DEEP TEXT

USING TEXT ANALYTICS
to Conquer Information Overload,
Get Real Value From Social Media,
and Add Big(ger) Text to Big Data

TOM REAMY

DEEP TEXT

Praise for *Deep Text*

"Remarkably useful—a must-read for anyone trying to understand text analytics and how to apply it in the real world."

—Jeff Fried, CTO, BA Insight

"A much-needed publication around the largely misunderstood field of text analytics … I highly recommend *Deep Text* as required reading for those whose work involves any form of unstructured content."

—Jim Wessely, President, Advanced Document Sciences

"Sheds light on all facets of text analytics. Comprehensive, entertaining, and enlightening. Reamy brings together the philosophy, value, and virtue of harnessing text data, making this volume a welcome addition to any professional's library."

—Fiona R. McNeill, Global Product Marketing Manager,
Cloud & Platform Technology, SAS

"One of the most thorough walkthroughs of text analytics ever provided."

—Jeff Catlin, CEO Lexalytics

"Reamy takes the text analytics bull by the horns and gives it the time and exposure it deserves. A detailed explanation of the complex challenges, the various industry approaches, and the variety of options available to move forward."

—Bryan Bell, Executive VP, Market Development, Expert System

"Written in a breezy style, *Deep Text* is filled with advice on the role of text analytics within the enterprise, from information architecture to interface. Practitioners differ on questions of learning systems vs. rule-based systems, types of categorizers, or the need for relationship extraction, but the precepts in Tom Reamy's book are worth exploring regardless of your philosophical bent."

—Sue Feldman, CEO, Synthexis

"A must read for anybody in the text analytics business. A lifetime's worth of knowledge and experience bottled up for us to drink at our own pace. Enjoy!"

—Jeremy Bentley, CEO, Smartlogic

DEEP TEXT

USING TEXT ANALYTICS
to Conquer Information Overload, Get Real Value From Social Media, and Add Big(ger) Text to Big Data

TOM REAMY

Information Today, Inc.
Medford, New Jersey

First Printing

Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value From Social Media, and Add Big(ger) Text to Big Data

Publisher's Note: The author and publisher have taken care in the preparation of this book but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book and Information Today, Inc. was aware of a trademark claim, the designations have been printed with initial capital letters.

Library of Congress Cataloging-in-Publication Data

Names: Reamy, Tom.
Title: 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.
Description: Medford, New Jersey : Information Today, Inc., 2016.
Identifiers: LCCN 2016010975 | ISBN 9781573875295
Subjects: LCSH: Data mining. | Text processing (Computer science) | Text files—Analysis. | Big data. | Social media. | Electronic data processing.
Classification: LCC QA76.9.D343 R422 2016 | DDC 006.3/12—dc23
LC record available at https://lccn.loc.gov/2016010975

Printed and bound in the United States of America

President and CEO: Thomas H. Hogan, Sr.
Editor-in-Chief and Publisher: John B. Bryans
Project Editor: Alison Lorraine
Production Manager: Tiffany Chamenko
Production Coordinator: Johanna Hiegl
Marketing Coordinator: Rob Colding
Indexer: Nan Badgett

Interior Design by Amnet Systems
Cover Design by Denise Erickson

infotoday.com

Contents

PART 1 TEXT ANALYTICS BASICS

PART 3 TEXT ANALYTICS DEVELOPMENT

PART 4 TEXT ANALYTICS APPLICATIONS

PART 5 ENTERPRISE TEXT ANALYTICS AS A PLATFORM

Foreword

"Text analytics" is commonly used to refer to a clutch of quite diverse computer-aided techniques for extracting insight out of large volumes of unstructured text. The techniques, and their practitioners, come from diverse disciplines—information science and information retrieval, data science, statistics, knowledge organization, taxonomy and indexing, and more. The applications are equally diverse, from large scale machine-assisted categorization of knowledge bases, to business intelligence insights from market data, to patent mining for R&D analysts, to sentiment analysis for marketeers, to e-Discovery techniques for litigators.

It is a vibrant, complex, and highly technical field, often appearing mysterious to nonpractitioners. And where there is mystery, there is room for magical thinking, sleight of hand, and unrealistic expectations about what the technology can do, accompanied by an underappreciation of the human analyst's role in guiding the algorithms toward robust and valid insights.

While there are several books available on automated text mining and text analytics techniques and applications, this is the first from a text analytics expert with deep experience in knowledge management and knowledge organization. Here's why this is important. The various machine techniques for analyzing the contents of text in unstructured documents understand parts of speech and semantic rules, and can be told about document structure within any given genre. They can compare documents for similarity and difference, and, using statistical techniques, they can identify keyword strings that may be used to characterize documents or clusters of documents. But there are two very important things they don't know how to do, namely, tracking problems of polysemy (the same terms being used to mean different things) and synonymy (different terms being used to mean the same thing).

Polysemy and synonymy usage occur in the habits and patterns of the subcultures of document writers and consumers, represented by their localized working languages. In genomics, for example, specialists commonly reuse common-language names for gene sequences without any centralized nomenclature control; in consequence, the same name can refer to entirely different gene sequences across different species. Another example might be the document author who habitually uses local short forms that she knows her audience will understand, but that a more distant reader may not.

Machine techniques on their own can be excellent at picking up meanings from within well defined corpora, but once you are trying to understand concepts across very diverse corpora, the machines need to be sensitized by their human handlers to the contextual cues that create significant differences in language and meaning. The varieties of expression need to be mapped to each other using content models, taxonomy structures, or other forms of vocabulary control. These structures are what overcome the problems of polysemy and synonymy, and they allow the analyst to navigate the landscape in a purposeful and flexible way.

This is the realm of the taxonomist and the knowledge organization professional, whose expertise is in understanding and framing differences in context, and in creating language models to enable insights to be extracted in robust and transparent ways across large and diverse sets of unstructured data, and diverse communities of content producers. This is what moves the machine from being a producer of descriptive strings that hold unpredictable value for the analyst to a producer of reliable meaning and insight. And this is what is needed for text analytics to be scalable, reliable, and effective in its support of business goals.

