# Text Analytics 2014: User Perspectives on Solutions and Providers

Seth Grimes Alta Plana

A market study sponsored by



# **Table of Contents**

Executive Summary	
Growth Drivers	4
Key Study Findings	-
About the Study and this Report	
Text Analytics Basics	7
Patterns	7
Structure	7
Metadata	8
Beyond Text	8
Applications and Markets	8
Solution Providers	
Market Progress	9
Demand-Side Perspectives	10
Study Context	
About the Survey	
The Data Mining Community	
Demand-Side Study 2014: Survey Findings	13
Understanding Survey Respondents	
Q2: Application Areas	
Q3: Information Sources	
Q4: Return on Investment	
Q5: Mindshare; Awareness	
Q6: Spending	
Q8: Satisfaction	
Q9: Overall Experience	
Q11: Provider Selection	-
Q12: Provider Likes and Dislikes	
Q13: Promoter?	-
Q14: Information Types	-
Q15: Important Properties and Capabilities	
Q16: Languages	-
Q17: BI Software Use	
Q18: Guidance	
Q19: Comments	
Additional Analysis	
Interpretive Limitations and Judgments	
Sponsor Solution Profiles	
Solution Profile: AlchemyAPI	
Solution Profile: Digital Reasoning	
Solution Profile: Lexalytics	
Solution Profile: Luminoso	
Solution Profile: RapidMiner	
Solution Profile: SAS	
Solution Profile: Teradata Aster Solution Profile: Textalytics	

### **Executive Summary**

Text analytics, applied to social, online, and enterprise data, aims to extract useful information and create usable insights for business, personal, government, and research ends. The technology is pervasive, even if not ubiquitous, which is to say that it is deployed in applications ranging in scale from device-installed to distributed, "total information awareness" data mining. Text analytics is utilized wherever big/fast text is found. And while not every analytical application directly involves text, in our emerging big-data world, every task – including analyses of "machine data" and transactional records – may be enriched by the inclusion of text-sourced information.

Text analytics has found its greatest success in four industry groupings: consumer-facing businesses, public administration and government, life sciences and clinical medicine, and scientific and technical research. Analysis of online news, social postings, and enterprise feedback is of special interest across groupings. Search-driven applications – search, advertising, e-discovery, customer service – is a crosscutting functional category, aimed at meeting front-line, operational needs.

Users seek to extract many sorts of information, with continued growing interest in automated identification of topics, entities and concepts, events, and personal attributes, coupled with feature-linked intent and sentiment – attitudes, emotions, and opinions – as well as other subjective information. User experiences with the technology and solutions remain mixed, however.

These points and more are brought out in Alta Plana's report "Text Analytics 2014: User Perspectives on Solutions and Providers." The report opens with an executive summary and includes three components: a narrative analysis of the text analytics market, findings of a market study, and sponsor solution profiles.

### The User Perspective

There is no single or typical text analytics user, application, technology, or solution. Users and uses vary by industry, business function, information source, and goal.

In 2011, we observed, "tools and solutions now cover the gamut of business, research, and governmental needs." The technology covers – or is capable of covering – the variety of human languages, text sources, information types, and industries, handling large-volume and streaming big data.

Data scientist users perform text analyses using one or more of the several available textcapable software-development or data-analysis environments. Business users naturally gravitate toward functional solutions with embedded text analytics.

One new point: Many text analytics users don't understand that they're doing text analytics, nor do they need to.

### The Provider Market

Growth in text analytics, *as a vendor market category*, has slackened, even while adoption of text analytics, *as a technique*, has continued to expand rapidly.

The 2014 market is fragmented, featuring software tools, natural language processing (NLP) services, text analysis workbenches, social analytics dashboards, integrated data analysis environments, and solution-embedded technologies. These options support the practice of text analytics, but a large, increasing segment do not place text analytics front and center. Rather, text analytics operates behind the scenes, often as just one of several analytics elements – leading to my observation about the text analytics market category.

Technology and solution providers run the gamut in size from small start-ups, to established technology and solution companies, to the largest, global information technology brands. Some offerings are tightly focused on a particular function – for instance, single-language event or

sentiment extraction – while others embed text analysis capabilities in a line-of-business solution and others offer robust business intelligence (BI) and analytics suites.

Innovation is constant, related to scale, scope, and velocity of analyses; deployment of techniques such as deep learning and unsupervised learning; availability of linguistic and semantic resources; refinement of proven, rule-based approaches; closer adaptation to industry needs; and internationalization. Business value resides in all parts of the market, even if not typically under the text analytics banner.

### **Growth Drivers**

Many of us operate on the notion that every aspect of life and business can and should be recorded, measured, and analyzed for predictive purposes and optimization – including our language-based communications. Text technologies and applications have advanced in response. Sensors, devices, and social computing have created an always-on, real-time imperative, and the availability of cloud and mobile options now facilitate adoption and deployment of new technologies. The rapid increase in computing power and data availability has enabled deployment of new algorithms that tackle challenges formerly beyond commodity computing's capacity.

Consider four technology-related growth drivers:

- **Open source** text analytics via data-acquisition, information-extraction, classification, and analytical components is stronger than ever. Open source lowers barriers both to technology adoption for researchers and more-sophisticated users, and to solution providers, who can focus on building higher-level and domain-adapted capabilities. Frameworks such as UIMA, Gate, Python, and R; Hadoop and other parallel processing technologies; and stream and graph processing, for instance, via Apache Spark and Apache Storm, are of particular interest.
- **The API economy** e.g., hosted, on-demand, via-API Web services similarly lowers entry barriers and provides enormous flexibility for adopters.
- **Data availability** creates analytics demand, and data has never been more available, whether directly collected or acquired via services such as DataSift, Gnip, Moreover, and Xignite.
- **Synthesis** will increasingly automate online commerce, customer support, healthservice delivery, and other applications as systems continue to mature and are joined by other **question answering** technologies. IBM Watson and Wolfram Alpha exemplify the synthesis trend.

There are notable market drivers as well:

- **Customer interactions:** Customer service and customer experience are strong text analytics growth drivers, deploying text analytics to data from traditional channels such as contact centers and extending coverage to move social initiatives from listening, to engagement, to service optimization.
- Omnichannel solutions: Natural language processing (NLP) capabilities are essential in competitive voice-of-the-customer (VOC) programs and for customer interaction analytics efforts that join text- and speech-derived data to operational data. Text, speech, and operational analyses, applied to survey, social media, news, warranty, chat, and voice sources, form part of *omnichannel* solutions that crunches data collected across the full set of customer touchpoints, from contact center, online, chat, social, email, and in-store interactions.
- Consumer and market insights: Text analytics application in the insights industry notably new or next-generation market research – has much in common with the interaction/experience analytics use case. Insights researchers also study VOC, most often via enterprise feedback management (EFM) programs that rely on surveys and

via social analyses. Social is increasingly viewed as a credible research source that can supplement surveys by delivering complementary insights.

- Search and search-based applications: Search has expanded far beyond enterprise and online information retrieval to provide a platform for high-value applications that include advertising, e-discovery and compliance, business intelligence (BI) in the form of unified information access (UIA), and customer self-service.
- Health care and clinical medicine: While life-sciences researchers were among the earliest text mining adopters, uptake for related areas involving mining beyond scientific literature has taken longer to advance. Analytics ranging from diagnostic systems to claims analysis have begun pacing a segment of the text analytics market.

#### The Survey

Alta Plana's 2014 text analytics market study combines a survey-based, quantitative and qualitative examination of usage, perceptions, and plans, with observations derived from numerous conversations with solution providers and users. It seeks to answer the question, **"What do current and prospective text analytics users think of the technology, solutions, and solution providers?"** Responses will help providers craft products and services that better serve users. Findings – both numerical tabulations and free-text verbatims – will guide users seeking to maximize benefit for their own organizations.

Alta Plana received 220 valid survey responses between January 18 and April 15, 2014, 193 of them during the first four weeks, when we actively publicized the survey. The 220 figure is four fewer than the 224 responses connected to the 2011 study. This document reports findings and, when appropriate, contrasts them with comparable numbers from Alta Plana's 2009 and 2011 text analytics market studies (available for free download at <a href="http://www.slideshare.net/SethGrimes/documents">http://www.slideshare.net/SethGrimes/documents</a>.)

Key Study Findings

The following are key 2014 study findings:

- The big news is not news at all: Social remains by far the most popular source fueling text analytics initiatives. Four of the top five information categories are social/online (as opposed to in-enterprise) sources:
  - blogs and other social media (61%)
  - news articles (42%)
  - comments on blogs and articles (38%)
  - online forums (36%)

Respondents chose an average of 5.6 sources, compared to 4.5 in 2011.

Direct customer feedback, in the form of customer/market surveys, rated 37%, squeezing in at fourth place. Interestingly, the percentage listing e-mail and correspondence as an information source dropped from 36% in 2009, to 29% in 2011, to 26% in 2014.

- All four **top capabilities** that users look for in a solution each garnering over 50% selection **relate to getting the most information out of sources**:
  - the ability to generate categories or taxonomies [which would include topic extraction] (64%)
  - the ability to use specialized dictionaries, taxonomies, ontologies, or extraction rules (54%)
  - broad information extraction capabilities (53%)
  - document classification (53%)

Deep sentiment/emotion/opinion extraction was chosen by 45% of respondents, down from 57% in 2011.

Low cost was important to 44% of respondents, up from 38% in 2011, but down from 51% of 2009 responses.

- Top business applications of text/content analytics for respondents are the following:
  - brand/product/reputation management (38%)
  - voice of the customer/customer experience management (39%)
  - research (38%)
  - competitive intelligence (33%)
  - search, information access, or question-answering (29%)
- Seventy-four percent of users are Satisfied or Completely Satisfied with text analytics and 22% are Neutral with only 4% Disappointed or Very Disappointed. Dissatisfaction is greatest, at 29%, with ease of use and with availability of professional services/support, with only 50% satisfied in each category.
- Only 48% of users are likely to recommend their most important provider, nearly unchanged from the 2011 figure. However, 36% would recommend against their most important provider, up from 28% in 2011.

### About the Study and this Report

Seth Grimes, an industry analyst and consultant who is a recognized authority on the text analytics marketplace – technologies, solutions, and providers – designed and conducted the study "Text Analytics 2014: User Perspectives on Solutions and Providers" and wrote this report.

The author is grateful for the support of the eight study sponsors, <u>AlchemyAPI</u>, <u>Digital</u> <u>Reasoning</u>, <u>Lexalytics</u>, <u>Luminoso</u>, <u>RapidMiner</u>, <u>SAS</u>, <u>Teradata</u>, and <u>Textalytics</u>. Their sponsorships allowed him to conduct an editorially independent study that should promote understanding of the text analytics market and of user-indicated implementation and operations best practices. The solution profiles that follow the report's editorial matter were provided by the sponsors and included with only minor editing to regularize their layout. Otherwise, the author is solely responsible for the editorial content of this report, which was not reviewed by the sponsors prior to publication.

This report opens with a text analytics and applications backgrounder that refreshes material from the previous study report.

### **Text Analytics Basics**

The term text analytics describes software and transformational processes that uncover business value in "unstructured" text. Text analytics applies statistical, linguistic, machine learning, and data analysis and visualization techniques to identify and extract salient information and insights. The goal is to inform decision-making and support business optimization.

#### Patterns

Text, images, speech, and video are all directly understandable by humans (although not universally: any given human language – English, Japanese, or Swahili – is spoken by a minority of people, and not everyone recognizes a Beethoven symphony in a sound file or Nelson Mandela in a photo). To benefit from the power of automation – to create machine understanding of human communications – we need three software capabilities:

- 1. the ability to recognize small- and large-scale patterns
- 2. the ability to grasp context and, from context, to infer meaning
- 3. the ability to create and apply models

Statistics provide a text-processing starting point, applied to detect words and terms and count their frequencies to infer message or document topics. We can create categories and classify text (a form of modeling) based on notions of statistical similarity. Statistical co-occurrences suggest associations.

Other steps take advantage of the linguistic structure of text – e.g., word form ("morphology"), arrangement (grammar and syntax), and lexical chains (word sequences), as well as higher-level narrative and discourse. Usage may be correct (as judged by editors, grammarians, and linguists) or not, whether the language is spoken, formally written, or texted or tweeted. The most robust technologies deal with text in the wild. We apply assets such as lexicons of "named entities"; part-of-speech resolution that can help identify subject, object, relationship, and attributes; and "word nets" that associate words to help in disambiguation to determine the contextual sense of terms.

Yet, in the words of artificial-intelligence pioneer Edward A. Feigenbaum,

"Reading from text, in general, is a hard problem because it involves all of common sense knowledge. But reading from text in structured domains, I don't think is as hard."

Reading implies not just processing, but also deeper understanding. Application of knowledge representations such as ontologies is one approach to deeper understanding, as is co-joining text to the metainformation and complementary profile, behavioral, and transactional data present in structured domains.

