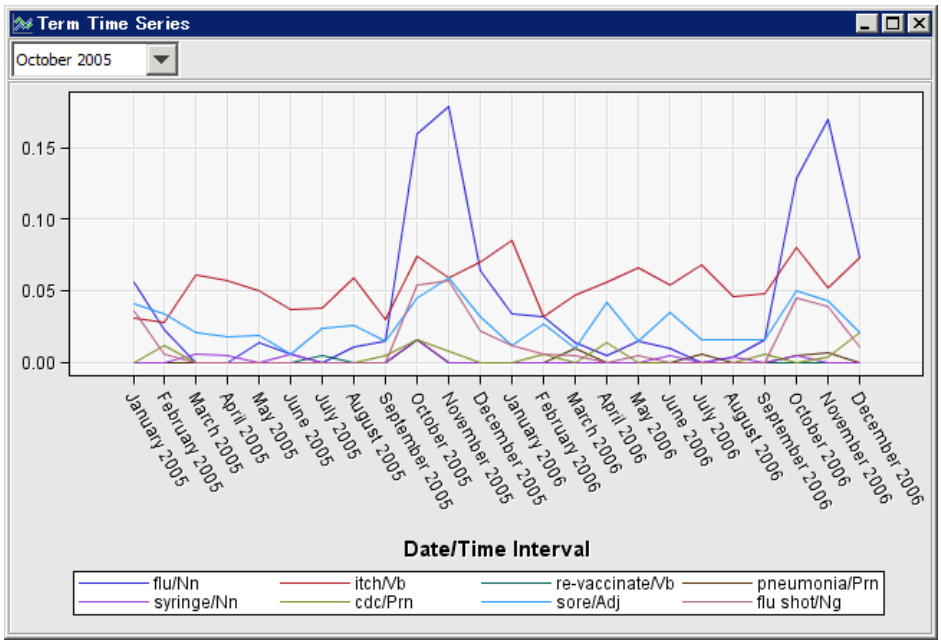
       この時点で、プロセスフローダイアグラムは次のようになります。

    i. **テキストプロファイル**ノードを選択し、**ビンの間隔**プロパティの値として**月単位**を選択します。

    j. ダイアグラムワークスペースで、**テキストプロファイル**ノードを右クリックし、**実行**をクリックします。表示される確認ダイアログボックスで**はい**をクリックします。

    k. 同ノードの実行完了後に表示される実行ステータスダイアログボックス内で**結果**をクリックします。

l. 語(時系列)ラインプロットを選択した後、メニューから `Oct 2005` を選択します。



この例では、インフルエンザが流行し始め、数カ月間はピーク状態を保つことを確認できます。

33. (オプション)**テキストプロファイル**ノードを順序尺度のターゲット変数とともに実行します。この結果、語(順序列)ラインプロットが作成されます。

15 章

テキストマイニングのヒント

大規模なドキュメントコレクションの処理

SAS Text Miner ノードを使用して大規模なドキュメントコレクションを処理する場合、非常に大量の計算時間とリソースが必要となることがあります。リソースが限られている場合、次に示すアクションのうちいずれか 1 つまたは複数を実施する必要があります。

- ドキュメントコレクションのサンプルを使用すること。
- 解析プロパティの一部を **No** または **None** に設定すること(**名詞グループ**や**エンティティの検索**など)。
- SVD 次元やロールアップ語の数を減らすこと。SVD アプローチでメモリ問題が発生している場合、特定の数の語をロールアップすると、残りの語は自動的に破棄されます。
- 名詞、固有名詞、名詞グループ、動詞以外のすべての品詞をオフにすることにより、解析を情報性の高い語に限定すること。
- 最良の結果を得るために、正しい文法、句読点、大文字の使用を含めて、センテンスを構造化すること。エンティティの抽出が常に適切な結果を生成するとは限りません。

長大なドキュメントの処理

SAS Text Miner は、ドキュメントを表示するために"bag-of-words"(さまざまな語を詰めた袋)式のアプローチを使用します。これは、ドキュメントが、各語が各ドキュメント内で現れる頻度を含むベクター(一次元配列)を使用して表されることを意味します。なお、語の順序は無視されます。このアプローチは、短いパラグラフサイズのドキュメントの場合には非常に有効ですが、長大なドキュメントの場合には有害な損失を引き起こすことがあります。各自のモデルで実際に使用するコンテンツを切り分けるためには、長大なドキュメントを前処理することを検討した方が良いでしょう。たとえば、論文を分析

する場合、概要のみの分析が最も良い結果をもたらすと分かることがあります。長大なドキュメントから関連するコンテンツを抽出する場合、Perl のような別のプログラミング言語を使用することを検討してください。

サポートされていない言語やエンコーディングのドキュメントを処理するには

サポートされていない言語やエンコーディングのドキュメント群を保有している場合でも、テキストを上手に処理し、有益な結果が得られる場合があります。次の操作を実行します。

