Scaling the Mountain of Unstructured Text

Deep text is an approach to text analytics that adds depth and intelligence to our ability to utilize unstructured text. In *Deep Text*, author Tom Reamy explains what deep text is and surveys its many uses and benefits. He provides best practices, discusses business issues including ROI, and offers guidance on selecting software and building a text analytics capability within an organization.

What Is Text Analytics?

Tom Reamy

And Why Should You Care?

So, what is text analytics? And why should you care? Well, the why part is pretty easy. Text analytics can save you tens of millions of dollars, open up whole new dimensions of customer intelligence and communication, and actually enable you to make use of a giant pile of what is currently considered mostly useless stuff: *unstructured text*.

The "what is" question is a little more complicated, but stick with me and I'll try to give you a good answer in 25 pages or less.

What Is Text Analytics?

About 90% of the time when I tell people what I do—*text analytics*—there is an awkward silence, followed by a kind of blank look. Then, depending on the personality of the person, there is often an "oh, what is that?" Or, there is a sort of muttered, "oh." And then, they start looking for the nearest exit. In other words, it's not a very good icebreaker or conversation starter.

Now, I'm not overly fond of precise definitions of an entire complex field of study, especially one as new and still morphing as text analytics. But I would like to be able to tell people what it is I do, and so I guess I'd better take a stab at defining it.

Actually it's not just the layperson on the street who could use a new definition of text analytics, but there seems to be a great deal of disagreement among those professionals who claim to do text analytics as to what exactly it is. Text analytics encompasses a great variety of methods, technologies, and

applications, so it shouldn't be too much of a surprise that we haven't quite nailed it down yet.

To make matters worse, there are all sorts of claimants for the title of "what I do is the REAL text analytics." For one, "text mining" often claims to deal with all things text. Then, the so-called "automatic categorization" companies will tell you that they do all you need to do with text. And finally, the "semantic technology" or the "semantic web" people not only claim the word semantic as their own but also that what they do is *the* essential way of utilizing unstructured text.

I'm also a firm believer in Wittgenstein's notion of family resemblances, that is, for any complex field, there is no one or two essential characteristics, but rather a family of overlapping characteristics that define what it is—yet another reason why I'm suspicious of attempts to define something as complex as text analytics in a one-sentence definition.

But, we still have to try.

Text Analytics Is …

In my view, the term *text analytics* should be defined in the broadest possible way. Almost anything that someone has described as text analytics belongs within the definition.

In essence, what we're trying to do is *add structure to unstructured/semi-structured text*—which includes everything from turning text into data, to diving down into the heart of meaning and cognition, through to making that text more understandable and usable.

My "big tent" definition of text analytics includes, for example:

- Text mining
- The latest mathematical, vector space, or neural network model
- The grunt work of putting together vocabularies and taxonomies
- The development of categorization rules, the application of those rules, advanced automated processing techniques—everything from your company's official anti-discrimination policy to the chaos of Twitter feeds
- The development and use of sophisticated analytical and visual front ends to support analysts trying to make sense of the trends in 20 million email threads, or the political and social rantings of millions of passionate posters, both evil and heroic (depending on your point of view)

So, with all those caveats (or quibbles) in mind, the essential components of "big tent" text analytics are:

- Techniques – linguistic (both computational and natural language), categorization, statistical, and machine learning
- Semantic structure resources – dictionaries, taxonomies, thesauri, ontologies
- Software – development environment, analytical programs, visualizations
- Applications – business intelligence, search, social media … and a whole lot more

We will go into each of these components in more detail, but one thing they all have in common: They are all used to process unstructured or semi-structured text. And so, the fifth essential component of text analytics is:

- Content – unstructured or semi-structured text, including voice speech-to-text

The output of all this text processing varies considerably. A short list includes:

- Counting and clustering words in sets of documents as a way of characterizing those sets
- Analyzing trends in word usage in sets of documents as part of broader analyses of political, social and economic trends
- Developing advanced statistical patterns of words and clustering of frequently co-occurring words, which can be used in advanced analytical applications—and as a way to explore document or results sets
- Extracting entities (people, organizations, etc.), events, activities, etc., to make them available for use as data or metadata, specifically:
  - Metadata to improve search results
  - Turning text into data, such that all our advanced data analytical techniques can be applied
- Identifying and collecting user and customer sentiment, opinions, and technical complaints to feed programs that support everything customer—customer relations, early identification of product issues, brand management … and even technical support