#### Structure

The text analytics process aims to generate machine-processable structure for source text or for text-extracted information. Natural language processing (NLP) outputs are typically expressed in the form of document annotations – that is, through in-line or external tags that identify and describe features of interest. These annotations and representations may or may not be visible to the system user.

Outputs may be mapped into machine-manageable data structures – such as key-value, graph, or relational database records – or distributed across HDFS nodes, whether represented in tabular, XML, JSON, RDF, or other form. (HDFS is the Hadoop Distributed File System. XML is Extensible Markup Language. JSON is JavaScript Object Notation. RDF is the World Wide Web Consortium's Resource Description Framework.)

RDF-represented data from text may form part of a linked data system, keyed by uniform resource identifiers (URIs). Text and text-derived information stored in a relational database or

HDFS may become part of a business intelligence system that jointly analyzes, for instance, DBMS-captured customer transactions and free-text responses to customer-satisfaction surveys. And text-extracted features – such as entities, topics, dates, and measurement units – in an Apache Lucene/Solr or other index may form the basis of an advanced semantic-search system.

#### Metadata

Metadata describes data properties that may include the provenance, structure, content, and use of data points, datasets, documents, and document collections. Content-linked metadata typically includes author, production and modification dates, title, topic(s), keywords, format, language, encoding (e.g., character set), rights, and so on. The *metadata* label extends to specialized annotations such as part-of-speech and data type.

Metadata may be created as part of content production or publication (for instance, the save date captured by a word processor, a geotag associated with a social update, camera information stored in an image file). It may be appended (for instance via social tagging, of people and of topics via hashtags), or extracted from content via text/content analysis. Whether stored internally within a data object (for instance via RDFa, rNews, or other microformats embedded in a Web page) or managed externally in a database or search index, metadata is fuel for a range of applications.

### **Beyond Text**

Beyond-text technologies for information extraction from images, audio, video, and composite media are advancing rapidly. Speech analysis technology – supporting indexing and search using phonemes capable of detecting emotion in speech – and transcription software is now widely adopted for contact center and others applications that include intelligence. Intelligence, along with consumer and social search, motivates work on image analysis, as do marketing and competitive-intelligence related studies of online and social brand mentions and use. Video analytics extends both speech and image analysis, with an added temporal aspect for security applications and for potential business uses such as study of customer in-store behavior.

Deep learning (multi-level machine learning), enabled by the availability of cheap, high-capacity computing resources, is being applied to model and discern features in the beyond-text media. Additionally, metadata is critically important for beyond-text media, as it is for text.

### Applications and Markets

Business users naturally focus on financial return and other forms of return on investment (ROI). They will find that text analytics has a place, that it can deliver positive ROI in any business domain, for any business function, and within any technology stack that involves significant text volume and velocity, as well as a variety of languages and sources.

Consider this telling quotation by Philip Russom of the Data Warehousing Institute from his 2007 report "BI Search and Text Analytics: New Additions to the BI Technology Stack" (http://tdwi.org/research/2007/04/bpr-2q-bi-search-and-text analytics.aspx),

"Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before."

TDWI's 2007 report recognized a role for text analytics in environments that had previously focused exclusively on capture and analysis of transactional data. That recognition now extends beyond TDWI's IT and data analyst audience. It extends to line-of-business staff handling functions that range from customer engagement and competitive intelligence to counterterrorism and medical diagnostic systems. And businesses now understand that it is not enough simply to know more. They need to *operationalize* knowledge gained to rework business processes in ways that transform insights into ROI.

Text analytics particulars – information sources, insights sought, processes, and ROI measures – will vary by industry and application.

### Solution Providers

The solution-provider spectrum features, as it has for years, constant emergence of new startups, growth (and occasional demise) of established vendors, and continued investment by software giants that have built or acquired text technologies. Open-source remains an attractive option, for both project use and as the technology foundation for commercial products and solutions.

The market continues to favors as-a-service analytics, whether in the form of online applications, cloud provisioned, or provided via Web application programming interfaces (APIs). This shift makes sense.

- The most in-demand information sources are online, social, and in the cloud.
- Use of as-a-service, cloud, and via-API applications means low up-front investment, faster time-to-use, and pay-as-you-go pricing without IT involvement.
- Certain providers offer as-a-service access to both historical and current data at attractive costs given the buy-once, sell-many-times economies they enjoy.
- Modern applications are designed to draw data via APIs, facilitating applicationinclusion of plug-in text and content analytics capabilities.

There is every expectation that the solution-provider market will continue to evolve to keep pace with user needs and broad-market business and technical trends.

### **Market Progress**

The 2011 market study that preceded this one included a text analytics market-size estimate of \$835 million globally for 2010, projecting sustained annual growth in the 25-40% range. (See *Text Analytics Demand Approaches* \$1 *Billion*, published May 12, 2011, in <u>InformationWeek</u>, <u>http://ubm.io/1tknWzg</u>.) This growth rate anticipated continued expansion of both a core text analytics market and a market for business solutions centered on text analysis, for instance for customer service/support and market research.

Given the variety of uses and diffusion of the technology in myriad directions, the 2014 market size surely exceeds the \$2 billion figure suggested by applying a compounded 25% annual growth rate to the 2011 figure. Producing an accurate estimate would be quite difficult, however, and of little purpose. It would be only suggestive of the true business value generated, which surely totals many times the \$2 billion figure.

### **Demand-Side Perspectives**

Alta Plana's "Text Analytics 2014: User Perspectives on Solutions and Providers" is built around the third instance of a survey run previously in 2009 and 2011, designed to collect raw material for an exploration of key text analytics market-shaping questions:

- What do customers, prospects, and users think of the technology, solutions, and vendors?
- What works, and what needs work?
- How can solution providers better serve the market?
- How will companies expand their use of text analytics in the coming years? Will spending on text analytics grow, decrease, or remain the same?

Current and prospective text analytics users wish to learn how others are using the technology. And solution providers, of course, need demand-side data to improve their products, services, and market positioning to boost sales, to better satisfy customers, and to stay on top of evolving needs. The Alta Plana study therefore has three goals:

- to raise market awareness and educate current and prospective users;
- to collect information of value to solution providers, both study sponsors and nonsponsors; and
- to understand trends and project future developments.

Survey findings, as presented and analyzed in this study report, provide a measure of the state of the market, a benchmark. They are designed to be of use to everyone who is interested in the commercial text/content analytics market.

### Study Context

The author has previously explored technology and market questions in numerous papers and articles. These range from *The Word on Text Mining*, published in December 2003; to white papers created for the Text Analytics Summit in 2005, *The Developing Text Mining Market*, and 2007, *What's Next for Text*; to yearly report articles, *Text Analytics in 2012*, *Text Analytics in 2013*, and *Text Analytics in 2014*. (Articles, papers, and reports are available via links at http://www.altaplana.com/writing.html.)

A systematic look at the demand side provides a good complement to provider-side views and to vendor- and analyst-published case studies, including the author's own. This understanding motivated the 2009 and 2011 studies <u>Text Analytics 2009</u>: <u>User Perspectives on Solutions and Providers</u> and <u>Text/Content Analytics 2011</u>: <u>User Perspectives on Solutions and Providers</u>.

### About the Survey

There were 220 valid responses to the 2014 survey, which ran from January 18 to April 15, 2014, 193 of them in the first four weeks, during which time we actively publicized the survey. (Contrast with 224 responses to the 2011 survey and 116 responses to the 2009 survey.)

#### Survey invitations

The author solicited responses via

- e-mail to the American Society for Information Science, BioNLP, ContentStrategy Corpora, Data Mining, Lotico, SentimentAI, SLA Knowledge Management Division, and TextAnalytics lists and the author's corporate list;
- invitations published in electronic newsletters: CMSWire, KDnuggets, and LT-Innovate;
- notices posted to LinkedIn forums and Facebook groups and on Twitter; and
- messages sent by sponsors to their communities.

#### Survey introduction

The survey started with a definition and brief description as follow:

Text analytics/content analytics is the use of computer software or services to automate:

- annotation and information extraction from text entities, concepts, topics, facts, events, and attitudes;
- analysis of annotated/extracted information;
- document processing retrieval, categorization, and classification; and
- derivation and communication of business insight from textual sources.

This is a survey of demand-side perceptions of text technologies, solutions, and providers.

Please respond only if you are a user, prospect, integrator, or consultant – or a manager or executive of staff in those roles. Researchers and developers who apply technologies/solutions to data analysis problems are also welcome to respond, but the survey is not intended otherwise for solution providers.

There are 21 questions. The survey should take you 5-10 minutes to complete.

For this survey, text mining, text data mining, content analytics, and text analytics are all synonymous.

I'll be preparing a free report with my findings. Thanks for participating!

Seth Grimes (grimes@altaplana.com, +1 301-270-0795)

The introduction ended with the text:

Privacy statement: This survey records your IP address, which we will use only in an effort to detect bogus responses. It is your choice whether to provide your name, company, and contact information. That information will not be shared with sponsors without your permission, and if shared with sponsors, it will not be linked to your survey responses.

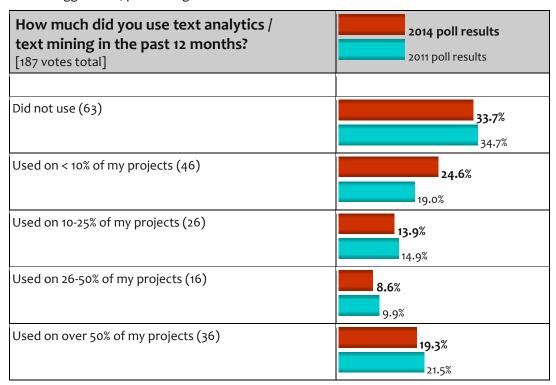
#### Survey respondents

The survey attracted a high proportion of experienced text analytics users: 85% of respondents who answered Q1, "How long have you been using Text Analytics?" are using or have used the technology (n=220 total responses). The other 15% are nonusers or are still evaluating.

By contrast, 68% who replied to Q7, "Are you currently using text/content analytics?" answered Yes (n=198). The same question had 78% Yes responses (current users) in 2011. I conclude that the 2014 respondents included many experienced, non-current users, who may of course return to the technology for later projects.

### The Data Mining Community

A February 2014 KDnuggets poll provides another data point (<u>http://bit.ly/toyHwRJ</u>). KDnuggets publisher Gregory Piatetsky reports that 66% of respondents say they used text analytics on projects in the preceding year (n=187, versus n=121 for a poll run in 2011 that found 65% use). Only 28% use text analytics on more than a quarter of projects. Piatetsky observes, "[T]he results don't show much change [since 2011] in the adoption of text analytics among KDnuggets readers."



The KDnuggets 2014 poll findings:

### Demand-Side Study 2014: Survey Findings

The subsections that follow tabulate and chart survey responses, which are presented without unnecessary elaboration.

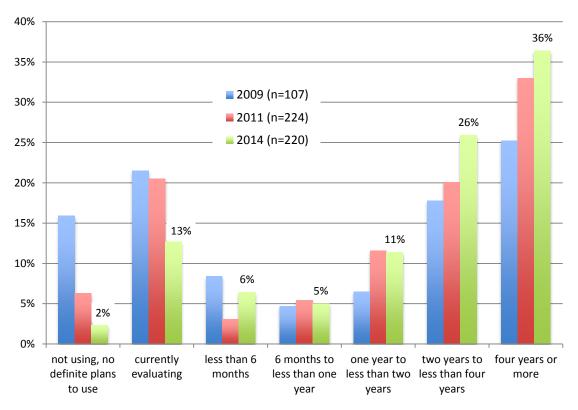
### **Understanding Survey Respondents**

Responses to questions 1, 10, and 20 help us understand survey respondents.

### Q1: Length of Experience

As in previous years, 2014 survey opened with a basic question –

### How long have you been using text/content analytics?



We see that 2014 responses skew to longer experience than was measured in 2011 and still longer than was recorded in 2009. Survey results were not based on a scientifically designed or measured population sample in any of the survey years, but because roughly the same outreach channels were used for all three years, I infer support for the view that text analytics, as the label for a market category, is growing much more slowly than solution-embedded uptake of text analytics technologies.

In any case, Q1 responses will prove illuminating in analyses of subsequent survey question, in studying how attitudes vary by length of text analytics experience.

### Q10: Providers

Question 10 asked, "Who is your provider? Enter one or more, separated by commas, most important provider first." Note that the survey asked, "Please respond only if you are a user, prospect, integrator, or consultant." There were 63 response records for Q10, listing providers (sorted and without counts):

AlchemyAPI, Attensity, BERCA Translator, Cirilab, Clarabridge, Continuum Semantic Technologies, Crimson Hexagon, GATE, IBM, Lex, Lexalytics, Leximancer, MALLET,

Medallia, Megaputer, MotiveQuest, NLTK, Natlanco bvba, NetValue Reeltwo, Nuance, ORNL, OdinText, OpenNLP, R, RapidMiner, SAP, SAS, Semantic-Knowledge, Semantria, Sematext, ServiceTick, Smartlogic, Stanford NLP, TEMIS, TEXTPACK, TextOre, UIMA, in-house, open source

GATE, Lex, MALLET, [Python] NLTK, OpenNLP, R, Stanford NLP, and UIMA are open source, and RapidMiner licenses older releases as open source.

