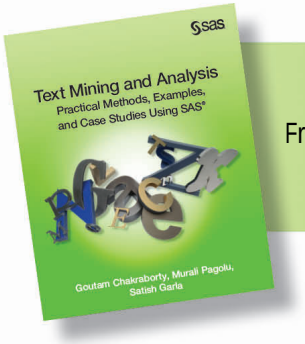
Text Mining and Analysis

Practical Methods, Examples, and Case Studies Using SAS®

Goutam Chakraborty, Murali Pagolu, Satish Garla



Contents

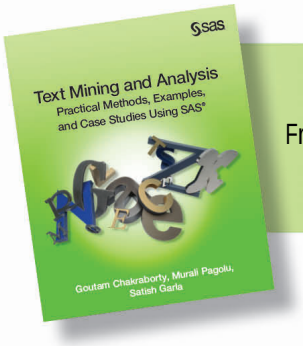
About This Book xi
About The Authors xv
Acknowledgments xvii
Chapter 1 Introduction to Text Analytics 1
Overview of Text Analytics 1
Text Mining Using SAS Text Miner 5
Information Retrieval 7
Document Classification 8
Ontology Management 9
Information Extraction 10
Clustering 11
Trend Analysis 12
Enhancing Predictive Models Using Exploratory Text Mining 13
Sentiment Analysis 14
Emerging Directions 15
Handling Big (Text) Data 15
Voice Mining 16
Real-Time Text Analytics 16
Summary 16
References 17
Chapter 2 Information Extraction Using SAS Crawler 19
Introduction to Information Extraction and Organization 19
SAS Crawler 20
SAS Search and Indexing 20
SAS Information Retrieval Studio Interface 20
Web Crawler 22
Breadth First 23
Depth First 24
Web Crawling: Real-World Applications and Examples 24
Understanding Core Component Servers 26
Proxy Server 26
Pipeline Server 27
Component Servers of SAS Search and Indexing 28
Indexing Server 28
Query Server 28

Query Web Server 29
Query Statistics Server 29
SAS Markup Matcher Server 29
Summary 39
References 39
Chapter 3 Importing Textual Data into SAS Text Miner 41
An Introduction to SAS Enterprise Miner and SAS Text Miner 41
Data Types, Roles, and Levels in SAS Text Miner 42
Creating a Data Source in SAS Enterprise Miner 43
Importing Textual Data into SAS 48
Importing Data into SAS Text Miner Using the Text Import Node 49
%TMFILTER Macro 57
Importing XLS and XML Files into SAS Text Miner 58
Managing Text Using SAS Character Functions 62
Summary 67
References 68
Chapter 4 Parsing and Extracting Features 69
Introduction 69
Tokens and Words 70
Lemmatization 70
POS Tags 71
Parsing Tree 71
Text Parsing Node in SAS Text Miner 73
Stemming and Synonyms 73
Identifying Parts of Speech 78
Using Start and Stop Lists 81
Spell Checking 84
Entities 86
Building Custom Entities Using SAS Contextual Extraction Studio 88
Summary 90
References 90
Chapter 5 Data Transformation 93
Introduction 93
Zipf's Law 94
Term-By-Document Matrix 96
Text Filter Node 97
Frequency Weightings 98
Term Weightings 98
Filtering Documents 102
Concept Links 106
Summary 108
References 108

Chapter 6 Clustering and Topic Extraction 111
Introduction 111
What Is Clustering? 111
Singular Value Decomposition and Latent Semantic Indexing 113
Topic Extraction 122
Scoring 130
Summary 130
References 131
Chapter 7 Content Management 133
Introduction 133
Content Categorization 134
Types of Taxonomy 136
Statistical Categorizer 139
Rule-Based Categorizer 141
Comparison of Statistical versus Rule-Based Categorizers 144
Determining Category Membership 145
Concept Extraction 146
Contextual Extraction 150
CLASSIFIER Definition 150
SEQUENCE and PREDICATE_RULE Definitions 155
Automatic Generation of Categorization Rules Using SAS Text Miner 157
Differences between Text Clustering and Content Categorization 159
Summary 160
Appendix 161
References 162
Chapter 8 Sentiment Analysis 163
Introduction 163
Basics of Sentiment Analysis 164
Challenges in Conducting Sentiment Analysis 165
Unsupervised versus Supervised Sentiment Classification 165
SAS Sentiment Analysis Studio Overview 166
Statistical Models in SAS Sentiment Analysis Studio 167
Rule-Based Models in SAS Sentiment Analysis Studio 172
SAS Text Miner and SAS Sentiment Analysis Studio 175
Summary 176
References 177
Case Studies 179
Case Study 1 Text Mining SUGI/SAS Global Forum Paper Abstracts to Reveal Trends 181
Introduction 181
Data 181
Results 189
Trends 190

Summary 194
Instructions for Accessing the Case Study Project 194
Case Study 2 Automatic Detection of Section Membership for SAS Conference Paper Abstract Submissions 197
Introduction 197
Objective 198
Step-by-Step Instructions 198
Summary 208
Case Study 3 Features-based Sentiment Analysis of Customer Reviews 209
Introduction 209
Data 209
Text Mining for Negative App Reviews 210
Text Mining for Positive App Reviews 217
NLP Based Sentiment Analysis 219
Summary 225
Case Study 4 Exploring Injury Data for Root Causal and Association Analysis 227
Introduction 227
Objective 227
Data Description 227
Step-by-Step Instructions 228
Part 1: SAS Text Miner 228
Part 2: SAS Enterprise Content Categorization 234
Summary 238
Case Study 5 Enhancing Predictive Models Using Textual Data 241
Data Description 241
Step-by-Step Instructions 241
Summary 249
Case Study 6 Opinion Mining of Professional Drivers' Feedback 251
Introduction 251
Data 251
Analysis Using SAS® Text Miner 251
Analysis Using the Text Rule-builder Node 258
Summary 272
Case Study 7 Information Organization and Access of Enron Emails to Help Investigation 273
Introduction 273
Objective 273
Step-by-Step Software Instruction with Settings/Properties 274
Summary 281
Case Study 8 Unleashing the Power of Unified Text Analytics to Categorize Call Center Data 283
Introduction 283
Data Description 284

Examining Topics 285
Merging or Splitting Topics 288
Categorizing Content 288
Concept Map Visualization 289
Using PROC DS2 for Deployment DEPLOYMENT 292
Integrating with SAS® Visual Analytics 293
Summary 294
Case Study 9 Evaluating Health Provider Service Performance Using Textual Responses 297
Introduction 297
Summary 311
Index 313



Chapter 1 Introduction to Text Analytics

Overview of Text Analytics 1
Text Mining Using SAS Text Miner 5
Information Retrieval 7
Document Classification 8
Ontology Management 9
Information Extraction 10
Clustering 11
Trend Analysis 12
Enhancing Predictive Models Using Exploratory Text Mining 13
Sentiment Analysis 14
Emerging Directions 15
Handling Big (Text) Data 15
Voice Mining 16
Real-Time Text Analytics 16
Summary 16
References 17

Overview of Text Analytics

Text analytics helps analysts extract meanings, patterns, and structure hidden in unstructured textual data. The information age has led to the development of a wide variety of tools and infrastructure to capture and store massive amounts of textual data. In a 2009 report, the International Data Corporation (IDC) estimated that approximately 80% percent of the data in an organization is text based. It is not practical for any individual (or group of individuals) to process huge textual data and extract meanings, sentiments, or patterns out of the data. A paper written by Hans Peter Luhn, titled "The Automatic Creation of Literature Abstracts," is perhaps one of the earliest research projects conducted on text analytics. Luhn writes about applying machine methods to automatically generate an abstract for a document. In a traditional sense, the term "text mining" is used for automated machine learning and statistical methods that encompass a bag-of-words approach. This approach is typically used to examine content collections versus assessing individual documents. Over time, the term "text analytics" has evolved to encompass a loosely integrated framework by borrowing techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management.

Text analytics applications are popular in the business environment. These applications produce some of the most innovative and deeply insightful results. Text analytics is being implemented in many industries. There are new types of applications every day. In recent years, text analytics has been heavily used for discovering trends

in textual data. Using social media data, text analytics has been used for crime prevention and fraud detection. Hospitals are using text analytics to improve patient outcomes and provide better care. Scientists in the pharmaceutical industry are using this technology to mine biomedical literature to discover new drugs.

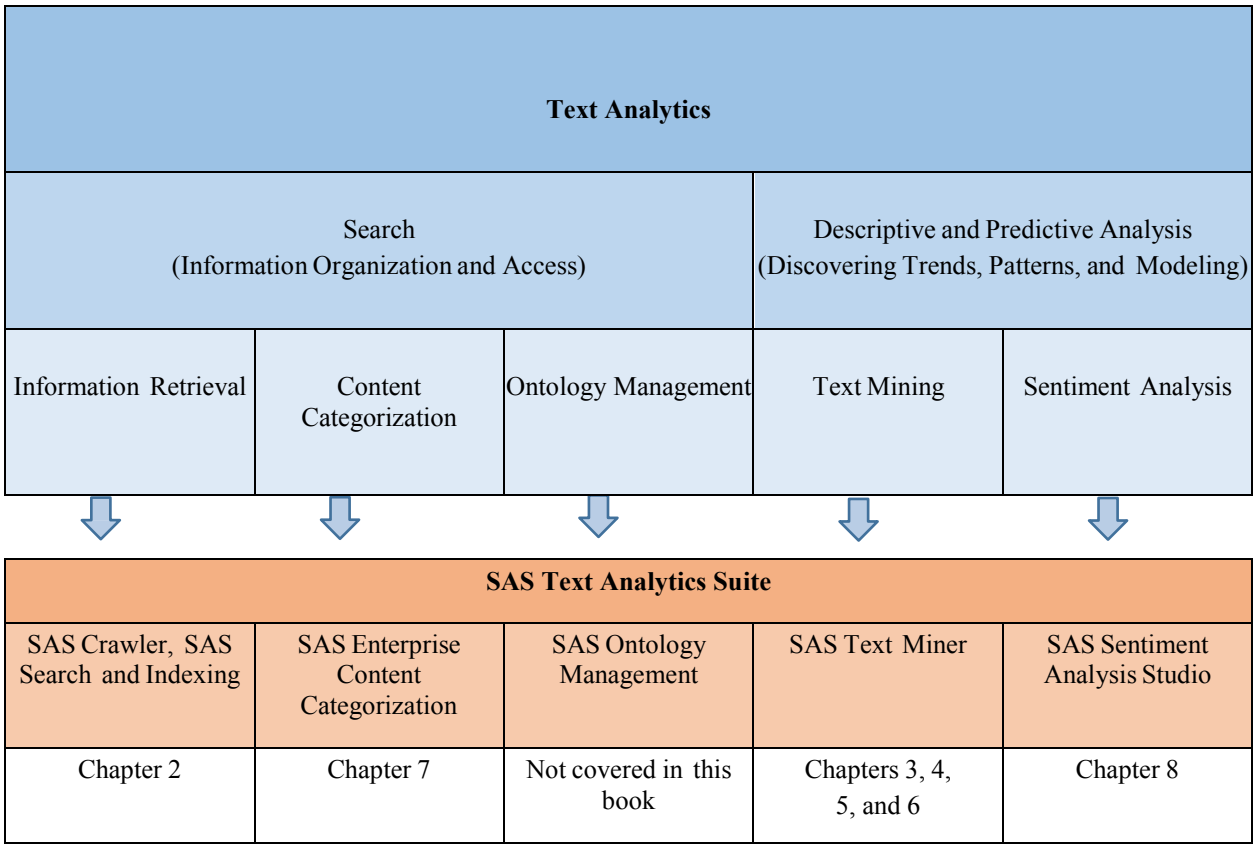
Text analytics incorporates tools and techniques that are used to derive insights from unstructured data. These techniques can be broadly classified as the following:

