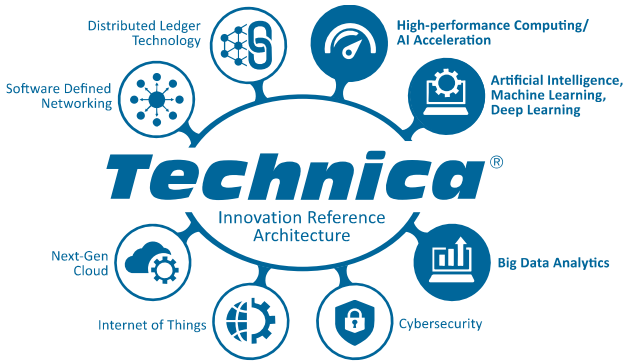
cleArInsights is a Technica AI solution that provides cognitive search and discovery capabilities beyond Enterprise Search. While useful, Enterprise Search is a mature technology, relying on a relatively crude system of recognizing keywords and returning exact matches.

The Technica Innovation Platform White Paper Series presents advanced topics that will drive competitive advantage for next-generation IT over the next three-to-five years.



cleArInsights FOR COGNITIVE SEARCH AND DISCOVERY

Table of Contents

When first envisioned in 2016, cleArInsights sought to extend enterprise keyword searches with discovery via graph analytic algorithms and graph visualizations. Over time, Artificial Intelligence (AI) capabilities were added, including Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) algorithms.

The amount of unstructured data (text, images, audio, etc.) vastly exceeds structured data (databases, tables, etc.). However, until recent advancements in ML, DL, and NLP, unstructured data sources have been off-limits to analytics software and—by extension—analysts.

cleArInsights' NLP methods bring structure and analytics to unstructured data. This is done by analyzing and extracting highly correlated concepts and relationships from vast document collections. Then, Technica's cognitive search and discovery AI solution automatically recognizes parts-of-speech, acronyms, compound words, sentences, and phrases to produce a variety of machine and human-readable visualizations for more in-depth insight.

In a nutshell, cleArInsights' capabilities consist of:

- **Enhanced NLP with configurable Taxonomies**
- **Cognitive discovery – using graph analytics and NLP to reveal synonyms, topics, keywords, and relationships; accessible via the discovery Application Programming Interface (API)**
- **Cognitive search – NLP with ML/DL to rank and return semantically similar documents—available via the search API**
- **Configurable visualizations, user interfaces (UIs), and dashboards**

cleArInsights' capabilities were incorporated into two APIs: the discovery API and the search API. The discovery API uses NLP and graph analytics and provides various graph visualizations. The search API leverages NLP and ML. Thus, cleArInsights is a software tool that offers several configurable AI algorithms

and visualizations that dramatically improve the effectiveness of knowledge discovery—extending traditional Enterprise Search.

Technica's solution uses sophisticated AI algorithms to analyze unstructured text documents. The discovery API is used to explore and discover the critical elements of a document or corpus of documents. The search API provides the ability to determine a document's relevance and recommends similar documents. Technica extended this functionality further by supporting geospatial metadata analysis for cognitive discovery of factual and contextual data within a geospatial and/or temporal cube, or within multiple geospatial boundaries.

cleArInsights' cognitive search and discovery occurs without the user needing any *a priori* knowledge of the data. The novel combination of graph analytics, NLP, ML, and advanced visualizations makes Technica's capability unique.

The current vision for cleArInsights is to allow machine and human interaction, with the machine augmenting human understanding. Over time, the goal is for the cognitive search and discovery tool to become more autonomous—with the increased merging of human and machine cognition—as well as a rudimentary ability of the machine to possess a "common sense" understanding of the text. This paper will discuss cleArInsights' background, current capabilities, architecture, use-cases, and touch briefly on future directions.

cleArInsights' discovery and search APIs provide automated discovery of themes, topics, named entities, facts, phrases, compound words, acronyms, etc. This quickly reveals content, context, and relevance to analysts. For example, if Technica's solution examined a corpus of text documents regarding "bombs," it would immediately extract keywords, themes, and topics of conversation such as, "explosive device," "detonate at a military checkpoint," and "high-value target." It would identify named entities such as "Army Officer" or "Sergeant." Important acronyms that could be processed, like "IED," would be translated into a phrase. Phrases can be treated as a single idea or synonym for a keyword.