The most notable provider not listed is HP Autonomy. Several notable providers that focus on life sciences, financial markets, and military/intelligence were not mentioned. They include Basis Technology, Digital Reasoning, Expert System, Linguamatics, NICE, Recorded Future, SRA, and RavenPack. Social intelligence providers not mentioned include Kana (Verint), NetBase, and Sysomos (MarketWired).

### Q20: What country do you work in (your base)?

Q20 is new to this year's survey and elicited 121 valid responses. In 79 other cases, the county could be found via IP address lookup. The tabulation shows a split between Europe and North America, with very few responses from Asia beyond the 7.5% from India.

Country	Count	Percentage
Australia	5	2.5%
Brazil	2	1.0%
Bulgaria	2	1.0%
Canada	6	3.0%
Croatia	2	1.0%
Czech Republic	4	2.0%
France	7	3.5%
Germany	10	5.0%
Greece	1	0.5%
India	15	7.5%
Iran	1	0.5%
Ireland	2	1.0%
Japan	4	2.0%
Malaysia	1	0.5%
Pakistan	2	1.0%
Poland	1	0.5%
Portugal	3	1.5%
Russia	4	2.0%
Spain	11	5.5%
Sweden	1	0.5%
Switzerland	1	0.5%
USA	93	46.5%
United Kingdom	22	11.0%

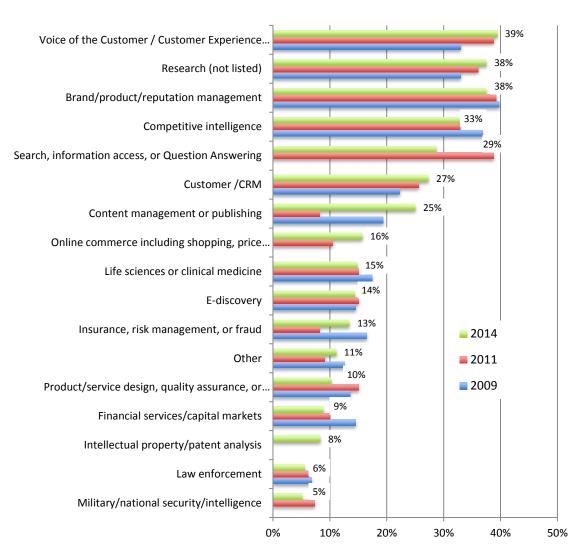
### **Applications and Sources**

The next series of responses describes information types and sources, regardless of tool choices.

### **Q2:** Application Areas

The 216 Q2 respondents chose a total of 748 primary applications – 725 categorized and 24 other – with an average of 3.4 primary applications per respondent. While there is some category overlap, it is notable that respondents are applying text analytics to multiple business needs.

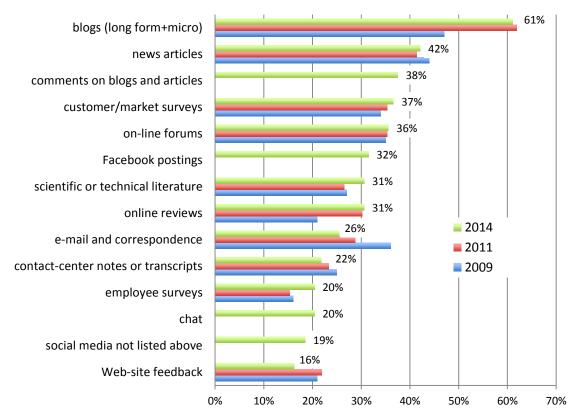
### What are your primary applications where text comes into play?



### Q3: Information Sources

The 216 respondents chose a total of 962 textual information sources, with an average of 5.6 sources per respondent, up from 4.5 in 2011. The big news is not news at all: Social sources are by far the most popular and six of the top eight categories – of the categories chosen by at least 30% of respondents – are social/online (as opposed to in-enterprise) sources.

### What textual information are you analyzing or do you plan to analyze?



### Q4: Return on Investment

Question 4 asked, "How do you measure ROI, Return on Investment? Have you achieved positive ROI yet?" There were 170 responses. Results are charted from highest to lowest values of the sum of "measure," "achieved," and "plan to measure."

Out of 170 respondents, 42% (n=71) report that they have achieved positive ROI according to some categorized measure. (*Other* responses are excluded from this count.) That figure is an increase from 38% in the 2011 survey. Those 71 respondents reported achieving ROI according to a total of 201 measures, that is, 2.83 ROI-achieved measures for each respondent who achieved positive ROI.

Out of 170 respondents, 29% (n=49) are measuring ROI but have not yet achieved positive ROI according to any measure.

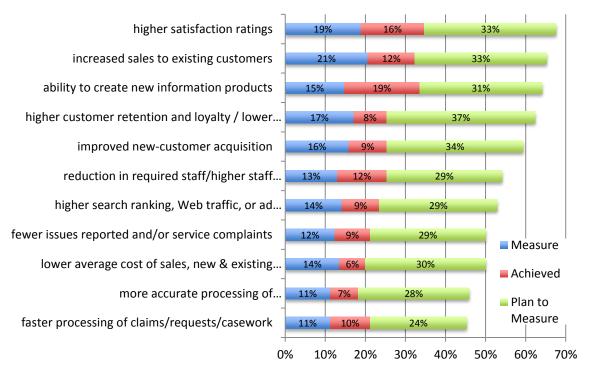
The 120 respondents who are measuring ROI (whether achieved or not) track a total of 477 measures, 3.98 measures per respondent.

The following are several of the Other responses, paraphrased:

- To obtain accurate data about respondents' impressions.
  - To avoid fraudulent sellers from social media exchanges.
- To confirm research strategies.
- To decrease communication time, increased communication accuracy, increase fit of

group membership.

- To achieve higher predictive precision and recall.
- To increase knowledge sharing and collaboration across the enterprise.
- To determine level of trust in online product reviews.
- To learn about new product innovation.
- To learn about new research discoveries.
- We're innovators. ROI only works in mature markets. We measure by competitive rankings and delivery consistency.



#### How do you measure ROI, Return on Investment?

### Q5: Mindshare; Awareness

A word cloud – we generated the one below at *Wordle.net* – seemed a good way to present responses to the query, "Please enter the names of companies that you know provide text/content analytics functionality, separated by commas. List up to the first eight that come to mind." There were 141 responses, many offering several companies. A bit of data cleansing was done to regularize names and remove inappropriate responses.



Respondent awareness of Lexalytics and of Semantria, which offers an as-a-service feature, via-API, grew significantly from 2011 to 2014, as did awareness of RapidMiner. Below is the 2011 word cloud (deliberately rendered smaller than the 2014 cloud, with no attempt to create sizing consistency between the two clouds) based on 129 response records:



And the 2009 cloud based on 48 response records:

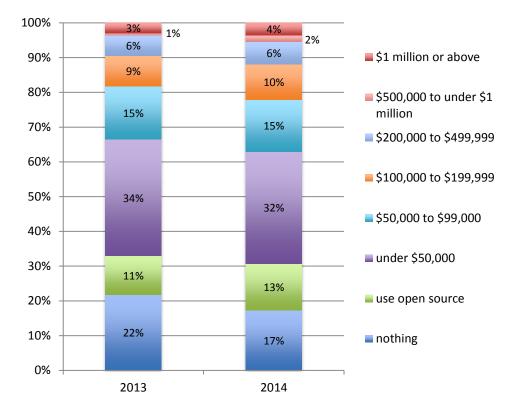


Note that IBM acquired SPSS in mid-2009. As part of 2011 and 2014 data preparation, SPSS and IBM mentions were consolidated.

# Q6: Spending

Question 6 asked about past-year spending and current-year expected spending.

# Amount spent in 2013 and amount of expected 2014 spending on text/content analytics



### Questions asked of only current text analytics users

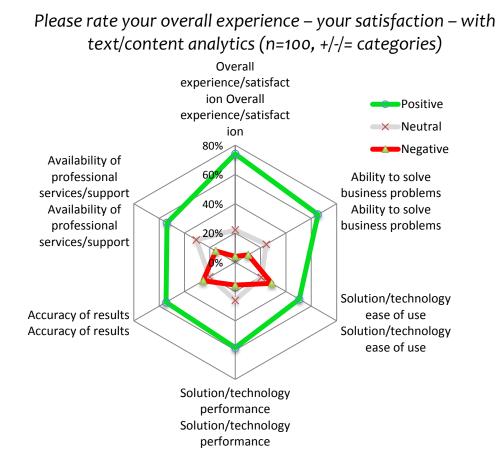
Questions 8 through 13 were posed exclusively to current text analytics users, to the 68% of the 198 respondents to Q7: Are you currently using text/content analytics, directly or as an essential part of a business application?

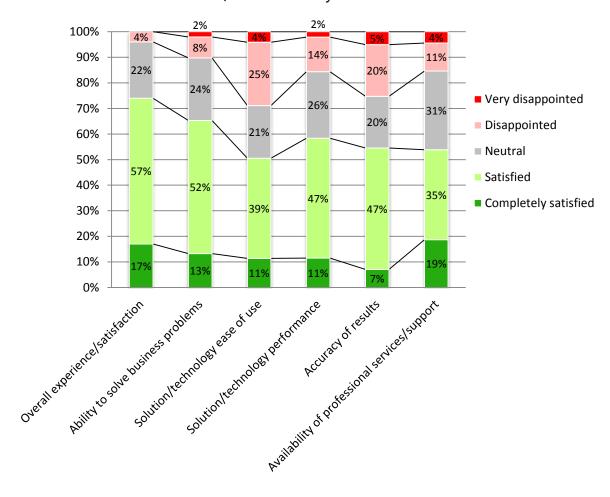
### Q8: Satisfaction

Question 8 asked, "Please rate your overall experience – your satisfaction – with text analytics." It offered five categories, listed here with response counts:

- Overall experience/satisfaction (n=100).
- Ability to solve business problems (n=100, 2 No experience /No opinion).
- Solution/technology ease of use (n=98, 1 NE/NO).
- Solution/technology performance (n=98, 2 NE/NO).
- Accuracy of results (n=99, 0 NE/NO).
- Availability of professional services/support (n=98, 7 NE/NO).

Overall, 74% of current-users respondents who had an opinion reported themselves *Satisfied/Completely Satisfied* even while the breakout-category counts totaled 65%, 51%, 58%, 55%, and 54% *Satisfied/Completely Satisfied*. Responses, excluding *No Experience/No Opinion* (NE/NO) numbers from the totals, are first reduced to *Positive/Neutral/Negative*, then at five levels.





Please rate your overall experience -- your satisfaction -- with text/content analytics

### **Q9: Overall Experience**

Question 9 asked, "Please describe your overall experience – your satisfaction – with text analytics." The following are 48 from among the 56 responses, categorized, lightly edited for spelling and grammar, and with the names of three products masked:

#### Satisfaction

Satisfied with the current state of solutions, but expecting it to evolve and improve in future.

We get very good results from huge amount of data. We are very satisfied.

It is a messy business, but invaluable if there is no other information available.

Very satisfied in my areas of application, but there is still some work to do making products user friendly whilst coping with complex tasks.

It gives as an overview of the data that we could not achieve without it.

Has led to dramatic findings in investigations, faster than previous manual/investigatorcentered methods. Frustrations stem from relative difficulty with using tools, most of which are self-developed and script-based, with equal frustration with the trade off, spending \$100k+ annually with a vendor.

I have been doing text analytics since 1984, and I have yet to find an environment that meets my requirements for knowledge extraction.

From a consultant perspective: High level of satisfaction, undervalued by most business functions, yet provides great insight and regular outcomes compared to many current ad hoc approaches that businesses deploy (word counting, manual reviews).

When applied properly and when its limits are understood, it works quite well.

Good; need more knowledgeable people.

With access to proper info, I can generate a PhD level analysis in one day.

It has been a learning process, but we get a lot of value.

I am a researcher, so I am never satisfied.

It is critical to my work so I look for tool flexibility and have been generally satisfied.

#### What Works

There is no [single] top solution; we have to mix approaches and different tools to achieve our goals.

We annotate incoming text against our taxonomy and then use the annotations as the basis of text analytics as well as search. We are currently generally satisfied, although we are trying to continuously improve.

Text/content analytics is quickly becoming an integral part of our overall analytics strategy. As with any "adolescent" technology, there is no single end-to-end product that finds, analyzes, and visualizes all available data sources.

We use our own technology, our own R&D. So we're continuously improving things whenever we're not satisfied. However, much still can be improved.

#### Accuracy

Overall maturity of the technology is increasing. However a high effort is still required in order to achieve acceptable accuracy. Acceptance mainly depends on accuracy. Lack of off-the-shelf libraries for specific domains /industries /use cases that can be used across different vendors.

Accuracy needs improvement. Tools need to be customized to specific business cases.

The costs, the shortcomings with accuracy, and the time needed to build and refine data dictionaries are frustrating at times.

Bad accuracy.

