A Review of Text Analytics Suppliers
The Text Analytics Market

Text analytics is still largely an immature science, and embraces several different approaches. Natural language processing (NLP) includes dozens of techniques for accomplishing tasks such as language translation, document categorization and tagging, extraction of meaningful terms and so on. Text mining on the other hand is primarily concerned with the extraction of meaningful metrics from unstructured text data so they can be fed into data mining algorithms for pattern discovery.

Some suppliers have applied text analytics to very specific business problems, usually centring on customer data and sentiment analysis. This is an evolving field and the next few years should see significant progress. Other suppliers provide NLP based technologies so that documents can be categorized and meaning extracted from them. Text mining platforms are a more recent phenomenon and provide a mechanism to discover patterns which might be used in operational activities. Text is used to generate extra features which might be added to structured data for more accurate pattern discovery. There is of course overlap and most suppliers provide a mixture of capabilities. Finally we should not forget information retrieval, more often branded as enterprise search technology, where the aim is simply to provide a means of discovering and accessing data that are relevant to a particular query. This is a separate topic to a large extent, although again there is overlap.

The terms ‘text mining’, ‘machine learning’ and ‘natural language processing’ have different meaning depending on who you speak with. For our purposes ‘text mining’ is the application of algorithms to text data for the purpose of finding exploitable patterns. ‘Machine learning’ is when a software system learns something so that a task can be performed more effectively next time around. NLP is a set of capabilities which allow humans to extract meaning from text data.

The following lists categorize suppliers by the functionality they offer.

Customer and Sentiment Oriented Text Analytics

- **AlchemyAPI** - sentiment analysis, marketing, content discovery, business intelligence.
- **Angoss KnowledgeREADER** - part of a broad suite of analytics tools and specifically addresses text analytics in the context of customer oriented and marketing applications.
- **Attensity** - majors on social analytics, but also offers a general purpose text analytics engine.
- **Clarabridge** - provides a text analytics solution with a customer experience focus.
- **Eaagle** provides text mining technology to marketing and research professionals.
- **ExpertSystem** majors on semantic analysis, employing a semantic analysis engine and complete semantic network.
- **FICO** provides sophisticated text analytics capability in its analytics tools and in the form of specific business solutions.
- **Lexalytics** is one of the forerunners in text analytics through its Salience text analytics engine.
- **Leximancer** uses 'concepts' as a primary analytic structure.
- **SAS Text Analytics** is part of the very broad analytics capability offered by SAS.

NLP

- **AlchemyAPI** - Innovative techniques using deep learning technologies (multi-layered neural networks) also go well beyond most of the competition.
- **Angoss KnowledgeREADER** - make use of the embedded Lexalytics text analytics engine.
- **Attensity** Text Analytics provides an underlying engine that embraces several unique NLP technologies.
A Review of Text Analytics Suppliers

- **Basis Technology** - delivers a variety of products and services based on multilingual text analytics and digital forensics.
- **Clarabridge** Intelligence Platform is essentially a natural language processing (NLP) engine.
- **Clustify** is built on proprietary mathematical models which measure the similarity of documents.
- **Connexor** provides a suite of text analytics tools which embrace a wide variety of NLP methods.
- **DatumBox** provides a cloud based machine learning platform with 14 separate areas of functionality.
- **ExpertSystem** majors on semantic analysis, employing a semantic analysis engine and complete semantic network.
- **FICO** provides sophisticated text analytics capability in its analytics tools and in the form of specific business solutions.
- **IBM** provides text analytics support through two products
- **IntelleXer** provides a family of tools for natural language search, document management, document comparison and the summarization and analysis of documents and web content.
- **KBSPortal** provides an NLP capability that can be implemented in-house or via a web service.
- **Lexalytics** is one of the forerunners in text analytics through its Salience text analytics engine.
- **Leximancer** uses ‘concepts’ as a primary analytic structure.
- **Linguamatics** provides a NLP capability with either in-house or cloud based implementation.
- **NetOwl** provides both text and entity analytics in the cloud and in private deployments.
- **Provalis** provides a suite of text analytics tools, each of which facilitates a particular type of text analysis.
- **AeroText** is a text extraction and text mining solution.
- **SAS Text Analytics** is part of the very broad analytics capability offered by SAS.