- information retrieval
- exploratory analysis
- concept extraction
- summarization
- categorization
- sentiment analysis
- content management
- ontology management

In these techniques, exploratory analysis, summarization, and categorization are in the domain of text mining. Exploratory analysis includes techniques such as topic extraction, cluster analysis, etc. The term "text analytics" is somewhat synonymous with "text mining" (or "text data mining"). Text mining can be best conceptualized as a subset of text analytics that is focused on applying data mining techniques in the domain of textual information using NLP and machine learning. Text mining considers only syntax (the study of structural relationships between words). It does not deal with phonetics, pragmatics, and discourse.

Sentiment analysis can be treated as classification analysis. Therefore, it is considered predictive text mining. At a high level, the application areas of these techniques divide the text analytics market into two areas: search and descriptive and predictive analytics. (See Display 1.1.) Search includes numerous information retrieval techniques, whereas descriptive and predictive analytics include text mining and sentiment analysis.

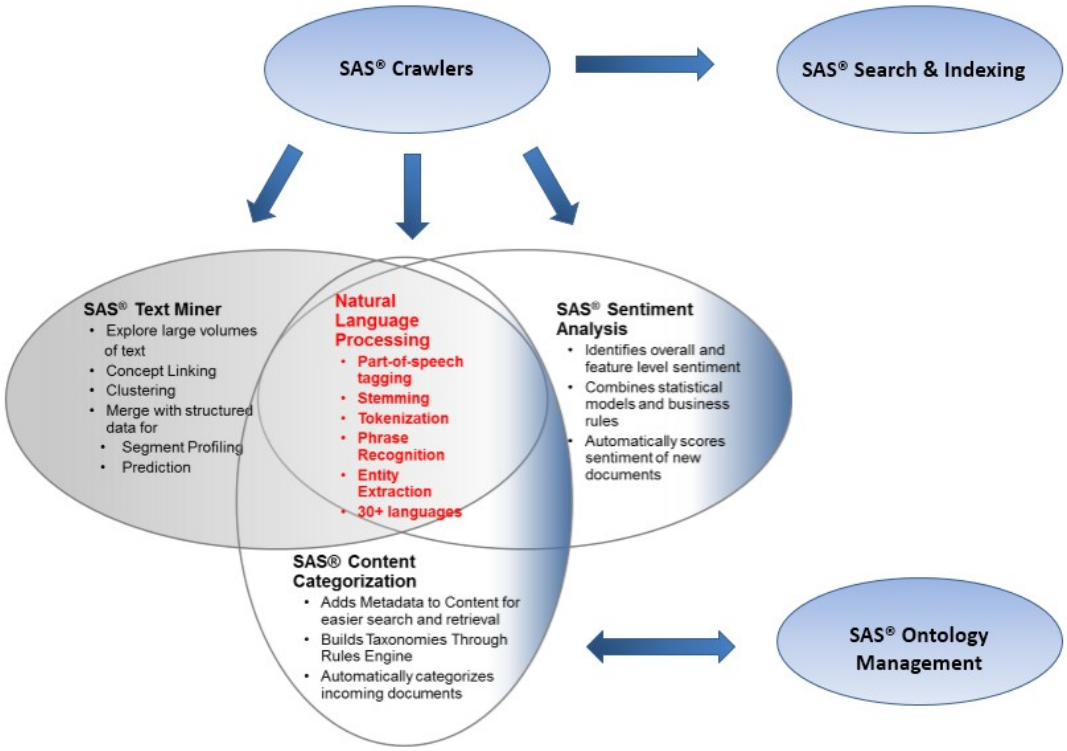
Display 1.1: High-Level Classification of Text Analytics Market and Corresponding SAS Tools



SAS has multiple tools to address a variety of text analytics techniques for a range of business applications. Display 1.1 shows the SAS tools that address different areas of text analytics. In a typical situation, you might need to use more than one tool for solving a text analytics problem. However, there is some overlap in the underlying features that some of these tools have to offer. Display 1.2 provides an integrated view of SAS Text Analytics tools. It shows, at a high level, how they are organized in terms of functionality and scope. SAS Crawler can extract content from the web, file systems, or feeds, and then send it as input to SAS Text Miner, SAS Sentiment Analysis Studio, or SAS Content Categorization. These tools are capable of sending content to the indexing server where information is indexed. The query server enables you to enter search queries and retrieve relevant information from the indexed content.

SAS Text Miner, SAS Sentiment Analysis Studio, and SAS Content Categorization form the core of the SAS Text Analytics tools arsenal for analyzing text data. NLP features such as tokenization, parts-of-speech recognition, stemming, noun group detection, and entity extraction are common among these tools. However, each of these tools has unique capabilities that differentiate them individually from the others. In the following section, the functionality and usefulness of these tools are explained in detail.

Display 1.2: SAS Text Analytics Tools: An Integrated Overview



The following paragraphs briefly describe each tool from the SAS Text Analytics suite as presented in Display 1.2:

- **SAS Crawler, SAS Search and Indexing** – Useful for extracting textual content from the web or from documents stored locally in an organized way. For example, you can download news articles from websites and use SAS Text Miner to conduct an exploratory analysis, such as extracting key topics or themes from the news articles. You can build indexes and submit queries on indexed documents through a dedicated query interface.
- **SAS Ontology Management** – Useful for integrating existing document repositories in enterprises and identifying relationships between them. This tool can help subject matter experts in a knowledge domain create ontologies and establish hierarchical relationships of semantic terms to enhance the process of search and retrieval on the document repositories.

  Note: SAS Ontology Management is not discussed in this book because we primarily focus on areas where the majority of current business applications are relevant for textual data.
- **SAS Content Categorization** – Useful for classifying a document collection into a structured hierarchy of categories and subcategories called taxonomy. In addition to categorizing documents, it can be used to extract facts from them. For example, news articles can be classified into a predefined set of categories such as politics, sports, business, financial, etc. Factual information such as events, places, names of people, dates, monetary values, etc., can be easily retrieved using this tool.
- **SAS Text Miner** – Useful for extracting the underlying key topics or themes in textual documents. This tool offers the capability to group similar documents—called clusters—based on terms and their frequency of occurrence in the corpus of documents and within each document. It provides a feature called "concept linking" to explore the relationships between terms and their strength of association.

  For example, textual transcripts from a customer call center can be fed into this tool to automatically cluster the transcripts. Each cluster has a higher likelihood of having similar problems reported by customers. The specifics of the problems can be understood by reviewing the descriptive terms explaining each of the clusters. A pictorial representation of these problems and the associated terms,

events, or people can be viewed through concept linking, which shows how strongly an event can be related to a problem.
SAS Text Miner enables the user to define custom topics or themes. Documents can be scored based on the presence of the custom topics. In the presence of a target variable, supervised classification or prediction models can be built using SAS Text Miner. The predictions of a prediction model with numerical inputs can be improved using topics, clusters, or rules that can be extracted from textual comments using SAS Text Miner.

- **SAS Sentiment Analysis** – Useful for identifying the sentiment toward an entity in a document or the overall sentiment toward the entire document. An entity can be anything, such as a product, an attribute of a product, brand, person, group, or even an organization. The sentiment evaluated is classified as positive or negative or neutral or unclassified. If there are no terms associated with an entity or the entire document that reflect the sentiment, it is tagged "unclassified."

  Sentiment analysis is generally applied to a class of textual information such as customers' reviews on products, brands, organizations, etc., or to responses to public events such as presidential elections. This type of information is largely available on social media sites such as Facebook, Twitter, YouTube, etc.

Text Mining Using SAS Text Miner

A typical predictive data mining problem deals with data in numerical form. However, textual data is typically available only in a readable document form. Forms could be e-mails, user comments, corporate reports, news articles, web pages, etc. Text mining attempts to first derive a quantitative representation of documents. Once the text is transformed into a set of numbers that adequately capture the patterns in the textual data, any traditional statistical or forecasting model or data mining algorithm can be used on the numbers for generating insights or for predictive modeling.

A typical text mining project involves the following tasks:

1. **Data Collection:** The first step in any text mining research project is to collect the textual data required for analysis.
2. **Text Parsing and Transformation:** The next step is to extract, clean, and create a dictionary of words from the documents using NLP. This includes identifying sentences, determining parts of speech, and stemming words. This step involves parsing the extracted words to identify entities, removing stop words, and spell-checking. In addition to extracting words from documents, variables associated with the text such as date, author, gender, category, etc., are retrieved.

The most important task after parsing is text transformation. This step deals with the numerical representation of the text using linear algebra-based methods, such as latent semantic analysis (LSA), latent semantic indexing (LSI), and vector space model. This exercise results in the creation of a term-by-document matrix (a spreadsheet or flat-like numeric representation of textual data as shown in Table 1.1). The dimensions of the matrix are determined by the number of documents and the number of terms in the collection. This step might involve dimension reduction of the term-by-document matrix using singular value decomposition (SVD).

