



CAAT Product Brochure

Content Analyst Analytical Technology



11720 Sunrise Valley Drive
4th Floor
Reston, VA 20191

Phone: 703-391-8700
Toll Free: 888-349-9442
Web: www.contentanalyst.com
Info: info@contentanalyst.com

The Challenge of Unstructured Text

The Challenge of Unstructured Text is that it comprises the bulk of today's business documents and correspondence: word documents, emails, PDFs, manuals – all are “unstructured” text. It is straightforward to mine and manage structured data such as databases, Excel spreadsheets, and financials, but unlocking the value in unstructured text is far more difficult. If you are developing or selling a software product that uses unstructured text, then you face this problem on a regular basis and the traditional “search engine” solution doesn't really address the issues.

Not About Words

Words are at the heart of unstructured text, which is mostly correspondence and documents. The trouble with words, however, is that the value isn't about the words themselves but what they are being used to say. In other words, it's about content and context. Looking for keywords – however specific or meaningful they may be – misses the mark and much of the real content.

The Challenge of Language

Language is complex. The same word can have different meanings in different context: think of the word “bank” – is it a financial institution or the edge of a river? The same thing can be expressed in different ways using different words: is it a layoff, a reduction in force, a

restructuring, or a consolidation? Language usage changes over time, often rapidly: Webster might have given its nod of approval to “unfriend” because of its widespread use thanks to Facebook, but educators and even Microsoft® Word still think it's a grammatical error. Language is morphing at an incredible rate, affecting words and content alike.

The Solution – Document Analytics

Document Analytics is a new class of software that approaches the challenge of words and language head-on, by going further to determine the concepts that underlie how words and language is used. CAAT from Content Analyst Company uses the mathematical underpinnings of language to create a Document Analytics platform that addresses all of these challenges.



CAAT – Designed for Integration

As an OEM or Integrator, you are solving a larger business challenge – unstructured text is one of your issues but not the only one. You are faced with the task of either solving the unstructured text issues on your own, or finding a suitable third-party product that solves these issues and lets you concentrate on your core product features. What would be ideal? A document analytics platform that was designed for integration and would provide all the features you require without you becoming an expert on search and analytics.

Document Analytics is Complicated

As soon as you start addressing the issue of document analytics, you realize just how complicated it is. Do you need specialized word lists? What formats do you need to support? What technologies will you need, and are they in the public domain – and how complete and stable are these technologies? How much programming overhead will this require?

Search Engines aren't the Answer

Many credible search engines can create a powerful word-based index for you, but this is only a small part of the document analytics challenge. If words alone aren't enough, how can a search engine solve that issue? Document Analytics goes well beyond search: in fact, the core struggle for anyone using a search engine is that they provide results versus knowledge. Analytics attacks the issue of how to find knowledge, relevant information (versus

keyword-responsive), through a variety of techniques that are much more powerful than search.

CAAT – Analytics designed for Integration

CAAT provides an entire platform of Analytics capabilities, all driven by the same mathematically-based engine, all accessible from the same API toolkit, and totally interoperable. You don't have to re-architect your application to take advantage of Analytics; you simply incorporate the desired calls to the CAAT platform – which runs within your environment – and CAAT provides the power of document analytics to the appropriate module in your application. To ensure your success, CAAT provides some of the best OEM documentation in the industry and a partner-centric ContentCare program that will support you from evaluation through sales and beyond.



Multi-function Document Analytics Platform

Content Analyst engineers, architects, and mathematicians continually evolve CAAT so its range of features are always increasing. Today, key partners focus on the following CAAT capabilities:



Dynamic Clustering – CAAT takes an entire collection of information and automatically sorts it into folders and sub-folders by conceptual topics, even creating titles for each folder

Benefit – quickly organizes information in a logical fashion based on what it’s about, not the words in it. Researchers and reviewers can narrow-in to only the information which is relevant to them, and extraneous information can be discarded before it consumes valuable time, space, and resources.



Concept-based Categorization – groups documents based on content, whether or not the same words are used to describe the same topics or concepts

Benefit – quickly locates information that is relevance-related, without being flooded with “keyword-responsive” documents that aren’t on-topic. This speeds the cost of legal review, streamlines enterprise content management, and sorts through social media content



Conceptual Search – CAAT searches the way people think: by topics or concepts, versus keywords. CAAT can use an entire sentence, phrase, or even a document to find other information which is conceptually similar

Benefit – 2/3’s of keyword searches fail because they are overly inclusive or don’t find the right information, but concept searches will always find the most relevant information to the query. Because queries are natural language, even cut and pasted from actual documents, searching is faster – accuracy increases several fold.