- Analyzing the deeper meaning and context around words to more deeply understand what the word, phrase, sentence, paragraph, section, document, and/or corpus is about—this is perhaps the most fundamental and the most advanced technique that is used for everything from search ("aboutness") to adding intelligence or context to every other component and application of text analytics

Content and Content Models

With a name like text analytics, it should come as no surprise that the primary content of text analytics is … *text!* But having said that, we haven't said much, so let's look a little more deeply. The stuff that text analytics operates on is all kinds of text from simple notepad text to Word documents and websites, blogger forum posts, Twitter posts, and so on. In other words, anything that can be expressed in words (and can be input into a computer one way or another) is fair game for text analytics.

What we don't deal with are things like video, although there are a number of applications that incorporate video into a text analytics application, either by generating a transcript of all the spoken words in a video and/or operating on any text metadata descriptions of the video.

Text analytics also does not deal directly with data, although again, there is an enormous amount of data incorporated into text analytics applications at a variety of levels.

This type of text is often referred to as *unstructured text*, but that is not really accurate. If it were really unstructured text, we wouldn't be able to make any sense out of it. A slightly more accurate description would be *semi-structured text*, which is what a lot of people call it.

However, this does not really capture the essence of the kinds of text that text analytics is applied to. Only someone raised in a world in which databases rule would come up with the term *semi-structured*. More accurate terms would be *multi-structured*, or even *advanced-structured* (OK, that's probably a bit much).

The reality is, this type of text is structured in a wide variety of ways, some fairly primitive and simple, and still others exemplifying the height of human intelligence.

Let's start with the primitive and simple structure of the text itself. In most languages, ranging from English to Russian to Icelandic, the first level of structure consists of *letters*, *spaces* and *punctuation marks*. We won't be dealing much at the level of letters, although in English and other similar

languages, spaces are how we define the second level of structure—*words*. Also, punctuation marks are important—particularly for the third level of structure, namely *phrases*, *clauses*, *sentences* and *paragraphs*—and this is where the concept of *meaning structures* comes into play.

For obvious reasons, words—the second level of meaning structure—are the basic unit that we deal with in text analytics, normally in conjunction with the third-level meaning structure of phrases, clauses, sentences, and paragraphs. We don't want to get too bogged down in linguistic theory, but we do use words, phrases, clauses, sentences, and paragraphs in text analytics rules.

For example, a standard rule would be to look for two words within the same sentence, and count them differently than finding those two words separated by an indeterminate amount of text. In other words, it is usually more important to find two words in the same sentence than two words in different sentences that happen to be within five words of each other.

The next level of meaning structure is that of *sections* within documents, which can be defined in a wide variety of ways and sizes, but this is where it gets really interesting in terms of text analytics rules. Structuring a document in terms of sections typically improves readability, but it can also lead to very powerful text analytics rules.

For example, in one application we developed rules that dynamically defined a number of sections, which included things like abstracts, summaries, conclusions, and others. The words that define these sections were varied and so had to be captured in a rule, but then that gave us the ability to count the words, phrases and sentences that appeared in those sections as more important than those in the simple body of the document.

Metadata—Capturing and Adding Structure

The last type of structure is *metadata*—data or structure that is added to the document, either by authors, librarians, or software. This includes things such as title, author, date, all the rest of the Dublin Core,[1] and more. Currently, the most popular and successful approach to metadata is done with what are called *facets*—or faceted metadata.

Metadata may not have the exalted meaning of metaphysics and the like, but nevertheless, it is a fundamental and powerful tool for a whole variety of applications dealing with the semantic structure of so-called unstructured text.

The Meaning of "Meta"

Whenever I write about metadata, I'm always struck by the variety of meanings that the word "meta" has accumulated over the centuries. These meanings range from the mundane—metadata is data about data—to the sublime of metaphysics and all the associated uses based on the fundamental meaning of something higher than normal reality.

On a more personal note, it always reminds me of weird little facts that we pick up. As an undergraduate student, I decided that rather than take the standard French or Spanish as my foreign language, I would study ancient Greek. I'm still not sure why I did, but my guess is it had something to do with the fact that I was also reading James Joyce's *Ulysses* at the time. Whatever the reason, I took two-and-a-half years of it!