Context, semantics, and sentiment allow the AI solution to strip out non-relevant documents. For example, many commercial solutions rely primarily on keyword occurrence and co-occurrence in documents. This can result in false positives in which the phrase "the bomb" is mistakenly interpreted as a destructive device when the phrase is employed as a term of praise. A taxonomy of terms can also be applied to incorporate domain-specific knowledge and terminology.

BACKGROUND

Enterprise Search has existed for over a decade in applications such as commercial public search engines, like Google, Bing, and Duck Duck Go. These solutions index documents and return a result-set in response to a specific keyword query. The difference from the products mentioned above is that Enterprise Search indexes enterprise content for quick retrieval based on keywords instead of publicly available Internet content. Open source tools like Apache Lucene, Apache Solr, and Elasticsearch have had great success in providing Enterprise Search capabilities, so much so that Google recently removed its proprietary Enterprise Search appliance from the market.

Technica has chosen to use Forrester Research's definition of cognitive search and discovery:

> *"The new generation of enterprise search solutions that employ AI technologies such as natural language processing and machine learning to ingest, understand, organize, and query digital content from multiple data sources."*

A downside to keyword-only-search is that organizations must formulate all the information they seek to know in the form of keywords, rephrasing their question multiple times until they hit on something that provides the answer—or something close to the answer. This process can iterate many times, wasting valuable resources.

The problem magnifies when the analyst does not have a working knowledge of the subject matter contained in the documents or needs search results targeted

for temporal or geospatial boundaries or relationships. According to a survey by Forrester Research, greater than half (54%) of information workers say their work is interrupted a few times or more per month to spend time looking for or trying to get access to information, insights, and answers.[1]

Cognitive search enhances traditional search indexing by incorporating different fields of AI known as NLP, ML, and DL to derive more meaning and insights from text. DL is a form of ML that loosely mimics the human brain through the creation of neural networks. cleArInsights currently implements doc2vec, a DL algorithm, for performing document similarity analysis.

Keyword-based, Enterprise Search will always provide a vital technique to surface relevant content. However, Technica foresaw the need for additional capabilities for the enterprise—especially with coming advancements in AI. cleArInsights was created to enhance traditional Enterprise Search with AI-infused, cognitive search and discovery.

Figure 1 graphically portrays the evolution of functionality of Technica's solution from a keyword search, to cognitive search and discovery—and beyond.

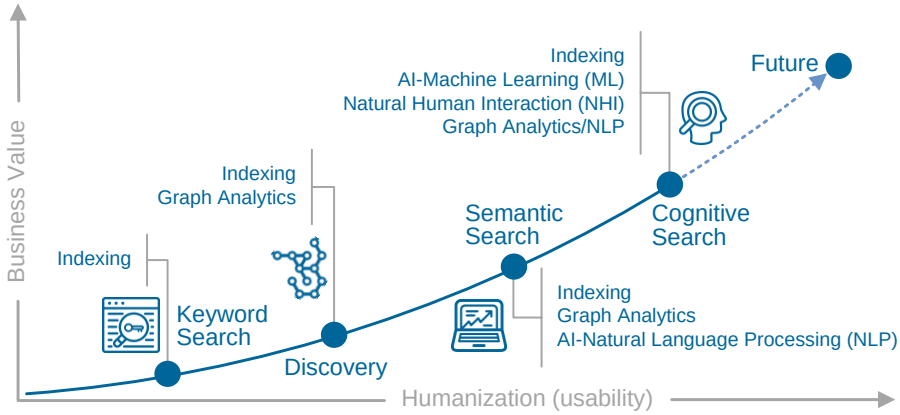


Figure 1 – Evolution of cleArInsights Functionality

cleArInsights has added functionality over time to improve both human usability (x-axis) and business value (y-axis). Keyword search, provided by Enterprise Search solutions like Elasticsearch, indexes documents, and other enterprise content for quick retrieval based on the appropriate selection of keywords. These search engines collect, index, parse and store data and metadata to facilitate fast and accurate recovery of information.