1. 言語を**英語**に設定します。
2. 次の解析プロパティをオフにします。
   - **品詞を区別する**
   - **名詞グループ**
   - **エンティティの検索**
   - **語のステミング**
3. **テキスト解析**ノードを実行します。

16 章

次なるステップ:追加機能の概要

%TEXTSYN マクロ

%TEXTSYN マクロは、SAS Text Miner と共に提供されます。**テキスト解析**ノードの実行完了後にこのマクロを使用することで、入力データソース内にあるスペルミスを検出して修正できます。英語でのみ使用することができます。

このマクロは、ユーザーが SAS Text Miner で利用できる類義語データセットを作成します。このデータセットには、スペルの間違った語や、正しいスペルの候補語が含まれています。このデータセットには、変数として term、parent、category が含まれています。また、オプション引数を使用することで、類義語データセットに、スペルの間違った語の使用例(最大で 2 つのドキュメントに含まれているもの)を含めるように指定できます。

%TEXTSYN マクロの詳細については、SAS Text Miner のヘルプを参照してください。

%TMFILTER マクロ

%TMFILTER マクロは、ファイルを SAS データセットに変換する SAS マクロです。%TMFILTER マクロは、SAS Text Miner と共に提供されます。同マクロのフィルタリング機能はすべてのオペレーティングシステムでサポートされており、クロール機能は Windows 上でサポートされています。%TMFILTER マクロは、Windows マシン上にインストールされ実行されている SAS Document Conversion Server を利用します。詳細については、SAS Document Conversion Server のマニュアルを参照してください。このマクロを使用して次のタスクを実行できます。

- 任意のサポートされているフォーマットで保存されているドキュメントコレクションをフィルタリングし、SAS Text Miner のデータソースの作成に使用できる SAS データセットを出力すること。
- Web クロールを実施し、SAS Text Miner のデータソースの作成に使用できる SAS データセットを出力すること。Web クロールは、開始 Web ページのテキストを取得し、同ページ内の URL リンクを抽出した後、それらのリンク先のページ内で同じ処理を再帰的に繰り返します。また、開始 URL のドメインに対するクロールを禁止す

ることや、開始 URL のドメイン内には存在しないリンク先のページをクロールすることもできます。クロールは、指定されたドリルダウンのレベル数に達するまで、またはドメイン制約を満たす Web ページが見つかるまで続けられます。Web クロール機能は、Windows オペレーティングシステムでのみサポートされています。

- コレクション内のすべてのドキュメントの言語を識別すること。

%TMFILTER マクロの詳細については、SAS Text Miner のヘルプを参照してください。

推奨資料

- *SAS Text Miner 13.2 (http://support.sas.com/documentation/onlinedoc/txtminer)* や *SAS Enterprise Miner 13.2 (http://support.sas.com/documentation/onlinedoc/miner)* で紹介されているその他の製品ドキュメントで取り上げられている多くの概念やトピックも、*SAS Text Miner 13.2* を使用する場合に役立ちます。

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

libref (ライブラリ参照名)
SAS ライブラリに一時的に関連付けられる名前。SAS ファイルの完全名は、ピリオドで区切られた 2 つの語から構成されます。最初の語はライブラリ参照名であり、これはライブラリを表します。2 番目の語は、特定の SAS ファイルの名前になります。たとえば、VLIB.NEWBDAY の場合、ライブラリ参照名 VLIB は、ファイル NEWBDAY が格納されているライブラリを表しています。ライブラリ参照名を割り当てるには、LIBNAME ステートメントを使用するか、またはオペレーティングシステムのコマンドを使用します。

SAS データセット
SAS 固有のいずれかのファイル形式で内容が格納されたファイル。SAS データセットには次の 2 種類があります。SAS データファイルと SAS データビューです。SAS データファイルは、データ値に加えて、そのデータに関連付けられているディスクリプタ情報を含みます。SAS データビューには、ディスクリプタ情報と、他の SAS データセットまたはソフトウェアベンダのファイル形式で格納されたファイルからデータ値を取り出すために必要となるその他の情報のみが含まれます。

エンティティ
SAS Text Miner が一般的なテキストから区別することができるタイプの情報。たとえば、SAS Text Miner は、名前(人名、地名、会社名、製品名など)、アドレス(番地、郵便番号、メールアドレス、URL など)、日付、単位、通貨量、およびその他多くのエンティティを識別できます。

解析
テキストを成分語、フレーズ、マルチワード語、句読点、およびその他のタイプの情報に分割する目的でテキストを分析すること。

学習データ
モデル学習に使用される入力値とターゲット値を含んでいる現在利用可能なデータ。

カタログディレクトリ
カタログの各メンバの名前、種類、説明、および更新ステータスに関する情報を格納して保持する、SAS カタログの一部。

クラスタリング
各グループ内のオブザベーションが可能な限り互いに近くなるように、かつ異なるグループが可能な限り互いに遠くなるように、1 つのデータセットを相互排他的な複数のグループへと分割する処理。SAS Text Miner では、クラスタリングには、特定のコレクションに含まれているドキュメントのうち、互いに類似しているグループ