Consider a collection of three reviews (documents) of a book as provided below: Document 1: I am an avid fan of this sport book. I love this book.

Document 2: This book is a must for athletes and sportsmen. Document 3: This book tells how to command the sport.

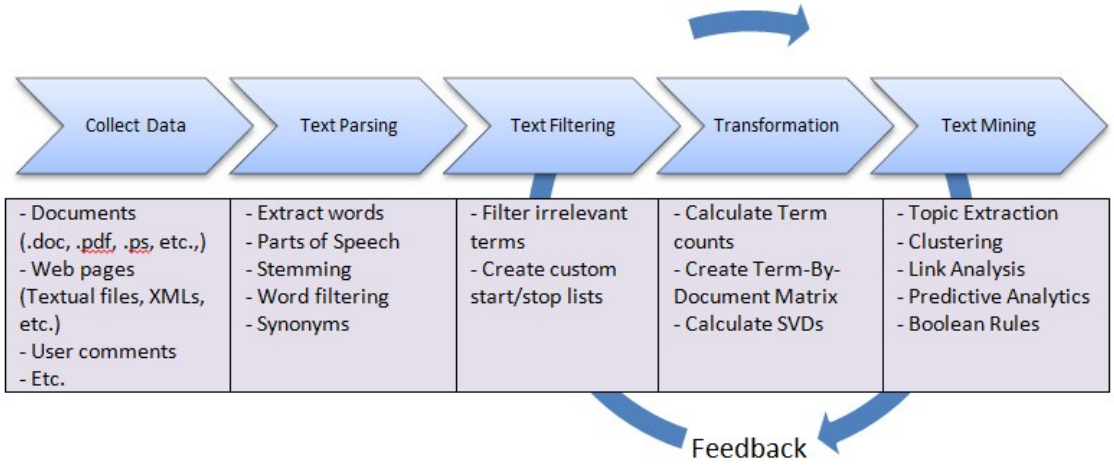
Parsing this document collection generates the following term-by-document matrix in Table 1.1:

Table 1.1: Term-By-Document Matrix

Term/Document	Document 1	Document 2	Document 3
the	0	0	1
I	2	0	0
am	1	0	0
avid	1	0	0
fan	1	0	0
this	2	1	1
book	2	1	1
athletes	0	1	0
sportsmen	0	1	0
sport	1	0	1
command	0	0	1
tells	0	0	1
for	0	1	0
how	0	0	1
love	1	0	0
an	1	0	0
of	1	0	0
is	0	1	0
a	0	1	0
must	0	1	0
and	0	1	0
to	0	0	1

3. **Text Filtering:** In a corpus of several thousands of documents, you will likely have many terms that are irrelevant to either differentiating documents from each other or to summarizing the documents. You will have to manually browse through the terms to eliminate irrelevant terms. This is often one of the most time-consuming and subjective tasks in all of the text mining steps. It requires a fair amount of subject matter knowledge (or domain expertise). In addition to term filtering, documents irrelevant to the analysis are searched using keywords. Documents are filtered if they do not contain some of the terms or filtered based on one of the other document variables such as date, category, etc. Term filtering or document filtering alters the term-by-document matrix. As shown in Table 1.1, the term-by-document matrix contains the frequency of the occurrence of the term in the document as the value of each cell. Instead, you could have a log of the frequency or just a 1 or 0 value indicating the presence of the term in a document as the value for each cell. From this frequency matrix, a weighted term-by-document matrix is generated using various term-weighting techniques.
4. **Text Mining:** This step involves applying traditional data mining algorithms such as clustering, classification, association analysis, and link analysis. As shown in Display 1.3, text mining is an iterative process, which involves repeating the analysis using different settings and including or excluding terms for better results. The outcome of this step can be clusters of documents, lists of single-term or multi-term topics, or rules that answer a classification problem. Each of these steps is discussed in detail in Chapter 3 to Chapter 7.

Display 1.3: Text Mining Process Flow

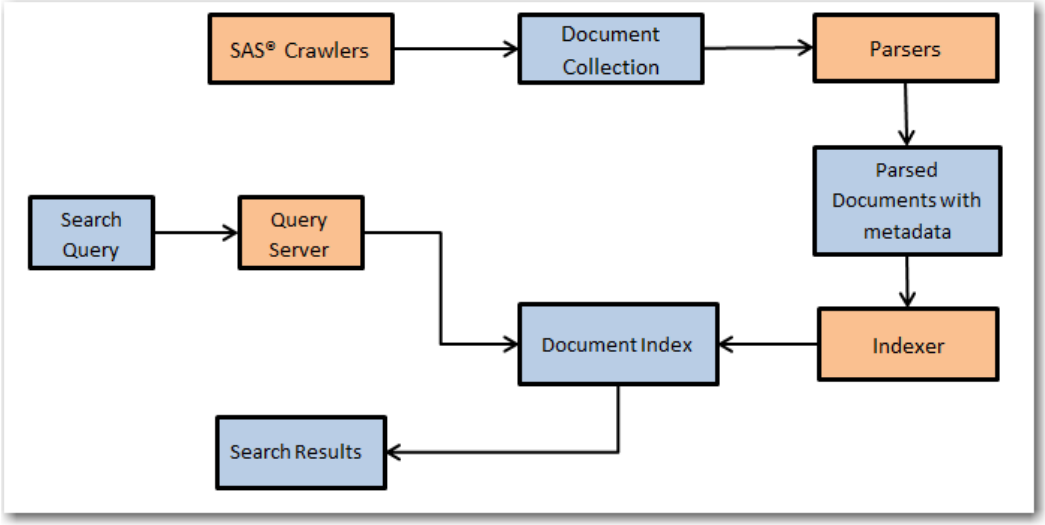


Information Retrieval

Information retrieval, commonly known as IR, is the study of searching and retrieving a subset of documents from a universe of document collections in response to a search query. The documents are often unstructured in nature and contain vast amounts of textual data. The documents retrieved should be relevant to the information needs of the user who performed the search query. Several applications of the IR process have evolved in the past decade. One of the most ubiquitously known is searching for information on the World Wide Web. There are many search engines such as Google, Bing, and Yahoo facilitating this process using a variety of advanced methods.

Most of the online digital libraries enable its users to search through their catalogs based on IR techniques. Many organizations enhance their websites with search capabilities to find documents, articles, and files of interest using keywords in the search queries. For example, the United States Patent and Trademark Office provides several ways of searching its database of patents and trademarks that it has made available to the public. In general, an IR system's efficiency lies in its ability to match a user's query with the most relevant documents in a corpus. To make the IR process more efficient, documents are required to be organized, indexed, and tagged with metadata based on the original content of the documents. SAS Crawler is capable of pulling information from a wide variety of data sources. Documents are then processed by parsers to create various fields such as title, ID, URL, etc., which form the metadata of the documents. (See Display 1.4.) SAS Search and Indexing enables you to build indexes from these documents. Users can submit search queries on the indexes to retrieve information most relevant to the query terms. The metadata fields generated by the parsers can be used in the indexes to enable various types of functionality for querying.

Display 1.4: Overview of the IR Process with SAS Search and Indexing



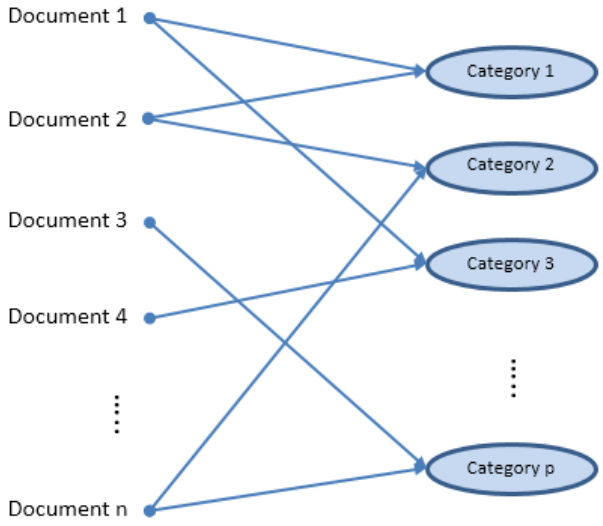
Document Classification

Document classification is the process of finding commonalities in the documents in a corpus and grouping them into predetermined labels (supervised learning) based on the topical themes exhibited by the documents. Similar to the IR process, document classification (or text categorization) is an important aspect of text analytics and has numerous applications.

Some of the common applications of document classification are e-mail forwarding and spam detection, call center routing, and news articles categorization. It is not necessary that documents be assigned to mutually exclusive categories. Any restrictive approach to do so might prove to be an inefficient way of representing the information. In reality, a document can exhibit multiple themes, and it might not be possible to restrict them to only one category. SAS Text Miner contains the text topic feature, which is capable of handling these situations. It assigns a document to more than one category if needed. (See Display 1.5.) Restricting documents to only one category might be difficult for large documents, which have a greater chance of containing multiple topics or features. Topics or categories can be either automatically generated by SAS Text Miner or predefined manually based on the knowledge of the document content.

In cases where a document should be restricted to only one category, text clustering is usually a better approach instead of extracting text topics. For example, an analyst could gain an understanding of a collection of classified ads when the clustering algorithm reveals the collection actually consists of categories such as Car Sales, Real Estate, and Employment Opportunities.

Display 1.5: Text Categorization Involving Multiple Categories per Document



SAS Content Categorization helps automatically categorize multilingual content available in huge volumes that is acquired or generated or that exists in an information repository. It has the capability to parse, analyze, and extract content such as entities, facts, and events in a classification hierarchy. Document classification can be achieved using either SAS Content Categorization or SAS Text Miner. However, there are some fundamental differences between these two tools. The text topic extraction feature in SAS Text Miner completely relies on the quantification of terms (frequency of occurrences) and the derived weights of the terms for each document using advanced statistical methods such as SVD.

On the other hand, SAS Content Categorization is broadly based on statistical and rule-based models. The statistical categorizer works similar to the text topic feature in SAS Text Miner. The statistical categorizer is used as a first step to automatically classify documents. Because you cannot really see the rules behind the classification methodology, it is called a black box model. In rule-based models, you can choose to use linguistic rules by listing the commonly occurring terms most relevant for a category. You can assign weights to these terms based on their importance. Boolean rule-based models use Boolean operators such as AND, OR, NOT, etc., to specify the conditions with which terms should occur within documents. This tool has additional custom-built operators to assess positional characteristics such as whether the distance between the two terms is within a distance of *n* terms, whether specific terms are found in a given sequence, etc. There is no limit on how complex these rules can be (for example, you can use nested Boolean rules).