It's as good as it can be right now. With the nature of language, text analytics will only ever be 80% accurate. Still need a human to interpret context, inference, etc.

#### **Product notes**

Software with the best capabilities is very awkward to use.

We use <product> in order to enable our users to read entire articles without having to leave the site. Our users love the improved reading experience and distraction free presentation of long form content. This wouldn't be possible without <product>.

We have an in-house developed tool. We tried a few others but we find ours simpler.

We are a vendor providing analytics over NLP capabilities. After trying out a few NLP engines/APIs, we ended up writing our own classifiers to get what we wanted. Gave us a first hand impression of how far NLP providers have to go before catering to client needs.

Some lack enterprise support and integration for industry best practices; they are more R&D suited. Capability and new features have grown at a good pace.

Great technology but the solution providers are disappointing

We currently use a very specialized point solution for social media monitoring. We are interested in and considering a project to evaluate solutions that would analyze/visualize news and other published content for competitive analysis.

There is always some analytic that the commercial service does not include that is necessary for a library; too much customization needs to be done. But SharePoint has a detailed analytic system that is satisfactory.

After years of investment and additional development on our end, the service is efficient and highly accurate. We are very dissatisfied with the products on the market, and so, are stuck with our current solution.

Powerful tool, but poor user interface makes me less confident about championing results.

#### **Observations**

Positive outcomes heavily depending on good understanding of tools and appropriate/correct setup. Some tools are difficult to use for non-specialists, whilst [specialists] have a less appropriate charging model.

The tools are hard to configure and take a good bit of research and testing before they yield results.

It is (relatively) easy to apply algorithms. It is difficult to assess the accuracy of the results or to translate them into strategic insight.

NLP and computer-based sentiment and text analysis will always be flawed. Language is infinitely creative and emotion only exists within the central nervous system of the reader/observer – not the words themselves. Crowd-sourced solutions for sentiment work well. Everything other solution is so bad they're not worth using.

Text analytics have come a long way but have a way to go. It is still difficult to create taxonomies and there is still a diminished level of accuracy when compared to quantitative data.

It requires an MSc in computing science with NLP courses (or equivalent).

It's not about the text analytics; it's about what the client reasonably expects that [analysis] can do and how you interpret the results that make [analyses] valuable.

OK for structured – not for unstructured

Exciting times.

#### **Futures**

Good for future.

An emerging field with enormous potential.

Text content analytics is in its early infancy, and there is a long road ahead.

It has been good information, however the market needs deeper insights and more models.

### Q11: Provider Selection

Question 11 asked, "How did you identify and choose your provider? (If more than one, limit response to your most important provider.)" The following is a selection of responses.

#### Noncompetitive

Limited choice: We went with the first (and only) one we could find.

I'm not sure. My co-founder stumbled upon <vendor> somehow; we reached out to them, and we've had a great working relationship ever since.

Google search led to provider's site where we experimented with their free API.

Online search, a trial phase to see if fits our business processes.

Through the Web.

Internet research.

Previous partnership.

Recommendation.

Word of mouth.

Past experience by developers.

Client choice.

Other providers.

Is a close company, with several common partners.

There are only a few companies providing real solutions in this space, so there is not as much choice as I would like.

#### Competitive

They were best of show in contextual analytics when we completed our comparative analysis a few years ago.

Market analysis of service providers taking into account price, service support, sustainability, and ease of use.

Data test/"bake off."

Running trials for feature set and performance then comparing the results.

Exhaustive search, vendor vetting, RFP.

Evaluation based on the following criteria: price/license model; service level in Australia; ability to integrate with other analytics tools, e.g., Tableau.

We researched white papers and networked with other users before asking companies to do onsite demos.

Comparison shopping for suitable product.

An RFP before I started. I was brought on to work on a better implementation with <vendor>.

RFP process followed by proof of concept.

Recommendations, review, and free trial.

Analysis off the market with ROI.

They were the largest and best provider in their class at the time.

Depends on client environment.

#### **Built Our Own**

Forced to write our own.

Have built our own product.

Due diligence, demos, discussions with vendors, more due diligence, etc. In the end, our owndeveloped tools do the same job as the vendor tools, but not as compliant with Daubert standards when in litigation environment.

Hired and developed.

It is open source, and I am now the provider.

We have our own tool, which gives us better control.

And from Q11, what criteria do respondents apply in their choices?

### **Choice Criteria**

Price, quality.

Accuracy, the best sentiment methodology, flexible within our system

Availability of data. Popularity of service among users.

Free bundled analytics.

Combination of functionality and price. We needed a flexible, powerful engine we could incorporate into existing processes. One of the mail goals was to replace manual coding of survey responses with automated coding in order to get less expensive, scalable results.

Experience in the industry.

Based on client licenses and use-case.

Depends on client environment.

Experience in semantic publishing implementation.

User experience.

### Q12: Provider Likes and Dislikes

Question 12 asked, "What do you like or dislike about your solution or software provider(s)?" Respondents were allowed to enter up to five points. Forty-seven individuals responded, entering a total of 132 points. A subset are reduced and classified here into four categories: *Tech/Solution Likes, Tech/Solution Shortcomings, Provider Likes, and Provider Shortcomings.* 

#### **Tech/Solution Likes**

Easy to use. (multiple mentions)

Accuracy of results. (multiple mentions)

Flexible/Very flexible. (multiple mentions)

Sentiment analysis. (multiple mentions)

Powerful. (multiple mentions)

Works with a variety of source data. (multiple mentions)

Powerful, state of the art.

Ease of customization.

Love the use of Wikipedia to group, [via] concept matrix.

Automated theme detection.

Flexible APIs.

Ontology-based analysis.

Data prep functionality.

Easier to integrate.

Configurable.

API source code included for modification and customization.

Cleaner data.

Availability of trained models.

Strong analysis.

Extremely multilingual.

Presents results in easy to digest format.

Advanced modeling functionality.

Data aggregation.

Visual analytics.

Availability of source data for trained models.

Vocabulary maintenance is not difficult.

Easy to integrate into another system.

More sophisticated analytical frameworks.

[Supports] comparative analyses.

Better than average abilities to customize ontologies.

#### **Tech/Solution Shortcomings**

Lack of disambiguation.

Difficult user interface, requires high involvement of analysts.

Horizontal focused APIs or capabilities, too general to be useful.

Length of time it takes to query very large data sets.

Steep learning curve.

Not focused on terminology management.

Lack of transparency of the solution, i.e. hard to customize.

Poor user interface.

Required data was not available in a package to download, and I had to collect data myself.

Visualization (or lack thereof).

Needs time to adjust.

Perceived shortfall in identifying negators.

Still need human researcher to make sense of or contextualize much of the data.

Don't like the need for a [Java Virtual Machine].

Sentiment accuracy is poor.

Lack of ability to go deep.

Requires senior programmers.

Limited functionality (e.g., machine learning for entities).

Could improve metrics.

[Need for] manual tuning.

Some of the required analytics cannot be accommodated.

Lack of accuracy.

Dreadful user interface.

Clumsy outputs unless integrated with something else, e.g., a BI tool.

No easy way to do document deduplication.

Few add-on features.

Some sub-technologies still need to get more mature.

Runtime performance (in terms of time and memory).

Lack of means to create good automatic reports.

Dislike cumbersome process (no graphical user interface).

#### **Provider Likes**

It is free/cheap.

Flexibility of in-house.

ROI.

Availability of source code.

Excellent customer service.

User community support

Customer support.

Low cost.

Small company.

Access to the provider.

Training courses offered.

Frequent upgrades.

High degree of service and flexibility regarding licenses, integration with other tools, training.

Large community to rely on.

Friendly approachable people.

Documentation, blog have good examples and extensive information.

Low initial cost.

Responsiveness of support.

**Provider Shortcomings** 

Lack of tech support.

Price and high level entry cost can't be justified to businesses [given that text analytics is] only one part of many to improve satisfaction, risk, etc.

Inability to provide bug-free updates.

Lack of online support material.

More training videos needed.

Two separate solutions [are required for] social media and surveys rather than one.

It is hard to stay current.

Expensive.

[Provider] bought out [another] but treats it as a low-priority product with little investment.

Price.

Flexibility of data dictionaries is poor.

The costs are high [solution and tech providers].

Recent price increases for non-English support.

Costing model not appropriate for our use.

[Provider] is moving most of its tools to... a more expensive platform.

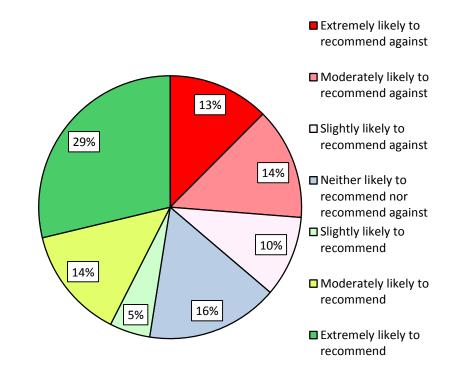
Annual price tag.

### Q13: Promoter?

Question 13 is a basic net-promoter type question, without the "net" part: "How likely are you to recommend your most important provider to others who are looking for a text/content analytics solution?"

#### Of 80 responses, 48% were positive, 16% were neutral, and 36% were negative.

How likely are you to recommend your most important provider?

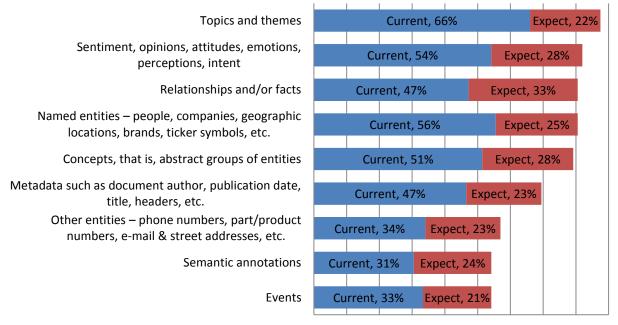


Promoters outweigh detractors by a 12 percentage points.

## Q14: Information Types

Information extraction is a key text/content analytics capability, so Question 14 asked, "On the technical front, do you need (or expect to need) to extract or analyze –" with a response count of 144. Responses are charted alongside 2011 (n=138) and 2009 (n=80) responses.

### Do you currently need (or expect to need) to extract or analyze -



 $0\% \quad 10\% \quad 20\% \quad 30\% \quad 40\% \quad 50\% \quad 60\% \quad 70\% \quad 80\% \quad 90\%$ 

There were eleven Other responses. They included

- synonyms
- genres, forms
- parts of speech with triples extraction
- emoticons
- motivations
- business related classes, e.g., public/private
- social and conceptual networks
- lexical variation

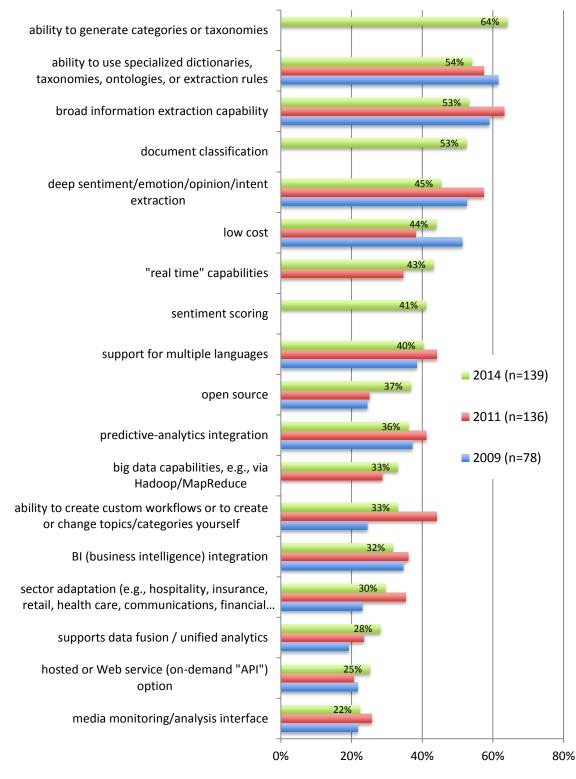
### Q15: Important Properties and Capabilities

For Q15, the 2011 and 2009 response-choice options were retained to create comparability, and several choices were added. There were 139 response records in 2014 to the question, "What is important in a solution?"

The most-chosen response, "ability to generate categories or taxonomies," is new to the 2014 survey. It was selected by 64% of Q15 respondents. Interpret it to include topic identification. (Not many tools have the ability to generate taxonomies, that is, hierarchical arrangements of categories and subcategories.)

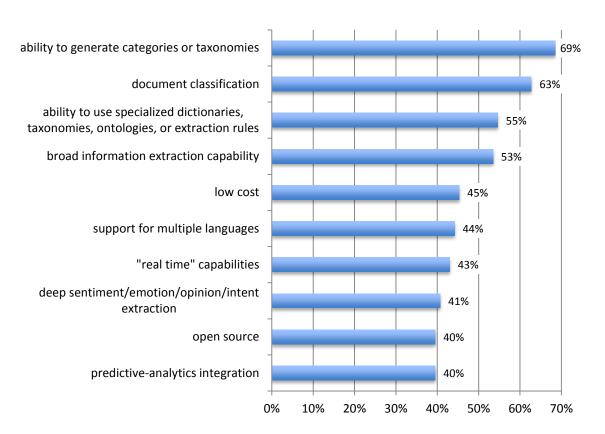
As in past years, the areas of greatest concern relate to core technical capabilities.