Discovery enhances keyword searches by utilizing graph analytic algorithms to represent key topics and themes within a document or database of documents and allows for configurable data visualizations. Technica combines its graph analytics with NLP to enable improved semantic understanding. NLP is a branch of AI that allows computers to understand and process text at a near human-level of cognition.

[1] Forrester Data Global Business Technographical Devices and Security Workforce Survey, 2016

Unlike keyword searches, the information provided by discovery does not require prior knowledge of search terms, enabling non-experts to execute effective, data-driven searches. As such, Technica's AI solution can identify non-obvious or even counter-intuitive relationships between topics. Thus, discovery can provide clear insights, even to experienced analysts.

cleArInsights' search capability leverages NLP, ML, and DL algorithms to compare documents to one another. Search provides the ability to perform side-by-side comparisons between selected documents and a pre-analyzed database of similar documents. This capability has several use-cases, discussed in the Section entitled, "Use-Cases."

Both cleArInsights' search and discovery capabilities are made available via APIs, i.e., the search API and the discovery API. This combination of APIs and the resultant visualizations allow the analyst to work collaboratively with the machine—a process we call Natural Human Interaction (NHI).

Our tool currently provides automated discovery of themes, topics, named entities, facts, phrases, compound words, acronyms, etc. to quickly reveal content, context, and relevance to analysts. For example, if Technica's cleArInsights solution examined a corpus of text documents regarding "bombs," it would immediately extract keywords, themes, and topics of conversation such as "explosive device," "detonate at a military checkpoint," and "high-value target." Context, semantics, and sentiment allow cleArInsights to strip out non-relevant documents.

For example, many commercial solutions rely primarily on keyword occurrence and co-occurrence in documents. This can result in false positives. For example, the phrase "a dog" can be mistakenly interpreted as a canine—when the phrase is being used to describe a poorly performing equity holding in a stock portfolio. A taxonomy can also be applied to incorporate subject-specific knowledge or terminology.

Taxonomies can be retrieved from publicly available sources or generated from existing enterprise data. The ability to consume taxonomies allows cleArInsights to aggregate terms into higher-level categories. For example, "cheddar" is a "cheese" and "cheese" is a "food." "Brie" is another type of "cheese." In graphical form, the taxonomy for this example can be seen in **Figure 2**.

Figure 2 – Example Taxonomy

cleArInsights can also use different types of geospatial information for added context. The solution can connect to a customer's mapping service and apply the available map data as part of the search and discovery processes. Technica's AI solution can further enrich the process by restricting or identifying relationships

Technica's admittedly aggressive objective for cleArInsights over time—by expanding the catalog of advanced NLP, ML and DL algorithms—is to fuse human and machine understanding to produce "common-sense" intelligence that can be leveraged enterprise-wide. We call this aspirational goal, "Neural Fusion." The following Section will provide more details on the key terms/ideas introduced in this Section and their associated capabilities.

based upon temporal data within a corpus. Such data-driven intelligence is useful when attempting to determine when and where certain conversations happened or what events the conversations might be referencing.

cleArInsights' NHI paradigm can be tailored to use-case-specific workflows. For example, cleArInsights used by a patent examiner will have a different UI and sequence of steps/capabilities than an implementation for National Security Intelligence analysis. The idea is to augment human analysts with the capabilities of machine intelligence to save time and surface new contextual insights that previously would have been unrecognized or non-intuitive. In other words, the goal is to provide analysts with the right data, delivered at the right moment to contextually inform actions.

cleArInsights' CAPABILITIES

cleArInsights performs many diverse functions designed to give analysts the right information, at the right time, within the appropriate context. cleArInsights is not a one-size-fits-all solution: NLP, discovery and cognitive search capabilities can be mixed and matched to meet the needs of specific use-cases.

