System Assessment and Validation for Emergency Responders (SAVER)

Text Mining and Analysis Software Market Survey Report

July 2016

Homeland Security
Science and Technology

U.S. Department of Homeland Security
SAVER
System Assessment and Validation for Emergency Responders

Prepared by the National Urban Security Technology Laboratory

The *Text Mining and Analysis Software Market Survey Report* was prepared by the National Urban Security Technology Laboratory for the U.S. Department of Homeland Security, Science and Technology Directorate.

The views and opinions of authors expressed herein do not necessarily reflect those of the U.S. Government.

Reference herein to any specific commercial products, processes, or services by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government.

The information and statements contained herein shall not be used for the purposes of advertising, nor to imply the endorsement or recommendation of the U.S. Government.

With respect to documentation contained herein, neither the U.S. Government nor any of its employees make any warranty, express or implied, including but not limited to the warranties of merchantability and fitness for a particular purpose. Further, neither the U.S. Government nor any of its employees assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed; nor do they represent that its use would not infringe privately owned rights.

FOREWORD

The U.S. Department of Homeland Security (DHS) established the System Assessment and Validation for Emergency Responders (SAVER) Program to assist emergency responders making procurement decisions. Located within the Science and Technology Directorate (S&T) of DHS, the SAVER Program conducts objective assessments and validations on commercially available equipment and systems and develops knowledge products that provide relevant equipment information to the emergency responder community. The SAVER Program mission includes:

- Conducting impartial, practitioner-relevant, operationally oriented assessments and validations of emergency response equipment
- Providing information, in the form of knowledge products, that enables decision-makers and responders to better select, procure, use, and maintain emergency response equipment.

SAVER Program knowledge products provide information on equipment that falls under the categories listed in the DHS Authorized Equipment List (AEL), focusing primarily on two main questions for the responder community: "What equipment is available?" and "How does it perform?" These knowledge products are shared nationally with the responder community, providing a life- and cost-saving asset to DHS, as well as to Federal, state, and local responders.

The SAVER Program is managed and executed by the National Urban Security Technology Laboratory (NUSTL). NUSTL is responsible for all SAVER activities, including selecting and prioritizing program topics, developing SAVER knowledge products, coordinating with other organizations, and ensuring flexibility and responsiveness to responder requirements.

NUSTL provides expertise and analysis on a wide range of key subject areas, including chemical, biological, radiological, nuclear, and explosive weapons detection; emergency response and recovery; and related equipment, instrumentation, and technologies. In support of this tasking, NUSTL developed this market survey report to provide emergency responders with information on text mining and analysis software, which fall under AEL reference number 13IT-00-DACQ, titled Data Acquisition.

For more information on the SAVER Program or to view additional reports on text mining and analysis software or other technologies, visit www.dhs.gov/science-and-technology/SAVER.

POINT OF CONTACT

SAVER Program
National Urban Security Technology Laboratory
U.S. Department of Homeland Security
Science and Technology Directorate
201 Varick Street
New York, NY 10014-7447

E-mail: nustl@hq.dhs.gov
Website: www.dhs.gov/science-and-technology/SAVER

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION

Text mining and analysis software is used by data analysts to scan large amounts of text from the internet, extract data from the text, and analyze and draw conclusions from the data. The software can assist first responders with collecting critical information from the internet, allowing them to respond to that information in a timely and appropriate manner. To provide emergency responders with information on commercially available text mining and analysis software, the System Assessment and Validation for Emergency Responders (SAVER) Program conducted a market survey.

This market survey report is based on information gathered between October 2015 and February 2016 from vendors, internet research, industry publications, and a government-issued Request for Information (RFI) that was posted on the Federal Business Opportunities website.[1] For inclusion in this report, the text mining and analysis software had to meet the following criteria:

- Able to scan large amounts of text from the internet in real-time
- Able to filter search results
- Able to set watch alerts for future warnings
- Commercial off-the-shelf (COTS) product.

Due diligence was performed to develop a report that is representative of products in the marketplace.

2. TEXT MINING AND ANALYSIS SOFTWARE OVERVIEW

2.1 General Overview

From a first responders' perspective, text mining and analysis software can be used to prevent incidents or better respond to incidents that have already occurred. The software can be used to extract critical information from newly published social media updates for situational awareness and evaluation in real-time, i.e., to quickly identify any potential or ongoing threats online. For example, first responders may be able to prevent an incident if there is a tweet concerning an imminent incident, or they can learn about traffic conditions around the scene of an emergency incident based on civilians' Facebook updates and images. Text mining and analysis software can provide first responders with an extra channel of information that can help them make better decisions.

Different text mining and analysis products require different levels of programming, statistical, and data analysis knowledge to use the products effectively. In general, commercial products tend to require minimal or even no knowledge in these fields, whereas free products often require intermediate or expert levels. Nonetheless, free products are mostly open source so they can quickly gain popularity among users and over time become very useful; also, most free products

[1] Federal Business Opportunities, RFI-16-01, *Text Mining and Analysis Software,* https://www.fbo.gov/index?s=opportunity&mode=form&id=0209e4df7a45196f69e3db24a69c4528&tab=core&_cview=0%20 (November 18, 2015)

do not require the user to have a high-performance computer in order to fully utilize their functionality. Some popular free products are R, Weka, and Rapidminer, which will be discussed below. Commercial products provide users with technical support; support is not provided with free, open-source products.