- 1. The desire for customizability remains high, in particular, "ability to use specialize dictionaries, taxonomies, ontologies, or extraction rules" at 54% selected. However, selection of "ability to create custom workflows" dropped from 44% in 2011 to 33% in 2014.
- Selection of information extraction choices remained high 53% for "broad information extraction capabilities" and 45% for "deep sentiment/emotion/ opinion/intent extraction" – although rates decreased markedly from those of previous years.
- 3. Low cost remained a factor, at 44%, with 37% choosing "open source," which of course implies lower software cost.



### What is important in a solution?

Contrast responses in the chart above with 2014 top experienced-user responses, selected by a third or more of users with two or more years experience with text analytics (n=86).



### What is important in a solution?

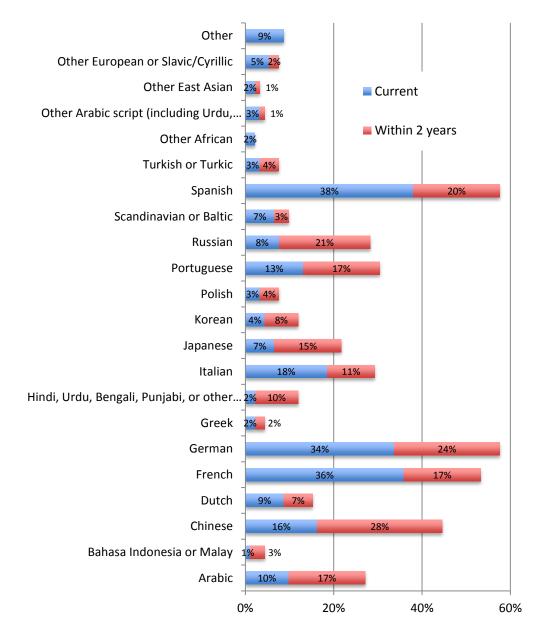
(Users with two or more years of text analytics experience, n=86.)

## Q16: Languages

Question 16 asked, "What written languages other than English do you currently analyze or seek to analyze? What languages do you see a need for within two years?" It offered, as responses, major languages and language groupings. It did not offer a *None* response; that is, it did not measure survey respondents who are not currently analyze languages other than English, and have no foreseeable (two-year) need to do so.

There were 92 responses, with 214 current languages given and 199 within two years, for an average of 2.3 languages per response for current, 2.2 languages for future, and 4.4 overall (noting a rounding effect).

# What languages other than English do you currently analyze? What languages do you see a need for within two years?



The response profile would surely have been different if the survey had had greater reach into military, intelligence, and international communities.

### Q17: BI Software Use

Question 17 asked, "What BI (business intelligence) or analytics software do you use if any? Please separate entries with commas, listing first the most important to you."

This question was posed because not infrequently, text analytics users seek to integrate results and analyses within numbers-focused BI tools. There were n=86 response records, reduced here to 34 choices. There was no None choice offered, so survey respondents who did not respond to this question may or may not be BI software users. Responses (sorted and without counts) were these:

Custom/in-house, Amazon Redshift, Coremetrics, Dedoose, Excel, Expedition, Hadoop, IBM Cognos, IBM Netezza, IBM SPSS/Modeler, JMP, Leximancer, Logi Analytics, Mallet, MicroStrategy, Microsoft SQL Server/SSIS/SSRS, NVivo, Oracle Business Intelligence, Oracle Endeca, Panoratio, Pentaho, PowerPivot, Python: numpy, scipy, QDA Miner/WordStat, QlikView, R, Rapid Insight, RapidMiner, SAP Business Objects, SAS, Statistica, Tableau, Verint, Weka

### Q18: Guidance

Question 18 asked, "What guidance do you have for others who are evaluating text/content analytics?" There were 79 responses. After editing for clarity and relevance, 57 are listed here – in five categories and arranged to create a sense of narrative progression.

### **Evaluating**

First identify key benefits and describe in business English (what it's for, what it does, not how it works). Then evaluate the solutions on a matrix in a sensible order, i.e. starting with the least expensive, easiest to use, etc. There are too many offers for a non-expert to make sense of, so "hats off" if you can do it for us. Key benefits will look very different for different types of businesses, i.e. individual consultants, corporate departments (by function), service provider companies, OEMs (solution sellers).

Identify [the specific] problem you need to solve first. Speaking to NLP vendors who claim to solve all your unstructured data needs leads nowhere.

Read up on the technical guidance beforehand; otherwise, you can end up with a lot of information that you can't make any sense of.

Take your time to evaluate different possibilities for your needs.

Have a defined strategy with measurable goals before embarking. Look for ease of customization.

Don't be dazzled by vendor sales pitch and hype. If they aren't willing to give you an unfettered demo using your data and not their carefully-crafted-to-work-all-the-time data, then run, don't walk, away from them. Also, open source Python tools will, with some effort and skill, yield the same results in my experience. Just validate, validate, validate if you use open source tools.

Make sure the clients/sponsors and your team members, understand and value [analyses] before you begin. Don't let the database people get into the project – they'll drive it on their tangent. Database folks typically don't understand text analytics or metadata. Get the enterprise architects involved: They need governed taxonomies for their "reference models."

Start experimenting, join forums, build up your knowledge on HOW things work (i.e. some theory).

Try a number of different vendors before committing to one. The difference in pricing, features, and support you can get out there can be overwhelming at first.

### **Proof of Concept**

It takes time to research and text solutions. Lots of tune up time is required to ensure that the results achieve precision and recall targets. Prototypes with a reputable service provider are usually required.

Get a proof of concept.

Focus on specific use cases first.

Test it on the obvious or easy texts and impossibly difficult ones.

Perform accuracy checks.

Define a test metric, and do cross validation.

Check how customized the solution is.

Look to see what can be done with free/open source before licensing a "solution." Make sure your data will work with the analysis you see in the demos. Make sure you understand what the software is doing, what's happening under the surface.

### Capabilities/Properties

Make sure a solution is not generic off-the-shelf, but suitable for your purpose

Reach out for the small, specialized vendors. They come with specialized solutions, often offering much greater quality than the big vendors.

Evaluate your content and your need for precision first.

The main point is the accuracy of the analysis.

Don't trust sentiment/coding accuracy claims of tools. The tools themselves are not accurate or inaccurate. It's the application of the tool to your problems that will be accurate or not.

Assess accuracy. Integrate/embed into existing processes.

Textual disambiguation and context are key. Language/slang/ abbreviation translation is very important. Filter out noise early.

Disambiguation of text and the ability for scaling and data aggregation [are important].

Understand what you're looking for. Vendors are great at blurring the lines, so you need to know: Do you need entity extraction? Fact/event detection? Classification? Sentiment?

Ensure that [a candidate solution] can search across apps and is customizable, that vendors bundle analytics with the application, and that no extra costs are incurred for customization

Improve your categorization to use across multiple data sets.

Treat sentiment analysis as a single-use case, and don't select a product simply for that.

Don't be confused into thinking that open source software is free.

Implementation / Guidance

Ensure you are clear on what you want to use text analytics for: To process and analyze large amounts of text and provide widespread access to top-level information/summaries or to support in-depth text analysis.

Have a clear purpose [in mind] and be aware that you will need people who have time to find the insights.

A key question that I've seen is whether to outsource text analytics or create the expertise/acquire tools internally?

It pays to be familiar with statistics, data mining, and NLP as well as have a solid knowledge of an open source software package that allows one to perform text analytics.

Have an NLP expert on staff.

Pay for training on the tool you select. Have real data for training.

Someone on the team has to understand the domain and the algorithms at the same time.

The computer is only as good as you tell it. It can get pretty close with content analysis but humans still need to drive it.

Ask for answers, not analytics.

Start simple, collect user data, enhance complexity.

The service/support and training component are important to help with set-up of taxonomies and improvement opportunities to get the best value out of the tools; rather than wait three years to get a budget of \$300-400K approved, start small, e.g., with niche provider (where possible) and show results/ROI that will help to get approval for a \$300K+ solution (where appropriate); Risk and compliance are easier to be convinced than marketing to invest into text analytics – then other functions can profit from the solution – marketing might already have social media monitoring and very, very basic text analytics (Wordle type) in place. Hence [there may be no] need to invest.

Start simple and powerful.

Most text analytic systems are designed for the single administration type of project, if you are running a VOC/CRM program that is ongoing you need to work out pricing and services up front to avoid sticker shock.

Make sure you understand the data quality and sources before you start crunching numbers.

APIs have a big risk. Use them to bootstrap, but then move to something built on top of open source libraries.

Think Agile - small deliverables, flexible solutions, easy to adjust scope and focus.

Don't underestimate the effort in getting a successful project off the ground. Whether it's domain expertise, ontologies and dictionaries, etc., there's a lot of effort involved. The purely statistical models (classification, sentiment analysis) are much easier to get started with – if you can create strong training sets, you're good to go. But entity/event detection takes a lot more.

Understand that it's about interpretation, not the software (unless the software's really, really bad).

### **Expectations**

It takes more time than you might think.

Be wary of vendor promises and their data.

Some value is better than no value.

Doing it open source is difficult, but seize opportunities and go with it.

Most companies over-promise and under-deliver, so set expectations accordingly.

It is an emerging technology; it won't be perfect in all areas. Focus on where it can save time and add insight.

Be prepared to work on refining these solutions for a long time and with a significant investment of resources. And don't let your expectations get too high.

Don't underestimate the investment of time up front to develop a well-trained model.

Will have to invest time, money, and energy to generate more insights, however with [all elements] in place, [text analytics] is effective.

## Q19: Comments

Last, Question 19 solicited comments. The following are 9 of the 21 responses.

Read major parts of your material before and after doing quantitative analyses.

Be prepared to spend some time figuring out how to best use the text analytics solution. At the moment there is scarce text analytics know-how, at least in the market research industry. It took us six months to tune our solution just to start even thinking about ROI.

For text analytics to grow, there is a need to educate businesses about the existence of this tool and what it offers, but also to educate new text analysts about where these skills can be applied and where the most value can be gained from text analytics. This includes all the tools that text mining offers as well as possible sources of data that can be analyzed via text mining.

Dialects of languages [that are] very extended geographically, such as English or Spanish, should be considered, for analytic purposes, as different. I think that this is especially true in the case of sentiment analysis. At least, in the case of Spanish, the same word can have a different sentiment in different geographical areas.

Australia is not very advanced in text analytics from a buyer's perspective. Cost/benefit ratio is still a hard sell and competing with other initiatives, e.g., crude net promoter score (operational, brand) surveys/customer feedback loops, and social media monitoring (word counting). That of course is also a function of the market size compared to the U.S. But many European countries may see similar issues.

Sentiment analysis is busted. Don't chase it.

Your last [Sentiment Analysis Symposium] conference made me a convert; however what I'm seeing is still too complex/expensive to incorporate into my qualitative research practice easily. I'm part of a QRCA (Qualitative Research Consultants Association) subgroup trying to do the above for our members. We see this as an ongoing activity.

Using unstructured data sources has become mainstream now versus a decade ago when I first began exploring the technology. While I'm infrequently involved in analysis of text with the big data automotive sector research company I work for (contract), I see it as critical in the future to blend customer attitudinal perspectives with the massive sources of behavioral data.

Most clients are looking to make business decisions not just crunch numbers. I think it is key to evaluate the solutions based on their actual usage of the output to make real decisions.

### Additional Analysis

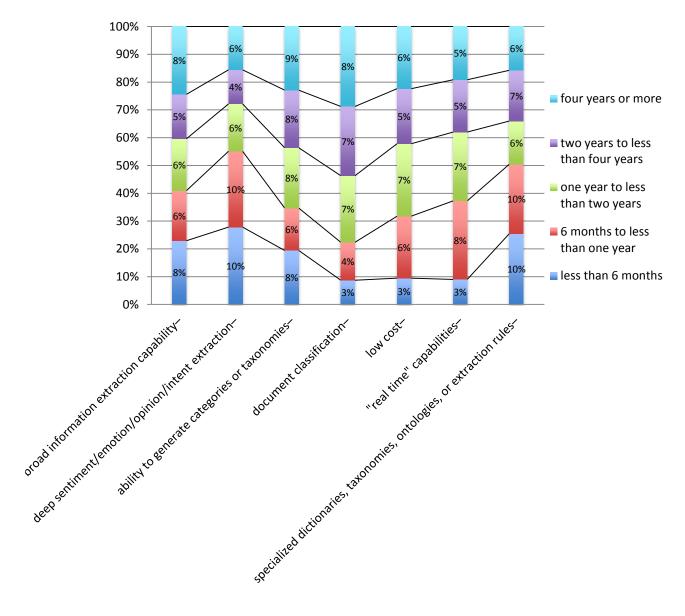
The survey was designed so that responses to questions would be immediately useful without elaborate cross-tabulation or filtering. Certain analyses are revealing, however – for instance whether a respondent is currently using text analytics or not and the length of time using text analytics, cross-tabulated against other variables.