Natural Language Processing

In the field of NLP, "tokenization" refers to the process of breaking apart documents into atomic pieces, called "tokens." These tokens can be acronyms, nouns, verbs, compound phrases, punctuation, or even pre-fixes or suffixes. An example of the tokenization process is shown in **Figure 3**.

This tokenization process involves multiple steps, including:

- **Part of Speech Tagging**
- **Named Entities Identification**
- **Acronym Identification**
- **Compound Word Identification**
- **Phrase Identification**
- **Semantic Relations Development**
- **Sentence and Document Neighbor Relationship Identification**
- **Taxonomy Application**

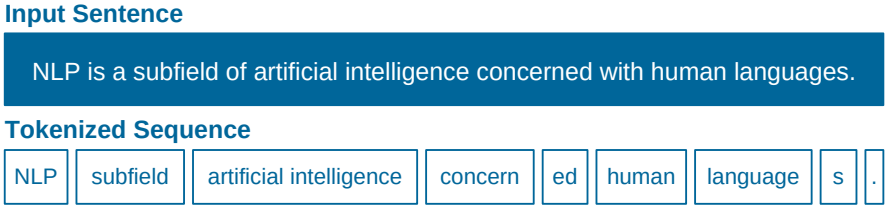


Figure 3 – Sample Tokenization

Identifying appropriate tokens (synonyms, acronyms, etc.) is an essential component of NLP. These tokens encode semantic information about the relationships between words. cleArInsights' NLP capability is used by both search and discovery to go beyond the obvious—analyzing documents based on meaning rather than word choices.

Discovery

After tokenization, cleArInsights executes graph-analytic algorithms—the key step within discovery—to compare and analyze tokens and how they are used across a database of documents. This identifies and reveals semantic relationships among tokens and documents. The analysis identifies semantically equivalent tokens (such as "Natural Language Processing" and "NLP"), significant keywords and phrases (represented as nodes), and constructs a graph connecting nodes according to relationships found in the document.

These relationships are represented as edges. Edges can be directed (including arrows), possibly indicating Topic A modifies Topic B; or undirected (without arrows), possibly indicating Topic A and Topic B mutually modify each other.

As an example, **Figures 4** and **5** provide a database of documents and the Summary Graph produced by analyzing it. This graph provides a visual representation of the discovery's understanding of the documents and topics contained within the database.

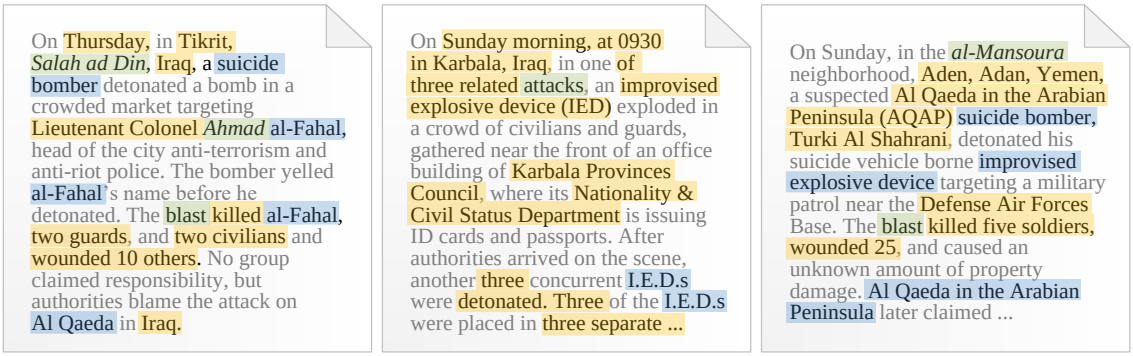


Figure 4 – Document Database

Figure 5 illustrates how cleArInsights analyzes and summarizes the contents of the database illustrated in **Figure 4** by using nodes and directed edges to reveal the most salient terms/phrases clustered and connected to reveal relationships between keywords and phrases.