A graphical user interface (GUI) allows the user to visually interact with the software using items such as icons, windows, and menus, as opposed to a command-line interface, where the user interacts with the software by typing command lines. GUIs are more user friendly and intuitive; command line interfaces, in contrast, give the user more freedom to modify the software features and specifications, but require a background in programming.

Under some circumstances, text mining and analysis software can be performance demanding, requiring the user to have a high-performance computer. However, if the product is web-browser based, no installation on the user's computer is required, computing is done in the cloud, and the user does not require a high-performance computer. Purchasers should review the system hardware requirements before purchase. The three key hardware components that need to be considered are the central processing unit, random-access memory, and hard-drive space.

A software development kit or software developer's kit (SDK) is a set of software development tools that allows additional features to be created within software. The SDKs mentioned in this report are sold separately by the vendors and require the user to have some background in programming in order to fully utilize the SDK.

There are no current standards or regulations for text mining and analysis software.

Some text mining software companies charge their users not only by the number of licenses they purchase, but also by the number of requests the software processes. Processing fees can easily accumulate to tens of thousands of dollars; thus, the user should be aware of all costs and fees before purchase. Some companies may not be willing to openly discuss their charges until the customer appears ready to purchase the product. Moreover, if the user requests special software features, specifications, or modifications, as frequently happens with government agencies, the software companies often charge additional fees.

2.2 Capabilities

2.2.1 Advanced Search Options

The amount of data produced on the internet each second is enormous, so if the software is unable to narrow down its search to certain time periods, websites, and geographical areas, many irrelevant results would be extracted, and first responders would have to further sieve the results. Results are often time-sensitive and require first responders to react immediately. Therefore, first responders should pay attention to how easily and quickly the user can configure the product's search options and configure the results, as this can greatly impact the value and interpretation of the results. For instance, if a first responder from New York is unable to set the search parameters to *within the last hour, Twitter,* and *New York City,* results from other times, social sites, and cities would also be retrieved, and the number of results could easily accumulate to hundreds of thousands.

2.2.2 Boolean Operations

Boolean operations are widely known as *and*, *or*, *true*, *false*, and *not*, and with the assistance of parenthesis, they allow users to further define their search parameters. For example, the user can conduct a search with these three keywords: *New York City and fire and 911*, where all three keywords must be present in order for a result to be included in the search. Alternatively, the user can define the parameters as *New York City and (fire or 911)*; in this case, a result that contains *fire* but not *911* would still be included in the search and vice versa.

2.2.3 Alarm Monitoring

Alarms can be set to alert responders to potential emergencies in order to prevent or mitigate emergencies. If an incident has already occurred, first responders can use alarms to continue monitoring the scene while additional information is collected. The software should be able to establish watch alarms as an autonomous tool. If a fire department, for example, is able to set an alarm to scan social updates with keywords such as *fire* or *smoke*, images from the scene might be available to the firefighters before they arrive on the scene.

2.2.4 Data Visualization

Most products, commercial or free, provide good functionality on data visualization. The user is allowed to create graphs, charts, and maps for the text data he/she collects, and many products even allow the visualization to be automatically updated on a regular basis.

2.2.5 Real-Time Calculations

In order for first responders to promptly react to emergency events, the software needs to be able to search text and filter the results in seconds. As previously mentioned, how easily the user can define the search parameters can have a great impact on the speed by which useful results are obtained.

2.3 Applications

2.3.1 Trend Algorithm

Some text mining and analysis products have the capability to track words and phrases to identify those that are "trending," i.e., becoming more popular than usual. Practically, this means that breaking news incidents are highlighted to decision-makers before they become widely known, giving first responders time to organize, plan, and respond. A trending function can act as an additional alarm, alerting first responders to emerging incidents in near real time.

2.3.2 Sentiment Analysis

It is fairly common for retailers and manufactures to scan through the comments and feedback they receive from their customers to determine if, overall, customers are content with their products and services. Words such as *disappointed* or *happy* are the major targets in sentiment analysis since they may reflect an emotional response. The software counts the frequency of such keywords, assigns them to different segments, and determines if a review is positive, negative, or neutral. From first responders' point of view, emotional keywords in social updates, along with other triggering keywords in alarm monitoring, can help first responders to better identify real threats and filter out false alarms.

2.4 Emerging Technologies

Users with a background in programming and data analysis can gain advantages if the product is open source. Experienced users may be able to write their own functions to meet specific needs or utilize functions created by the software community. Many analysts tend to modify useful functions rather than start from scratch. For instance, an open source data analytics product may not be designed for text analytics, but since there are users who have written functions for text mining and analysis, new users can simply download these packages and enjoy the features that were not originally included in the product.