Ontology Management

Ontology is a study about how entities can be grouped and related within a hierarchy. Entities can be subdivided based on distinctive and commonly occurring features. SAS Ontology Management enables you to create relationships between pre-existing taxonomies built for various silos or departments. The subject matter knowledge about the purpose and meaning can be used to create rules for building information search and retrieval systems. By identifying relationships in an evolutionary method and making the related content available, queries return relevant, comprehensive, and accurate answers. SAS Ontology Management offers the ability to build semantic repositories and manage company-wide thesauri and vocabularies and to build relationships between them.

To explain its application, consider the simple use case of an online media house named ABC. (The name was changed to maintain anonymity.) ABC uses SAS Ontology Management. ABC collects a lot of topics over a period of time. It stores each of these topics, along with metadata (properties), including links to images and

textual descriptions. SAS Ontology Management helps ABC store relationships between the related topics. ABC regularly queries its ontology to generate a web page for each topic, showing the description, images, related topics, and other metadata that it might have selected to show. (See Display 1.6.) ABC uploads the information from SAS Ontology Management to SAS Content Categorization, and then tags news articles with topics that appear in the articles using rules that it's created. All tagged articles are included in a list on the topic pages.

Display 1.6: Example Application of SAS Ontology Management from an Online Media Website

Information Extraction

In a relational database, data is stored in tables within rows and columns. A structured query on the database can help you retrieve the information required if the names of tables and columns are known. However, in the case of unstructured data, it is not easy to extract specific portions of information from the text because there is no fixed reference to identify the location of the data. Unstructured data can contain small fragments of information that might be of specific interest, based on the context of information and the purpose of analysis. Information extraction can be considered the process of extracting those fragments of data such as the names of people, organizations, places, addresses, dates, times, etc., from documents.

Information extraction might yield different results depending on the purpose of the process and the elements of the textual data. Elements of the textual data within the documents play a key role in defining the scope of information extraction. These elements are tokens, terms, and separators. A document consists of a set of tokens. A token can be considered a series of characters without any separators. A separator can be a special character, such as a blank space or a punctuation mark. A term can be a defined as a token with specific semantic purpose in a given language.

There are several types of information extraction that can be performed on textual data.

- Token extraction
- Term extraction or term parsing
- Concept extraction
- Entity extraction

- Atomic fact extraction
- Complex fact extraction

Concept extraction involves identifying nouns and noun phrases. Entity extraction can be defined as the process of associating nouns with entities. For example, although the word "white" is a noun in English and represents a color, the occurrence of "Mr. White" in a document can be identified as a person, not a color. Similarly, the phrase "White House" can be attributed to a specific location (the official residence and principal workplace of the president of the United States), rather than as a description of the color of paint used for the exterior of a house. Atomic fact extraction is the process of retrieving fact-based information based on the association of nouns with verbs in the content (i.e., subjects with actions).

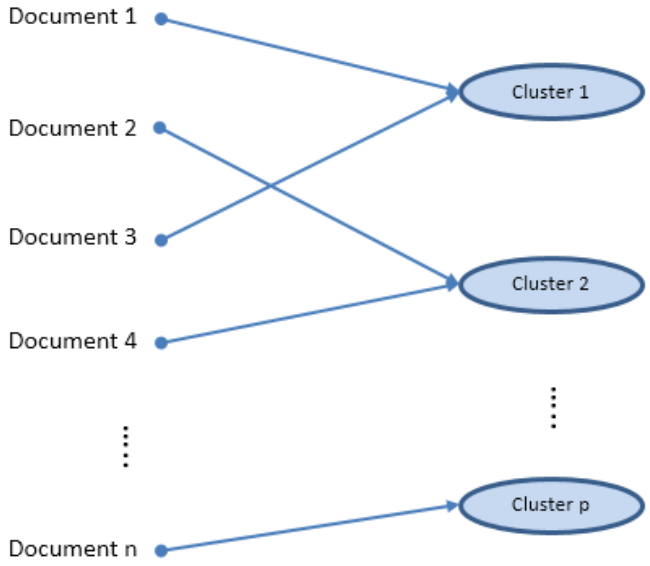
Clustering

Cluster analysis is a popular technique used by data analysts in numerous business applications. Clustering partitions records in a data set into groups so that the subjects within a group are similar and the subjects between the groups are dissimilar. The goal of cluster analysis is to derive clusters that have value with respect to the problem being addressed, but this goal is not always achieved. As a result, there are many competing clustering algorithms. The analyst often compares the quality of derived clusters, and then selects the method that produces the most useful groups. The clustering process arranges documents into nonoverlapping groups. (See Display 1.7.) Each document can fall into more than one topic area after classification. This is the key difference between clustering and the general text classification processes, although clustering provides a solution to text classification when groups must be mutually exclusive, as in the classified ads example.

In the context of text mining, clustering divides the document collection into mutually exclusive groups based on the presence of similar themes. In most business applications involving large amounts of textual data, it is often difficult to profile each cluster by manually reading and considering all of the text in a cluster. Instead, the theme of a cluster is identified using a set of descriptive terms that each cluster contains. This vector of terms represents the weights measuring how the document fits into each cluster. Themes help in better understanding the customer, concepts, or events. The number of clusters that are identified can be controlled by the analyst.

The algorithm can generate clusters based on the relative positioning of documents in the vector space. The cluster configuration is altered by a start and stop list.

Display 1.7: Text Clustering Process Assigning Each Document to Only One Cluster



For example, consider the comments made by different patients about the best thing that they liked about the hospital that they visited.

1. Friendliness of the doctor and staff.
2. Service at the eye clinic was fast.
3. The doctor and other people were very, very friendly.
4. Waiting time has been excellent and staff has been very helpful.
5. The way the treatment was done.
6. No hassles in scheduling an appointment.
7. Speed of the service.
8. The way I was treated and my results.
9. No waiting time, results were returned fast, and great treatment.

The clustering results from text mining the comments come out similar to the ones shown in Table 1.2. Each cluster can be described by a set of terms, which reveal, to a certain extent, the theme of the cluster. This type of analysis helps businesses understand the collection as a whole, and it can assist in correctly classifying customers based on common topics in customer complaints or responses.

Table 1.2: Clustering Results from Text Mining

Cluster No.	Comment	Key Words
1	1, 3, 4	doctor, staff, friendly, helpful
2	5, 6, 8	treatment, results, time, schedule
3	2, 7	service, clinic, fast

The derivation of key words is accomplished using a weighting strategy, where words are assigned a weight using features of LSI. Text mining software products can differ in how the keywords are identified, resulting from different choices for competing weighting schemes.

SAS Text Miner uses two types of clustering algorithms: expectation maximization and hierarchical clustering. The result of cluster analysis is identifying cluster membership for each document in the collection. The exact nature of the two algorithms is discussed in detail in "Chapter 6 Clustering and Topic Extraction."

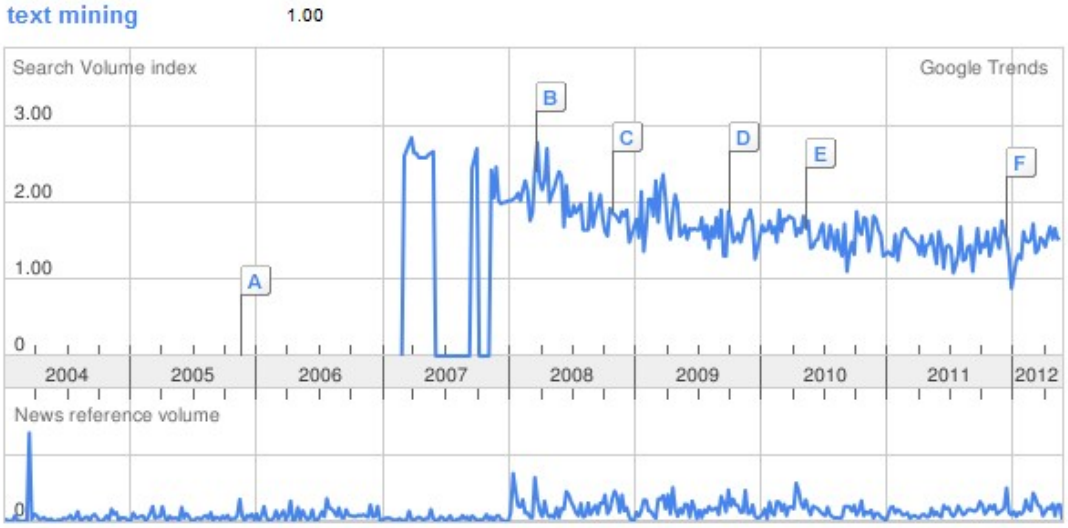
Trend Analysis

In recent years, text mining has been used to discover trends in textual data. Given a set of documents with a time stamp, text mining can be used to identify trends of different topics that exist in the text. Trend analysis has been widely applied in tracking the trends in research from scientific literature. It has also been widely applied in summarizing events from news articles. In this type of analysis, a topic or theme is first defined using a set of words and phrases. Presence of the words across the documents over a period of time represents the trend for this topic. To effectively track the trends, it is very important to include all related terms to (or synonyms of) these words.

For example, text mining is used to predict the movements of stock prices based on news articles and corporate reports. Evangelopoulos and Woodfield (2009) show how movie themes trend over time, with male movies dominating the World War II years and female movies dominating the Age of Aquarius. As another example, mining social networks to identify trends is currently a very hot application area. Google Trends, a publicly available website, provides a facility to identify the trends in your favorite topics over a period of time. Social networking sites such as Twitter and blogs are great sources to identify trends. Here is a screenshot of the trend for the topic "text mining" from Google Trends. It is clearly evident that the growth in search traffic and online

posts for the term "text mining" peaked after 2007. This is when the popularity of text mining applications in the business world jump-started.

Display 1.8: Trend for the Term "text mining" from Google Trends



The concept linking functionality in SAS Text Miner helps in identifying co-occurring terms (themes), and it reveals the strength of association between terms. With temporal data, the occurrence of terms from concept links can be used to understand the trend (or pattern) of the theme across the time frame. Case Study 1 explains how this technique was applied to reveal the trend of different topics that have been presented at SAS Global Forum since 1976.

