Text Mine Your Big Data: What High Performance Really Means
Introduction

Let’s face it. It’s getting harder and harder to keep up with rapidly changing (and repetitive) modeling requirements from analytic professionals. They seem to continually need new data sources, more sophisticated data preparation, increased computing power to test new ideas and new scenarios – and the list goes on.

In fact, more time and effort is spent provisioning, supporting and managing existing analytics infrastructures than extending capabilities to meet new demands. Yet all this time and effort still doesn’t guarantee predictable, repeatable or enhanced performance. As a result, bottlenecks occur and cause delays that affect business performance and damage IT’s reputation and perceived value to the business.

The situation will most likely get worse. With expectations that 90 percent of the digital universe will comprise unstructured data over the next 10 years,¹ the pressure on IT to drive better performance can only continue to increase.

Even if we only account for unstructured text data, these volumes can be staggering. Consider that:

- Google processes 11.4 billion queries each month.
- Twitter processes half a billion tweets each day.
- Facebook reached 1 billion active users in 2012 and averages more than 1 billion status updates daily.

With this scale of activity, opportunities abound to analyze, monitor and predict what customers and constituents are saying about your organization, doing with your products and services, and believing about your competitors. Yet because of the massive amounts of Web and internally generated data, you must spend significantly more time and computing power to perform analytical tasks. The unstructured content from forums, blogs, emails and product review sites certainly provides abundant input for analysis. But you need new strategies for computational efficiency so that you can analytically process all this data. Only then can you quickly derive meaningful conclusions that have a positive impact on your business.

With SAS® High-Performance Analytics solutions, you can analyze big data from structured data repositories as well as unstructured text collections. This enables you to derive more accurate insights in minutes rather than hours, helping you to make better-informed, timely decisions.

By allowing complex analytical computations to run in a distributed, in-memory environment that removes computational restrictions, SAS High-Performance Analytics provides answers to questions you never thought to ask. Now you can examine big data in its entirety – no more need for sampling.

This suite of products includes distinct high-performance analytical capabilities to address statistics, optimization, forecasting, data mining and text mining analysis. These new products use a highly scalable, distributed in-memory infrastructure designed specifically for analytical processing. The result? Your organization gets faster insights so you can be very responsive to customers, market conditions and more. The result could be a complete transformation of your business as you confidently make fact-based decisions.

SAS High-Performance Analytics helps you to:

- Quickly and confidently identify and seize new opportunities, detect unknown risks and make the right choices.
- Use all your data, employ complex modeling techniques and perform more model iterations to get more accurate insights.
- Derive insights at breakthrough speed so you can make high-value, time-sensitive decisions.
- Furnish a highly scalable and reliable analytics infrastructure for testing more ideas and evaluating multiple scenarios to make the absolute best decision.

SAS High-Performance Text Mining has revolutionized the way in which large-scale text data is used in predictive modeling for big data analysis, for both model building and scoring processes. It provides full-spectrum support for deriving insight from text document collections and operates in both symmetric multiprocessing (SMP) systems as well as massively parallel processing (MPP) environments – harnessing the power of multithreaded and distributed computing, respectively. Both SMP and MPP implementation strategies execute the sophisticated analytic processing associated with parsing, term weighting, dimensionality reduction with singular value decomposition (SVD) and downstream predictive data mining tasks distributed in memory.

High-performance text mining operations are defined in a user-friendly interface, similar to that of SAS Text Miner, so there is no requirement for SAS programming knowledge. It also supports various multicore environments and distributed database systems. As a result, you can further boost performance with distributed, in-memory processing, which brings computational processing to your data rather than the other way around.

How It Works

SAS High-Performance Text Mining contains three components for processing unstructured text data, which lead to the automatically generated term-by-document matrix that forms the foundation for computing SVD dimensions. These SVD dimensions constitute the numeric representation of the text document collection and are formatted to be directly used in predictive analysis that includes text-based insights. These three components are:
• **Document parsing**, which applies natural language processing (NLP) techniques\(^2\) to extract meaningful information from natural language input. Specific NLP operations include document tokenizing, stemming, part-of-speech tagging, noun group extraction, default setting or stop/start list-definition processing, entity identification and multiword term handling.