#### Length of experience with text analytics correlates with solution needs.

The following chart cross-tabulates Q1 Length of experience (this time for respondents with any experience) by Q15 What is important in a solution. The top six Q15 responses are included.

Bars are ordered by the choices of the most experienced (four-years-or-more) users, who also represented a large proportion of the respondents. It is notable that the top choices for both those users and for all users involve capabilities linked to information extraction, classification, and disambiguation.

# What is important in a solution? (Proportion for each of top six picks, chosen by >43% of respondents, by length of text analytics experience.)



# Other interesting points come out by contrasting respondents who are already using text analytics with respondents who are still in planning stages.

The 216 Q3 respondents chose an average of 5.6 selections per respondent. (Q3 was What textual information are you analyzing or do you plan to analyze?) Only 194 of those respondents also answered Q7 Are you currently using text/content analytics? Crossing the responses to those two questions, we learn that

- current text analytics users choose an average of 6.15 sources and
- prospective text analytics users choose an average of 4.53 sources.

The following table which lists the Top 6 responses to Q3 by Q7 values. Note, in particular, the 19-point and 16-point gaps between current and prospective users for news articles and for long-form blogs as sources, versus smaller gaps for other sources.

(Because the survey was not based on a scientifically designed sample, we do not evaluate the statistical significance of the difference.)

Textual Information Source	Current User (n=134)	Not Yet Using (n=64)	All (weighted)
Twitter, Sina Weibo, or other microblogs	51%	40%	46%
Blogs (long form), including Tumbr	48%	32%	43%
News articles	50%	31%	42%
Comments on blogs and articles	39%	34%	36%
Customer/market surveys	42%	34%	37%
Online [discussion] forums	40%	27%	36%

### Textual Information Sources by Current/Prospective User

### Interpretive Limitations and Judgments

The number of survey respondents was not large enough to support further useful crosstabulation of variables beyond the analyses above.

In interpreting presented findings, do keep in mind that the survey was not designed or conducted scientifically – that is, with the intention or the actuality of creating a random sample or a statistically robust characterization of the broad market. Findings surely reflect selection bias due to 1) the outlets where the survey was advertised and 2) a likelihood that those individuals who are unaware of text/content analytics or the potential for the technologies or solutions to help them solve their business problems would not respond to the survey. Findings therefore over-represent current users.

### About the Study

**Text Analytics 2014:** *Users Perspectives on Solutions and Providers* reports the findings of a study conducted by Seth Grimes, president and principal consultant at Alta Plana Corporation. Findings were drawn from responses to an early-2014 survey of current and prospective text analytics users, consultants, and integrators. The survey ran from January 18 to April 15, 2014, with the vast majority of responses in the first four weeks. It asked respondents to relay their perceptions of text analytics technology, solutions, and vendors. It asked users to describe their organizations' usage of text and content analytics and their experiences.

### Sponsors

The author is grateful for the support of eight sponsors – <u>AlchemyAPI</u>, <u>Digital Reasoning</u>, <u>Lexalytics</u>, <u>Luminoso</u>, <u>RapidMiner</u>, <u>SAS</u>, <u>Teradata</u>, and <u>Textalytics</u> – whose financial support enabled him to conduct the current study and publish study findings. The sponsors provided the content of the sponsor solution profiles.

The survey findings and the editorial content of this report do not necessarily represent the views of the study sponsors. The sponsors did not review this report prior to publication, with the exception of the solution profiles they provided.

### Media Partners

The author acknowledges assistance received from three media partners in disseminating invitations to participate in the survey, <u>CMSWire</u>, <u>KDnuggets</u>, and <u>LT-Innovate</u>.

### Seth Grimes

Author Seth Grimes is an information technology analyst and analytics strategy consultant. He founded Washington DC-based Alta Plana Corporation in 1997.

Seth consults, writes, and speaks on information-systems strategy, data management and analysis systems, industry trends, and emerging analytical technologies. He is the leading industry analyst covering the text analytics and sentiment analysis providers, solutions, and markets.

Seth is long-time contributor to publications that include InformationWeek, CustomerThink, and Social Media Explorer; he is founding chair of the <u>Sentiment Analysis Symposium</u> (<u>@SentimentSymp</u> on Twitter) and the Text Analytics Summit (2005-13).

Seth can be reached at grimes@altaplana.com, 301-270-0795. Follow him on Twitter at @sethgrimes.

# **Sponsor Solution Profiles**



### Solution Profile: AlchemyAPI

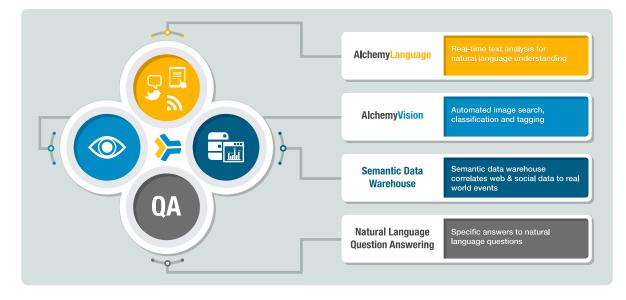
#### Powering the New Artificial Intelligence Economy



At AlchemyAPI, we are democratizing the latest breakthroughs in deep learning to power the new AI economy. We provide the world's most popular cloud platform and community for easily creating unstructured data applications that make sense of the oceans of web pages, documents, tweets, chats, email, photos and videos produced every second.

You can rely on AlchemyAPI as the #1 cloud provider of unstructured data analysis to give you everything you need:

- Speed at small-to-massive scale: Process billions of pieces of text and images monthly
- **Proven ease-of-use:** Over 40,000 developers across 36 countries use our services
- Global reach: Understands 8 languages, and recognizes 80 more
- Tailor to your needs: Includes custom taxonomies, sentiment and image libraries



#### AlchemyLanguage

We deliver the most comprehensive set of knowledge extraction web services for text. Classify any text passage into a taxonomy, and then identify its named entities, keywords, high-level concepts, sentiment, intent, facts-and-relations and author. Add integrated web crawling and text cleaning along with AlchemyAPI's famous speed, accuracy and ease-of-use, and it's no surprise why AlchemyLanguage is the most popular choice to power today's smart, real-time applications in commerce, customer engagement, digital marketing, finance, publishing and social media.

#### AlchemyVision

The new Visual Web escalates the need to extract knowledge from photos. AlchemyVision employs our deep-learning innovations to understand a picture's content and context so you can automatically tag and classify photos or search for like-images. Unlike today's image systems that are limited to recognizing specific objects like faces, logos or labels, AlchemyVision *sees* complex visual scenes in their entirety—without needing any textual clues—in order to understand the multiple objects and surroundings in common smartphone photos and web images.

**Semantic Data Warehouse** (currently in AlchemyLabs, ask for a demo) AlchemyAPI's semantic data store\_brings analysts an entirely new array of business insights. It is a massively scalable analytical data warehouse that enriches and organizes unstructured data like web pages, tweets, email, chats or social posts. Users can analyze and correlate this rich data to real-world events taken from stock markets, CRM and financial systems, weather, crime statistics, etc. to enable a new world of predictive analysis built up from unstructured data.



"AlchemyAPI gives us a huge addition to our data set that we don't have otherwise. The data isa lot cleaner and there's language support that competitors just don't have."

- Nate Weiner, Founder & CEO at Pocket

**Natural Language Question Answering** (currently in AlchemyLabs, ask for a demo) Like IBM's Watson, our question answering service provides confidence-based responses to natural language questions; but it delivers these innovative question-answering features at lower cost, complexity and risk. AlchemyAPI's question answering service shortens time-to-decision because it extracts and ranks specific answers drawn from across the world's web sites and your company's unique knowledge bases.



"The team we worked with at AlchemyAPI has been outstanding in helping us to understand NLP technology and how we could economically implement it."

- Anthony lacovone, CEO at AdTheorent

### What drives AlchemyAPI's speed, accuracy and cost advantages?

Our deep learning expertise and unique computing infrastructure gives us the ability to process at a massive scale with zero human intervention and no human-curated training data. This lets us deliver a host of advantages that enable you to create smarter applications and services:

- More accurate text and image algorithms that greatly deepen a computer's understanding of human language and vision.
- Our gigantic knowledge graphs create an intelligence that learns how every thing in the world is related. This deep understanding of context delivers superior disambiguation of named entities and recognizes higher-level concepts the text is discussing.
- High-performance infrastructure meets the real-time demands of your processing pipelines.
- Secure, easy-to-use cloud services eliminate your costs to install, integrate, protect and maintain advanced language and vision solutions.

### Learn more: <u>www.alchemyapi.com/Content-Analytics-Solutions</u> Reach out to sales: 1-877-253-0308 (toll-free), or email at sales@alchemyapi.com

### Solution Profile: Digital Reasoning



Digital Reasoning is a leader in cognitive computing. We provide organizations with a new kind of software that intelligently analyzes information and reveals previously unknown relationships, risks, activities and assertions. Digital Reasoning's open and cloud-ready machine learning-based analytics platform, Synthesys, provides a proven approach for analyzing human language data at scale and helping organizations make better and more timely decisions. Digital Reasoning does this by incorporating a more human-like intelligence approach — by semantically analyzing information and intelligently revealing patterns and behaviors that are important to the organization.

"Digital Reasoning's ability to analyze and understand human communications at scale is a valuable solution for customers looking to unify the management of their structured and unstructured information."

Tim Stevens, vice president for business and corporate development at Cloudera

Used by the Intelligence Community, Defense agencies, global financial institutions and other organizations, the Synthesys® platform provides a proven suite of intelligence analytical capabilities that organizes important and valuable information into a private knowledge graph. Synthesys seamlessly integrates a set of knowledge services in the form of a rich Application Programming Interface (API) to access the "knowledge graph" abstracted from the data. Synthesys provides an entity-centric approach to data

analytics, focused on uncovering the interesting facts, concepts, events, and relationships defined in the data rather than just filtering down and organizing a set of documents that may contain the information being sought by the user.

In order to assemble this rich knowledge graph from data, Synthesys ingests both unstructured and structured data, and performs three basic phases of analysis: Read, Resolve, and Reason.

The "Read" phase ingests the data and performs

Synthesys DATA INGESTION DATA DISCOVERY  $\rangle\rangle$ Entity Extrac Fact Extraction UNSTRUCTURED DATA Intelligence Reports KNOWLEDGE API +++ & Websites Real-Time Query e-[]-e 3rd Party Apps Y -P STRUCTURED DATA Export Formats Country Profiles eadership Profiles Watchlist Data Navigation Data Glance

Natural Language Processing (NLP), entity extraction, and fact extraction. The "Resolve" phase assembles, organizes, and relates the results from the "Read" phase to perform global concept resolution (i.e. entity resolution) and detect synonyms (i.e. synonym generation) and closely related concepts. The final "Reason" phase applies spatial and temporal reasoning and uncovers relationships that allow resolved entities to be compared and correlated using advanced graph analysis techniques. These three phases of analysis are performed in a distributed processing environment, and their results are stored into a unified entity storage architecture called a Knowledge Base (KB).

Synthesys' learning algorithms efficiently transform data into knowledge that can be trusted as the basis for human analysis and decision-making. Synthesys detects patterns of potentially relevant activity and then highlights behaviors based on combinations of patterns that suggest some sort of risk. Based on those behaviors Synthesys alerts the analyst and produces a profile, which can be explored through Synthesys "Digital Reasoning applies AI to understand human communication to ferret out suspicious activity. Over time, this class of service may become indispensable."

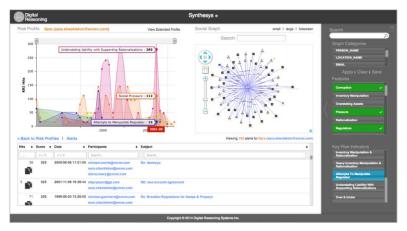
Gartner 2014 Cool Vendor Report for Smart Machines

Glance, a powerful web application.

How does Synthesys differ from other data solutions? That's simple—it learns. Synthesys finds order in data and understands patterns in language like a human. Synthesys has a number of unique abilities that enable it to read and understand your human-generated data better than any other software.