And that is where I came across this weird little fact about the word "meta:" In Greek, "meta" has a few basic meanings, but these meanings really took off after a librarian in Alexandria attempted to categorize all of Aristotle's works. He had just finished the volume/scroll on physics, and the next work he picked up was this strange work on the nature of reality. And so the story goes: He didn't know what to call it, so he called it *metaphysics*, which in Greek simply meant "the volume that came after the volume on physics." A humble beginning for a word that has come to mean so much.

What text analytics does in the area of metadata is twofold. First, it incorporates whatever existing metadata there is for a document into its own rules. For example, if there is an existing title for a document, then a text analytics rule can count the words that appear in the title as particularly significant for determining what the document is all about.

The second role for text analytics is to overcome the primary obstacle to the effective use of metadata—actually tagging documents with

good metadata values. In particular, this is an issue for faceted metadata applications, which require massive amounts of metadata to be added to documents.

We will explore this topic in more detail in Chapter 10, Text Analytics Applications, but the basic process that has had the most success is to *combine human tagging with automatic text analytics-driven tagging*. This hybrid approach combines the intelligence of the human mind with the consistency of automatic tagging—the best of both worlds.

Text analytics is also ideally suited to pulling out values for facets, such as "people" and "organizations," that enable users to filter search results more effectively (see Chapter 10 for more on facets and text analytics). Text analytics can also pull out more esoteric facets, such as for one project where we developed rules to pull out all the mentions of "methods"—everything from analytical chemical methods to statistical survey methods.

However, the most difficult (but also the most useful) metadata are *keywords* and/or *subject*—in other words, what the document's key concepts are and what the document is about. This is where text analytics adds the most value.

Subject and keywords metadata are typically generated by the text analytics capability of auto-categorization, which we will more fully discuss later in the chapter.

Text analytics uses a variety of meaning-based resources to implement auto-tagging and other metadata assignments. The basic resource is some type of controlled vocabulary, which can be anything, from a simple list of allowed values (names of states or countries) to fully-developed taxonomies.

There is a rich literature on taxonomies (see the bibliography), but the basic idea is that a *taxonomy is a hierarchical structure of concepts* (or events, actions or emotions) used to add a dimension of meaning to the analysis of text documents. Taxonomies are typically used in text analytics to provide a structure for sets of rules that can be applied to the text documents, where each node in the taxonomy will contain rules that categorize the documents as belonging to that node or not.

We will deal with how text analytics utilizes and creates content structure in Chapter 7, Text Analytics Development, and Part 5—Enterprise Text Analytics as a Platform.

Technology / Text Analytics Development Software

Theoretically, text analytics could be done by hand with teams of librarians or indexers, but the reality is it's only possible to do with some fairly sophisticated

technology in the form of software. This software operates on a number of levels. The initial stage is simply structuring all the text into words, words into sentences, and finally into paragraphs. In most languages, including English, this is very simple: Words are defined by spaces, sentences by end-of-sentence code (period), and paragraphs by end of paragraph codes (hard return, followed by new indented text).

All text analytics software is able to perform these basic processes. In addition, the other basic process that virtually all text analytics software includes is part of speech characterization—characterizing words as articles, prepositions, nouns, verbs, and so on.

There are many books on the underlying technology used to accomplish these analytical tasks, so we won't be covering that level in this book.

One of the amazing things about the field of text analytics is that you can actually build a great many extremely valuable applications just on this very, very simple set of capabilities. Some applications, for example, build characterizations of document types based on simple word counts of various parts of speech. In fact, one of the most advanced applications I've seen uses the patterns and frequencies of articles and prepositions (so-called "function words") to build very sophisticated models that can do things, like determine the gender of the writer, and establish the power relationship of the writer to the addressee.[2]

However, in addition to these basic text processing capabilities, the field of text analytics has recently added a range of capabilities, including noun phrase extraction, auto-categorization, analyzing the sentiment of documents, and more.