I am very excited that Tom Reamy has at last brought order and clarity to this field using the tools and insights of knowledge organization. This book is written with the practitioner in mind and is full of practical examples and wisdom born of deep experience. In the process, Reamy takes the field of text analytics from a fragmentary clutch of diverse techniques, with very little methodological consistency, toward what he calls a "deep-text" approach—a framework that integrates well-established knowledge organization methodologies with a portfolio or toolkit approach to the use of methods and techniques, and that is focused on getting repeatable value from a text analytics infrastructure, as distinct from the ad hoc project-driven approaches that are so typical today.

Patrick Lambe
April 2016

Patrick Lambe is the author of *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness* (Chandos 2007) and co-author of *The Knowledge Manager's Handbook* (Kogan Page 2016). He is a founder of the Innovations in Knowledge Organisation conference (www.ikoconferen.org) and he consults, researches, and teaches in the field of knowledge organization and knowledge management. Patrick is based in Singapore.

Acknowledgments

After publishing a number of articles over the years, I thought it would not be too difficult to make the step up to writing a book. Boy, was I wrong! It turned out to be more difficult than I imagined, and it would not have been possible at all without the help of a great many people. This help came in many guises and consisted of both large and small contributions. I want to deeply thank all of the people who have helped write this book, one way or another, and if I don't mention you by name, forgive me, but I can only plead the usual rushing to meet a deadline.

First, I want to thank all of the following people for agreeing to be interviewed, and for providing extremely valuable insights into their companies, the overall market for text analytics, and ideas about its future direction:

Trevor Carlow, SAP
Margie Hlava, Access Innovations
Richard Mallah, Cambridge Semantics
Fiona McNeill, SAS
Sartendu Sethi, SAS
Fiona Mitchell, SAS
Meta Brown, Consultant
Jeff Caitlin, Lexalytics
Jeremy Bentley, Smartlogic
Catherine Havasi, Luminoso
Mary McKenna, Textwise
Jim Wessely, Consultant
Tom Anderson, OdinText
Daniel Mayer, Temis
Bryan Bell, Expert System
David Schubmehl, IDC
Boris Evelson, Forrester
Seth Grimes, Market Analyst—Text Analytics
Sue Feldman, Synthexis

I learned a great deal from all of you, especially those of you who disagreed with me. In addition to the interviews, I've had many discussions with a number of you at various conferences, including my own, Text Analytics World—and mostly over a glass or two of fine wine. Thank you.

In addition, I want to thank the consultants who have participated in one or more projects for my company, the KAPS Group. Thanks for making us successful and for putting up with my management style. I especially want to thank Marcia Morante and Jim Wessely, who have been with me on a number of projects. I also want to thank Evelyn Kent, Heather Hedden, Barbara Deutsch, Wendi Pohs, and Michael Kilgore for their help during the company's critical phases. Thanks as well to Laurie Wessely, Diana Bradley, Marlene Rockmore, Deborah Hunt, Heather Dubnik, Barbara Brooker, Donna Cohen, Kyle Nicholls, Deborah Plumley, Cheryl Armstrong, and Melanie Reamy.

And, of course, thanks to all the clients who hired us and allowed us to develop our skills, along with new ideas and solutions. I also want to express my gratitude for all the conferences, especially Information Today and Text Analytics World. Thank you for inviting me to speak and allowing me to try out lots of ideas, including developing a three-hour workshop on text analytics that was the initial basis for this book (with a lot more additions than I thought).

In addition, I would like to thank those people without whom the book would never have any readers—the ITI staff: John B. Bryans, Tiffany Chamenko, Beverly Michaels, Alison Lorraine, Johanna Hiegl, Rob Colding, and Denise Erickson.

I also want to thank the following for their direct contributions in reviewing a number of chapters in the book. These reviews covered everything from grammar and syntax (not my strong suit), and more importantly, the flow of the presentation. They provided a number of ideas for content, as well. These reviewers include Jim Wessely, Evelyn Kent, and Melanie Reamy. And a special thanks to Stefanie Mittelstadt, who did a wonderful job reviewing multiple chapters and providing invaluable suggestions for both flow and content.

Finally, I have to thank my wife, Melanie Reamy, without whose help and support this book would not have been written. She not only contributed in multiple ways but also provided the material and emotional support that I needed. And she put up with too many obsessive-compulsive and just plain cranky days. Thank you.

Introduction

It has become a common assertion that 80% of valuable business information can be found in unstructured text. Exactly how this was determined is a little unclear, but it's huge and very obviously growing with the explosion of social media. Considering the dramatic increase in the amount of unstructured text we are witnessing, I would not be surprised to learn the figure is actually closer to 90%.

We can get a good idea of the value that organizations place on their ability to handle the 10%–20% of information in structured data by the amount of research and activity in the area of data and databases. According to the Bureau of Statistics, there were 118,700 database administrators in the United States in 2012.[1] According to the same source, employment of database administrators is projected to grow 15% by 2022.

In addition, a Google search of database management degree programs turned up 114 different programs—and that doesn't include the vastly more numerous certification and training programs for database administrators and/or programmers.

So if organizations are spending that amount of money and devoting that many resources to handle their structured data—which only contains 10%–20% of valuable business information—imagine how much they're spending on getting the maximum value out of all that unstructured text. It must be a gigantic number, right?

Not exactly. The truth is we have absolutely no precise idea how much money and resources are being devoted to dealing with unstructured text. Nobody has considered that it is significant enough to track—*yet*.

But I do remember a statistic saying that companies devoted 0.5 people (on average) to support their efforts for search, which is one of the primary unstructured text applications.

Text analytics, in its broadest sense, is *the* major tool for dealing with unstructured text other than the human brain—and there are only a handful of degree programs that cover unstructured text in any depth and they focus mainly on text mining as part of a computer science degree. Instead of large database associations with thousands of members, we have a few small groups, such as Text Analytics Summit and Text Analytics World, with their relatively tiny conferences (as compared to the attendance for database conferences).

Baffling, Isn't It?

So, what could explain this baffling lack of attention and effort to get value from all that unstructured text? There are a few possible factors.

First, most of the people that I talk to don't know what text analytics is, and those who actually do recognize the term think that it's basically tracking nice and negative things that people say on Twitter about their company or their products.