- **Knows how people talk** Synthesys has an astounding understanding of human language. It understands what people have said and can deal with the ambiguity of words—for instance, when different names are referring to same thing (Joe, Joseph, Joey). It not only understands the words being said but also learns from the in-context usage, exposing the real human meaning in the data.
- Accumulates context Synthesys doesn't just identify the names of people, places and organizations; it relates them to the real-world entities they represent. Once identified, it accumulates knowledge about those entities to include attributes (DOB, POB, date founded, HQ location, etc.), relationships and facts, creating rich knowledge profiles. These profiles are the fundamental building blocks needed to make relevant predictions about future behaviors of employees, customers or potential threats.
- **Keeps a watchful eye** Most solutions index data and expect you to know what you're looking for. As a result, key information gets overlooked. Synthesys offers a smarter way by providing organizations the ability to build and configure pattern detection logic to identify key risk/performance indicators. Synthesys is now on the lookout for these patterns, proactively detecting activities of interest.
- Learns and gets smarter The best thing about Synthesys is that its knowledge graph gets smarter and grows with you. It never forgets. It's persistent and pervasive. Synthesys teaches itself to draw conclusions based on what you're looking for in your data. That means, like a great employee, Synthesys becomes even more intelligent and valuable over time.
- Knowledge

visualization Synthesys Glance<sup>™</sup> is a new Web application that enables analysts to interactively browse and analyze knowledge profiles of people and organizations to discover valuable patterns and relationships that were previously difficult to detect. Glance provides organizations the ability to identify and respond to



threats and opportunities more quickly than ever before.

Synthesys can be deployed in the cloud or in-house. Synthesys Cloud, currently available through the Amazon Web Services (AWS) Marketplace, provides a secure and flexible way for your organization to gain the advantages of Synthesys without investing in additional hardware or placing additional demands on your network. Synthesys Enterprise can be installed within your data center, providing you with full operational control of the platform, including full adherence to your corporate access and data control policies. Synthesys Enterprise can also be customized to facilitate easier integration with your other applications and process workflows.



### Solution Profile: Lexalytics

### Introduction

Since 2003, <u>Lexalytics</u> has provided the most popular text and sentiment analysis solutions for everyone from individual users through service providers. Our engine <u>Salience</u>, and SaaS service <u>Semantria</u>, are the go-to text analytics systems for companies in media monitoring, <u>business intelligence</u>, voice of customer, customer experience management, <u>Excel analysis</u>, predictive analysis, and more.

Hundreds of F1000 companies rely on results processed through Salience, with applications ranging from tweaking consumer goods packaging to ensuring guest comfort or monitoring pharmaceutical side effects.

Our partners process over 3 billion new documents per day

### What Products Do We Offer?

Salience is a high-performance text analysis engine, packaged as a set of software libraries. Integrate Salience once and let us handle staying current on all the gnarly technology necessary to convert that unstructured text into structured data. (Check out the amount of thought we've put into just one problem – social media specifics.)

Semantria is the SaaS version of Salience. Perfect for cloudbased applications, with an available Excel Add-In for users that don't like to program. through Salience, extracting meaning and distilling emotion, each partner displaying and providing those results in their unique way.

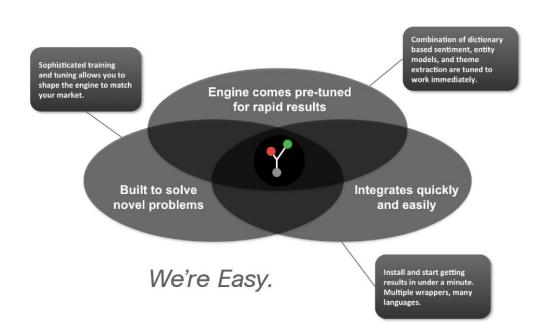
#### All of our partners are

- Enabled by an engine integrated in just 3 hours to 3 days.
- Precisely serving their widely varying markets with easy customizability.
- Focusing more on their core competencies, and leaving the text to Lexalytics.

We offer 9 different languages, covering well over 2 billion speakers. Natively multi-lingual means that you will have an automatic quality advantage over any of your translationbased competitors.

Already an NLP expert? Great! <u>Work with a set of power tools</u> that will let you get your ideas and projects off the ground and into operation in days rather than weeks (or months!)

Come try our SaaS and get 10,000 free api calls now. Use them with the Excel add-in if you want!



### What Can You Do With Our Products?

You can build a media monitoring service around Salience, like our partners MediaVantage, Vocus, and Bottlenose.

You can make business intelligence apps more intelligent, like our partner Oracle.

You can teach Excel to read, like our partner <u>Semantria</u>. (They also provide a SaaS version of Salience.)

You can provide additional metadata to search (for real knowledge management), like our partners <u>Coveo</u> and <u>Playence</u>.

You can analyze surveys and understand customer feedback like our partners  $\underline{SMG}$  and  $\underline{evolve24}$ .

You can predict the future, like our partner Angoss.



### Specifications

Yes, our products probably do what you want.

<u>Salience will</u>: strip HTML, detect the language, tokenize, PoS tag, chunk, sentence break, **extract named entities**, automatically detect important topics, **summarize**, **categorize**, classify, **score sentiment**, and more.

<u>Semantria</u> provides the most common functions of Salience in a convenient SaaS package, and will get anyone up and



running with text analytics in less than 3 *minutes*, by using the Excel Add-In.

Both products work great out of the box: *you don't have to do anything for most cases.* But they're both highly customizable too: if you have a special case you want to extract, a rich toolbox is ready and waiting for you.

32/64bit. Windows or Linux. Java, PHP, Python, .NET, C

Founded in 2003, Lexalytics was first to market with a sentiment analysis engine. Our text analytics software, Salience<sup>™</sup>, and online service Semantria<sup>™</sup> are engineered for easy integration into third-party applications, and are a critical component in many high-volume content processing services for industries such as social media monitoring, voice of customer, online media, knowledge management and more.

Lexalytics Salience Engine<sup>™</sup> provides on-premise and OEM text analysis functionality, and Semantria<sup>™</sup> provides SaaS text mining and a no-programming Excel add-in.

In other words, we've been around for over 10 years, and we're the go-to company for dealing with text, regardless of whether you're an enterprise or a single user.

http://www.lexalvtics.com http://www.semantria.com

### Solution Profile: Luminoso

**Overview** 

# 🍊 LUMINOSO

Luminoso is a cloud based multilingual text analytics solution that provides businesses, researchers, and decision-makers with the ability to quickly analyze and derive insights from large amounts of unstructured text, especially survey responses, consumer feedback, and social media. Luminoso enables Fortune 1000 companies to tap into customer experience, fine-tune marketing strategy, elevate customer service and evolve product offerings. Our multi-lingual capabilities enable global companies to understand their customers worldwide. The intuitive, easy to use nature of the Dashboard, paired with the core technology's unrivaled ability to understand human language, makes it essential for efficient and impactful text analysis. Customers include Aegis Media, ConAgra, General Mills, MARS, Mass Challenge, Scotts Miracle-Gro, and SONY.



In the current state of text analytics, vendors offer solutions based on systems that require the manual construction of intensive dictionary or keyword sets, as well as the preparation of data and grooming of results. Luminoso is a new breed of text analytics, one that requires neither ontologies nor constant maintenance in order to understand and interpret unstructured text correctly. Because Luminoso has harnessed the conceptual understanding gained from over 15 years worth of research at MIT's Media Lab, it adapts and learns new words

without the encumbrance of manual updates. Luminoso's API services facilitate the embedding of text understanding within existing systems, while its intuitive analytics dashboard enables rapid extraction of insights by analysts with free-form data.

# Why Lumínoso?

**Common Sense:** Because Luminoso has harnessed the power of the MIT Media Lab's Concept Net, it adapts and learns new words as you or I would

**Time to Value**: It takes only minutes to upload data in the form of a "CSV" file into Luminoso's Insight Engine dashboard

**No Training Required**: Luminoso does not require any pre-supplied or pre-built ontologies, keyword lists, or human generated categories in order to perform analyses, even with domain specific language

**Multilingual:** Luminoso's software automatically understands cultural connotations and idiomatic associations in English, Spanish, French, German, Italian, Portuguese, Japanese, and Chinese

**Deep Sentiment:** Luminoso gives users the ability to understand simple sentiment in conversation (positive/negative/mixed/neutral), as well as what drives deeper sentiment (uncomfortable/ convenient/ sleek)

Flexibility/ Ease of Use: The dashboard is intuitive and simple to use, making the harvesting of insights easier and faster than ever

**Seamless Integration of Multiple Data Sources:** Whether it's social media, surveys, focus group responses, emails, or call transcripts, our software can analyze and weigh each piece of data automatically

### Solution Profile: RapidMiner

### Predict to Act



As more and more data is created and collected, data is increasingly perceived as an asset. But it's not the data itself that brings an advantage, it's the information and insights derived from data that provide additional value for a business.

Who are the prospects that are the most likely to respond to your marketing campaign? Which customers are the most likely to churn and what would be the best option to prevent that? Being able to answer questions like these can give you and your business a competitive edge.

RapidMiner provides business analysts with the means to jumpstart a project and be productive much faster than is possible with programming languages and legacy tools. The inherent ease-of-



use helps users, from the seasoned data scientist to the line of business executive, work together collaboratively.

## RapidMiner Benefits

### **High Connectivity**

Data not only comes in different sizes and shapes, the variety of data storage and formats is also unimaginably broad.

RapidMiner embraces this variety and provides numerous connectors. RapidMiner offers connectors to common data file formats like comma-separated values (CSV), tab delimited, and other text-based formats but also those formats written by Microsoft Excel, IBM SPSS, and SAS. In addition to tabular data, RapidMiner can also read contents from text files, PDF files, and HTML files. As more and more external data needs to be analyzed along with internal data, RapidMiner also offers connectors to Web sources through web crawlers, HTTP connectors, and RSS feed readers.

"RapidMiner deserves to be better known and more widely used for predictive analytics and text mining. It's easier to use than other products, and the company support staff is very knowledgeable, helping us with troubleshooting, and designing processes to fix any problems we might have."

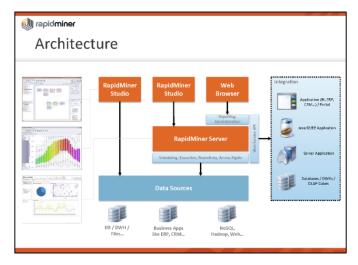
-- Brian Tvenstrup, Chief Analytics Officer at Modern Marketing Concepts. Inc. (MMC)

### **Unstructured Data Capability**

RapidMiner can access and load unstructured data including texts, images, and audio tracks,

and can also extract information from these types of data and transform the unstructured into structured data. Moreover, RapidMiner can mash-up structured data with unstructured data and analyze both types of data in combination to fully harness all the data available for generating descriptive insight and deriving information for predictions.

As an example, text can be loaded from various file types like plain text, HTML, PDF, or RTF, and transformed by a huge set of different filtering



techniques including tokenization, stemming, stop-word filtering, part-of-speech tagging, ngrams, dictionaries, and many more. Information contained in the text can be structured by extracting statistical metrics like term occurrences or frequencies and TF/IDF values. Alongside other tabular data, this data can be analyzed through all statistical approaches available in RapidMiner such as classification, regression, and clustering techniques.

- Text analytics:
  - Operators necessary for statistical text analysis
  - Load texts from different data sources or from your data sets including plain texts, HTML, PDF, RTF, and many more.
  - Transform texts by a huge set of different filtering techniques including tokenization, stemming, stop word filtering, part of speech tagging, n-grams, dictionaries, and many more.
  - Connection to WordNet and other services to clean up texts before processing
- Web data processing:
  - Access to internet sources like web pages, RSS feeds, and web services
  - Specific operators for handling the content of web pages
  - Extend structured data with data from the web and combine those data sources for getting new insights and detect chances and risks

More than 50 extensions are available for all types of input formats, data transformations, and analysis including extensions for image mining, audio mining, time series analysis, automatic system creation and many more.

RapidMiner USA	RapidMiner UK	RapidMiner Germany
10 Fawcett St	Quatro House, Frimley Road	Stockumer Str. 475
Cambridge, MA 02138	Camberley GU16 7ER	44227 Dortmund
United States	United Kingdom	Germany
E-mail contact-	E-mail contact-	E-mail contact-
us@rapidminer.com	uk@rapidminer.com	de@rapidminer.com
Phone +1 - 617 - 401 - 7708	Phone +44 1276 804 426	Phone +49 - 231 - 425 786 9-0

## Solution Profile: SAS



Since 1976, SAS® has helped organizations transform large amounts of data into knowledge they can use. With an extensive suite of analytics solutions and services, SAS makes it easier to solve complex business problems, drive sustainable growth and anticipate change.

Why is this so important? Because data – both structured and unstructured – is streaming into businesses at unprecedented speed, and organizations not managing it well are missing a major opportunity to glean value from it.

Big data, which is the term assigned to this massive proliferation of data, can be complex and unwieldy. And new technologies (such as Hadoop) that are being developed to store big data are not necessarily intuitive. This is where SAS comes into play.

SAS solutions help you:

- Improve, integrate and govern data so it's reliable and ready for action.
- Use visual analytics to identify business trends, optimize processes and forecast opportunities.
- Delve into Hadoop for fast, accurate insights that drive strategic decisions.

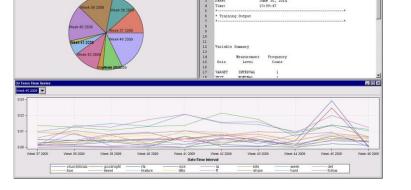
- Establish a more complete view of your customers so you can nurture relationships, improve marketing and drive revenue.
- Distill important insights from large, diverse content sources with text analytics.