We will cover those capabilities in finer detail in the next chapter, but will first take a look at the software development environment that is used to build on these basic text processing capabilities. The basic development processes are mostly the same for both text mining and text analytics at the initial stages. The differences show up at the end/analytical stage. The overall process is shown in the following list:

Basic Development Processes for Both Text Mining and Text Analytics:

1. Variety of text sources – web, email, document repositories, etc.
2. Document fetching/crawling processes
3. Preprocessing – categorization, feature/term extraction, sentiment, etc.
4. Processed document collection – machine processing

Text Mining:

5. Apply various algorithms, refine – pattern discovery, trend analysis
6. Basic user functionality – filters, query, visualization tools, GUI, graphing, etc.

Text Analytics:

7. Application – search, sentiment analysis, variety of application front ends

While text mining (TM) and text analytics (TA) share a lot of the initial processing stages and functions, different applications normally call for different approaches to those processing steps. For example, while both TM and TA employ categorization rules, they are typically different types of rules. TM categorization rules are almost always statistical, machine-based rules while TA rules often add explicit, human-created rules. As the title of the book implies, we will be focusing on TA in this book.

The following screenshot (Figure 1.1) shows one development environment for text analytics software. Most text analytics development environments share the majority of functions but, of course, being separate and competing

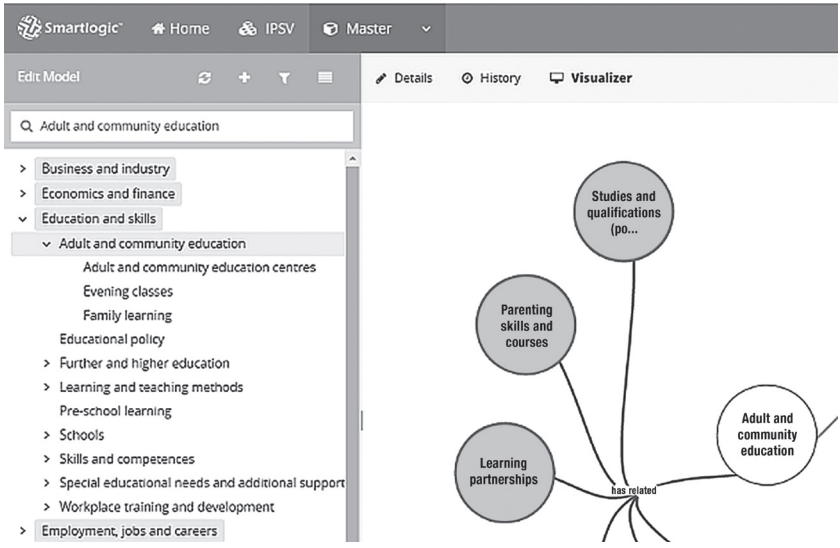

Figure 1.1 Development Environment 1

companies, they all do it slightly differently. This makes life interesting for those of us who work with and partner with multiple text analytics companies.

Figure 1.1 shows the vocabulary and taxonomy (or ontology) management functions that most text analytics development environments utilize. There is a taxonomy on the left, and associated with each node are broader, narrower and related terms. In this example, the phrase "adult and community education" is a narrower term of "education and skills," and the phrases, "adult and community education centres" and "evening classes" are its narrower terms. Of course, each software vendor uses different terminology to refer to the various parts—and typically have slightly different components—but all have the same basic features.

In addition, there are various standard features for basic project development, such as project functions and rudimentary editing functions.

Typically, there is also a variety of features for loading test text files and/or source text files. These text file collections can be used for developing initial categorization rules as well as for testing and refining those rules. This is usually done by running various tests that apply categorization or extraction rules to sets of text files and presenting the results in a variety of screens, showing pass/fail, scores, and other analytical results that also tend to vary by vendor.

The following screen from a different vendor (Figure 1.2) shows a development environment having rules associated with the taxonomy nodes.