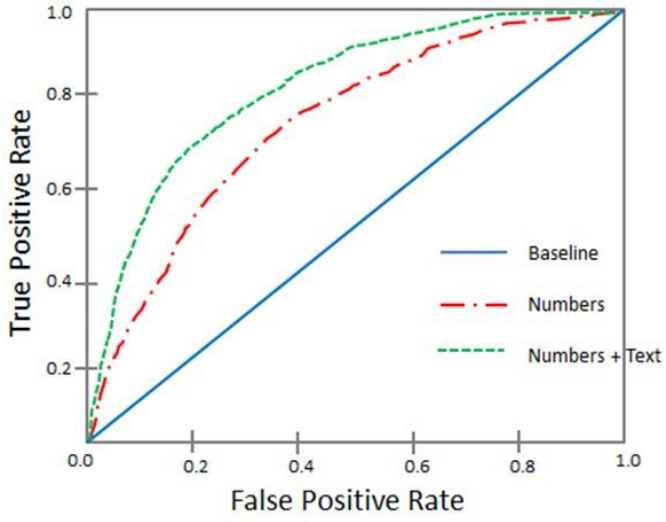
Enhancing Predictive Models Using Exploratory Text Mining

Although text mining customer responses can reveal valuable insights about a customer, plugging the results from text mining into a typical data mining model can often significantly improve the predictive power of the model. Organizations often want to use customer responses captured in the form of text via e-mails, customer survey questionnaires, and feedback on websites for building better predictive models. One way of doing this is to first apply text mining to reveal groups (or clusters) of customers with similar responses or feedback. This cluster membership information about each customer can then be used as an input variable to augment the data mining model. With this additional information, the accuracy of a predictive model can improve significantly.

For example, a large hospital conducted a post-treatment survey to identify the factors that influence a patient's likelihood to recommend the hospital. By using the text mining results from the survey, the hospital was able to identify factors that showed an impact on patient satisfaction, which was not measured directly through the survey questions. Researchers observed a strong correlation between the theme of the cluster and the ratings given by the patient for the likelihood for the patient to recommend the hospital.

In a similar exercise, a large travel stop company observed significant improvement in predicting models by using customers' textual responses and numerical responses from a survey. Display 1.9 shows an example receiver operating characteristic (ROC) curve of the models with and without textual comments. The ROC curve shows the performance of a binary classification model. The larger the area under the curve, the better the model performance. The square-dashed curve (green), which is an effect of including results from textual responses, has a larger area under the curve compared to the long-dashed-dotted curve (red), which represents the model with numerical inputs alone.

Display 1.9: ROC Chart of Models With and Without Textual Comments



With the widespread adoption by consumers of social media, a lot of data about any prospect or customer is often available on the web. If businesses can cull and use this information, they can often generate better predictions of consumer behavior. For example, credit card companies can track customers' posts on Twitter and other social media sites, and then use that information in credit scoring models. However, there are challenges to using text mining models and predictive models together because it can be difficult to get textual data for every member in the data mining model for the same time period.

Sentiment Analysis

The field of sentiment analysis deals with categorization (or classification) of opinions expressed in textual documents. Often these text units are classified into multiple categories such as positive, negative, or neutral, based on the valence of the opinion expressed in the units. Organizations frequently conduct surveys and focus group studies to track a customer's perception of their products and services. However, these methods are time-consuming and expensive and cannot work in real time because the process of analyzing text is done manually by experts. Using sentiment analysis, an organization can identify and extract a customers' attitude, sentiment, or emotions toward a product or service. This is a more advanced application of text analytics that uses NLP to capture the polarity of the text: positive, negative, neutral, or mixed. With the advent of social networking sites, organizations can capture enormous amounts of customers' responses instantly. This gives real-time awareness to customer feedback and enables organizations to react fast. Sentiment analysis works on opinionated text while text mining is good for factual text. Sentiment analysis, in combination with other text analytics and data mining techniques, can reveal very valuable insights.

Sentiment analysis tools available from SAS offer a very comprehensive solution to capture, analyze, and report customer sentiments. The polarity of the document is measured at the overall document level and at the specific feature level.

Here is an example showing the results of a sentiment analysis on a customer's review of a new TV brand:

The TV is **wonderful**. **Great size**, **great picture**, **easy interface**. It makes a **cute** little song when you **boot** it up and when you shut it off. I just want to point out that the **43" does not** in fact play videos from the USB .This is really **annoying** because that was one of the major perks I wanted from a new TV. Looking at the product description now, I realize that the feature list applies to the X758 series as a whole, and that each model's capabilities are listed below. Kind of a **dumb** oversight on my part, but it's equally **stupid** to put a description that does not apply on the listing for a very specific model.

In the previous text, green color represents positive tone, red color represents negative tone, and product features and model names are highlighted in blue and brown, respectively. In addition to extracting positive and negative sentiments, names of product models and their features are identified. This level of identification helps identify the sentiment of the overall document and tracks the sentiment at a product-feature level, including the characteristics and sub-attributes of features.

"Chapter 8 Sentiment Analysis" discusses sentiment analysis using SAS Sentiment Analysis Studio through an example of tracking sentiment in feedback comments from customers of a leading travel stop company.

Emerging Directions

Although the number of applications in text analytics has grown in recent years, there continues to be a high level of excitement about text analytics applications and research. For example, many of the papers presented at the Analytics 2011 Conference and SAS Global Forum 2013 were based on different areas of text analytics. In a way, the excitement about text analytics reminds us of the time when data mining and predictive modeling was taking off at business and academic conferences in the late 90s and early 2000s. The text analytics domain is constantly evolving with new techniques and new applications. Text analytics solutions are being adopted at the enterprise level and are being used to operationalize and integrate the voice of the customer into business processes and strategies. Many enterprise solution vendors are integrating some form of text analytics technology into their product line. This is evident from the rate of acquisitions in this industry. One of the key reasons that is fueling the growth of the field of text analytics is the increasing amount of unstructured data that is being generated on the web. It is expected that 90% of the digital content in the next 10 years will be unstructured data.

Companies across all industries are looking for solutions to handle the massive amounts of data, also popularly known as big data. Data is generated constantly from various sources such as transaction systems, social media interactions, clickstream data from the web, real-time data captured from sensors, geospatial information, and so on. As we have already pointed out, by some estimates, 80% of an organization's current data is not numeric!

This means that the variety of data that constitutes big data is unstructured. This unstructured data comes in various formats: text, audio, video, images, and more. The constant streaming of data on social media outlets and websites means the velocity at which data is being generated is very high. The variety and the velocity of the data, together with the volume (the massive amounts) of the data organizations need to collect, manage, and process in real time, creates a challenging task. As a result, the three emerging applications for text analytics will likely address the following:

1. Handling big (text) data
2. Voice mining
3. Real-time text analytics

Handling Big (Text) Data

Based on the industry's current estimations, unstructured data will occupy 90% of the data by volume in the entire digital space over the next decade. This prediction certainly adds a lot of pressure to IT departments, which already face challenges in terms of handling text data for analytical processes. With innovative hardware architecture, analytics application architecture, and data processing methodologies, high-performance computing technology can handle the complexity of big data. SAS High-Performance Text Mining helps you decrease the computational time required for processing and analyzing bulk volumes of text data significantly. It uses the combined power of multithreading, a distributed grid of computing resources, and in-memory processing. Using sophisticated implementation methodologies such as symmetric multiprocessing (SMP) and massively parallel processing (MPP), data is distributed across computing nodes. Instructions are allowed to execute separately on each node. The results from each node are combined to produce meaningful results. This is a cost-effective and highly scalable technology that addresses the challenges posed by the three **V**s. (variety, velocity, and volume) of big data.

SAS High-Performance Text Mining consists of three components for processing very large unstructured data. These components are document parsing, term handling, and text processing control. In the document parsing component, several NLP techniques (such as parts-of-speech tagging, stemming, etc.) are applied to the input text to derive meaningful information. The term handling component accumulates (corrects misspelled terms using a synonyms list), filters (removes terms based on a start or stop list and term frequency), and assigns weights to terms. The text processing control component manages the intermediate results and the inputs and outputs generated by the document parsing and term handling components. It helps generate the term-by-document matrix in a condensed form. The term-by-document matrix is then summarized using the SVD method, which produces statistical representations of text documents. These SVD scores can be later included as numeric inputs to different types of models such as cluster or predictive models.

Voice Mining

Customer feedback is collected in many forms—text, audio, and video—and through various sources—surveys, e-mail, call center, social media, etc. Although the technology for analyzing videos is still under research and development, analyzing audio (also called voice mining) is gaining momentum. Call centers (or contact centers) predominantly use speech analytics to analyze the audio signal for information that can help improve call center effectiveness and efficiency. Speech analytics software is used to review, monitor, and categorize audio content. Some tools use phonetic index search techniques that automatically transform the audio signal into a sequence of phonemes (or sounds) for interpreting the audio signal and segmenting the feedback using trigger terms such as "cancel," "renew," "open account," etc. Each segment is then analyzed by listening to each audio file manually, which is daunting, time-intensive, and nonpredictive. As a result, analytical systems that combine data mining methods and linguistics techniques are being developed to quickly determine what is most likely to happen next (such as a customer's likelihood to cancel or close the account). In this type of analysis, metadata from each voice call, such as call length, emotion, stress detection, number of transfers, etc., that is captured by these systems can reveal valuable insights.

Real-Time Text Analytics

Another key emerging focus area that is being observed in text analytics technology development is real-time text analytics. Most of the applications of real-time text analytics are addressing data that is streaming continuously on social media. Monitoring public activity on social media is now a business necessity. For example, companies want to track topics about their brands that are trending on Twitter for real-time ad placement. They want to be informed instantly when their customers post something negative about their brand on the Internet. Less companies want to track news feeds and blog posts for financial reasons. Government agencies are relying on real-time text analytics that collect data from innumerate sources on the web to learn about and predict medical epidemics, terrorist attacks, and other criminal actions. However, real time can mean different things in different contexts. For companies involved in financial trading by tracking current events and news feeds, real time could mean milliseconds. For companies tracking customer satisfaction or monitoring brand reputation by collecting customer feedback, real time could mean hourly. For every business, it is of the utmost importance to react instantly before something undesirable occurs.