#### What is SAS® Text Analytics?

Any forward-thinking organization knows that text is much more than words or figures on a page. It's untapped potential for business value. With the right technology, you can glean

insights from text that reveal patterns or truths you wouldn't have otherwise noticed.

With SAS Text Analytics, you can put processes in place that automatically mine structured and unstructured text, turning information overload into a competitive advantage. You can get a better read on social media. Analyze your digital content in online publications. Gauge citizen intelligence in government. And get it done quickly, accurately and in realtime.



### SAS Text Analytics Offerings

### SAS® Text Miner

It's no longer necessary to limit your analysis to legacy data. With SAS Text Miner, you can dynamically retrieve pertinent content from the web and other sources – content that helps you discover new ideas and insights hidden in huge document collections. Technologies such as machine learning and natural language processing make manual activities automatic.

Interactive GUIs make it easy to identify relevance, modify algorithms and group materials into meaningful aggregations.

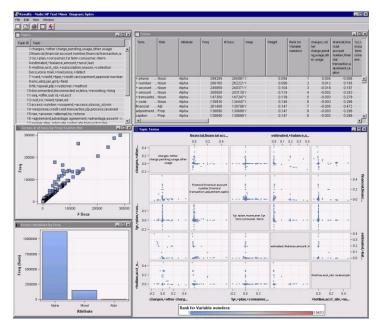
With high-performance text mining, which is designed for big text data, you can develop models that examine millions and tens of millions of documents – and run the models in minutes or seconds.

#### **SAS®** Contextual Analysis

With contextual analysis technology from SAS, you can add easily add structure to unstructured data. Text analysis becomes more efficient by automating document classification and eliminating the need for manual document reviews. As a result, you can find emerging issues and themes in your unstructured data and use these insights to classify document collections or use them for other analytics projects.

#### SAS® Enterprise Content Categorization

The data that floods your organization holds little value if it's not organized. SAS **Enterprise Content** Categorization applies natural language processing and advanced linguistic techniques to identify key topics and phrases in electronic text so you can more easily categorize large volumes of multilingual content. The solution parses, analyzes and extracts content for entities. facts and events as a way to isolate relevant information and generate metadata tags that index documents.



### SAS® Sentiment Analysis

Sentiment analysis technology from SAS helps you automatically pinpoint sentiment from the web and internal electronic documents, making it easier to understand trends or identify emotional changes. The solution automatically scores input documents as they're received, providing real-time updates about sentiment changes that deliver quantified insights into the overall impressions people have of products, services and brands.

### SAS® Ontology Management

Taking a consistent approach to text analytics is important when it comes to managing semantically related content, and SAS Ontology Management helps you accomplish this across the enterprise. By creating ontologies that are centrally managed, you can examine content through a sophisticated, informed lens and maximize content value.

In addition to specialized text analytics, SAS provides text processing and analysis capabilities with solutions such as SAS® In-Memory Statistics for Hadoop, SAS® Visual Analytics and SAS® Event Stream Processing Engine to seamlessly manage big data jobs.

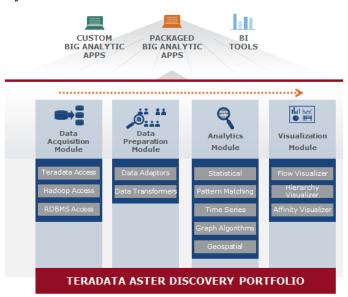
### For more information, please visit <u>sas.com</u>.

# Solution Profile: Teradata Aster

# TERADATA, ASTER

The Teradata® Aster® Discovery Platform is the market-leading big data analytics solution. This analytic platform embeds MapReduce analytic processing for deeper insights on new data sources and multi-structured data types to deliver analytic capabilities with breakthrough performance and scalability.

Teradata Aster Discovery Portfolio provides a suite of ready-to-use SQL, SQL-MapReduce<sup>®</sup> and Graph functions for fast and easy discovery of business insights from big data. Discovery Portfolio is part of the Teradata Aster Discovery Platform that also includes the Teradata Aster Database. The Discovery Portfolio consists of four modules of SQL, SQL-MapReduce and SQL-GR<sup>™</sup> functions: Teradata Aster Data Acquisition Module. Teradata Aster Data Preparation Module, Teradata Aster Analytics Module, and Teradata Aster Visualization Module. All this is possible on a single platform and



through a single query that integrates all the discovery process steps.

**Enable Richer Analytics:** The Discovery Portfolio contains 100+ advanced analytic functions that address use cases across a variety of industry verticals. The use cases include text, fraud and risk analysis, churn, golden path analysis, marketing attribution, sentiment extraction, and more. Discovery Portfolio enables *multi-genre analytics* (e.g., nPath<sup>™</sup> and Naïve Bayesian functions that determine churn likelihood) and introduces a rapid iterations approach to decision making that is founded on hypothesis testing.

**Complementary Integration:** Teradata Aster has integrations with solutions from leading partners including SAS®, Tableau®, MicroStrategy® and more. The typical user moves seamlessly among data access, extraction, analytics, and visualizations. Aster supports indatabase text and sentiment analysis via Attensity<sup>™</sup>, in-database R analytics, and in-database PMML execution via Zementis®.

Aster Prebuilt Text Functions: Aster's text analytic functions enable SQL users to create applications that examine unstructured information in large volumes with a high degree of flexibility. These functions allow users to remove artifacts and information content that has low information value and to create higher-level functions like fuzzy matching, text cleansing, sentiment analysis and naïve Bayes model prediction. Most importantly, these functions leverage the powerful massively parallel processing (MPP) architecture to create a highly performing and scalable solution. Below is a short list of text analytic foundation functions:

- **nGram:** Tokenizes (or splits) an input stream and emits n multi-grams based on the specified delimiter and reset parameters.
- Levenshtein Distance: Computes the number of edits (distance) needed to transform one string into the other.
- Sentiment Extraction: Deduces an opinion (positive, negative, neutral) from text-based

#### content.

- **Chinese Text Segmentation**: Identifies the sequence of words in Chinese language text which contain *hanzi* characters to mark boundaries (segments) in appropriate places.
- Named Entity Recognition (NER): Finds mentions of specified entities in text.
- **Text Parser:** Tokenizes an input stream of words, optionally stems them (getting to the root of a word. For example, hunt is the root of hunting, hunted.), and outputs the counts for each word.

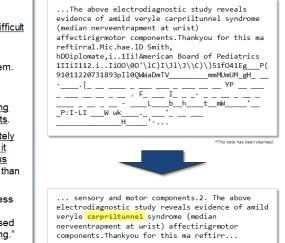
Aster Text Claims Processing Use Case: A classic Aster use case involves our ability to ingest OCR scanned information which is notorious for having unwanted artifacts. This was a two-week proof-of-value. Using functions like Levenshtein distance, nGram and text analysis, Aster was able to cleanse and filter large amounts of claim information to highlight only the information needed by the review staff to determine claim validity.

Finally, to enable this technology for non-technical users (nursing staff, clinicians, etc.), we

created a simple HTML browser application using our ODBC/JDBC interface that staff members could use to quickly input search word criterion and level of accuracy to produce a summary snapshot of relevant claims. The application included hyperlinks so that staff could quickly locate and view the original claim documentation.

#### SUMMARY

- OCR-generated data contains artifacts that make the content difficult to analyze.
  Stored COL office are simple
- Standard SQL offers no simple, repeatable solution to this problem.
- <u>SQL-MapReduce enables or</u> signficantly improves Regular Expression capability for matching keywords in spite of OCR artifacts.
- By focusing on the text immediately surrounding matched keywords, it becomes possible to simply focus upon small chunks of text rather than the entire document.
- SQL-MapReduce can also address this issue through a variety of methods including dictionary-based fuzzy matching and "word-splitting."



**Bottom Line Impact:** The new text analytic functionally decreased average time to review a case by 86%. Nursing staff increased average cases reviewed per hour per resource over 6X compared to the current manual process.

#### For More Information

To find out how the Teradata Aster Discovery Portfolio can help you take advantage of big data volumes in a fast, efficient, and cost-effective way while you improve your decision-making capabilities and grow a stronger, more productive business or to learn more about Teradata Aster Discovery Platform, contact your local Teradata representative or visit **Teradata.com**.

Teradata, Aster, the Teradata Logo and SQL-MapReduce are registered trademarks and SQL-GR and nPath are trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide. SAS is a registered trademark of SAS Incorporated. Tableau is a registered trademark of Tableau Software, Inc. MicroStrategy is a registered trademark of MicroStrategy Incorporated. Attensity is a trademark of Attensity Group. Zementis is a registered trademark of Zementis Incorporated Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions, and operations described herein may not be marketed in all parts of the world.

## Solution Profile: Textalytics



### **Plug and Play Semantics for your Business Needs**

Textalytics is a new generation of semantic APIs in the cloud, specifically adapted to diverse applications and industries, which help you embed semantics into your applications in an easy, safe and inexpensive way.

Semantic technology paves the way for **extracting the business value** from large volumes of information in all types of internal content and social media. But adapting these technologies to your processes and applications is not free from difficulties.

There is a gap difficult to bridge between what business-focused developers need and the usually horizontal and basic semantic technologies that require a high level of adaptation and customization to be productive.

### Textalytics' Semantic APIs speak the language of your business needs

Textalytics is much more than a Semantic API: it provides a framework based on the cloud that makes the integration of semantic processing into your applications almost a plug-and-play experience.

In addition to APIs with standard and granular functionalities that can be combined by the user, Textalytics offers **vertical APIs** optimized for different industries and application scenarios. These APIs include dictionaries, taxonomies, process pipelines and functionalities specific to these applications, which build a layer of high added value on top of standard, basic semantic functionality.

For example, Textalytics combines general sentiment analysis, topics extraction and classification technologies to offer a corporate reputation API completely oriented to this business need.



Presently, vertical APIs for Media Analysis and Semantic Publishing are available, and a Voice of the Customer API is coming soon (June 2014). With Textalytics you have all the power of a fully granular and standard semantic functionality, which can be enhanced with several vertical APIs to boost productivity.

### Customer Story: Unidad Editorial

Unidad Editorial is the leading media group in the sector of global communication in Spain, with mastheads such as El Mundo, Marca and Expansión. Daedalus has embedded its semantic APIs into the new content manager of the Group to incorporate information extraction and content organization and enrichment capabilities. Thus, editors can tag, classify and automatically link their pieces, while clients receive personalized content according to their preferences. For Unidad Editorial, the result is an improvement in its offering's quality and greater audience engagement and satisfaction.

### Higher productivity, shorter time-to-market

You don't need to be an API architecture or natural language processing expert to obtain the maximum benefit from Textalytics. Vertical APIs provide a functionality optimized for any application and a great ease of use. In addition, customized resource management capabilities allow users to easily incorporate

their own semantic resources (dictionaries, taxonomies) to adapt the operation of the system to their needs. Then, SDKs and plug-ins increase the convenience and integrability of the APIs in the most common environments and platforms. The result is a fast learning curve and a short time to obtain results.

### **Main Features**

### Customer Story: WhoGotFunded

The advanced extraction of information allows identifying not only entities or concepts in a text, but also detecting the relationships between them and the meaning of these ones. WhoGotFunded is a project developed with the French company Digimind, specialized in competitive intelligence. It is an online service that applies Textalytics' technology to monitor and analyze more than a thousand online sources (media, blogs, Twitter, etc.) in various languages, to detect which startups have received funding, by which investor, in which amount and other data related to the operation. WhoGotFunded not only detects and alerts about this particular type of relationships ("financing events"), but also extracts and structures their most relevant data for a better query and analysis.

### Functionality

- Media Analysis API (for marketing agencies, press clipping services and social media analysis providers): brand monitoring / organizations / people, topics, opinion analysis, both in social and traditional media.
- Semantic Publishing API (for media companies, publishers and content providers): tagging, enrichment, proofreading and standardization of content.
- Voice of the Customer API (June 2014): UGC moderation, customer journey, buying signals, corporate reputation, perceptual mapping, CRM-oriented semantic processing.
- Core API (for developers in general who want to build their own pipelines and resources): topics extraction, sentiment analysis, classification, relationships detection, demographic profiling, connection with linked data.

Languages: English, Spanish, French, Italian.

- Technology
- ✓ Standard, Rest-based APIs.
- ✓ Hybrid technology combining deep language analysis and machine learning.
- ✓ SDKs for several programming languages: Java Python, PHP
- ✓ Plug-ins for personal productivity tools, CMSs and other enterprise applications.
- ✓ Integration with NLP environments.

#### Why use Textalytics?

- Vertical APIs and integration support features provide a quick learning curve, high productivity and a short time-to-market.
- Advanced, tested Core technology deployed in mission-critical scenarios- provides high functionality, robustness and scalability.
- > Pay-as-you-go, no investment approach (including Free Plan) reduces costs and risks.

Register at <u>textalytics.com</u> and start enjoying our Free Plan up to 500,000 processed words per month.

Textalytics is a brand of Daedalus, S. A.