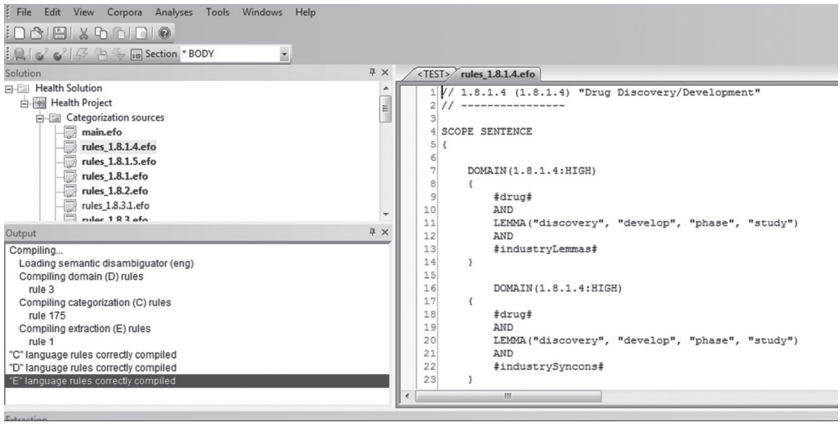

Figure 1.2 Development Environment 2

These rules can be simple lists of terms that you would expect to show up in documents about that topic, but not in other related topics. Or, they may include various advanced rules.

In many ways, these rules are essentially saved searches that can be applied to sets of documents that in turn can be used in a search application to help find specific documents. But they can also be used for a variety of other applications, where the goal is not to find a specific document, but to categorize sets of documents which can then be fed into applications, which might, for example, analyze the overall sentiment expressed in that document set.

Also, these categorization rules are typically orders of magnitude more complex and sophisticated than those of almost all searchers, with the possible exception of professional librarian searchers.

The following screen (Figure 1.3) shows one of those advanced rules as well as other common features in text analytics development environments.

On the left side of the screen is an area labeled "categorizer," which is used to manage a taxonomy that will provide the structure of the set of rules to be applied to documents. Below "categorizer" is an area labeled "concepts," where rules for extracting specific text or types of texts are developed and managed. The area to the right contains the actual rules that are used to categorize and/or extract from the target documents.

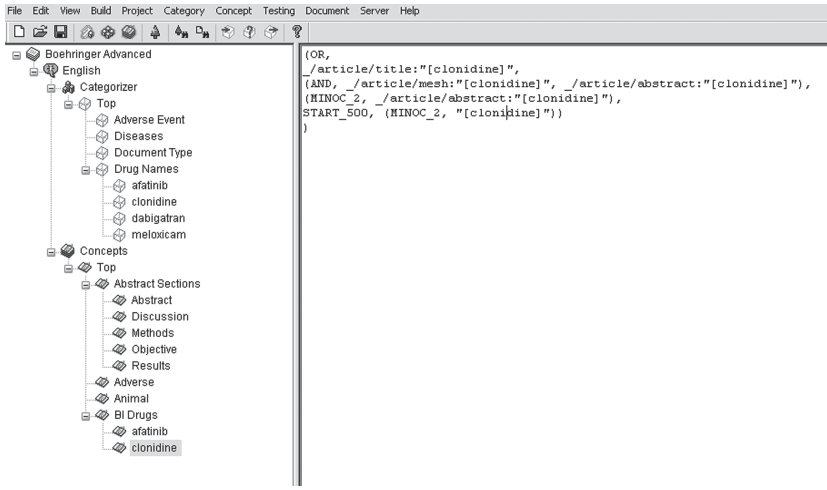

Figure 1.3 Categorization Rules

In addition to these basic development features, text analytics software typically includes functions to generate rules from a set of training documents. These rules can be statistical and/or sets of terms. In addition, some software has functions that attempt to automatically generate subcategories of the particular taxonomy.

Another basic set of functions enables developers to run and manage the testing environment, where rules are tested against a variety of documents and the results can then be analyzed. These tests then become the means to refine the rules.

Text Analytics Applications

Text analytics by itself provides no real value—it is only the range of applications that can be built with text analytics that can help companies deal with the ever-growing mess of unstructured content. In fact, a couple of representatives I interviewed for this book took exception to my characterization of them as a text analytics vendor. One expressed that text analytics is not really a field, but only a component found within various applications. Another representative explained that their company sold applications and services—*not* text analytics—even though those applications and services depended entirely on their text analytics capabilities.

While I agree that text analytics does not provide direct value (except for those of us who find it fun and occasionally profitable), I disagree with the notion that text analytics is not a field (more on that in the Conclusion). In fact, it seems to me that one factor slowing the development of text analytics is an underappreciation of the uniqueness of the skills and capabilities that go into successful text analytics.

A good way to look at text analytics is that text analytics is a platform for building applications, and one thing is certainly true—text analytics applications continue to grow in both number and in the value that organizations are realizing from those applications. At this stage, the only limits seem to be the creativity of application designers and the still unconquered difficulty of intelligently processing all that messy unstructured text—the linguistic messiness problem.