Some text mining and analysis products are designed to run on a web-browser. The four major web-browsers people commonly use are Microsoft Internet Explorer, Mozilla Firefox, Google Chrome, and Apple Safari. The user should first consult with the vendor to see if the product he/she intends to purchase is web-browser based and, if so, which browsers and versions it supports. The immediate advantage of web-browser- based products is that they do not require users to download and install the software; thus, web-browser-based products do not require a high-performance computer, can easily run on a laptop, and can be accessed with any computer with internet access. Moreover, since such products can be accessed with any computer, a login credential is needed to ensure only authorized users have access. Users should be aware that technical issues in using the product could arise if a new browser update is installed.

3. PRODUCT INFORMATION–COMMERCIAL SOFTWARE

This section provides information on seven text mining and analysis products that vary in price and functionality. Product information was obtained directly from manufacturers' responses to the RFI, and from manufacturer and vendor websites. The information has not been independently verified by the SAVER Program.

A summary table of commercial text mining and analysis software products is provided in Appendix A.

Product characteristics in Table 3-1 are defined as follows, listed in column order:

Vendor is the developer and distributor of the product.

Product is the name of the particular product.

Version Number is a unique identification number assigned to a specific release of a software program.

Cost is the cost for using the product. Table 3-1 gives the minimum cost; more detailed cost information can be found in the sections 3.1-3.7 or from vendors' websites.

Real-Time indicates whether results can be achieved near real-time.

Event Monitor indicates whether the software includes an event monitor that allows automated/unsupervised monitoring.

Graphical User Interface indicates whether the product uses a graphical user interface (user-friendly).

Programming Language is a unique vocabulary and set of rules for writing computer programs, indicating the programming languages compatible with the product.

Algorithm indicates if the product can apply a data analysis algorithm (a set of problem solving operations) employed by the user.

Web-browser Base is a feature that allows the product to be accessed using a web-browser and be easily run on a laptop.

Demo Availablility is whether vendor allows user to try the product at no cost before purchase.

Table 3-1. Commercial Text Mining and Analysis Software

Vendor	Product	Version Number	Cost	Real Time	Event Monitor	Graphical User Interface	Programming Language	Algorithm	Web-Browser Based	Demo Availability
Provalis Research	WordStat & QDA Miner	WordStat 7.1 & QDA Miner 4.1	$2,995	With SDK	With SDK	✓	C++, C#, Visual Basic .NET, Delphi	✓	No	No
Meaning Cloud LLC	Meaning Cloud	2.0	Varies*	✓	✓	✓	Python, Java, PHP, Visual Basic	✓	✓	✓
Babel Street	Babel X	N/A	$14,640/user/year†	✓	✓	✓	N/A	N/A	✓	✓
Basic Technology	Rosette Text Analytics	7.14 for SDK	Per call basis†	✓	N/A	N/A	R, Python, C++, Java	✓	✓	✓
RepKnight Ltd.	RepKnight	N/A	About $2,800/month‡	✓	✓	✓	N/A	✓	✓	No
Expert System USA	Cogito	13.9	$100,000	✓	✓	✓	Python, C++, Java, PowerShell, Visual Basic	✓	No	✓
Averbis GmbH	Information Discovery	4.5	$100,000/year/server	✓	✓	✓	Java	✓	✓	✓

Notes:
*—no cost for usage below 42,000 requests/month
†—more details are listed in individual product descriptions, sections 3.1 to 3.7
‡—this product is priced at £2,000/month, which is equivalent to about $2,800/month based on the ratio of £1:$1.41
✓—system is equipped with corresponding feature
N/A—not applicable
SDK—software development kit, a set of software development tools that allows the creation of applications; often requires additional fees
Information in the table is based on data gathered from vendors and their websites from October 2015 to February 2016.

3.1 Provalis Research, WordStat & QDA Miner

Description

QDA Miner is a qualitative data analysis software that experts can use to manually code and annotate documents. It provides the document and database management tools that are needed for creating projects that can be analyzed by WordStat. Some features have been designed specifically for criminal justice or military applications (geocoding, mapping, and timeline, etc.); as a result, it is currently being used by some military, intelligence, criminal justice agencies such as the U.S. Special Operations Command, UK Ministry of Defense, and UK Royal Air Force.

Image courtesy of Provalis Research

WordStat is an add-on text mining and content analysis desktop tool that can be added to QDA Miner for analysts searching for a flexible tool to extract topics, identify trends, and classify or categorize documents. WordStat provides a wide range of interactive visualization tools and allows the user to build categorization and classification models. It focuses more on analysis than data collection.

A software development kit (SDK) is also available (sold separately), allowing users to apply categorization and classification models to text from outside the desktop tool (in a data collection system, for example). The SDK also allows the user to transform unstructured text into numerical data for further processing (reporting, alerts, etc.). Such a transformation can be based either on a content analysis dictionary (to measure concepts or topics) or on an automatic document classification model, both developed and validated in the WordStat desktop application. The SDK may also be used as a component to a surveillance or monitoring system when integrated with a data collection system.