Another factor is that for data there are established and common artificial languages, like SQL, that create a platform for both learning and developing applications, and thus developers can build on the work of hundreds of others. Yes, there are variations, but compared to the vendor-specific chaos within text analytics, these are minor.

But the real answers lie with the complexity of text—unstructured text is orders of magnitude more complex than simple data. In addition, text analytics (the primary tool for dealing with unstructured text) has yet to develop a systematic, explicit set of techniques that enable organizations to extract the value and meaning out of their unstructured text. Right now, text analytics is more of an art than a science. It requires programming … and poetry. It requires the ability to write in an artificial language—like SQL or R—and at the same time understand the nuances of meaning in natural languages that range from relatively formal business-speak to the vitriolic rantings found in much of social media. It also requires the ability to apply all of that within useful business contexts for the purpose of creating applications that actually produce value.

Because of the complexity of text, text analytics will probably never be as easy to learn as database design and programming. But we can build a better foundation for text analytics to improve not only how we train people in text analytics, but how we understand the business value of text analytics as a whole.

I hope that one step in the right direction is the writing of this book.

Text Analytics and Text Mining

I suppose at this point I ought to offer at least a preliminary definition of what text analytics really is. There is only one problem: *There's no real consensus about what text analytics really is.*

First of all, there's a lot of confusion over the terms *text mining* and *text analytics*. That modern arbiter of usage, Wikipedia, considers text analytics as a synonym for text mining (I disagree). In fact, the Wikipedia entry is entirely dominated by a mathematical, text mining approach:

> Text mining, also referred to as *text data mining*, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text … The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.[2]

Text analytics doesn't even merit its own entry! Instead, it is currently under text mining's subheader, *text mining and text analytics*:

> The term is roughly synonymous with text mining … The latter term is now used more frequently in business settings while "text mining" is used in some of the earliest application areas, dating to the 1980s, notably life-sciences research and government intelligence.
>
> The term text analytics also describes that application of text analytics to respond to business problems … It is a truism that 80% of business-relevant information originates in unstructured form, primarily text. These techniques and processes discover and present knowledge—facts, business rules, and relationships—that is otherwise locked in textual form, impenetrable to automated processing.[3]

I find it interesting that, even as they tout the similarity of text mining and text analytics, the two entries for "text mining" and "text mining and text analytics" are significantly different. "Text mining" is fundamentally "about extracting data from unstructured text" while "text mining and text analytics" is about how to "discover and present knowledge" locked within text.

You don't have to be a doctrinaire knowledge management (KM) devotee of the data-information-knowledge-wisdom hierarchy to see that there is a big difference between "extracting data" and "discovering knowledge." So, even the entries that claim that they are "roughly synonymous" point to clear differences. I guess it revolves around just how rough the "roughly synonymous" really is—because it looks positively mountainous to me!

When you dive into the specifics of these definitions, the first thing to note is how data-oriented the first one is. For example, the phrase "text data mining"—it's almost as if they can't bring themselves to only talk about text, something I've seen all too often.

Text data mining *is* a good description of one particular technique—extracting entities from unstructured text, then applying data processing methods for a variety of purposes. However, even in a strict text mining approach, counting and discovering patterns in text is often more important than extracting data from the text.

And when you look at the broader concept of text analytics, extracting data *is* an important part of text analytics—but it's only one part. What's missing is the concept of *meaning*—and two of the other biggest application areas within text analytics: search and sentiment. Or, I should say that it *does* include them, but only the simplest approaches to them, ignoring the whole richness (and messiness) of language. The dangers of this approach can be seen in early sentiment analysis applications, which just took dictionaries of positive and negative words and counted them up.

The results? Barely above chance.

As it turns out, to do a good job of extracting data from text, you have to take into account the linguistic and cognitive elements of text. In other words, you have to capture the context around those entities to use them in more intelligent and sophisticated ways. These contexts can be anything, from simple negation, "I didn't love the product," to more complex conditionals, such as "I would have loved this product if it had not been for the keyboard."

The early applications just saw "love" and "product" near each other—and so rated them as positive.

In addition to the differences in approach to the nuances of language, text mining and text analytics have a couple of other major differences. The first is that the applications built with the two techniques tend to be very different. Text mining applications really are much more data oriented, while text analytics powers applications, like search and e-Discovery, which deal with language and meaning far more than data.

The second difference is in terms of the skills and backgrounds of the practitioners. Text mining has more to do with math and statistics, while text analytics builds on language and taxonomies. You find very different people with different skills and experiences in the two fields.

What Is Text Analytics?

I don't want to get bogged down in terminology wars regarding what text analytics actually is, but I think it is important to emphasize the need for two terms that, while they can often be used together to build better applications, refer to very different techniques and different applications. However, the two fields do share a number of common elements, which complicates things. Perhaps what we need is a new term that encompasses both? If you have any candidates, please send them to me and maybe we'll have a contest.

In the meantime, it seems to me that *text analytics* is the broader term, so I will use it to refer to the entire field, except where noted. I will discuss text mining as a part of text analytics, but it is not something that we will devote much time to (see Chapter 1 for more on this).

So here goes: Briefly and generally, text analytics is the use of software and content models (taxonomies and ontologies) to analyze text and the applications that are built using this analysis.

To add more detail to the definition, let's take a look at some of the basic within text analytics. One basic is to simply count the words of various types and discover patterns within the text (text mining). These counts and patterns can be used to gain a variety of insights into the documents as well as the people writing the documents. These counts and patterns can be used to characterize such things as the educational or expertise level of the writer, or even whether the writer is telling the truth—people who are lying tend to use the word "I" less often than people telling the truth.

In addition, text analytics can be used to extract data of all kinds, such as people, companies, products, events, etc., from unstructured text. That data can then be incorporated into all the structured data techniques and applications we've developed. In other words, Big Text can make Big Data even bigger! Extraction can be used as part of text mining patterns, but can also be used directly in applications such as search, where extraction is used to capture *multiple metadata values*—and which in turn support one of the successful approaches to search, *faceted navigation*.

Another set of techniques can be used to generate summaries of large sets of documents to provide a much more manageable and useful approach than having to read through them all. These summaries can range from simple general characterizations (often used in search applications) to complex one- to five-page characterizations of all the key information contained in each 200-page document.