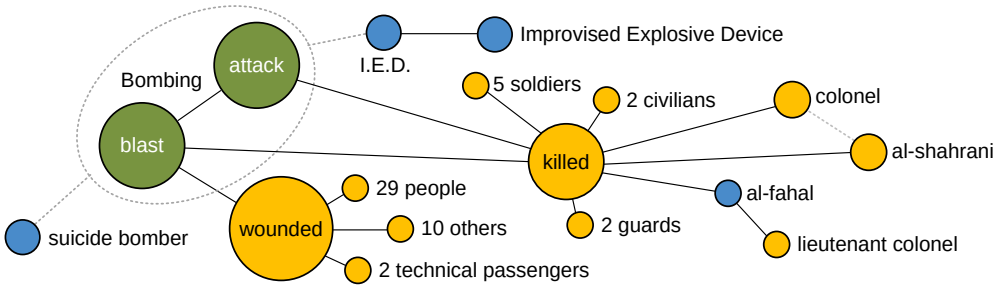


Figure 5 – Summary Graph

The nodes (circles) illustrated in **Figure 5**, depict important keywords or phrases identified in the database. A PageRank algorithm[2] is used to weight the nodes by relevance to the summary of the database, and a Louvain algorithm[3] is used to cluster nodes into broader Search Results Communities (topics). The colors in **Figure 4** and **Figure 5** indicate nodes within a topic; the nodes, in turn, define the topic, e.g., the green cluster is defined simply as the cluster containing the words "blast" and "attack."

For example, if the analyst needed to conduct discovery on a corpus of financial documents, a financial term taxonomy could provide further structure for typically unstructured text data. In this manner, specifically tailored taxonomies can be used to provide further insight into keywords and keyword relationships.

By analyzing the database in this way, discovery builds an internal understanding of the content, becoming a Subject Matter Expert on the topic within a matter of minutes. Discovery can also be configured to leverage domain-specific taxonomies, temporal, and/or geospatial data to provide further information about relationships among keywords and topics. This combination provides highly targeted awareness and actionable intelligence far beyond current Enterprise Search solutions.

[2] Originally developed by Google to rank websites in their search engine results.
[3] The Louvain Method for community detection is a method to extract communities from large networks.

Through cleArInsights' discovery capability, a human analyst can quickly find important keywords and previously unrecognized relationships. These insights are also made available to other software programs (e.g., cleArInsights' cognitive search) through discovery's API, facilitating automated downstream workflows and programmatic interactions.

Cognitive Search

cleArInsights' cognitive search API leverages NLP and ML/DL algorithms to enhance traditional search by deriving new insights. Documents retrieved as part of search and discovery are ranked by relevance according to the doc2vec ML algorithm, as described below.

Cognitive search begins with applying an ensemble of preprocessing NLP techniques to the database of documents. These techniques include identifying compound words as a single term, generalizing terms to associated categories, and parts-of-speech filtering. These techniques produce a version of each document that highlights different syntactic and semantic features. The resulting versions of the documents are then analyzed using doc2vec, an unsupervised ML algorithm that produces an "embedding vector"—a set of numerical features representing the document. A vector is a mathematical construct that has size and direction. These embedding vectors allow documents to be compared quantitatively using numerical methods such as ensemble voting or Word Mover's Distance.

The doc2vec algorithm is comprised of two stages:

Stage One - the unsupervised training occurs on the word vectors. The input consists of randomly selected fixed-length sequences of words within the document with the objective of predicting the next word in the sequence.

Stage Two - the document vectors are created with the word vectors already fixed. At each step, the document id and the word vectors of a randomly selected part of the document are input, intending to predict the next word in the document.