The future of text analytics will surely include the next generation of tools and techniques with increased usefulness for textual data collection, summarization, visualization, and modeling. Chances are these tools will become staples of the business intelligence (BI) suite of products in the future. Just as SAS Rapid Predictive Modeler today can be used by business analysts without any help from trained statisticians and modelers, so will be some of the future text analytics tools. Other futuristic trends and applications of text analytics are discussed by Berry and Kogan (2010).

Summary

Including textual data in data analysis has changed the analytics landscape over the last few decades. You have witnessed how traditional machine learning and statistical methods to learn unknown patterns in text data are now replaced with much more advanced methods combining NLP and linguistics. Text mining (based on a traditional bag-of-words approach) has evolved into a much broader area (called text analytics). Text analytics

is regarded as a loosely integrated set of tools and methods developed to retrieve, cleanse, extract, organize, analyze, and interpret information from a wide range of data sources. Several techniques have evolved, with each focused to answer a specific business problem based on textual data. Feature extraction, opinion mining, document classification, information extraction, indexing, searching, etc., are some of the techniques that we have dealt with in great detail in this chapter. Tools such as SAS Text Miner, SAS Sentiment Analysis Studio, SAS Content Categorization, SAS Crawler, and SAS Search and Indexing are mapped to various analysis methods. This information helps you distinguish and differentiate the specific functionalities and features that each of these tools has to offer while appreciating the fact that some of them share common features.

In the following chapters, we use SAS Text Analytics tools (except SAS Ontology Management, which is not discussed in this book) to address each methodology discussed in this chapter. Chapters are organized in a logical sequence to help you understand the end-to-end processes involved in a typical text analysis exercise. In Chapter 2, we introduce methods to extract information from various document sources using SAS Crawler. We show you how to deal with the painstaking tasks of cleansing, collecting, transforming, and organizing the unstructured text into a semi-structured format to feed that information into other SAS Text Analytics tools. As you progress through the chapters, you will get acquainted with SAS Text Analytics tools and methodologies that will help you adapt them at your organization.

References

Albright, R., Bieringer, A., Cox, J., and Zhao, Z. 2013. "Text Mine Your Big Data: What High Performance Really Means". Cary, NC: SAS Institute Inc. Available at: http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/text-mine-your-big-data-106554.pdf

Berry, M.W., and Kogan, J. Eds. 2010. *Text Mining: Applications and Theory.* Chichester, United Kingdom: John Wiley & Sons.

Dale, R., Moisl, H. and Somers, H. 2000. Handbook of Natural Language Processing. New York: Marcel Dekker.

Dorre, J. Gerstl, P., and Seiffert, R. 1999. "Text Mining: Finding Nuggets in Mountains of Textual Data".

KDD-99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, New York: Association for Computing Machinery, 398-401.

Evangelopoulos, N., and Woodfield, T. 2009. "Understanding Latent Semantics in Textual Data". M2009 12th Annual Data Mining Conference, Las Vegas, NV.

Feldman, R. 2004. "Text Analytics: Theory and Practice". *ACM Thirteenth Conference on Information and Knowledge Management (CIKM) CIKM and Workshops 2004*. Available at: http://web.archive.org/web/20041204224205/http://ir.iit.edu/cikm2004/tutorials.html

Grimes, S. 2007. "What's Next for Text. Text Analytics Today and Tomorrow: Market, Technology, and Trends". Text Analytics Summit 2007.

Halper, F.,Kaufman, M., and Kirsh, D. 2013. "Text Analytics: The Hurwitz Victory Index Report". Hurwitz & Associates 2013. Available at: http://www.sas.com/news/analysts/Hurwitz_Victory_Index-TextAnalytics_SAS.PDF

H.P.Luhn. 1958. "The Automatic Creation of Literature Abstracts*". IBM Journal of Research and Development*, 2(2):159-165.

Manning, C. D., and Schutze, H. 1999. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press.

McNeill, F. and Pappas, L. 2011. "Text Analytics Goes Mobile". *Analytics Magazine*, September/October 2011. Available at: http://www.analytics-magazine.org/septemberoctober-2011/403-text-analytics- goes-mobile

Mei, Q. and Zhai, C. 2005. "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining". KDD 05: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining,* 198 – 207.

Miller, T. W, 2005. Data and Text Mining: A Business Applications Approach. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Radovanovic, M. and Ivanovic, M. 2008. "Text Mining: Approaches and Applications". *Novi Sad Journal of Mathematics.* Vol. 38: No. 3, 227-234.

Salton, G., Allan, J., Buckley C., and Singhal, A. 1994. "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts". *Science*, 264.5164 (June 3): 1421-1426.

SAS® Ontology Management, Release 12.1. Cary, NC: SAS® Institute Inc.

Shaik, Z., Garla, S., Chakraborty, G. 2012. "SAS® since 1976: an Application of Text Mining to Reveal Trends". *Proceedings of the SAS Global Forum 2012 Conference.* SAS Institute Inc., Cary, NC.

Text Analytics Using SAS® Text Miner. Course Notes. Cary, NC, SAS Institute. Inc. Course information: https://support.sas.com/edu/schedules.html?ctry=us&id=1224

Text Analytics Market Perspective. White Paper, Cary, NC: SAS Institute Inc. Available at: http://smteam.sas.com/xchanges/psx/platform%20sales%20xchange%20on%20demand%20%202007%20sessions/text%20analytics%20market%20perspective.doc

Wakefield, T. 2004. "A Perfect Storm is Brewing: Better Answers are Possible by Incorporating Unstructured Data Analysis Techniques." DM Direct Newsletter, August 2004.

Weiss S, Indurkhya N, Zhang T, and Damerau F. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer-Verlag.

Index

A

accessing project data for Case Study 1: Text Mining SUGI/SAS Global Forum Paper Abstracts to Reveal Trends 194–196
Activated statistical model 171
ADDRESS default entity 86
advanced searches, in Query Web Server 29
algorithms, clustering 112–113, 119
AND operator 143
applications of content management 134
aspect level 165
aspect-level sentiment classification 166
association-based methods 112
attribute level 165
AUTHOR concept name 89
automatic detection 198–207
automatic generation
  of categorization rules using SAS Text Miner 157–159
  of rules in SAS Content Categorization Studio 140–141, 157–159
"The Automatic Creation of Literature Abstracts" (Luhn) 1

B

bag-of-words method, of text analysis 69
Bayes classifier 166
big data, handling 15–16
binary frequency weighting 98
Binary level 43
black box model
  See statistical categorizer
Boolean approach 143–144
breadth first mode 23–24
Brown tag set 71
Buckley, C. 98

C

_c context marker symbol 89
case studies 179
  See also specific case studies
Case Study 1: Text Mining SUGI/SAS Global Forum Paper Abstracts to Reveal Trends
  about 181
  accessing project 194–196
  data 181–189
  results 189–190
  trends 190–193
Case Study 2: Automatic Detection of Section membership for SAS Conference Paper Abstract Submissions
  about 197–198
  data 198
  objective 198
  step-by-step instructions 198–207
Case Study 3: Features-based Sentiment Analysis of Customer Reviews
  about 209
  data 209–210
  NLP based sentiment analysis 219–224
  text mining for negative app reviews 210–216
  text mining for positive app reviews 217–218
Case Study 4: Exploring Injury Data for Root Causal and Association Analysis
  about 227
  data 227
  objective 227
  step-by-step instructions 228–238
Case Study 5: Enhancing Predictive Models Using Textual Data 241–248
Case Study 6: Opinion Mining of Professional Drivers' Feedback
  about 251
  analysis using SAS Sentiment Analysis Studio 264
  analysis using SAS Text Miner 251–258
  analysis using Text Rule-builder Node 258–264
  building a rule-based model 268–271
  building a statistical model 265–268
  data 251
Case Study 7: Information Organization and Access of Enron Emails to Help Investigation
  about 273
  data 274
  objective 273–274
  step-by-step instructions 274–281
Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data
  about 283
  categorizing content 288–289
  concept map visualization 289–292
  data 283–284
  examining topics 285–288
  integrating with SAS Visual Analytics 293–294
  merging splitting topics 288
  step-by-step instructions 284–285
  using DS2 procedure for deployment 292–293
Case Study 9: Evaluating Health Provider Service Performance Using Textual Responses 297–310
Castellanos, M. 164
Categories pane (SAS UTAI) 288–289
categorization
  about 137–138
  automatic generation of rules using SAS Text Miner 157–159
  of content in Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data 288–289
category membership, determining 145–146
category rules, scoring 156

C_CONCEPT rule 153, 173
child page 23
Chomsky, N. 69
classifier concept 88
CLASSIFIER rule 150–152, 173
classifiers 147–148
cluster analysis
    See clustering
clustering
    about 11–12, 111–112
    algorithms 112–113, 119
    hierarchical 257
    similarity metrics 112
    Text Cluster node 119–122, 123, 243
COMPANY default entity 86
COMPBL function 62
component servers, of SAS Search and Indexing 28–38
COMPRESS function 62
concept 88
concept definitions, scoring 156
concept extraction 145–150
concept links 106–108
concept map visualization, in Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data 289–292
CONCEPT rule 152–153, 173
CONCEPT_RULE rule 153–154, 173
Consumer Product Safety Commission (CPSC) 227
content categorization
    about 134–136
    compared with text clustering 159–160
Content Categorizer 239
content extraction
    classifiers 147–148
    grammar 149–150
    regular expressions 148–149
content management
    about 133–134
    automatic generation of categorization rules using SAS Text Miner 157–159
    C_CONCEPT rule 153, 173
    CLASSIFIER rule 150–152, 173
    concept extraction 145–150
    CONCEPT rule 152–153, 173
    CONCEPT_RULE rule 153–154, 173
    content categorization 134–136, 159–160
    contextual extraction 134, 150
    determining category membership 145–146
    differences between text clustering and content categorization 159–160
    NO_BREAK rule 154
    PREDICATE_RULE definition 155–156, 173
    REMOVE_ITEM 154–155
    rule-based categorizer 141–145
    SEQUENCE rule 155–156
    statistical categorizer 139–141, 144–145
    types of taxonomy 136–139
content tagging 134
contextual extraction 134, 150
core component servers
    Pipeline Server 21, 27
    Proxy Server 21, 26
correlation-based methods 112
CPSC (Consumer Product Safety Commission) 227
CURRENCY default entity 86