Finally, text analytics can be used to characterize the content of unstructured text by subject matter (major and minor topics) and by positive and negative sentiment. This is often the most difficult thing to do but also the most valuable if done correctly. Frequently referred to as "autocategorization," it can be used to do far more than categorize a document.

This functionality is really the heart and brain of text analytics. It is the technique that adds depth and intelligence to our analysis of text by capturing and utilizing the context around individual words for better text mining, extraction, sentiment characterization, and all the other potential uses of text analytics. Understanding context is at the heart of the human understanding of meaning, and without it, text analytics is severely restricted in scope and value.

Is Deep Learning the Answer?

Deep learning is the current hot topic in artificial intelligence and is associated with successes like IBM's Watson. It is an approach that is also being applied to the analysis of unstructured text of all kinds.

However, the technique is actually based on neural networks, which is something that we explored starting in the '80s. The difference now is primarily scale—the huge jumps in processing speed and storage enable us to build much more complex neural networks. This scale is what has led to some major successes in AI and text analysis, but it also has some major limitations. It is very good for discovering patterns, both perceptual and linguistic, but it has not been as successful with conceptual problems.

With all the hype about deep learning, I used to think that I was one of the very few that had doubts about how far you could go with this approach. However, I just came across a very good article in the *MIT Technology Review* titled "Can This Man Make AI More Human?"[4]

The article discusses Gary Marcus, who is taking a different approach to AI by trying to apply how children learn. Children don't learn by being exposed to millions of examples and discovering the common patterns within those millions. They basically learn by developing a rule and generalizing it—and then learning how to apply it better by learning the exceptions to the rule. And the key to those exceptions is that it is not a *what* in the form of simply a list of exceptions, but a *why*, as in why the exception needs to be recognized. The article asks the question, "But is deep learning based on a model of the brain that is too simple?" And the answer is *yes*. The article goes on to state:

> These systems need to be fed many thousands of examples in order to learn something …
>
> In contrast, a two-year-old's ability to learn by extrapolating and generalizing—albeit imperfectly—is far more sophisticated … Clearly the brain is capable of more than just recognizing patterns in large amounts of data: It has a way to acquire deeper abstractions from relatively little data.

The article goes on to describe a situation that is very similar to what I found in text analytics, which is that deep-learning methods are extremely powerful (and better than human) in identifying things, like faces and even the spoken words in audio recordings, but they are not even close to human capabilities when dealing with the meaning of those words. As the article puts it:

> A deep-learning system could be trained to recognize particular species of birds in images or video clips, and to tell the difference between ones that can fly and ones that can't. But it would need to see millions of sample images in order to do this, and it wouldn't know anything about why a bird isn't able to fly.

And in this case, what is true for AI is also true for text analytics: *Deep learning is not enough by itself.* On the other hand, it is definitely a powerful technique that can be combined with other techniques, such as those found within text analytics. To continue to use the brain example, "In other words, the brain uses something like a deep-learning system for certain tasks, but it also stores and manipulates rules about how the world works so that it can draw useful conclusions from just a few experiences."

Text Analytics, Deep Text, and Context

For text analytics, a better approach is what I call *deep text* (yes, I'm somewhat stealing the term, but it's a good one). Like Marcus' approach to AI, deep text is an approach that focuses on developing rules and applying those rules to unstructured text. Of course, the rules that we develop in a deep-text approach are nowhere near as sophisticated as our two-year-old human—at least not yet, but it is a much more powerful approach than some text analytics vendors take with their overuse of example documents and statistics.

A text analytics rule might be something like this: "If you see any of the sets of words that represent a particular concept within the same paragraph or section and you don't see any of another set of words that would change the meaning, then this indicates this concept is present and is an important one for this document."

This approach focuses on rules (and the exceptions to rules) to characterize the meaning in text, as well as doing a more sophisticated job of extracting key entities and concepts. For example, if a human brain is reading a paper on pharmaceutical companies and sees the word "pipeline," it automatically (or unconsciously) knows that it means a product pipeline, not an oil and gas pipeline. Since we still don't know how to construct a human mind except the old fashioned way, we have to tell our AI and text analytics applications how to look at the context around a word and use that context to determine the meaning of that word for this instance. This context (and the rules that characterize it) needs to look at multiple contextual dimensions or layers, from the words immediately before and after "pipeline," to words in the same sentence or paragraph—and within an overall document context as well.

Context and rules are the key concepts in a deep-text approach to text analytics. In succeeding chapters, we'll dive more deeply into what this approach means and how to do text analytics in a way that goes beyond simple text mining. However, since there is really not a good overall book on text analytics, our focus will be on thoroughly understanding all the different approaches. Hopefully by the end of the book, you will have a good general knowledge of what text analytics is all about and also be able to see how a deep-text approach to text analytics is key to its success.

We will also cover everything, from how to get started in text analytics and how to develop text analytics, to the kinds of applications you can build with text analytics, wrapping it all up with the best way to approach incorporating text analytics within an enterprise.

Who Am I?

Text analytics is still a young and developing field and it is not clear what the best background to have really is (though most of the consultants I hire have library science backgrounds). The reality is people in text analytics can come from a variety of backgrounds and experiences. My particular path is one that I would not necessarily recommend to anyone else, but it worked out well for me.

My academic experience was long and varied as I was one of those professional students who really didn't want to graduate because they were having too much fun learning. When I did finally get an undergraduate degree, I had the requirements and credits for three degrees—English, philosophy, and the history of ideas. My official degree was in the field of history of ideas, in which I earned a master's and ABD ("all but dissertation").

I was hard at work writing my dissertation on a somewhat obscure German philosopher/thinker named Ernst Cassirer when I was seduced by the work that was going on in the Stanford AI program—and by their promise that they were only a couple of years away from modeling common-sense knowledge. I sold my car and a few other possessions, moved out of too-expensive Palo Alto, and bought my first computer and began to explore AI and programming.

It didn't take too long to realize that they were not two years away—or even two decades as it turned out. But it was too late, and I embarked on a career of consulting, mostly in educational software, though my first two "products" were two science-fiction computer games that I designed and programmed (they didn't sell all that well, but they did win a couple of awards). This period ended with one of the few full time jobs I've ever had programming educational software.