The doc2vec algorithm builds on the popular word2vec algorithm, embedding whole documents rather than just individual words. There is a computational advantage of doc2vec (or any ML-based embedding approach) over keyword search. The vector embeddings can be pre-computed and cached for run-time comparison. This allows computation at search time to remain light, requiring only fixed-length vector comparison over the database.

Document comparison and relevancy discovery can also be achieved using cleArInsights. For example, consider a patent examiner who needs to verify the novelty of a new patent application. This task requires the examiner to ensure all the claims in the patent application are semantically distinct from all existing and past patents, i.e., exchanging words in the claims for synonyms is insufficient to prove novelty. In this use-case, cleArInsights can quickly discover and rank highly related patents based on semantic content, thereby accelerating time-to-decision.

Visualization

Several data visualization techniques have been developed for cleArInsights. More can be added as necessary. These visualizations are designed to reveal significant topics, concepts, keywords, and relationships discovered within the database. A few are described below.

- **Full-Text View**

  Sometimes the best is also the simplest. The Full-Text View (or "raw" view for other data types) surfaces the actual text being analyzed by cleArInsights.

1. A method of defeating an RPG, the method comprising: attaching a frame to a vehicle or structure in a spaced relationship with respect to the vehicle or structure; attaching a net made of synthetic line to the frame, the net having a mesh size and configured such that when an RPG ogive impacts the net by passing through a net mesh, the net material collapses the RPG ogive rendering the RPG inoperable; and whereby when an RPG impacts the net material by passing through a net mesh, the net material collapses the RPG ogive rendering the RPG inoperable without explosive interactions.

Figure 6 – Full-Text View

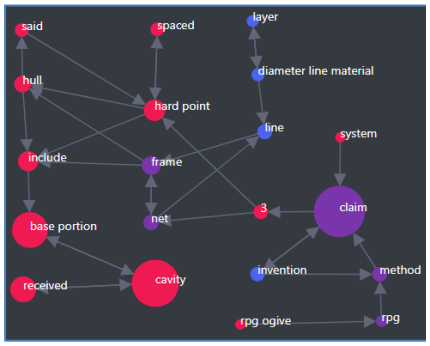


Figure 7 – Match Graph

- **Summary Graph**

  The Summary Graph identifies the main topics of conversation occurring in a document or across a database. See **Figure 5** above.

- **Match Graph**

  The Match Graph is used for document comparison. In **Figure 7**, two documents are being compared: nodes distinct to the query document are displayed in red, while nodes distinct to the comparison document are displayed in blue. Nodes common to both are shown in purple.

- **Seriated Topics Graph**

  The Seriated Topics Graph[4] is a more structured way of displaying information like the Summary Graph. "Seriation" refers to the algorithm that orders nodes within a topic so that the associated tokens can be read sequentially. **Figure 8** depicts the analysis of a resume. The columns are arbitrarily named topics. For instance, Topic 8 (orange) is defined by the nodes (rows) "data," "management," and "developed," indicating this individual may have developed data management processes or possibly previously managed a data-driven software development team.

Additionally, we can see from **Topics 0 (blue)** and **14 (green)** that this individual has a background in:

- **Physics**
- **Outreach**
- **Product**
- **Owner**
- **Project**

Lastly, we notice that the "data" circle—and bar on the right—is substantially larger than the others, indicating the highest significance/prevalence within this resume.