Compatible Programming Languages

WordStat is incompatible with other programming languages. However, the SDK allows a programmer to integrate the text analytics models developed in the desktop application into any data collection or other type of document management software. The SDK is a 32-bit or 64-bit Windows dynamic-link library, so it can be integrated in any application that can access such a library (C++, C#, VB.Net, and Delphi, etc.).

Supported Data Algorithms

WordStat supports a wide variety of data algorithms. For text pre-processing, it supports stemming, lemmatization, and phrase extraction; for topic extraction, it supports clustering, latent semantic analysis, and multidimensional scaling; for dictionary (or taxonomy) construction, it supports integrated thesaurus, named-entity extractions, and misspelling categorization; for comparison and relationship identification, it supports cross tabulation, correspondence analysis, heat maps, and geographic information system; and for classification, it supports naive Bayes and k-nearest neighbor.

Focus Search on Certain Geographical Areas?

Yes, providing that the data imported in the WordStat project contains geographic area information, it is possible to filter geographic information. WordStat 7.1 contains a geocoding feature that can identify the geographic coordinates from other information, including IP addresses.

Focus Search on Social Media?

No, WordStat focuses on analyzing data in documents. The company will, however, release in 2016 a web collector for Twitter and RSS feeds, and potentially for Facebook.

Level of Professional Knowledge

In order to use WordStat effectively, the user needs beginner statistical and data analysis knowledge, but no programming knowledge.

Cost

The WordStat 7.1 and QDA 4.1 package costs $2,995 per license, which includes perpetual licenses for the software, feature upgrades, and maintenance releases (bug fixes). Quantity discounts apply when more than one license is purchased; for example, the cost for five users is $2,396 per user and $2,096 per user for ten users. The SDK is sold separately for $15,000. An optional maintenance plan can be purchased, which includes priority support, advanced access to new features, and free upgrades for 18 percent of the initial cost.

Technical Support

There is no need for a trained technician to install the software package, except for administration rights or special installations. Time between major new versions is every 2.5 to 3 years, while one or two minor feature updates might be released between the major upgrades.

Minimum Hardware Requirements

The minimum hardware requirements are Windows XP, 120 MB of hard-drive space, and 2 GB of RAM (recommended 8 GB).

3.2 MeaningCloud LLC, MeaningCloud

Description

Image courtesy of MeaningCloud LLC

MeaningCloud is a set of multilingual text analytics application programming interfaces (APIs) that are customizable and integratable into application scenarios. The product can be used as an Excel add-in to extract keywords from tweet, social posts, and opinions in forums, surveys, news, conversations in contact centers, and other multilingual content. MeaningCloud can extract relevant information from texts, such as the names of specific people, locations, organizations, and concepts, as well as other important data (dates, time expressions, monetary quantities, etc.). Additionally, the product can perform sentiment analysis, classify documents by topics, and tag the names of people, places, and organizations. Lastly, MeaningCloud can automatically detect the language of texts obtained from any type of source.

Compatible Programming Languages

MeaningCloud is compatible with Python, Java, PHP, and Visual Basic.

Supported Data Algorithms

MeaningCloud supports named entity, concept recognition, text classification, text clustering, and sentiment analysis.

Level of Professional Knowledge

In order to use MeaningCloud effectively, the user requires an intermediate programming and data analysis knowledge but beginner statistical knowledge.

Cost

MeaningCloud 2.0 charges users the following rates: no fee if usage is less than 42,000 requests per month, with a limit of two requests per second; $99 per if monthly usage is less than 120,000 requests per month, with a limit of five requests per second; $399 per if monthly usage is less than 700,000 requests per month, with a limit of ten requests per second; $999 per if monthly usage is less than 4,200,000 requests per month, with a limit of 15 requests per second.
A MeaningCloud request corresponds to the analysis of any text up to 500 words. If the text exceeds this number of words, an extra request will be charged for every additional 500 words. More details can be found at www.meaningcloud.com/products/pricing.

Technical Support

MeaningCloud provides full technical support and monthly software updates.

Minimum Hardware Requirements

There is no minimum hardware requirement as MeaningCloud is a web-browser based product.

3.3 Babel Street Inc., Babel X

Description

Image courtesy of Babel Street Inc.

Babel X is an open source, intelligence monitoring platform designed specifically to meet the needs of the Intelligence Community (IC) and federal law enforcement in developing Open Source Intelligence (OSINT). The product can search for keywords and keyword combinations, phrases, hash tags, and names. It can handle requests across more than 25 social media platforms and millions of URLs and deep/dark web data. Babel X can perform cross-lingual searches across more than 200 languages, allowing users to enter terms in English and return foreign language results. The product can also build sophisticated filtering options for catered feeds and render various data visualizations. Lastly, Babel X can discover social media networks with social media link analysis and provide user anonymity in viewing collected results. The company and its products were reviewed and approved by the US Department of Justice as part of their Privacy Impact Assessment.

Level of Profession Knowledge

Babel X requires no programming, statistical, and data analysis knowledge to use the product effectively.

Event Monitoring?

Yes, users of Babel X can customize E-mail alerts, setting for daily or instant notification via E-mail; additionally, it can alert multiple people at once.

Focus Searches on Certain Geographical Areas?