• **Term handling**, which supports term accumulation, term filtering and term weighting. This entails quantifying each distinct term that appears in the input text data set/collection, examining default or a customized synonym list, as well as filtering (removing terms based on frequencies or stop lists) and weighting the resultant terms.

• **Text processing control**, which supports core and threading control processing, manages the intermediate results, controls input and output, and uses the results that are generated by document parsing and term handling to create the term-by-document matrix, stored in a compressed form.

---

\(^2\) Natural language processing (NLP) is the use of computer programs to understand human speech as it is written.
Figure 2: Processing text documents in a massively parallel processing (MPP) environment.

MPP processing is similar to that of SMP and is illustrated in Figure 2. In this case, documents are loaded by the control component onto each computing node (aka worker node) and parsed using multiple threads. After the document-parsing step, the term-handling component sends the terms that are identified on each computing node to the master node (aka general node) to create a global terms table. This table is sent back to each computing node, which the control component uses along with the tokenized documents to create a document-by-term table on each node using multiple threads. As in the SMP mode, the steps of document parsing and document-by-term matrix creation are effectively parallelized.

SAS® High-Performance Text Mining

SAS High-Performance Text Mining is an add-on to SAS® High-Performance Data Mining. It can be accessed using the high-performance text mining node. This node calls two different high-performance procedures, HPTMINE and HPTMSCORE, respectively, and provides functionalities that include: text parsing, term accumulating, filtering and weighting, document-by-term table creation, SVD computation, and topic generation. It processes and scores the text data, enabling you to develop repeatable and shareable process flows for predictive modeling that use your unstructured text data as part of your data mining efforts.

Within the high-performance text mining node, you can set a host of parameters to represent much of that provided in traditional SAS Text Miner nodes, including text parsing, text filter and text cluster nodes.

---

3 Customized term weighting, running SVD separately from parsing, and topics creation will be supported in upcoming releases.
As illustrated in Figure 3, these parameters (i.e., properties) provide a number of configurable options that define the analysis of the text collection. The parameters include:

- **Different Parts of Speech** specifies whether or not to identify each term’s part of speech. For example, within a text mining analysis, noun forms may be of interest whereas verbs may not be. So you would first need to define each term as a noun or a verb before it could be eliminated from downstream analysis.4
- **Find Entities** directs the system to automatically extract entities in text parsing. There are 17 different entities that are predefined to the system (such as location, telephone number, date, company, title, etc.). If you select “Yes,” both PROC HPTMINE and PROC HPTMSCORE apply automatic entity extraction when they parse the documents.
- **Multiword Terms** refers to a data set that contains (case-sensitive) multiple-word terms for text parsing. A sample data set is included with the product.

---

4 A term may take on one of 16 different parts of speech as automatically identified by the system.
• **Synonyms** specifies a data set that contains user-defined synonyms that are used in text parsing; it can include both single and multiword synonyms. A default and customizable synonym list is provided.

• **Stop Lists** enables you to control which terms are used in a text mining analysis. A stop list is a data set that contains a list of terms to be excluded in the parsing results.

• **Minimum Number of Documents** defines a cutoff value to include terms that pass this threshold. During parsing, if a term has a frequency less than the cutoff value, it is excluded from the term-by-document matrix.

• **SVD Resolution** indicates the resolution to be used when computing the SVD dimensions. The software has the capability to automatically determine the number of SVD dimensions to compute. When the value of Max SVD Dimensions is fixed, a higher resolution will lead to more SVD dimensions being calculated.

• **Max SVD Dimensions** specifies the maximum number of SVD dimensions to be calculated and needs to be defined as greater than or equal to two dimensions.\(^5\)

• **Number of Terms to Display** specifies the maximum number of terms to display in the results view associated with high-performance text mining processing. The default value for this property is 20,000; however, it can be set to All. Because the term list is usually very lengthy for a large text corpus, loading all terms into the viewer may take significant time, even in a high-performance implementation.

In addition to the parameter settings used to refine models, you can also view the imported and exported data sets, decide which variables you want the model to train on, and examine the node status – which assesses, among other things, the processing duration of the high-performance text model.

**SAS® High-Performance Text Mining in Action**

Developing predictive models that include variables derived from text mining often leads to better models, simply because they include additional richness about the business scenario stemming from text-based assessments (i.e., the numeric, SVD representations of the text collection).