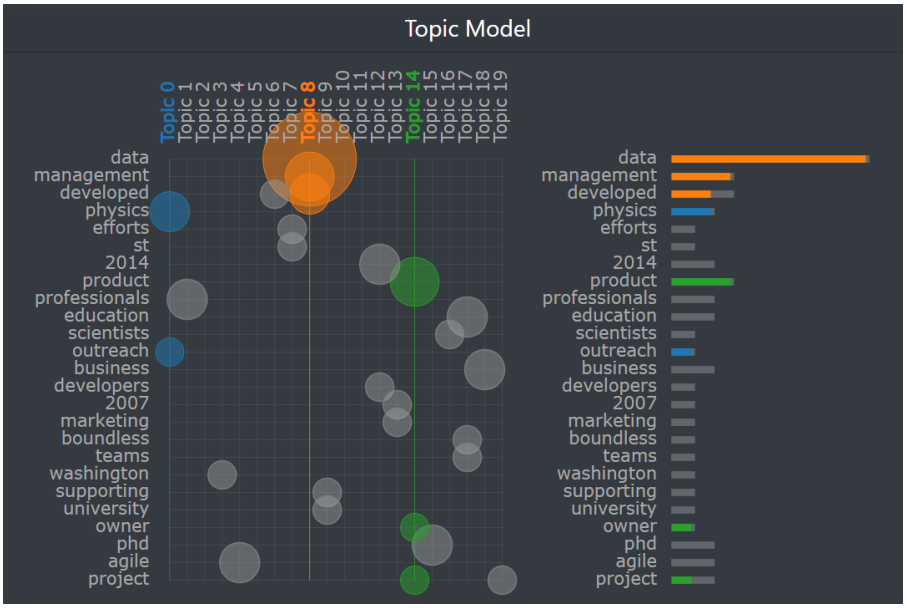


Figure 8 – Seriated Topics Graph

[4] Chuang, J. Manning, C. Heer, J. Termite: Visualization Techniques for Assessing Textual Topic Models. 2012.

The graphs are intended to accelerate analyst workflows by revealing the most salient keywords, topics, and relationships within the data. Additional visualizations and subject-specific visualization can easily be incorporated to facilitate a more natural presentation of the data for each use-case.

ARCHITECTURE

cleArInsights currently provides two UIs which can (and should) be tailored to the specific use-case: a Graphical User Interface (GUI), shown in **Figure 9**, and a REST API to facilitate programmatic workflows. Additionally, cleArInsights can be integrated into open source/proprietary UI technologies like Kibana or Sitscape.

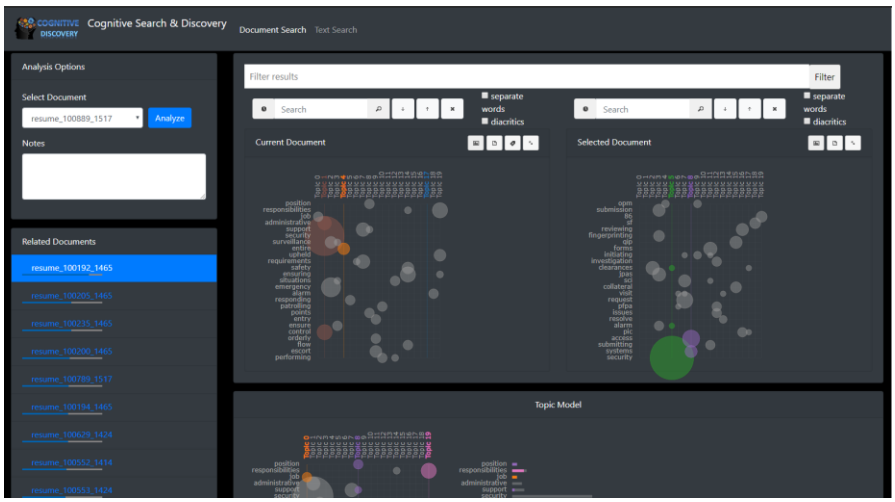


Figure 9 – cleArInsights GUI

The GUI illustrated in **Figure 9** was developed to support a Human Resources use-case, described in greater detail in the Section entitled, "Use-Cases." This GUI allows the user to select a resume from a database (**top-left** module) and identify other individuals with similar resumes within the database (**bottom-left** module). Once a resume is selected, it is analyzed by discovery and visualizations that are presented to the right. If no comparison document is selected, only the visualization for the analyzed document is presented (**bottom-right**). If a document from the bottom-left is selected for comparison, side-by-side graphics are presented for easy visual analysis (**top-right**). **Figure 9** is just one example of how the cleArInsights GUI might be designed. cleArInsights' UI architecture is modular, extensible, and configurable to meet the needs of the use-case.

cleArInsights operates as a lightweight analytics suite that can be seamlessly integrated with Solr, Cassandra, Elasticsearch, and other databases. **Figure 10** provides a notional architectural diagram, highlighting its modularity.