Text Mining

- **Angoss KnowledgeREADER** - More advanced analysis can be achieved by combining text with structured data and text can be used to generate additional features for data mining activities.
- **DatumBox** provides a cloud based machine learning platform with 14 separate areas of functionality.
- **Eaagle** provides text mining technology to marketing and research professionals.
- **FICO** provides sophisticated text analytics capability in its analytics tools and in the form of specific business solutions.
- **PolyAnalyst** from Megaputer is a data and text mining platform which embraces the complete analytics lifecycle.
- **Provalis** provides a suite of text analytics tools, each of which facilitates a particular type of text analysis.
- **AeroText** is a text extraction and text mining solution.
- **SAS Text Analytics** is part of the very broad analytics capability offered by SAS.
- **Statistica** Text Miner is part of the extensive Statistica statistical analysis and data mining product set.
AlchemyAPI
AlchemyAPI provides cloud based text analytics services to support sentiment analysis, marketing, content discovery, business intelligence, and most tasks where natural language processing is needed. An on-site capability can also be provided if needed.

The capabilities offered by AlchemyAPI go beyond those most large organizations could build in-house, and not least because the training set used to model language is 250 times larger than Wikipedia. Innovative techniques using deep learning technologies (multi-layered neural networks) also go well beyond most of the competition, and AlchemyAPI distinguishes itself by using the technology for image recognition in addition to text analytics.

The functionality is broad and includes:

- Named entity recognition for identifying people, places, companies, products and other named items.
- Sentiment analysis with sophisticated capabilities such as negation recognition, modifiers, document level, entity, keyword and quotation level sentiment.
- Keyword extraction to identify content topics.
- Concept tagging, which is capable of identifying concepts not explicitly mentioned in a document.
- Relation extraction where sentences are parsed into subject, action and object.
- Text categorization to identify most likely categories.
- Other functionality such as author extraction, language detection and text extraction.

Alchemy API was founded in 2005 and is based in Denver Colorado. Pricing plans for the cloud based services are based on transaction per day and start with a free Starter subscription.

Angoss KnowledgeREADER
KnowledgeREADER from Angoss is part of a broad suite of analytics tools and specifically addresses text analytics in the context of customer oriented and marketing applications. It majors on visual representation including dashboards for sentiment and text analysis, and also provides a somewhat unique map of the results of association mining to display words that tend to occur together.

Many of the advanced features make use of the embedded Lexalytics text analytics engine - widely recognized as one of the best in class. Entity, theme and topic extraction are supported along with decision and strategy trees for profiling, segmentation and predictive modelling. Sentiment analysis supports the visual graphing of sentiment trends and individual documents can be marked up for sentiment.

Angoss provides its technology through the cloud or on-site implementation. High levels of end-user functionality are claimed with much of the functionality available to business users. More advanced analysis can be achieved by combining text with structured data and text can be used to generate additional features for data mining activities.

Obviously this is a sophisticated product best suited to the needs of large organizations in the main, although the cloud based access will suit the needs of some mid-sized organizations too. Overall this is well suited to customer and marketing text analytics needs where text is used to gain insight into sentiment and customer behaviour.
Attensity
Attensity majors on social analytics, but also offers a general purpose text analytics engine. Four major components define the offering:

- Attensity Pipeline collects data from over one hundred million social sources as input for analysis.
- Attensity Respond provides a mechanism for responding to social comment.
- Attensity Analyze allows text in emails, call-center notes, surveys and other sources of text to be analyzed for sentiment and trend.
- Attensity Text Analytics provides an underlying engine that embraces several unique NLP technologies and a semantic annotation server for auto-classification, entity extraction and exhaustive extraction. It comes with good integration tools too so that the results of text analytics can be merged with structured data analytics.

Three horizontal solutions are offered for marketing, customer service and IT.

Basis Technology
Basis Technology delivers a variety of products and services based on multilingual text analytics and digital forensics. The Rosette platform provides morphological analysis, entity extraction, name matching and name translation in fields such as information retrieval, government intelligence, e-discovery and financial compliance.

The Rosette search and text analytics technology comes in five distinct functional units:

- RLI - Rosette Language Identifier - automatic language and character encoding identification.
- RBL - Rosette Base Linguistics - many search engines have used RBL to provide essential linguistic services such as tokenization, lemmatization, decomposing, part-of-speech tagging, sentence boundary detection, and noun phrase detection. Currently supports 40 languages.
- REX - Rosette Entity Extractor - finds entities such as names, places, organizations and dates.
- RNI - Rosette Name Indexer - matches the names of people, places and organizations written in different languages against a single, universal index.
- RNT - Rosette Name Translator - provides multilingual name translation through a combination of dictionaries, linguistic algorithms and statistical inference.