を検出する機能が含まれています。クラスタの決定時に、そのクラスタ内にあるワードを検証することで、同クラスタのフォーカスが明らかになります。特定のドキュメントコレクション内でクラスタを形成することにより、各ドキュメントを読まなくとも、そのコレクションの内容を理解し要約できるようになります。クラスタを形成することで、当該コレクションにより強調されている中心テーマやカギとなる概念を明らかにできます。

検証データ
学習データを使用して開発されたデータモデルの適合性の検証に使用されるデータ。学習データセットと検証データセットの両者には、ターゲット変数値が含まれています。学習データ内のターゲット変数値は、モデルの学習に使用されます。検証データセット内のターゲット変数値は、学習モデルの予測値を既知のターゲット値と比較するために使用されます。これにより、そのモデルを使用して新しいデータをスコアリングする前に、同モデルの適合性を評価できます。

コンセプトのリンク付け
語テーブル内の選択された語に概念的に関連付けられている語を検索し表示する機能です。

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処理。スコアリングは、データマイニングで実行される最後の処理です。

ステミング
語の原形を見つけて戻す処理。たとえば、語 grind、grinds、grinding、ground の原形は grind になります。ステミングは英語でのみ使用することができます。

セグメント化
1 つの母集団を、同様の要素を含む複数のサブ母集団へと分割する処理。セグメント化は、スーパーバイザーモードで実行することもできれば(ターゲット変数と、デシジョンツリーのような各種の手法を組み合わせて使用)、スーパーバイザー権限なしでも実行できます(クラスタリングまたは Kohonen ネットワークを使用)。

ソースレベルデバッガ
開発中のプログラム内の論理エラーを検出し解決するために使用される SAS システムの対話環境。デバッガは、複数のウィンドウと一群のコマンドから構成されます。

ダイアグラム
プロセスフローダイアグラムを参照。

データソース
Java ベースの Enterprise Miner GUI 環境において SAS データセットを表すデータオブジェクトです。データソースには、Enterprise Miner がデータマイニングのプロセスフローダイアグラムでデータを使用するために必要とする SAS データセットに関するすべてのメタデータが含まれています。SAS Enterprise データソースの作成に必要となる SAS データセットのメタデータには、同データセットの名前と場所、そのライブラリパスの定義に使用される SAS コード、およびデータマイニング処理で使用される変数役割、測定レベル、関連付けられている属性が含まれています。

停止リスト
テキストマイニング分析から除外したい情報に乏しい無関係な語の単純なコレクションを含んでいる SAS データセット。

テストデータ
学習時には使用されないが、一般化やモデルの比較に使用される入力値とターゲット値を含んでいる現在利用可能なデータ。

特異値分解(SVD)
高次元データを低次元データに変換する手法。

ノード
(1) SAS Enterprise Miner のユーザーインターフェイスにおける、プロセスフローダイアグラム内のデータマイニングタスクを表すグラフィカルオブジェクト。データマイニングタスクを実行する統計ツールは、データマイニングのプロセスフローダイアグラム内に配置された時点でノードと呼ばれます。各ノードは、分析および予測データモデルのコンポーネントとして、数学的操作やグラフィカル操作を実行します。(2) ニューラルネットワークにおける線形または非線形のコンピューティング要素であり、1 つまたは複数の入力を受け取り、入力関数を計算し、オプションでその結果を 1 つ以上のニューロンに振り向けます。ノードはニューロンまたはユニットとも呼ばれます。(3) ツリーダイアグラム内のリーフ(葉)。リーフ、ノード、セグメントという用語は密接に関連しており、これらはツリー内の同じ部分を指す場合があります。

プロセスフローダイアグラム
データマイニング分析時に、個々の Enterprise Miner ノードにより実行される各種のデータマイニングタスクをグラフィカルに表現したもの。プロセスフローダイアグラムは、データマイナーが希望する対応する統計的操作の実行順に接続された、2 つ以上の個別ノードから構成されます。省略形は PFD です。

分割
利用可能なデータを、学習(training)、検証(validation)、テスト(test)の各データセットに分割すること。

変数
SAS データセットまたは SAS データビュー内の列。各変数のデータ値は、すべてのオブザベーションの単一の特性を表します。各 SAS 変数は、名前、データタイプ(文字または数値)、長さ、出力形式、入力形式、ラベルという属性を持ちます。

モデル
入力から出力を計算する公式またはアルゴリズムです。データマイニングモデルには、入力変数が与えられた場合、ターゲット変数の条件付き分布に関する情報が含まれています。

ロールアップ語
ドキュメントコレクション内で最も大きく重み付けされている語。

キーワード