This was followed by working with two colleagues to start an educational training software company. Unfortunately, we launched our first product just as California was going into a major recession. End of that story.

So it was back to consulting, with more of a focus on helping companies design their approach to information, rather than programming. There was also another short full-time job with Charles Schwab organizing their intranet and developing a corporate taxonomy. This, too, fell victim to a recession.

That was followed by creating my second company (and the one that I still am working for), KAPS Group. The group was me and a number of other consultants that I met over the years. The initial focus was developing taxonomies and consulting on their use in search and other applications. While my company still does taxonomy consulting, the focus has shifted to text analytics, with an early project working with software from Inxight.

Since then, I've worked on a large number of projects for clients in business and government, during which I and my consultants have learned how to do text analytics of all kinds. I've also been speaking about text analytics at a variety of conferences, including Taxonomy Boot Camp, Enterprise Search Summit, KMWorld, Semantic Technology, and others.

In addition to talks on specific aspects of text analytics, I've also been offering a three-hour workshop on what text analytics is and what you can do with it. Since 2013, I've served as program chair for a conference devoted to text analytics, Text Analytics World.

My company now partners with many leading text analytics vendors. Having a wide range of different vendor technologies as well as working on projects from news aggregation to analyzing social media for sentiment about different phones and plans was exciting, and it also led me to realize that the field of text analytics needed something.

I decided that what the field needed was a firmer foundation, both theoretical and practical, and so I decided to write this book. There were other reasons as well, chiefly that most people I meet (outside of conferences) don't know what it is and don't have any idea of what you can do with it.

In addition, I've seen too many people that have been charged with doing text analytics—often with no training or background—do it pretty badly. And then they blame the software, or they come to the conclusion that text analytics is not worth doing, and so they drop the whole thing.

I'm distrustful of claims that some new technology will "revolutionize everything," but I do think that text analytics done well (deep text) has the potential to take the chaos and messiness of unstructured text and turn it into an incredibly valuable asset for business, government, academic research, and, who knows, perhaps even art.

Plan of the Book

The book is divided up into five parts with three chapters each.

Part 1: Text Analytics Basics

The first part, Text Analytics Basics, lays the foundation for the rest of the book. In Chapter 1 (What is Text Analytics?), I present a broad definition of text analytics that includes text mining, auto-categorization, and sentiment analysis among other features and capabilities. In this chapter, I also discuss the importance of content models (taxonomies and ontologies, etc.) and metadata as a basic component of adding structure to unstructured text. I also briefly discuss the technology behind text analytics (which we will go into more detail in Chapter 2). The chapter closes with a quick look at broad text analytics application areas, including enterprise search, social media (including voice of the customer), and a variety of applications that can be built with text analytics.

In Chapter 2, we take a much deeper look at the actual functionality on which text analytics rests. This includes text mining, but we don't go into as much depth on this functionality, both because it is significantly different from all the others and it has been well covered in other books. Text mining is more mathematical than linguistic.

Entity extraction is one of the basic functionality areas of text analytics in which software is used to basically extract data from all that unstructured text. The software is set up to extract noun phrases of various kinds, such as people, organizations or companies, and a whole range of other types of entities. In addition to simple entities, the software can also be used to extract facts which consists of sets of entities and relationships between them. For example, you might extract a person's name and then also their address and phone number. More advanced facts might include the relationships between two companies, for example, they are competitors and/or merger candidates.

Another basic functionality that we cover in this chapter is summarization in which software can be set up to summarize large documents into more manageable and general form. Currently this is an underutilized functionality, but that could very well change in the near future. A summary could be a simple two- or three-sentence, high-level summary that might be used in a search results list to take the place of the snippet (the first 100 or 200 words of the document). On the other hand, a summary could be a five- to 10-page document that captures all the essential information in large documents, enabling editors or analysts to quickly scan large collections in relatively short periods of time.

Sentiment analysis uses text analytics software to determine or capture the positive or negative sentiment being expressed in social media, such as Twitter or blogs. This has been one of the hottest areas in text analytics for the last few years, and the field has matured lately with more advanced rule-based analysis that replaced simple positive and negative vocabulary dictionaries. These early efforts were barely above chance, but it led to an explosion of companies that offered an easy development path. However, as we shall see, easy text analytics is basically a contradiction in terms—if you want meaningful results.

The last, and in many ways the most important and complex, functionality is normally called "auto-categorization." This name came from the early focus on search applications in which the software characterized the primary topics—or the *aboutness*—of documents in a search results list. However, this functionality is really at the heart of advanced text analytics, and thus can be used to make all the other pieces smarter. This functionality can be based on everything, from categorization by comparing targets with sample documents, to building sophisticated rules using each vendor's proprietary language.

In Chapter 3, we switch gears to talk about the business value of text analytics, which is often analyzed in terms of the return on investment, or ROI of text analytics. Text analytics is really a foundation technology that supports a variety of ways to get value from that most underutilized resource, unstructured text. This makes doing ROI calculations a bit complex, but still doable. In fact, when you do ROI calculations, the only real problem is the numbers are so high that it is sometimes hard to believe for many people—believe it!

We look at the basic benefits of text analytics in three main areas: enterprise search, social media applications, and a range of text analytics–based applications. We then explore the question, "If text analytics is so powerful and produces so much value, why isn't everybody doing it?" As you might expect, the answers are many and varied. We end this chapter with the discussion of how best to sell the benefits of text analytics by, in some cases, going beyond the numbers in order to make the case at the C-level. Success stories are still a great way to convince people at all levels.

Part 2: Getting Started in Text Analytics

In this part we cover how to get started in text analytics with research into both the current text analytics software market and research into your organization's information environment and needs.

Chapter 4, Current State of Text Analytics Software, covers the early history and current state of the text analytics market. This market is characterized by a number of themes, including that the market is a very fragmented one with no dominant leader. Another theme looks at the factors contributing to the growth of the market as well as the variety of applications and offerings by various vendors. We look at current trends in the market, and we also look at what the obstacles are that are holding back an even more dramatic growth in the market.

Chapter 5, Text Analytics Smart Start, discusses the issues that organizations should pay attention to when getting started in text analytics. We then discuss the methodology that my company has developed to help people do both the necessary research and an initial evaluation of which software will work best in their organization. This methodology can be applied to both enterprise text analytics and social media applications.