A Rosette plug-in is available for Lucene and Solr search technologies and Basis Technology provides solutions for government, social media monitoring, financial compliance, e-discovery and enterprise search.

Clarabridge
Clarabridge provides a text analytics solution with a customer experience focus. This embraces various sources of customer information including surveys, emails, social media and the call centre.

The technology addresses three essential steps in the analysis of textual information. It supports the aggregation of information from most sources imaginable, allows the information to be processed for linguistic content and the creation of categories, and finally provides a rich user interface so the results of analysis can be seen. There are three main areas of functionality:

- Clarabridge Analyze comes with the ability to tune classification models and the way sentiment is scored, and provides various reports and visualizations.
• Clarabridge Act provides a customer engagement environment for all customer facing employees by providing real-time dashboards and the mechanisms to address customer feedback.
• Clarabridge Intelligence Platform carries out analysis and is essentially a natural language processing (NLP) engine. Connections to other applications in the organization can be facilitated by Clarabridge Connect, and includes out-of-the-box connectors for salesforce, Radian 6, Lithium and other applications.

Mobile workers are well catered for by Clarabridge Go - a mobile application providing various reports and visuals. A variety of horizontal (product management, customer care, operations management, sales and marketing, human resources) and vertical solutions are also available.

**Clustify**
Clustify, used mainly by legal firms, groups related documents into clusters, providing an overview of the document set and aiding with categorization. This is done without preconceptions about keywords or taxonomies — the software analyzes the text and identifies the structure that arises naturally. Clustify can cluster millions of documents on a desktop computer in less than an hour, bringing organization to large projects.

Clustify identifies important keywords used for clustering and reports frequency information so that clusters can be browsed which contain a set of specified keywords. It also identifies a representative document for each cluster, allowing decisions to be made on other documents in the same cluster.

Uses of Clustify include taxonomy development, search engine enhancement, litigation and ad targeting. The technology is built on proprietary mathematical models which measure the similarity of documents.

**Connexor**
Connexor provides a suite of text analytics tools which embrace a wide variety of NLP methods. These include metadata discovery, name recognition, sentiment detection, language identification, automatic document summarization, document classification, text cleansing, language analysis (10 European languages) and machine translation.

Connexor's Machinese libraries transform text into linguistically analyzed structured data. This includes Machinese Phrase Tagger which splits text into word units, Machinese Syntax which shows the relationship between words and concepts and Machinese Metadata which will extract information in 10 languages.

Solutions are offered for organizations operating in defence and security, life sciences and media, and Connexor works with a wide variety of organizations (software houses, businesses, systems integrators etc.) to deliver NLP capability.

**DatumBox**
DatumBox provides a cloud based machine learning platform with 14 separate areas of functionality, much of which is relevant to text analytics. The various functions are called via a REST API and address the following types of application:

• Sentiment Analysis - classifies documents as positive, negative or neutral.
• Twitter Sentiment Analysis - specifically targeted at Twitter data.
• Subjectivity Analysis - classifies documents as subjective (personal opinions) or objective.
• Topic Classification - documents assigned to 12 thematic categories.
A Review of Text Analytics Suppliers

- Spam Detection - documents labeled as spam or nospam.
- Adult Content Detection.
- Readability Assessment - based on terms and idioms.
- Language Detection.
- Commercial Detection - commercial or non-commercial based on keywords and expressions.
- Educational Detection - based on context.
- Gender Detection - written by or targeting men/women based on words and idioms.
- Keyword Extraction.
- Text Extraction - extraction of important information from a web page.
- Document Similarity - to detect web page duplicates and plagiarism.

Eagle
Eagle provides text mining technology to marketing and research professionals. Data is loaded into Eagle and a variety of reports and charts are returned showing relevant topics and words, word clouds, and other statistics. Both online and Windows based software is offered. The Windows offering is called Full Text Mapper with good visuals to explore topics and various word statistics.

ExpertSystem
ExpertSystem majors on semantic analysis, employing a semantic analysis engine and complete semantic network for a complete understanding of text, finding hidden relationships, trends and events, and transforming unstructured information into structured data. Its Cogito semantic technology offers a complete set of features including: semantic search and natural language search, text analytics, development and management of taxonomies and ontologies, automatic categorization, extraction of data and metadata, and natural language processing.