Yes, Babel X can focus its search to specific geographical areas.

Focus Search Social Media?

Yes, Babel X can ingest geo-tagged social media users from numerous social media platforms.

Cost

Babel X is sold on a yearly subscription per user basis and the plans are priced at Basic-$14,640, Pro-$22,200, and Enterprise-$41,640. Discounts are available as is enterprise pricing. More details can be found at www.babelstreet.com/Product_Babel_X.aspx.

Technical Support

Babel provides full technical support, and it provides major updates quarterly and minor updates approximately bi-weekly. Additional data sources are added approximately monthly.

Minimum Hardware Requirements

There is no minimum hardware requirement as Babel X is a web-browser based product.

3.4 Basic Technology, Rosette Text Analytics

Description

Image courtesy of Basic Technology

Rosette Text Analytics offers Java, web service, or native-code text analytics for around 55 human languages. Its capabilities include language identification, tokenization of words, identification of word classes or lemmas, recognition or extraction of named entities, relationship identification, name entity indexing and matching, and name entity resolution to an authoritative source. Its specialties include analytics of short text such as tweet and broad, multilingual coverage. Its SDK includes a web service option so it can be employed on a web-browser. When used with the SDK on the premises, no internet connection is required.

Compatible Programming Languages

Rosette Text Analytics is compatible with R, Python, C++, and Java. Its SDK is available in native C++, Java package, or web service to work with any programming language.

Supported Data Algorithms

Rosette Text Analytics is based on a variety of human language technology algorithms including finite state transducers, regular expressions, gazetteers, and statistical models.

Level of Professional Knowledge

In order to use Rosette Text Analytics effectively, the user is required to have intermediate programming knowledge but no statistical and data analysis knowledge.

Real-Time Calculation?

Yes, Rosette Text Analytics can process texts in real-time for reasonable document sizes.

Focus Search on Certain Geographical Areas?

Yes, the place analytics of Rosette Text Analytics uses search engines for geographical search.

Focus Search on Social Media?

Yes, Rosette Text Analytics is not a search engine but is used with search engines, particularly for social media analytics.

Cost

The web service API is available on a per-call basis over the open internet, with a cumulative plan at $1,000 per month. On-premise deployments depend on products and languages used. Perpetual license cost information can be found at www.gsaadvantage.gov/ref_text/GS35F0422N/0P5GRV.38USHK_GS-35F-0422N_BASISDEC.PDF (page 34). With perpetual licenses, maintenance and support is available for an additional 20 percent of the license fee per year.

Technical Support

Rosette Text Analytics provides full technical support, including a 6-month update cycle.

Minimum Hardware Requirements

There is no minimum hardware requirement as Rosette Text Analytics is a web-browser based product.

3.5 RepKnight, RepKnight

Description

RepKnight is a secure, real-time cyber intelligence system that has been designed specifically to help government agencies manage risks. The product is capable of rapidly surfacing any threat or vulnerability that has a digital footprint. RepKnight captures, analyzes, and reports on web data in near real-time, including social media, in multiple languages (UF0054-8 character set). The product displays the data on a user-friendly, visually intuitive dashboard. The solution provides data analysis in under 10 seconds, according to the developer. Data is captured to an evidential standard, with exhibits easily downloaded. RepKnight can also customize keyword searches (Boolean operations), visualize the aggregated data in charts and graphs, and automate customized reports in a mobile-app or via E-mail. Data can be imported in multiple formats such as CSV, XLS, or PDF. Lastly, the product can identify popular influencers on social media and categorize them in a simple color flag system.

Image courtesy of RepKnight

Supported Data Algorithms

RepKnight's proprietary algorithms include sentiment analysis, link analysis with central weighted projection visualization, entity extraction, and trend analysis.

Level of Professional Knowledge

In order to use RepKnight effectively, the user is required to have beginner data analysis knowledge but no programming and statistical knowledge.

Real-Time Calculation?

Yes, RepKnight ingests data and displays the results analyzed for users to view in near- to real-time (within 10 seconds of posting).

Focus Search to Certain Geographical Areas?

Yes, RepKnight can segment geographic locations and search by actual location.

Event Monitoring?

Yes, users can be alerted to spikes in traffic around their chosen search terms through in-app notifications, E-mail, or text messages.

Cost

Each set of 2 million data requests costs about $2,800 per month. As an example, if the user sends 4.7 million data requests in a month, the cost would be about $8,400 (3 sets). The minimum contract term for RepKnight is 6 months with no limitation of users.

Technical Support

RepKnight provides full technical support; it has a 2-week agile development schedule with fixes to serious issues deployed as soon as possible.

Minimum Hardware Requirements

There is no minimum hardware requirement as RepKnight is a web-browser based product.

3.6 Expert System USA, Cognito

Description

Image courtesy of Expert System USA

The vendor claims Cogito can understand the meaning of words in context much the way that people do, and the product can read and process unstructured content at real-time speed. Within Cogito is Sensigrafo, a map of the language, an ontology that contains millions of word definitions and related concepts with many more millions of relationships. This combined capability provides automated disambiguation, classification, entity extraction, and metadata. Unlike traditional text mining and analysis software, this product uses keyword and statistics based technologies to analyze text, Cogito interprets the meaning of text before analyzing it. The product has been applied to the needs of government and business in areas such as security and intelligence, automated self-service, social media, knowledge management, taxonomy management, linked content, compliance.