D

Dale, R. 69
data
    for Case Study 1: Text Mining SUGI/SAS Global Forum Paper Abstracts to Reveal Trends 181–189
    for Case Study 2: Automatic Detection of Section membership for SAS Conference Paper Abstract Submissions 198
    for Case Study 3: Features-based Sentiment Analysis of Customer Reviews 209–210
    for Case Study 4: Exploring Injury Data for Root Causal and Association Analysis 227
    for Case Study 6: Opinion Mining of Professional Drivers' Feedback 251
    for Case Study 7: Information Organization and Access of Enron Emails to Help Investigation 274
    for Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data 283–284
    collecting in text mining 5
    exploration and analysis of 276–277
    importing into SAS Text Miner using Text Import node 49–57
    retrieving from Web 54–57
    types of in SAS Text Miner 42–43
data levels, in SAS Text Miner 42–43
data roles, in SAS Text Miner 42–43
data sources, creating in SAS Enterprise Miner 43–48
DATA step 62–67
data transformation
    about 93–94
    concept links 106–108
    filtering documents 102–106
    frequency weightings 98
    term weightings 98–102
    term-by-document matrix 96–97, 112
    Text Filter node 97–98
    Zipf's law 94–96
DATE default entity 86
deploying models 292
depth first mode 24
Depth property (Text Import node) 56
Diagram workspace (SAS Enterprise Miner/SAS Text Miner) 41
diagrams, creating in SAS Enterprise Miner 45–46

dictionaries 85
Different Parts of Speech property 80, 82
document cutoff 126
document level 164
document processors 27
document tagging 134
document topic weight 126
document_converter 27
documents
  classifying 8–9
  filtering 102–106
Documents pane (SAS UTAI) 286–288
DS2 procedure, using in Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data 292–293
Duraidhayalu, H. 164

E

EM (Expectation-Maximization) cluster algorithm 113, 119, 256
emerging directions
  about 15
  big data 15–16
  real-time text analytics 16
  voice mining 16
enabling
  Virtual Indexing Server 40
  Virtual Query Server 40
English dictionary, creating 85
ENGSTOP default stop list 82
entities 86–90
entity level 165
entropy 99
Euclidean distance 112
Evangelopoulos, N. 12
Expectation-Maximization (EM) cluster algorithm 113, 119, 256
exporting content to CSV or XML files with SAS Markup Matcher server 34–38
export_to_files 27
Extensions property (Text Import node) 51, 55
extract_abstract 27
extracting
  See also parsing; text parsing
  content from HTML web pages with SAS Markup Matcher server 34–38
  content from websites with SAS Web Crawler 30–34
  text from PDF files 49–54

F

FA (factor analysis) 113
feature level 165
features-based sentiment analysis 209–225
Feed Crawler 21, 25
feeds 25–26
File Crawler 21, 25, 274–275
file formats, supported by Text Import node 49
File Import node 58
FILTERED variable 73
filtering
  documents 102–106
  words 185
Find Entities property (Text Parsing node) 87, 88
F-measure 136
forward slash (/) 232
frequency weighting 98
frequency-based relevancy type 145
functions
  See specific functions

G

Gillin, Paul
  The New Influencers: A Marketer's Guide to the New Social Media 163
global weight 98–102
Goodnight, J. 88
Google Trends 12
grammar concept 88
grammar concept definitions 149–150
Grover, S. 164

H

Help panel (SAS Enterprise Miner/SAS Text Miner) 41
heuristic_parse_html 27
hierarchical algorithm 119
hierarchical clustering 257
hierarchical taxonomy 137, 165
hub-and-spoke structure 106

I

identifying
  feeds 25–26
  parts of speech 78–81
IDF (inverse document frequency) 99–100
Ignore Parts of Speech property 80
IMPORT procedure 58–59
importing
  data into SAS Text Miner using Text Import node 49–57
  textual data into SAS Text Miner 41–67
  XLS and XML files into SAS Text Miner 58–62
INDEX function 62–67
indexing
  about 279–281
  content from HTML web pages with SAS Markup Matcher server 34–38
  content from websites with SAS Web Crawler 30–34
Indexing Server 21, 28
information extraction
  about 10–11, 19–20

using SAS Crawler 19–39
information organization
about 19–20
case study on 273–281
information parsing 274–276
information retrieval (IR) 7–8, 274–276
interactive filter viewer 102–103, 105, 239, 304
interactive topic viewer 126–128, 216, 218
International Standardization for Organization (ISO) 87
INTERNET default entity 86
Interval level 43
inverse document frequency (IDF) 99–100
IR (information retrieval) 7–8, 274–276
ISO (International Standardization for Organization) 87

J

Joshi, M. 164
JSON procedure 61–62

K

keywords, using Perl regular expressions (PRX functions) to extract text between 66–67
k-means 113
Kohonen, T. 113

L

Language property (Text Import node) 51, 55
latent semantic analysis (LSA) 113–122
latent semantic indexing 113–122
LDC (Linguistic Data Consortium) 71
lemmatization 70
lexicon-based approach 166
libraries, creating in SAS Enterprise Miner 44–45
Lieberman, H. 25
linguistic approach 141–143
Linguistic Data Consortium (LDC) 71
literal string 147–148
LITI classifier definitions 151–152
Liu, B. 164
Liu, J. 164
LOCATION default entity 86
log frequency weighting 98
LSA (latent semantic analysis) 113–122
Luhn, Hans Peter
"The Automatic Creation of Literature Abstracts" 1

M

Mahalanobis distance 112
managing
See also content management
big data 15–16
text using SAS character functions 62–67
Mandelbrot, B. 95–96
Manning, C.D. 69, 94
Maximum Number of Terms property 103
McCallum, A.K. 24
MEASURE default entity 86
menu bar (SAS Enterprise Miner/SAS Text Miner) 41
Metadata Node 243–244, 307–308
Minimum Number of Documents property 103
Model Comparison node 247–248
models, deploying 292
modes of crawl 23
modifying rules 206–207
mutual information 100

N

Nagarajan, D. 164
National Electronic Injury Surveillance System (NEISS) 227
natural language processing (NLP) 165, 219–224
negative app reviews, text mining for 210–216
negative precision 169
NEISS (National Electronic Injury Surveillance System) 227
The New Influencers: A Marketer's Guide to the New Social Media (Gillin) 163
NLP (natural language processing) 165, 219–224
NLP based sentiment analysis 219–224
NO_BREAK rule 154
nodes (SAS Enterprise Miner) 42
See also specific nodes
normalization 70, 169
NOT operator 143
Noun Groups property 80

O

object level 165
objectives
for Case Study 2: Automatic Detection of Section membership for SAS Conference Paper Abstract Submissions 198
for Case Study 4: Exploring Injury Data for Root Causal and Association Analysis 227
for Case Study 7: Information Organization and Access of Enron Emails to Help Investigation 273–274
ontology management 9–10
OpenOffice.org (website) 85
operator-based relevancy type 146
opinion mining
See sentiment analysis
OR operator 143
Ordinal level 43
ORGANIZATION default entity 86
overall precision 169

P

Pang, B. 166

Pantangi, A. 164
parsing
See also text parsing
content from HTML web pages with SAS Markup Matcher server 34–38
content from websites with SAS Web Crawler 30–34
parsing tree 71–73
partial least squares (PLS) 113
part-of-speech feature 78–81
parts-of speech (POS) tagging 71, 219
PCA (principal component analysis) 113
PDF files, extracting text from 49–54
Penn Treebank 71
PERCENT default entity 86
Perl regular expressions (PRX functions)
managing text using 63–64
using to extract text between keywords 66–67
using to match exact text 64–65
using to replace exact text 65–66
PERSON default entity 87
PHONE default entity 87
Pipeline Server 21, 27
PLS (partial least squares) 113
POS (parts-of speech) tagging 71, 219
positive app reviews, text mining for 217–218
positive precision 169
PREDICATE_RULE rule 155–156, 173
predictive models
enhancing using textual data 241–248
enhancing with exploratory text mining 13–14
PREFIX concept name 89
principal component analysis (PCA) 113
probabilistic clustering 113
process flows 41–42
Project panel (SAS Enterprise Miner/SAS Text Miner) 41
projects, creating in SAS Enterprise Miner 43–44
Properties panel (SAS Enterprise Miner/SAS Text Miner) 41
PROP_MISC default entity 87
Proxy Server 21, 26
PRX functions (Perl regular expressions)
managing text using 63–64
using to extract text between keywords 66–67
using to match exact text 64–65
using to replace exact text 65–66
PRXCHANGE function 65–66
PRXMATCH function 64–65, 66–67
PRXNEXT function 67
PRXPAREN function 67
PRXPARSE function 64–67
PRXPOSN function 67

Q

Query Server 21, 28–29
Query Statistics Server 22, 29
Query Web Server 22, 29

R

real-time text analytics 16
real-world applications and examples 24–26
receiver operating characteristics (ROC) curve 13–14
REGEX rule 148, 173
Regression node 245–247
regular expressions 148–149, 151–152
Relative weight of positive rules in rule-based model 171
REMOVE_ITEM 154–155
results, for Case Study 1: Text Mining SUGI/SAS Global Forum Paper Abstracts to Reveal Trends 189–190
retrieving documents 103–106
ROC (receiver operating characteristics) curve 13–14
rule-based categorizer
about 141
Boolean approach 143–144
compared with statistical categorizer 144–145
linguistic approach 141–143
rule-based models
building 268–271
compared with statistical models 225
in SAS Sentiment Analysis Studio 172–176
Rule-Builder node 260–262
rules
See also specific rules
adding 290–292
creating 221
editing 290–292
modifying 206–207
Rules pane 290–292