At the heart of Cogito is the Sensigráf, a rich and comprehensive semantic network, which enables the disambiguation of terms, a major stumbling block in many text analytics technologies. Sensigráf allows Cogito to understand the meaning of words and context (Jaguar: car or animal?; apple: the fruit or the company?) - a critical differentiator between semantic technology and traditional keyword and statistics based approaches.

Sensigráf is available in different languages and contains more than 1 million concepts, more than 4 million relationships for the English language alone, and a rich set of attributes for each concept. The Cogito semantic network includes common words, which comprise 90% of all content, and rich vertical domain dictionaries including Corporate & Homeland Security, Finance, Media & Publishing, Oil & Gas, Life Sciences & Pharma, Government and Telecommunications, providing rich contextual understanding that improves precision and recall in the process information retrieval and management.

The technology has found uses in CRM applications, product development, competitive intelligence, marketing and many activities where knowledge sharing is critical.

FICO
FICO provides sophisticated text analytics capability in its analytics tools and in the form of specific business solutions. At the heart of the offering is its Data Management Integration Platform (DMIP) addressing the complex issues associated with accessing diverse data sources and supporting a variety of analytics tools. Linguistic analysis supports tagging, dependency analysis, named entity extraction and intention analysis. Model Builder supports most forms of text analysis with parsing, indexing, stop words, n-grams, stemming, 'bag of words' analysis and so on. Some particularly
sophisticated text analytics solutions are offered for fraud detection, employing Latent Dirichlet Allocation as a method of categorizing customers. In its traditional banking, insurance and financial services markets FICO utilizes text analytics to provide additional features in its scorecard technology.

A cloud based solution text analytics solution will soon be available. While Model Builder is a large sophisticated product, the cloud based offering will provide a much easier user interface when it is launched later in the year.

IBM
IBM provides text analytics support through two products. IBM Content Analytics is primarily an extension of enterprise search technologies that adds several useful visualizations to discover structure within text data. LanguageWare on the other hand leverages natural language processing (NLP) to facilitate several types of text analysis.

A major component within IBM Content Analytics is IBM Content Analytics with Enterprise Search. This supports the visualization of trends, patterns within text and relationships. Facets feature highly in the analysis. These are categories which are derived from text analysis. For example documents on infectious diseases might be categorized by a 'hepatitis' facet. The facet-pair view shows how facets (categories) are related to each other, and a dashboard facility allows several analyses to be viewed simultaneously. A connections view displays relationships between various facet values and a sentiment view allows the sentiment behind facets to be displayed. Other components in IBM Content Analytics are targeted at specific applications including healthcare and fraud. Content Classification supports the organization of unstructured content.

LanguageWare uses NLP techniques at the document level. This includes entity and concept recognition, knowledge/information extraction and textual relationship discovery.

As always with IBM these capabilities are offered within the context of supporting infrastructure and services and will primarily be of interest to larger organizations. There is nothing particularly interesting here, and it is likely that less costly and more capable solutions will be available for many text analytics needs.

Intelexer
Intelexer provides a family of tools for natural language search, document management, document comparison and the summarization and analysis of documents and web content. Nine solutions are offered, all reasonably priced:

- Name recognition - extracts names (named entities) and defines relations between them.
- Summarizer - extracts main ideas in a document and creates a short summary.
- Categorizer - for automatic document categorization.
- Comparator - compares documents and determines the degree of proximity between them.
- Question-answering - looks for documents which answer a natural language query.
- Natural language interface - generates Boolean queries for any application.
- Related Facts - is an IE plugin for Google search and selects 5 main topics and supplements them with related facts.
- Summarizer plug-in for IE - summarizes web pages and extracts concepts.
- PDF Converter - to incorporate PDF documents into text processing.
KBSPortal
KBSPortal provides an NLP capability which includes tagging and categorizing user submitted web site content, text summarization, document linking by entities, vulgarity detection, sentiment rating and association of sentiment with products and people. This functionality is available as a web service or through purchase of source code for in-house deployment.