Figure 4 illustrates a predictive modeling example that uses the high-performance text mining node. In this example, the input text data set is first partitioned to generate training and validation data sets by using the high-performance data partition node.\(^6\) The data sets are then fed into the high-performance text mining node for parsing documents, extracting the SVD dimensions and creating an output data set by merging the extracted SVDs with the original variables in the input data sets. The result is then used in both the high-performance regression node and the high-performance neural node to build two different predictive models. Finally, the model comparison node compares the performance of the two derived models, producing various measures and illustrations that assess the relative benefits of each.

---


\(^6\) The training data set is used by the system to develop the model, while the validation data set is used to test the model on a hold-out sample of data.
To process the high-performance text mining node in a multicore environment, you must first specify the associated information in the Project Start Code property of your project. And for your input data set, the high-performance text mining node also requires you to specify a variable that has the role TEXT as well as a unique ID variable that has the role KEY.

Just as in the traditional text miner node (illustrated in Figure 5), the high-performance text mining node provides various statistics about the terms that appear throughout the document collection. These diagrams are generated from a subset of all the terms identified in your text data (20,000 by default) unless you change the Number of Terms to Display property in the high-performance text mining node’s property sheet.

---

For a detailed description of how this is done, see the SAS technical paper Enabling High-Performance Text Mining Procedures: Effectively translate parsing and SVD procedural code to take advantage of SAS multi-core, in-memory server processing. The paper is available at support.sas.com/resources/papers/Tips-High-Performance-Text-Mining-Procedures.pdf.
Performance Observations

Replicating the same model flow shown in Figure 4, but substituting the high-performance text mining node with the corresponding traditional SAS Text Miner nodes, we can compare the run time of the two approaches. As illustrated in Figure 6, the processing times associated with the nhtsa_cmp data\(^8\) are compared with the entire model flow; but they are also compared with just the text mining aspects, just the predictive modeling aspects and just the other nodes. For the traditional nodes, the results were obtained using a Windows server with two Intel Xeon E5-2667 CPUs and 128GB of memory. For the high-performance nodes, the results were from a cluster system containing 16 computing nodes, with each node having two Intel Xeon X5670 CPUs\(^9\) and 64GB of memory.

The results show that for both text mining and predictive modeling, the high-performance nodes are roughly an order of magnitude faster than the traditional nodes, noticeably increasing the time to results. Specifically, the massive run-time improvement demonstrated using the high-performance text mining node over the traditional text mining node affords inclusion of large-scale text documents to potentially improve predictive modeling processes. Figure 6 also shows that summarizing results and generating reports (the “others” category in Figure 6) require a similar amount of time (about one minute) from both traditional nodes and high-performance nodes, given that these operations are performed using single threads in both cases.

\(^8\) The nhtsa_cmp data set is based on the National Highway Traffic Safety Administration (NHTSA) safety-related defect complaints data set: nhtsa.gov/NCSA and contains 683,076 records, wherein each record is about a paragraph in length.

\(^9\) Each Intel Xeon X5670 CPU has six cores and 12M cache, and its maximum turbo frequency is 3.33GHz.
SMP and MPP Run-Time Comparisons

You can expect to see significant performance improvements when you transition from traditional text mining procedures to the new high-performance text mining procedures. The benefit is not only because you’re moving from a single core to a multicore environment. It is also because high-performance text mining procedures are multithreaded, which provides a significant improvement even on a single machine. The following example takes the same nhtsa_cmp data and compares processing times on a single machine with multiple threads, as well as on a clustered grid with multiple machines (MPP).

Figure 7 compares the run times for the data using the traditional text mining procedures relative to the run times for the high-performance text mining procedure (HPTMINE) on a single machine. The results were obtained using a Windows server with two Intel Xeon E5-2667 CPUs\(^1\) and 128GB of memory. In this comparison, the HPTMINE procedure was tested in the SMP mode with both one thread and eight thread scenarios.

When it uses only one thread, the HPTMINE procedure is about 35 percent faster than traditional text mining procedures. This speed increase is largely because traditional text mining procedures need to communicate with each other by writing their intermediate results out as SAS data sets, whereas the HPTMINE procedure does not. When it uses eight threads, however, the HPTMINE procedure is about seven times faster than traditional text mining procedures due to the enhanced power associated with parallel processing.