We will take a deeper look at these applications and the role of text analytics in their development in Part 3, but generally, text analytics-based applications fall into four major areas:

1. Search, particularly enterprise search

2. Voice of the customer and other types of social media analysis
3. Search-based applications
4. Embedded applications

Enterprise Search

In the area of enterprise search, text analytics improves search by improving the quality of metadata, improving the efficiency of metadata generation, and lowering the cost of generating all that metadata. One thing has become very clear in the last 10-plus years since just after the turn of the century—enterprise search will never get better without more metadata and better metadata—and this is what text analytics brings to the table.

The three elements of text analytics that are used to improve search are *summarization*, *extraction*, and *auto-categorization*.

Summarization can enhance search results displays by providing a better characterization of the document in the results list than simple snippets. Snippets, the first 50 or so words of a document, can sometimes provide a reasonable clue as to the content of the document, but just as often will be almost meaningless gibberish. There are other more complex kinds of summarization that can be anything from an automatically-generated table of contents to descriptions of all the important concepts and entities in the document.

Extraction can be used to generate large amounts of metadata for each document, and this metadata can be used to develop the one approach to enterprise search that has shown promise—*faceted search*, or faceted navigation. Faceted search works very well, particularly compared with traditional enterprise search with its rather woeful relevance ranking, but one limit is simply the effort to generate all the metadata needed for the various facets. Text analytics changes that by lowering the cost to automate or semiautomate the process while also improving the quality and consistency of the metadata.

Extraction can feed traditional facets like "people" and "organization," where the software extracts all the names of people and organizations, enabling users to filter search results based on those facets. The text analytics-generated metadata can also be combined with other types of metadata that could normally be generated in a content management system, such as "author," "date," and other system metadata.

Auto-categorization can be used to populate the most difficult (and in many ways the most important) facet, which is "topic" or "subject." "Subject" can be used for both what the document is about and/or the major ideas within

that document. This is where auto-categorization is primarily used and it can generate this metadata much more cheaply (and consistently) than hiring a team of out-of-work librarians or part-time taggers.

In addition, auto-categorization can also be used to improve the quality of the metadata generated by extraction through disambiguation rules and the ability to utilize context in much more sophisticated ways than simple catalog-based extraction.

Companies and organizations have spent millions of dollars buying one new search engine after another—and the results continue to disappoint. It is text analytics that has the promise of actually making enterprise search work.

Social Media—Voice of the Customer

Early social media applications consisted primarily of counting up positive and negative words in social posts of all kinds. These words were simply read out of dictionaries of positive ("good," "great," etc.) and negative terms ("terrible," and the ever popular "sucks"), which made the applications very easy to develop.

Unfortunately, it also made these applications rather stupid—and, if not useless, certainly much less valuable than the early bandwagon enthusiasts claimed. On the other hand, it was the beginning of what would become a major new avenue for enterprises to monitor and capture customer (and potential customer) feedback, along with their mindset to better meet their needs.

Thus it is text analytics in the broadest sense that makes this entire field possible—imagine trying to hire enough people to go through hundreds of thousands to millions of Twitter and/or blog posts per day!

While it was possible to get some value from the early simplistic approaches, the field only began to deliver real value when more sophisticated text analytics capabilities were applied. As was true of extraction, social media analysis needed the added intelligence of the full suite of text analytics capabilities to disambiguate, as well as to otherwise take into account the rich context, within which sentiment—or the voice of the customer—was being expressed. For example, with early approaches, the phrase, "I would have really loved this new laptop if it wasn't for the battery," would very likely have been classified as a positive sentiment—"love" and "new laptop" are within a few words of each other—and there are no sentiment words next to "battery."

Fortunately, we are currently in a more mature stage of social media applications, and while they are more difficult to develop, they deliver much more value. The applications include *voice of the customer*—monitoring social

posts for positive and negative customer reactions to basic product features, the features of new product releases, new marketing campaigns, and much more.

Search-Based Applications

Search-based applications is a term that basically refers to using search as a platform for building a whole variety of different applications. These applications include things like e-Discovery, business intelligence, and developing rich dashboards for everything from marketing to scientific research. The idea is to build on search engines' capability of dealing with unstructured text to enrich applications that previously could only utilize structured data.