Finally in this chapter, we discuss the design of the text analytics team that will do the initial selection and form the basis for future development.

Chapter 6, Text Analytics Software Evaluation, discusses the unique requirements for text analytics software, which is unlike most traditional software in that it deals with that messy language stuff. The unique nature of text analytics software calls for a two-part process, which consists of a fairly traditional market evaluation, but culminates in a proof of concept (POC). The POC is the essential part of the methodology, both for making the best purchasing decision and for creating the foundation for future development. We end the chapter by looking at two example evaluation projects.

Part 3: Text Analytics Development

In Part 3, we consider how to build on the aforementioned foundation; in other words, we ask "What is the actual development process that goes into text analytics applications?"

Chapter 7, Enterprise Development, focuses on developing categorization capabilities, as that is what poses the most complex development challenges. This development starts with analyzing enterprise content and any existing content structure or taxonomies that exist within the enterprise. This preliminary phase also includes developing a powerful understanding of users and what their information needs and behaviors are. We then cover the actual development process, which typically involves a number of cycles of develop-test-refine until you are ready for prime time. The chapter also covers a major topic—maintenance and governance—which too often is treated almost as an afterthought. The chapter ends by looking at the development process for entity and/or fact extraction, which, while much simpler than categorization, has a number of specific issues and techniques.

Chapter 8, Social Media Development, analyzes what is required to do advanced sentiment analysis and other social media development processes. In this chapter we use a number of example projects, including a project that looked at the extremely varied and creative ways of expressing positive and negative sentiments in a number of social media sites dedicated to expressing our love-hate for our phones. We also cover the basic development process, which is very similar to the process for categorization, but with a number of specific differences. We end this chapter with discussing what the current major issues are in social media development.

Chapter 9, Development: Best Practices and Case Studies, adds a level of concreteness to the discussion by looking at a number of specific projects that my company has worked on and tries to capture the best practices lessons that came out of those projects. The very first project we did, a news aggregation application, was both a successful project and an incredible learning experience as we developed a number of best practices and learned which ones to discard. We then look at two enterprise projects that were very similar—while one a major success, the other a major failure. No one likes to talk about failures, but you can learn a great deal about what does *not* work in text analytics by studying them.

Part 4: Text Analytics Applications

In this part, we cover the full range of text analytics applications, or as many as we can since the number and variety of applications continues to dramatically grow. We'll cover three main areas of applications: enterprise search, which will focus on faceted navigation, a broad range of text analytics applications, or *InfoApps* as they're also referred to, and lastly, social media applications.

Chapter 10, Text Analytics Applications—Search, starts with a description of the futility of trying to make search work—*without* dealing with the whole dimension of meaning. Without text analytics, search engines are stuck with dealing with words as essentially "stupid chicken scratches."

There have been two major advances in search—aside from technical advances, which do little to improve the quality of search—Google and faceted navigation. Unfortunately, within the enterprise, Google's PageRank algorithm doesn't work—and that leaves us with faceted navigation. Faceted navigation has the potential to dramatically improve search, but it has one major flaw—it requires an enormous amount of metadata. What this means is that most successful implementations have been on commercial websites, where they could use their product catalogs to supply all the necessary metadata. There is, however, a solution for search within the enterprise, which is using text analytics to generate all the necessary metadata to make faceted navigation really function.

Enterprise taxonomies are another approach that had a great deal of potential but ran into a major problem—the gap between the taxonomy and the content. In other words, having a well-designed taxonomy is only half the solution—someone has to apply the taxonomy to documents and there the success story is much more problematic. One approach is to combine text analytics with content management to develop a hybrid of automatic and human tagging.

We conclude the chapter by looking at how text analytics can enable us to look beyond individual documents and incorporate the characteristics of an overall corpus of documents, as well as dividing up documents into smaller and more meaningful sections.

Chapter 11, Text Analytics Applications—InfoApps, discusses some of the most interesting and most valuable applications that can be built with text analytics. These applications include things like business intelligence, e-Discovery, and fraud detection, all of which incorporate text analytics to some degree. Often with these applications, the text analytics components are largely hidden, as vendors offer the application and do the text analytics development themselves.

In addition to these hidden text analytics applications, there are a number of other applications in which the text analytics components are more explicit. One area that has seen a lot of success is the analysis of documents for characterizing expertise levels to build applications that can range from knowledge management expertise location applications to enhancing HR's evaluation of potential employees.

A less dramatic, but economically powerful application is using text analytics to uncover all the duplicates and near-duplicate documents that are typically found within enterprises today. Finally, we talk about different kinds of automatic and semiautomatic summarization that can be done with text analytics. These summaries can be used to enhance the use of unstructured documents in the enterprise, both by reducing the reading burden of wading through multiple large documents (most of which have little importance), and by creating richly-structured summarizations of documents that can support new applications.

Chapter 12, Text Analytics Applications—Social Media, discusses the huge and growing number of applications that are being built to discover sentiment and other insights into customers and competitors. We start out by exploring the unique characteristics of the social media world of extremely wild, and at times incoherent text found in Twitter and other social media posts. Getting beyond the poor quality of text is one of the major challenges in this area. As with most text analytics application areas, the other main challenge is to develop more in-depth models to support more sophisticated understanding of what this text means. For example, even categorizing something as short as a tweet as either positive or negative oversimplifies the ideas and sentiment in those tweets, almost to the point of uselessness.

If done well, however, social media provides a rich environment for applications, such as customer relationship management (CRM), and in particular voice of the customer (VOC) applications. As a second-generation set of applications are being developed, they are moving beyond simple positive and negative to much deeper psychological characterizations. In addition, another area of huge value is behavior prediction in which, for example, it is possible to distinguish customers who are likely to cancel from those simply trying to get something from you.

Part 5: Enterprise Text Analytics as a Platform

In the final part, Enterprise Text Analytics as a Platform, we discuss the best overall approach to text analytics. The question is whether to approach text analytics as a series of independent applications, or to develop an enterprise text analytics platform that can support all those applications.