S

Salton, G. 98
Sample, Explore, Modify, Model, and Assess (SEMMA) 42
Sarkar, M. 164
SAS Add-In for Microsoft Office 135
SAS character functions, managing text using 62–67
SAS Code node 263
SAS Content Categorization Single User Servers12.1: Administrator's Guide 156
SAS Content Categorization Studio
about 4, 9
automatic rule generation 140–141, 157–159
C_CONCEPT rule 153, 173
CLASSIFIER rule 150–152, 173
concept extraction 145–150
CONCEPT rule 152–153, 173
CONCEPT_RULE rule 153–154, 173
content categorization in compared with text clustering in SAS Text Miner 159–160
contextual extraction 150
determining category membership 145–146

features of 134–136
NO_BREAK rule 154
PREDICATE_RULE rule 155–156, 173
REMOVE_ITEM 154–155
rule-based categorizer 141–145
SEQUENCE rule 155–156
statistical categorizer 139–141
types of taxonomy 136–139
SAS Contextual Analysis, case studies using 283–295
SAS Contextual Extraction Studio, building custom entities using 88–90
SAS data sets, creating 49–54, 54–57
SAS Document Conversion Server 22
SAS Document Server 49
SAS Enterprise Content Categorization Studio
case studies using 197–225, 227–239, 273–281
exploring injury data for root causal and association analysis with 234–238
SAS Enterprise Miner
about 41–42
creating data sources in 43–48
Score node 130
SOM/Kohonen method 113
SAS Global Forum 99
SAS Information Retrieval Studio: Administration Guide 24
SAS Information Retrieval Studio Interface
about 20–21
case studies using 273–281
core components of 21
Query Server 28
SAS Markup Matcher Server
about 22, 29–34
exporting content to CSV or XML files with 34–38
extracting, parsing, and indexing content from HTML web pages with 34–38
SAS Ontology Management 4
SAS Rapid Predictive Modeler 16
SAS Search and Indexing
about 4, 20
component servers of 28–38
components of 21–22
indexing server 28
Query Server 28–29
Query Statistics Server 29
Query Web Server 29
SAS Markup Matcher server 29–38
SAS Sentiment Analysis Studio
See also sentiment analysis
about 5, 166–167
analysis using 264
case studies using 209–228, 251–272
rule-based models in 172–176
SAS Text Miner and 175–176
statistical models in 167–172
SAS Sentiment Analysis Studio: User's Guide 173
SAS Sentiment Analysis Workbench 167, 171
SAS Text Miner
about 4–5, 41–42
analysis using 251–258
automatic generation of rules using 157–159
case studies using 181–225, 227–239, 241–249, 251–281, 297–310
clustering in 112
data types, roles and levels in 42–43
exploring injury data for root causal and association analysis with 228–234
importing textual data into 41–67
importing XLS and XML files into 58–62
SAS Sentiment Analysis Studio and 175–176
Text Cluster node 123
text clustering in compared with content categorization in SAS Content Categorization Studio 159–160
Text Filter node 97–98
text mining with 5–7
Text Parsing node in 73–88
Text Topic node 123–126
%TMFILTER macro 57
Zipf's law and 95
SAS UTAI 283
SAS Visual Analytics, integrating with in Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data 293–294
SAS Web Crawler
about 4, 20, 54
components of 21
extracting, parsing, and indexing content from websites with 30–34
information extraction with 19–39
SAS XML Mapper 59–61
SCAN function 62–67
Schutze, H. 69, 94
Score node 130
scoring 130, 156
scree plot 117
searching documents 103–106
self-organizing map (SOM) 113
SEMMA (Sample, Explore, Modify, Model, and Assess) 42
sentence level 164
sentence-level sentiment classification 166
sentiment analysis
See also SAS Sentiment Analysis Studio
about 2, 14–15, 163–165
case study 251–271
challenges in conducting 165
unsupervised *versus* supervised 165–166
sentiment mining
See sentiment analysis
SEQUENCE rule 155–156

shortcut buttons (SAS Enterprise Miner/SAS Text Miner) 41
similarity metrics 112
simple searches, in Query Web Server 29
singular value decomposition (SVD)
  about 113–114
  mathematics of 114
  numerical example of 114–118
SOM (self-organizing map) 113
Somers, H. 69
SOM/Kohonen method 113, 256
Sparck Jones, K. 98
SPEDIS function 84
spell check feature (Text Filter node) 84–86, 229–230, 239
SSN default entity 87
start lists 81–83
statistical categorizer
  about 139–140
  compared with rule-based categorizer 144–145
statistical models
  building 265–268
  compared with rule-based models 225
  in SAS Sentiment Analysis Studio 167–172
stemming 70, 73–78
step-by-step instructions
  for Case Study 2: Automatic Detection of Section membership for SAS Conference Paper Abstract Submissions 198–207
  for Case Study 4: Exploring Injury Data for Root Causal and Association Analysis 228–238
  for Case Study 7: Information Organization and Access of Enron Emails to Help Investigation 274–281
  for Case Study 8: Unleashing the Power of Unified Text Analytics to Categorize Call center Data 284–285
stop lists 81–83
SUBSTR function 62–67
SUGI/SAS Global Forum paper abstracts, text mining to reveal trends in 181–196
supervised sentiment classification 165–166
support vector machines (SVMs) 165–166
SVD (singular value decomposition)
  about 113–114
  mathematics of 114
  numerical example of 114–118
SVMs (support vector machines) 165–166
synonyms 73–78
syntactic-pattern based approach 166

T

taxonomy
  building a 277–279
  hierarchical 137, 165
  types of 136–139
term cutoff 126
term parsing 69–70
term topic weight 125
Term Weight property 101
term weightings 98–102
term-by-document matrix 96–97, 112
testing, in categorization 138–139
text
  extracting from PDF files 49–54
  managing using Perl regular expressions (PRX functions) 63–64
  managing using SAS character functions 62–67
  using Perl regular expressions (PRX function) to extract between keywords 66–67
  using Perl regular expressions (PRX functions) to match exact 64–65
  using Perl regular expressions (PRX function) to replace exact 65–66
  text analytics 1–5
  Text Cluster node 119–122, 123, 243
  text clustering, compared with content categorization 159–160
  Text Filter node
  about 97–98
  interactive filter viewer 102–103, 239
  properties panel 260, 300
  spell check feature 229–230, 239
text filtering, in text mining 6
Text Import node 49–57
text mining
  about 2, 6, 181
  collecting data in 5
  enhancing predictive models with exploratory 13–14
  for negative app reviews 210–216
  for positive app reviews 217–218
  process flow 7, 302, 307
  SUGI/SAS Global Forum paper abstracts to reveal trends 181–196
  using SAS Text Miner 5–7
text normalization 169
text parsing
  about 69–70
  parsing tree 71–73
  POS tags 71, 219
  in text mining 5–6
  tokens 70–73
  words 70–73
Text Parsing node 53, 73–88, 242
  See also text parsing
Text Rule Builder feature (SAS Text Miner) 157–159
Text Rule Builder node 175–176, 258–264
Text Size (Text Import node) 55
Text Topic node 123–126, 128–130, 233, 239, 253–255
text transformation, in text mining 5–6
%TEXTSYN macro 84–85

textual data
  enhancing predictive models using 241–248
  importing into SAS Text Miner 41–67
textual responses, evaluating performance using 297–310
threshold value 142
TIME default entity 87
TIME_PERIOD default entity 87
TITLE default entity 87
TKCAT source code sample 292–293
%TMFILTER macro 57, 73, 198
tokens 70–73
Tom Sawyer (Twain) 94
tools (SAS Text Analytics) 4–5
tools palette (SAS Enterprise Miner) 42
topic extraction
  about 111, 122–123
  interactive topic viewer 126–128
  scoring 130
  Text Topic node 123–126
  user-defined topics 128–130
topics 285–288
Topics pane (SAS UTAI) 285–286
training, in categorization 138–139
TRANSLATE function 62
TRANSTRN function 62
TRANWRD function 62
trends
  analyzing 12–13
  for Case Study 1: Text Mining SUGI/SAS Global Forum Paper Abstracts to Reveal Trends 190–193
  text mining SUGI/SAS Global Forum paper abstracts to reveal 181–196
Twain, Mark
  Tom Sawyer 94

U

Unary level 43
unified text analytics 283–295
United States Consumer Product Safety Commission (CPSC) 227
unsupervised sentiment classification 165–166
User Topics property (Text Topic node) 128
user-defined topics 128–130

V

VEHICLE default entity 87
Virtual Indexing Server 22, 40
Virtual Query Server 22, 40
voice mining 16

W

web crawler
  about 21, 22–23
  breadth first 23–24
  depth first 24
  real-world applications and examples 24–26
  tasks of 54
websites
  dictionaries 85
  Linguistic Data Consortium (LDC) 71
  OpenOffice.org 85
  regular expressions guide 148
  resources for research purposes 174
  retrieving data from 54–57
Weight of statistical model in hybrid model 171
Weighted Linguistic Rule Relevancy Score 142–143
Woodfield, T. 12
words 70–73, 185

X

XLS files, importing into SAS Text Miner 58–62
XML files, importing into SAS Text Miner 58–62

Z

Zipf, George 94
Zipf's law 94–96
zone-based relevancy type 146

Symbols

/ (forward slash) 232
character 148–149

About The Authors

Dr. Goutam Chakraborty has a B. Tech (Honors) in mechanical engineering from the Indian Institute of Technology, Kharagpur; a PGCGM from the Indian Institute of Management, Calcutta; and an MS in statistics and a PhD in marketing from the University of Iowa. He has held managerial positions with a subsidiary of Union Carbide, USA, and with a subsidiary of British American Tobacco, UK. He is a professor of marketing at Oklahoma State University, where he has taught business analytics, marketing analytics, data mining, advanced data mining, database marketing, new product development, advanced marketing research, web-business strategy, interactive marketing, and product management for more than 20 years.

Murali Pagolu is a Business Analytics Consultant at SAS and has four years of experience using SAS software in both academic research and business applications. His focus areas include database marketing, marketing research, data mining and customer relationship management (CRM) applications, customer segmentation, and text analytics. Murali is responsible for implementing analytical solutions and developing proofs of concept for SAS customers. He has presented innovative applications of text analytics, such as mining text comments from YouTube videos and patent portfolio analysis, at past SAS Analytics conferences. He currently holds six SAS certification credentials.

Satish Garla is an Analytical Consultant in Risk Practice at SAS. He has extensive experience in risk modeling for healthcare, predictive modeling, text analytics, and SAS programming. He has a distinguished academic background in analytics, databases, and business administration. Satish holds a master's degree in Management Information Systems at Oklahoma State University and has completed the SAS and OSU Data Mining Certificate program. He is a SAS Certified Advanced Programmer for SAS 9 and a Certified Predictive Modeler using SAS Enterprise Miner 6.1.

Learn more about these authors by visiting their author pages, where you can download free chapters, access example code and data, read the latest reviews, get updates, and more:

http://support.sas.com/chakraborty

http://support.sas.com/pagolu

http://support.sas.com/garla