It is very interesting that in the early discussions of search-based applications, the need for text analytics was included as one component. Unfortunately, as search engine companies jumped on the idea as a natural way to extend their value, they tended to downplay the need for text analytics as something that emphasized the need for an element besides the search engine itself. Software companies seemed loathe to admit that something apart from their product was needed to really make search work. This is perhaps one reason why the idea of search-based applications has not made as much progress as it could have.

Incorporating unstructured text into this class of applications requires that we move beyond simple search results lists, which, as a number of people have discovered, is something that you need text analytics for. Even if you incorporate the results of a search engine's output into other applications, those results are still based on very simplistic relevance ranking calculations. Just as text analytics is needed to make enterprise search work, it is also needed to make search-based applications work. In fact, a better term might be "text analytics–based applications."

Embedded Applications / InfoApps

In addition to using text analytics (with or without a search engine) as a platform for applications, one new trend is to embed text analytics directly into them. These applications, which are somewhat of a second generation of applications built on top of search-based applications, have been termed *InfoApps* by Sue Feldman, who has a wonderful way with nomenclature.

Since the output of text analytics is normally simple XML, it is relatively easy to integrate these capabilities into other applications. The first example of that integration was with enterprise search itself. The second

early integration was with content management software to help generate metadata.

The new generation of InfoApps takes the output from text analytics and embeds them directly into applications that are similar to the search-based applications, but doesn't require an actual search engine platform. Some examples of this type of application:

- Use text analytics for processing a few hundred thousand proposals to pull out all the important facts, like names of the bidder (architects, subcontractors), key dates, project costs, addresses and phone numbers, etc., and make that data available for a wide range of applications.
- Use text analytics to analyze tens of millions of emails between vendors and suppliers to uncover key information, which can be used for anything from buttressing a legal claim to discovering unclaimed discounts owed by the supplier.
- Use business intelligence and customer intelligence for combining data and text processing in order to gain a more complete picture of what is going on in a particular market and/or what specific competitors are doing. This is often paired with sentiment to look into how customers are reacting to new products or marketing campaigns.
- Use your imagination! If you have a lot of unstructured text (and who doesn't?), there will likely be a way for text analytics to do anything from improving your current processes to creating whole new application areas.

We will be taking a deeper look at these kinds of applications in Part 4, but the basic situation is that unstructured text continues to constitute 80%–90% of valuable business information—and the only real way to get good value out of all that text is with text analytics. Text analytics is basically a foundation or platform capability that can be integrated with a whole variety of other application areas and other fields (like semantic technology or Big Data).

As we shall see in later chapters, text analytics can, or should be, a rare combination of approaches and skills—one that incorporates standard IT programming skills with deep academic linguistic skills and a deep appreciation for the complexity of language and actual day-to-day communication.

With that in mind, let's start to take a deeper look at all the elements of this rich and dynamic new field in the next chapter.

Endnotes

1. "DCMI Home: Dublin Core® Metadata Initiative (DCMI)." Dublincore.org, 2015.
2. Pennebaker, James W. *The Secret Life of Pronouns: What Our Words Say about Us.* New York: Bloomsbury Press, 2011.

About the Author

Tom Reamy is currently the chief knowledge architect and founder of the KAPS Group, a group of knowledge architecture, text analytics, and taxonomy consultants, and has 20 years of experience in information projects of various kinds. He has published a number of articles in a variety of journals and is a frequent speaker at knowledge management, taxonomy, and text analytics conferences. He has served as the program chair for Text Analytics World since 2013.

For more than a decade, Tom's primary focus has been on text analytics and helping clients select the best text analytics software as well as doing text analytics development projects that include applications such as call support, voice of the customer, social media analysis, sentiment analysis, enterprise search, and multiple enterprise text analytics–powered applications.

Tom's academic background includes a master's in the history of ideas, research in artificial intelligence and cognitive science, and a strong background in philosophy, particularly epistemology.

When not writing or developing text analytics projects, he can usually be found at the bottom of the ocean in Carmel, photographing strange critters.

This chapter originally appeared in Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value From Social Media, and Add Big(ger) Text to Big Data*, by Tom Reamy. For more information visit the book's webpage. (books.infotoday.com/books/Deep-Text.shtml)*