Summarization – CAAT uses its own notions of concepts to evaluate an entire document, sentence-by-sentence, and find the most relevant sentences to the overall document, presenting them in a summary form

Benefit – identifies what a document is about from its content, versus titles and author-provided summaries which are often misleading. Researchers and reviewers can quickly determine if a document is relevant to their queries, and if so, which parts are most relevant.



Near Duplicate Detection – CAAT’s analytics includes statistical capabilities to derive a number of duplicate and near-duplicate conditions for documents and text. These include exact duplicates, duplicates that vary only in composition (the traditional “near-duplicate”) and, most significantly, documents that are conceptually near-duplicates.

Benefit – identifying information that is nearly duplicate can be more significant than finding exact matches; this information can clog-up information sources, distort search results, and waste valuable reviewers’ time. By grouping documents that are very closely matched – even if they differ only slightly – users can identify all information that is closely related earlier in their workflows.



Language Analytics – CAAT is language-agnostic – it can perform analytics on any Unicode-B information. CAAT can determine the actual languages in documents, and can operate in a cross-lingual manner, allowing users to query or organize information in one language and locate relevant information in different languages without translation.

Benefit – information comes in many languages – not just English – and users can’t rely on inexact machine translations or expensive (and lengthy) human translations only to decide the information wasn’t worthwhile anyway. CAAT’s language analytics gets users a view of information that’s relevant – regardless of language – so they can make informed decisions on multi-lingual information.



Email Analytics – CAAT’s analytics includes a number of email features: these include thread identification, metadata tracking, segment analysis, and tracking statistics. CAAT can even identify where emails should be that are missing from a string or collection.

Benefit – email is the language of commerce – and CAAT can identify not only who is communicating but what they are communicating about, and if there are other similar email strings within communications. This allows reviewers to quickly narrow-in on only the most relevant conversations among only the most appropriate recipients, and by grouping similar strings and topics, can find relevant information in a fraction of the time they might spend going through chronological email trails.

Latent Semantic Indexing – the power behind CAAT

Content Analyst Company is the original patent-holder for Latent Semantic Indexing or LSI, the mathematical technology used by CAAT.

Content Analyst Company holds over a dozen patents (either awarded or pending) covering the application of unique technology across a variety of markets and solving a number of challenges with unstructured text.

Latent Semantic Indexing or LSI takes a generally-acknowledged fact – that language is essentially

patterned or mathematic – and expands this into a powerful set of patented algorithms that apply vector mathematics to the challenge of identifying concepts.

Content Analyst’s application of LSI is very different than a traditional search engine, which uses a tagging methodology to identify documents during indexing. CAAT’s indexing includes a powerful training component, during which the LSI algorithms automatically begin detecting mathematical patterns

or concepts. Because CAAT will train itself during this process, it can identify concepts that the users didn’t know beforehand – uniquely powerful for forensics, litigation and intelligence.

Content Analyst further extended CAAT’s training capability (in another of its patents) to add language capabilities: CAAT can train itself on sets of identical, translated documents, and gain a basic knowledge of language, i.e. how the same thing would be said in French and English.



ContentCare – an Essential Part of Content Analyst’s Partner Strategy

Content Analyst Company goes way beyond merely licensing our SDK toolkit to partners who develop innovative, compelling solutions using analytics. Our ContentCare® program is a holistic suite of services that has been recognized for its singular focus on ensuring our partners’ success with CAAT analytics.

ContentCare begins with an evaluation methodology but continues through a partner’s rollout and deployment of their analytics-powered solution:

Evaluation Support

{ Structured Evaluation Plan, Evaluation Criteria, and Evaluation Timeline

Lets you focus on where you expect CAAT can provide the greatest value

Integration Support

{ Detailed Integration Strategy, Ongoing Technical and Strategic Consultation

Ensures your integration is rapid and efficient

Solutions Support

{ Solution Architects who work in conjunction with your team on early projects

Ensures the success of your initial CAAT deployment

Analytics Certification



{ In-depth technical training and Certification on using CAAT

Provides solid working knowledge for your field technical team

Sales Training

{ Sales Tactics and Strategies to sell the value of Analytics

Provides your sales team the tools to successfully sell CAAT-powered solutions