Chapter 13, Text Analytics as a Platform, starts with a discussion of different approaches to text analytics and comes to the conclusion that—while the application-infrastructure dichotomy is really more of the spectrum—for most organizations, looking at text analytics as a semantic infrastructure is really the best approach. We present the arguments in favor of a strategic versus a tactical approach, and/or a platform or a project approach. The key argument is that even if you tend to favor tactical and project approaches, you still need to develop an overall strategic vision of what text analytics can do for your organization in order to come up with the best decisions.

We then present a number of arguments in favor of an infrastructure platform approach. One key factor is that unstructured text is growing in size and complexity, and it is found throughout the organization, not in any one department, but everywhere. The other key argument is that we have tried the tactical/project approach for years, and it made little headway in gaining real value out of all the unstructured text within and outside the enterprise.

Chapter 14, Enterprise Text Analytics—Elements, describes what an infrastructure platform approach to text analytics would look like in most organizations. We describe what the main features of enterprise text analytics (ETA) are and introduce the concept of a semantic infrastructure. Virtually all modern organizations have developed a technical infrastructure for dealing with information, which is a necessary component, but without a semantic infrastructure, the ability to deal with information will be extremely limited. Briefly, a semantic infrastructure consists of all the elements that deal with and support the use of language or semantics for the organization. This can include taxonomies and ontologies as well as other communication models.

We then present a kind of thought experiment that describes what an enterprise text analytics department might look like and where that ETA department might be located within the organization. This chapter also describes what the basic skill requirements are for doing text analytics and the best way to obtain these skills. We also discuss who the best candidates are for text analytics training.

The chapter then expands that discussion to include who are the best and natural partners and contributors to text analytics in other parts of the organization. This chapter describes some of the backgrounds of these partners and contributors. These key backgrounds can include cognitive science as well as various language studies. In addition, training departments often can play a key role, both in developing text analytics and integrating it within the organization. Finally, we look at the contributions of IT and the data—or structured information— groups.

This chapter continues with a quick look at the technology of an ETA group. This technology model includes not only the text analytics software itself but also enterprise search, enterprise content management, and of course, SharePoint. The chapter concludes with a look at what the range of services are that this ETA group could offer to the organization, and how you might create an ETA department or group.

Chapter 15, Developing ETA—Semantic Infrastructure, dives more deeply into what a semantic infrastructure is and how it can deliver value to the organization. We start with how important an understanding of the content within the organization is, how to create a content map of what you have, and why and how it is used. We also discuss how to develop content models, including taxonomies that can power text analytics. We also discuss what the implications of text analytics are for those content models and taxonomies, specifically that neither simple, one-dimensional taxonomies nor five-level taxonomies with thousands of nodes are very useful.

Mapping the content in the organization is one key step, but the second key component is to map all the various communities, both formal and informal, within the organization. It is important to understand the information needs and behaviors of all these communities in order to develop an ETA solution. We use the metaphor of the neocortical community model in which most communication takes place within small communities, but there is also a need for some intercommunity communication (communities can be anything from a group within a department to an informal special-interest group).

The chapter concludes by looking at how to add depth and intelligence to our understanding of these communities with a cognitive deep research effort. The final point is that when dealing with something like semantics and information, it is extremely practical to pay attention to this cognitive theoretical depth.

Conclusion

In the conclusion, we review and discuss a number of the major themes of the book, including the relationship of text mining and text analytics, the importance of integration for doing text analytics—in choosing your methodology as well as the kinds of content—and some of the important themes for development and application of text analytics. We end with a look at the future of text analytics in general, and specifically how text analytics and cognitive computing can mutually enrich each other.

We also explore the idea of *deep text* as the key to doing advanced text analytics. There are three essential characteristics of deep text:

- Linguistic and cognitive depth
- Integration of multiple techniques, methods, and resources
- Platform/infrastructure

These three characteristics are essential for doing text analytics in a way that goes beyond simple word counting. They are the keys to adequately modeling the rich complexity of natural language. They are also the key to the future of text analytics, whether it will continue to be a mildly interesting set of techniques, or whether it will fulfill its potential to dramatically improve our ability to integrate unstructured text into an ever-growing range of applications.

Endnotes

1. Bureau of Labor Statistics, U.S. Department of Labor, *Occupational Outlook Handbook, 2014-15 Edition*, Database Administrators.
2. "Text mining." Wikipedia. https://en.wikipedia.org/wiki/Text_mining.
3. "Text mining and text analytics." Wikipedia. https://en.wikipedia.org/wiki/Text_mining#Text_mining_and_text_analytics.
4. Knight, Will. "Can This Man Make AI More Human?" *MIT Technology Review*, January/February, 2016. www.technologyreview.com/featuredstory/544606/can-this-man-make-ai-more-human/.

PART 1

Text Analytics Basics

CHAPTER 1

What Is Text Analytics?

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.

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

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.

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 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

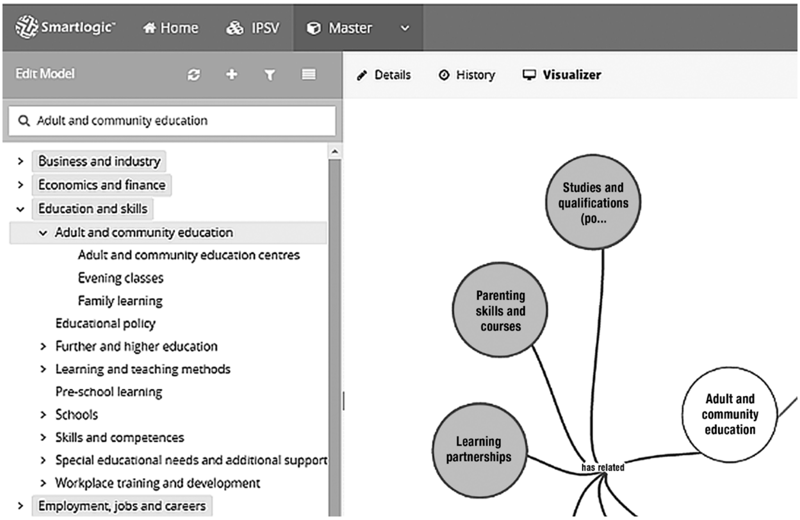

Smartlogic
Home
IPSV
Master
Edit Model
Details
History
Visualizer
Adult and community education
Business and industry
Economics and finance
Education and skills
Adult and community education
Adult and community education centres
Evening classes
Family learning
Educational policy
Further and higher education
Learning and teaching methods
Pre-school learning
Schools
Skills and competences
Special educational needs and additional support
Workplace training and development
Employment, jobs and careers
Studies and qualifications (po...
Parenting skills and courses
Adult and community education
Learning partnerships
has related