\(^{10}\) Each machine also uses multiple threads to process its data.

\(^{11}\) Each Intel Xeon E5-2667 CPU has six cores and 15M cache, and its maximum turbo frequency is 3.5 GHz.
This comparison also shows that when processing the text data set, the document parsing phase dominates the run time. During document parsing, NLP techniques are used to extract meaningful information from the text inputs. The high-performance text mining procedure effectively parallelizes this necessary and compute-intensive phase associated with any text analysis.

Extending the processing environment to an MPP mode (using the same data) further improves run-time performance. As illustrated in Figure 8, the same HPTMINE procedure was considered in a clustered grid system with 16 compute nodes, illustrating the performance differences when the number of grid nodes is varied.

Each computing node has two Intel Xeon X5675 CPUs and 98GB of memory. The HPTMINE procedure uses all the cores on a computing node for processing. Figure 8 shows that when only one node is used the HPTMINE procedure finishes processing in about three minutes. When four nodes are used, the HPTMINE procedure finishes in less than one minute. And with eight nodes, it finishes in just 37 seconds. The results also indicate that for this data, beyond eight nodes, the incremental performance improvements are not significantly different. The increase is limited because the computing nodes communicate and synchronize their results when they process text data in a multicore environment. When the input data set is not big enough, and the number of computing nodes is large, these communication costs start to offset the computing power gains by using more nodes. Therefore, to maximize hardware efficiency, you need to coordinate the number of computing nodes that are used for processing text data with the size of your big data problems.

---

12 Each Intel Xeon X5675 CPU has six cores and 12M cache, and its maximum turbo frequency is 3.46 GHz.
These testing results indicate that a 30-minute processing task for traditional text mining procedures can easily be reduced to less than a minute in a multicore computing environment. This reduced run time means that you can be more efficient in building predictive models that support the organization. With high-performance text mining, you can readily examine very large data sets – on the order of billions of documents – taking advantage of their volume to obtain the most reliable models. Models can also be more easily retrained using different parameters, so that you can optimize model performance in a shorter period of time.

**High-Performance Text Mining Deployment**

Considering that SAS High-Performance Text Mining provides significant processing efficiencies for big text data in a distributed, multicore processing computing environment, let’s examine the analytic architecture involved in a successful deployment. The functionality described is supported by a range of components operating at different levels and in a variety of host and hardware environments. Despite its power, it can be accessed using a point-and-click GUI in addition to a programmable interface. Figure 9 illustrates many different aspects of deployment.
Three categories of high-performance text mining components can be associated with deployment architectures.

- **Functional components.** These components perform the primary text mining functions, such as document parsing, term weighting and filtering, term-by-document matrix creation and dimensionality reduction. These components are fully threaded and perform distributed parallel computing by using the message passing interface (MPI) with the support of the foundation component, the SAS MPP In-Memory Processing. To do this, the text-processing components also need to call a SAS natural-language processing library for text parsing in local languages and the SAS/ACCESS® engines13 for loading data from different platforms.

- **SAS programming interface components.** This wraps functional components into an easy-to-use interface to the native programming environment.

- **SAS Enterprise Miner user interface component.** The high-performance text mining node is available from within the environment of SAS Enterprise Miner by simply dragging the icon onto the project template. This node uses the functions provided by the HPTMINE and HPTMSCORE procedures, conveniently bringing text data into high-performance data mining process flows.

---

13 More information regarding SAS/ACCESS capabilities can be found at: [sas.com/software/data-management/access/index.html](http://sas.com/software/data-management/access/index.html)
Conclusion

SAS High-Performance Text Mining addresses the ever-increasing need to process greater amounts of unstructured text data in less time to improve predictive modeling. This revolutionary text analysis capability enables organizations to continue to work in the SAS analytical environments they are familiar with and, at the same time, benefit from the increased computation power provided by distributed, in-memory analytic processing. The NHTSA’s Consumer Complaints example illustrates the options available to define text analysis in an easy-to-use GUI environment, or as code specifications. The run-time gains affiliated with processing data in a high-performance environment ensures that no big data is too big.

For More Information

Learn how SAS can help your organization get blazing fast insights from all your big data, structured and unstructured alike. Visit: sas.com/hptextmining
About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.