Supported Data Algorithms

Cogito supports its own patented algorithm, where the system can extract relations between concepts included in electronic texts. Details can be found at www.google.com/patents/US7899666.

Level of Professional Knowledge

In order to use Cogito effectively, the user is required to have intermediate programming knowledge but no statistical and data analysis knowledge.

Real-Time Calculation?

Yes, the vendor provides an open real-time online demonstration that allows user to try Cogito with any content he/she enters - www.intelligenceapi.com/demo/.

Event Monitoring?

Yes, Cogito can monitor real-time events on social media and other data sources in various ways. A demonstration can be requested from Expert System USA.

Compatible Programming Languages

Cogito is compatible with Python, C++, Java, PowerShell, and Visual Basic. The library of the company's APIs allows programmers to submit text to the Cogito analytic engine and retrieve the resulting metadata.

Cost

The cost of Cognito starts at $100,000 a year with a limitation of 10MB per day, which is approximately 1,000 documents or 75,000 tweets. This license fee increases for additional data volume. If requested, the cost of any associated professional services, such as inclusion of archived data and customization of organizational taxonomies, would be additional.

Technical Support

Expert System USA provides full technical support. On average, minor updates and bug fixes are provided every 6 months and major updates about every 2 years. The company also provides 3 days of support for software integration.

Minimum Hardware Requirements

The minimum requirements to use Cogito are 4-core CPU, 8GB RAM, and 100 MB hard-drive space.

3.7 Averbis GmbH, Information Discovery

Description

Image courtesy of Averbis GmbH

Information Discovery is a text analytics and data exploration platform that allows users to analyze both structured and unstructured data and explore information in a more flexible way. The product collects and analyzes documents, such as patents, research literature, databases, websites, and other enterprise repositories. By parsing and analyzing content and creating a searchable index, Information Discovery helps

users perform text analytics across relevant data on the internet and makes data available for analysis and search. This product allows users to explore facts and relationships across sources that would otherwise be hidden in unstructured data. Information Discovery also allows users to use data in other ways, for example integrate heterogeneous data from various sources in a single application, analyze big data in a short time, structure unstructured content, drill down results using advanced filters, discover hidden facts and relations in data, and develop specific data-driven applications.

Supported Data Algorithms

Information Discovery supports vector machines, conditional random fields, naïve Bayes, maximum entropy, and decision trees.

Level of Professional Knowledge

Information Discovery requires no programming, statistical, and data analysis knowledge to use the product effectively.

Real-Time Calculation?

Yes, Information Discovery uses both text mining techniques and a search engine in its search. While text mining runs at indexing time, searching and querying happens real-time.

Compatible Programming Languages

Information Discovery is compatible with Java as it is written in Java.

Cost

The cost is about $100,000 per year per server, and this cost includes technical support and maintenance.

Technical Support

Information Discovery provides full technical support. The company provides biweekly software updates, quarterly version releases, and continuous bug fixes.

Minimum Hardware Requirements

The minimum requirements of Information Discovery are 64-bit Windows or Linux, 16 GB RAM, and 100 GB hard-drive space.

4. PRODUCT INFORMATION–OPEN SOURCE SOFTWARE

This section provides general product specifications on five free, open source text mining and analysis products. Specifications presented in Table 4-1 were obtained from internet and industry publications. The information has not been independently verified by the SAVER Program.

Product characteristics in Table 4-1 are defined as follows, listed in column order:

Developer is the product developer team.

Product is the name of the particular product.

Version Number is a unique identification number assigned to a specific release of a software program.

Real-Time indicates whether results can be achieved near real-time.

Graphical User Interface indicates whether the product uses a graphical user interface (user-friendly).

Programming Language is a unique vocabulary and set of rules for writing computer programs, indicating the programming languages compatible with the product.

Algorithm indicates if the product can apply a data analysis algorithm (a set of problem solving operations) employed by the user.

Supported Format is the list of formats the product accepts.

Table 4-1. Open Source (Cost Free) Text Mining and Analysis Software

Developer	Product	Version Number	Real Time	Graphical User	Programming	Algorithm	Supported Format
Univeristy of Sheffield	General Architecture for Text Engineering (GATE)	8.1	✓	✓	Java	✓	TXT, HTML, XML, Doc, PDF
Team NLTK	Natural Language Toolkit (NLTK)	3.1	✓	No	Python	✓	HTML, TXT, PDF, Doc
R Core Team	R	3.2.3	✓	No	C, Fortan, R	✓	TXT, R-documentation
Rapidminer	Rapidminer	6.1	✓	✓	Java	✓	40+ file types, including XSLX, CSV, SAS,
University of Waikato	Weka	3.6.13	✓	✓	Java	✓	ARFF, XRFF, CSV

Notes:

✓—system is equipped with corresponding feature

Information in the table is based on data gathered from October 2015 to February 2016.