Figure 1.1 Development Environment 1

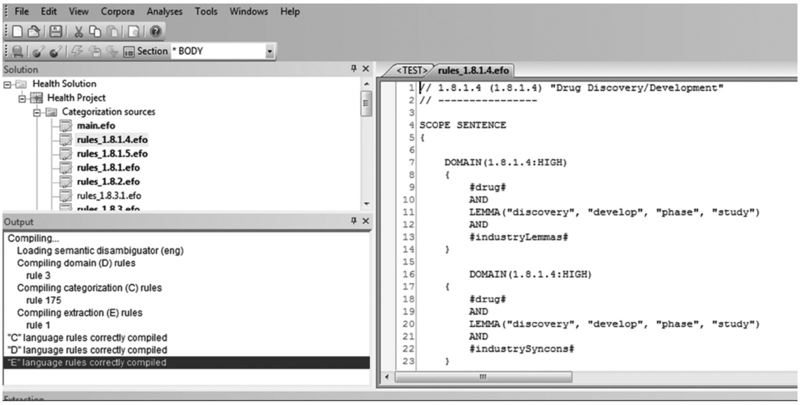
and community education" is a narrower term of "education and skills," and the phrases, "adult and community education centres" and "evening classes" are its narrower terms. Of course, each software vendor uses different terminology to refer to the various parts—and typically have slightly different 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. 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,



```
Compiling...
  Loading semantic disambiguator (eng)
  Compiling domain (D) rules
    rule 3
  Compiling categorization (C) rules
    rule 175
  Compiling extraction (E) rules
    rule 1
"C" language rules correctly compiled
"D" language rules correctly compiled
"E" language rules correctly compiled
```

```
// 1.8.1.4 (1.8.1.4) "Drug Discovery/Development"
// ----------------

SCOPE SENTENCE
{

    DOMAIN(1.8.1.4:HIGH)
    {
        #drug#
        AND
        LEMMA("discovery", "develop", "phase", "study")
        AND
        #industryLemmas#
    }

        DOMAIN(1.8.1.4:HIGH)
    {
        #drug#
        AND
        LEMMA("discovery", "develop", "phase", "study")
        AND
        #industrySyncons#
    }
```

Figure 1.2 Development Environment 2

which might, for example, analyze the overall sentiment expressed in that document set.

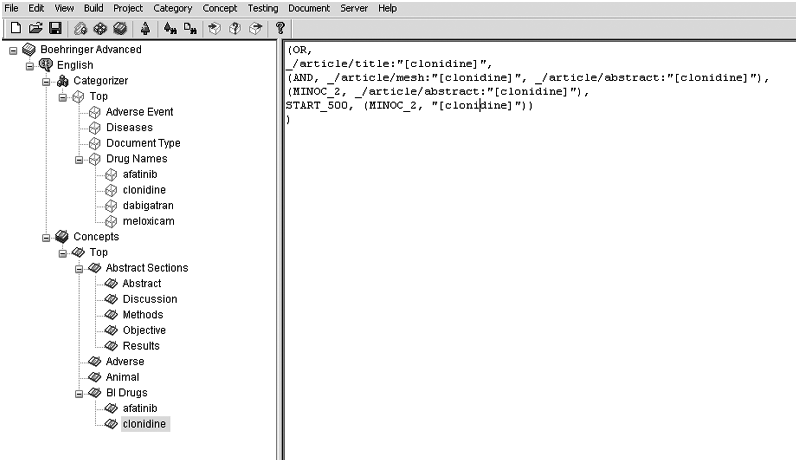
Also, these categorization rules are typically orders of magnitude more complex and sophisticated than those of almost all searchers, with the possible exception of professional librarian searchers.

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.

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.



```
(OR,
_/article/title:"[clonidine]",
(AND, _/article/mesh:"[clonidine]", _/article/abstract:"[clonidine]"),
(MINOC_2, _/article/abstract:"[clonidine]"),
START_500, (MINOC_2, "[clonidine]"))
)
```

Figure 1.3 Categorization 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

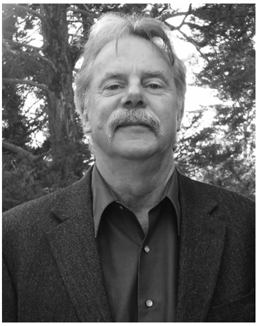


Photo by Melanie Reamy

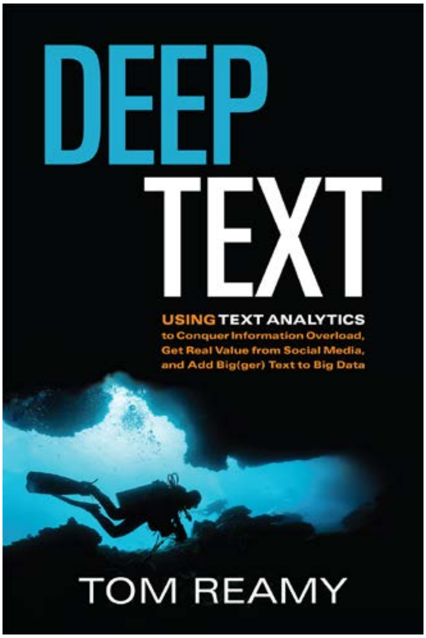
Tom Reamy is currently the chief knowledge architect and founder of the KAPS Group, a group of knowledge architecture, text analytics, and taxonomy consultants, and has 20 years of experience in information projects of various kinds. He has published a number of articles in a variety of journals and 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.

If you enjoyed reading this chapter of *Deep Text: Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Bigger Text to Big Data*, you can order it from the following online retailers.


DEEP
TEXT
USING TEXT ANALYTICS
to Conquer Information Overload,
Get Real Value from Social Media,
and Add Big(ger) Text to Big Data
TOM REAMY

amazon.com

Information Today, Inc.