Lexalytics
Lexalytics is one of the forerunners in text analytics and its Salience text analytics engine is used in market research, social media monitoring, survey analysis/voice of customer, enterprise search and public policy applications. The functionality offered by Salinece includes sentiment analysis, named entity extraction, theme extraction, entity-level sentiment analysis, summarization and facet and attribute extraction. The Salience engine can be integrated into other business applications via a flexible set of APIs, and can be tuned for very specific tasks and high levels of performance.

Another essential component in the Lexalytics approach is data directories. This effectively provides a parameter driven environment with files to set up relationship patterns, sentiment analysis, and the creation of themes. Non-English support is provided through this mechanism. Each directory can be configured to support a particular task delivering considerable flexibility and power.

Leximancer
Leximancer uses ‘concepts’ as a primary analytic structure, and these are automatically identified by the software without need for existing structures such as taxonomies or ontologies. Analysis is presented through a variety of useful visualizations, with drilling down to individual documents. It is used in survey analysis, market research, social media monitoring, customer loyalty and forensic analysis.

Leximancer Enterprise runs on a multi-user server providing users with a browser interface, and also provides a REST web services interface for application integration. A desktop version is available as a stand-alone environment, or users can access the LexiPortal via a web browser for a web based service (charging based on usage). Moderately priced academic versions are also available.

Linguamatics
Linguamatics provides a NLP capability with either in-house or cloud based implementation. A search engine approach to mining text comes with a good query interface and the ability to drill down to individual documents. A domain knowledge plug-in supports taxonomies, thesauri and ontologies.

The technology is widely used in life sciences and healthcare and the on-line service provides access to content in this domain. A web services API supports most programming languages.

Luminoso
Luminoso is a cloud based text analytics service that calls upon a multi-lingual capability. Many of the current problems associated with text analytics (ambiguity for example) are at least partly solved by Luminoso. A variety of useful reports and visualizations provide users with a particularly good interface.

Megaputer
PolyAnalyst from Megaputer is a data and text mining platform which embraces the complete analytics lifecycle. Megaputer provides two separate software packages for text analysis. PolyAnalyst performs linguistic and semantic text analysis and coding, clustering and categorization of
documents, entity extraction, visualization of patterns, automated or manual taxonomy creation, text OLAP, and generating interactive graphical reports on results. TextAnalyst provides a list of the most important keywords in a document, a set of related keywords for each word, and the ability to automatically summarize a document or perform natural language queries.

**NetOwl**
NetOwl provides both text and entity analytics in the cloud and in private deployments. Text analytics includes Extractor to perform entity extraction, DocMatcher which compares and categorizes documents according to user defined concepts, and TextMiner for mining large amounts of text. Entity analytics is used to accurately match and identify names - important in many areas, including CRM, anti-fraud and national security. This includes NameMatcher to identify name variants from large multicultural and multilingual name databases. EntityMatcher performs identity resolution on similar databases.

**Provalis Research**
Provalis provides a suite of text analytics tools, each of which facilitates a particular type of text analysis. QDA Miner (available in a free Lite version) supports qualitative analysis with coding, annotation, retrieval and analysis of document and image collections. WordStat on the other hand supports the extraction of themes and trends, taxonomy and ontology creation, clustering and proximity analysis, and machine learning tools for document classification. SimStat, as the name suggests provides statistical analysis tools for text analysis. These three components can be purchased separately or as ProSuite, and all components are integrated with each other.

**Rocket AreoText**
AeroText is a text extraction and text mining solution that derives meaning from content contained within unstructured text documents. AeroText is capable of discovering entities (people, products, dates, places, products) and the relationships between them, as well as event discovery (contract data, PO information etc.) and subject-matter determination. AeroText is also capable of resolving ambiguities, such as relative time references, 'one and the same' matches and semantic analysis, based on context at the document, paragraph or sentence-level.

**SAS Text Analytics**
SAS Text Analytics is part of the very broad analytics capability offered by SAS. Several modules are provided including:

- SAS Contextual Analysis - for the creation of document classification models.
- SAS Enterprise Content Categorization - for automated content categorization, and various add-on modules add extra capability as needed.
- SAS Ontology Management - to define semantic relationships.
- SAS Sentiment Analysis
- SAS Text Miner - use of various supervised and unsupervised techniques.

**Statistica Text Mining**
Statistica Text Miner is part of the extensive Statistica statistical analysis and data mining product set. Extensive pre-processing options are available with stemming and stub lists for most European languages. ‘Bag of words’ type analysis can be carried out with input to the data mining capabilities of Statistica.