4.1 University of Sheffield, General Architecture for Text Engineering

General Architecture for Text Engineering (GATE) is an open-source text processing tool written in Java, which was developed by the GATE Research Team in the Computer Science Department of Sheffield University. GATE can run on various operating systems, including Windows and MacOS. The product can interpret 12 languages, including English and Spanish. GATE accepts input in various formats; the most common ones are TXT, HTML, PDF, Doc, and XML documents. Within GATE there is ANNIE (A Nearly-New Information Extraction System), an information extraction system that can perform various text/data algorithms, such as sentence splitter, named entities transducer, tokenizer, etc. GATE operates on a GUI and can process queries in near real-time. More information relating to GATE, including a user guide and license information, can be found on gate.ac.uk/.

4.2 Team NLTK, Natural Language Toolkit

Natural Language Toolkit (NLTK) is a symbolic and statistical natural language processing tool written in Python, which was developed by Team NLK. NLTK supports five text mining algorithms: word and text tokenizer, n-gram and collocations, part-of-speech tagger, tree model and text chunker for capturing, and named-entity recognition. This open-source product has been successfully used as a teaching and individual studying tool, and it includes graphical demonstrations, sample data, and an official free online book to help the user to grasp the concepts in natural language processing. NLTK uses command line interface rather than GUI and can yield results in near real-time. It takes common document formats such as HTML, TXT, PDF, Doc, etc. More information about this product can be found at www.nltk.org/.

4.3 R Core Team, R

R is a data analysis tool that can perform statistical analysis, data visualization, and predictive modeling. This open-source product is primarily written in R, C, and Fortan (R itself is a programming language). R contains a wide variety of functions that can perform most text-mining algorithms and can also run on various operating system such as Windows or MacOS. R is open source not only in being cost free, but its source code is open to the public so users can examine the code and learn the coding techniques; this includes the code of its functions. While R has a command line interface, there are several GUI packages (add-ons) available. More information regarding R can be found www.r-project.org/.

4.4 Rapidminer, Rapidminer

Rapidminer is a data/text mining and machine learning tool written in Java, which was developed by the company of the same name. The product accepts more than 40 file types, including XSLX, CSV, and PDF; it also works with popular databases such as MySQL, Oracle, and Microsoft SQL Server. Rapidminer can take Twitter as a data source. Besides common data analytics functions that exist among open source text mining products, Rapidminer has extensive data management functions that allow users to blend and cleanse data. The basic edition of Rapidminer is available at no cost and the professional edition starts at $1,999 per year. The official site for the product is rapidminer.com/.

4.5 University of Waikato, Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular text mining tool written in Java, which was developed at theUniversity of Waikato. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, such as data preprocessing, clustering, classification, regression, and feature selection. The product works with its own GUI called Explorer and provides access to SQL databases. The user should be aware that all algorithms in Weka assume the data exists in one single file. Like R, Weka allows users to install packages for additional functionalities. The default file format for Weka is Attribute-Relation File Format (ARFF), but it can also take other formats such as CSV. More information about Weka can be found on weka.wikispaces.com/.

5. VENDOR CONTACT INFORMATION

Additional information on the text mining and analysis software included in this market survey report can be obtained from the vendors listed in Table 5-1.

Table 5-1. Vendor Contact Information

Vendor	Phone Number/Address	Website/E-Mail Address
Averbis GmbH	+49 761 203 97690 (Philipp Daumke, Developer) Tennenbacher Strasse 11 Freiburg, Baden- Württemberg 79106 Germany	www.averbis.com philipp.daumke@averbis.com
Babel Street Inc.	(703) 956-3572 (Brett Sutch, Sales)	www.babelstreet.com scoakley@babelstreet.com bsutch@babelstreet.com
Basic Technology	(617) 386-2000 (703) 727-0204 (Christopher Biow, Developer) 2553 Dulles View Drive, Suite 450 Herndon, VA 20171	www.basistech.com cbiow@basistech.com
Expert System USA	(703) 567-2255 (703) 732-8472 (Chris Hughes, Developer) 908 King Street, Suite 200/201 Alexandria, VA 22314	www.expertsystem.us chris.hughes@expersystem.us
MeaningCloud LLC	(646) 403-3104 +34 9133 24301 (Antonio Matarranz, Sales) 1030 Salem Road Union, NJ 07083	www.meaningcloud.com amatarranz@meaningcloud.com bgalego@meaningcloud.com info@meaningcloud.com

Vendor	Phone Number/Address	Website/E-Mail Address
Provalis Research	(514) 899-1672 (514) 899-1672, ext. 501 (Normand Peladeau, Developer) 1202-1255 Robert Bourassa Boulevard Montreal, Quebec H3B3W9 Canada	www.provalisresearch.com peladeau@provalisresearch.com
RepKnight	+44 2890 826 226 +44 7912 566786 (Laura Miller, Sales) 6B Weavers Court Linfield Road Industrial Estate Belfast, BT12 5GH Northern Ireland	www.repknight.com laura.miller@repknight.com david.henderson@repknight.com

6. SUMMARY

This market survey report provides information on 12 text mining and analysis products, including 7 commercial and 5 free, open-source products.

The commercial products differ in cost, applicable programming languages, supported data algorithms, required user knowledge, hardware requirements, whether or not they are web-browser based, and demonstration availability. All commercial products can calculate results in near real time. Six of the commercial products can monitor emergency events on social media. Five commercial products provide a graphical user interface for easy usage. Five of the commercial products support various programming languages, while six of the commercial products support common data algorithms. Five of the commercial products are web-browser based, providing easy access and minimal hardware requirements. Six of the commercial products provide free demonstrations/trails on their websites. Different products require various levels of related knowledge, but no product requires an expert level. Three of the commercial products are designed specifically for government agencies and are used or have been used by government agencies. More specific information regarding these commercial products can be obtained from the vendors.

Free text mining and analysis products can also calculate results in near real-time. All free products are open source and support various common data algorithms. However, the file formats and programming languages supported by the free products are quite different from each other. Lastly, two free products use a command line rather than a graphical user interface; they give user more freedom but require some knowledge of programming.

An important consideration in the selection of text mining and analysis software is cost and the amount of text scanned. As previously mentioned, the cost is ultimately determined by the amount of data the product handles and if any special modifications are requested.

Emergency responder agencies that consider purchasing text mining and analysis software should carefully research each product's overall capabilities and limitations in relation to their agency's operational needs.

APPENDIX A. SUMMARY TABLE OF COMMERCIAL SOFTWARE

Vendor/Product	Product Features	Price
Averbis GmbH Information Discovery INFORMATION discovery	• Analyzes both structured and unstructured data • Parses and analyzes content and creates a searchable index • Helps user perform text analytics across relevant data on the web and make data available for analysis and search • Allows user to explore facts and relationships across sources that would otherwise be hidden in unstructured data • Integrates heterogeneous data from various sources in a single application • Analyzes big data in a short time • Structures unstructured content • Drills down results using advanced filters • Discovers hidden facts and relations in data • Develops specific data-driven applications • Written in Java	$100,000/year /server
Babel Street Inc. Babel X BABEL X	• Designed specifically for federal law enforcement • Web-browser based • Performs searches for keyword, phrases, and names etc. • Handles requests across more than 25 social media platforms and millions of URLs • Performs cross-lingual searches across more than 200 languages • Quickly builds sophisticated filtering options for catered feeds • Renders various data visualizations • Discovers social media networks with social media link analysis • Provides user anonymity in viewing collected results • Reviewed and approved by the US Department of Justice as part of their Privacy Impact Assessment • Relatively low hardware requirement	$14,640
Basic Technology Rosette Text Analytics Rosette® BIG TEXT ANALYTICS	• Offers text analytics for 55 human languages • Search is conducted in a database of search engine • Capable of language identification; continue with tokenization of words; identification of word classes or lemmas; recognition or extraction of named entities; relationship identification; text indexing and matching • Functions with programming langauges of R, Python, C++, and Java • Supports algorithms of Finite State Transducers, regular expressions, and gazetteers etc. • The SDK covers other related products of the company	Calls/basis*
Expert System USA Cognito EXPERT SYSTEM USA	• Understands the meaning of words in context much the way that people do • Reads and processes unstructured content at real-time speed • Contains millions of word definitions and related concepts with many more millions of relationships • Provides automated disambiguation, classification, entity extraction, and metadata • Has been applied to the needs of government in areas such as security intelligence, social media, and compliance, etc. • Provides online demonstrations, allowing user to insert and test any content	$10,000/year

Vendor/Product	Product Features	Price
MeaningCloud LLC Meaning Cloud	• Web-browser based • Can be used as an Excel add-in • Extracts the meaning from social media and opinions in forums • Extracts relevant information from texts • Performs sentiment analysis and classifies by topics from spreadsheets and without programming • Tags names of people, places, or organizations • Assigns categories to texts, allowing to filter, sort, or group texts • Relatively low hardware requirement	Varies†
Provalis Research Word Stat 7.1 & QDA Miner 4.1	• Desktop software • Extract topics, identify trends, and classify documents or categorize them • Able to build categorization and classification models • Focuses more on analysis than data collection • Software development kit (SDK) is also available • Some features designed specifically for criminal justice or military applications • Special licenses for military secured computers, where the computer is no longer connected to the internet and provides no output capabilities	$2,995
RepKnight Ltd. RepKnight	• Designed specifically to help government organizations manage risks • Capable of surfacing any threat that has a digital footprint • Captures, analyzes, and reports on data in near real-time from the web • Searches in multiple languages • Displays data in dashboard • Delivers analyzed data in under 10 seconds • Customizes keyword searches • Provides data visualization • Automates customized reports • Compiles data in multiple formats • Identifies influencers of social media	About $2,800/ month/set*

Notes:
*—More details are listed in individual product descriptions, sections 3.1-3.7
†—No cost for usage below 42,000 requests/month
Information in the table based on data gathered from October 2015 to February 2016.