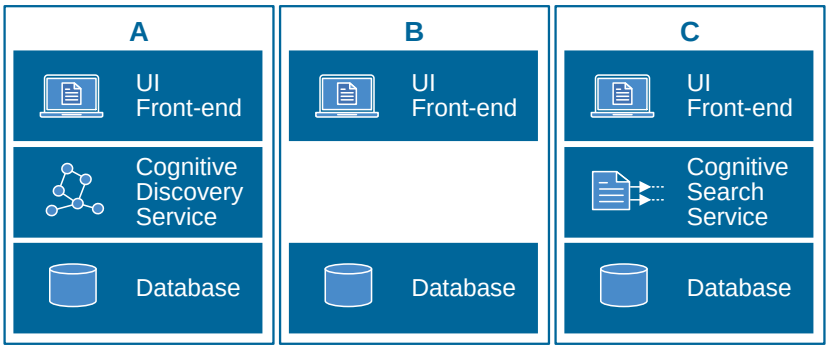


Figure 10 – Notional cleArInsights Architecture

Once a use-case is identified, Technica will work to develop the requirements, architecture, and implementation that achieves the desired objectives.

Other Use-Cases:

- **Text Mining** — aggregate relevant and accurate content from diverse data repositories to reduce costs and surface actionable intelligence.
- **Compliance & Privacy Audits** — through applicable taxonomies, ensure Personally Identifiable Information (PII) is protected according to regulatory guidelines, like Health Insurance Portability and Accountability Act (HIPPA).
- **Improved Collaboration** — improve the effective use of institutional knowledge and best practices, improving "findability" and effectiveness with information sharing in a controlled and secured manner.
- **Enhanced Classification & Taxonomies** — Utilize Discovery to surface poorly tagged content and derive more effective classification, tagging, and taxonomies. Additionally, cleArInsights can identify data that can be labeled into multiple classifications, like patent documents.

Figure 10 assumes that the repositories of enterprise data (bottom of **A, B,** and **C** above) have already been populated with content. These repositories can be pre-existing customer databases or new databases populated with customer data. cleArInsights can work with any type and number of data repositories, including Solr, Elasticsearch, MongoDB, and other SQL/NoSQL databases. Message brokering technologies like Apache Kafka or Logstash can be used to exchange and re-format data between the databases and the cleArInsights backend services: cognitive search (middle of **C**) and discovery (middle of **A**).

The workflow depicted in **Figure 10 (A)** allows the user to visualize document analyses. The cognitive discovery service can either be used to analyze a newly uploaded document and return its visualizations or return the visualizations of documents within the database. The user may access cognitive discovery through the UI Front-end or the API directly.

The workflow depicted in **Figure 10 (B)** allows the user to retrieve the full text of documents for human analysis. This workflow also supports keyword searches for document retrieval.

The workflow depicted in **Figure 10 (C)** allows the user to identify documents that are semantically similar to either a newly uploaded document or documents within the database. The cognitive search service employs doc2vec to analyze the uploaded/selected document and return the most semantically similar documents within the database.

USE-CASES

cleArInsights is not a one-size-fits-all commercial off-the-shelf (COTS) solution. In this section, we present several use-cases to inform and illustrate how cleArInsights could be extrapolated to additional use-cases.

Patent Examination

Patent examiners must evaluate the novelty of patent applications concerning prior art such as pre-existing patents, current patent applications, and already productized inventions. The current process for patent examiners consists largely of keyword searches on the internet and enterprise patent repositories (patent applications and granted patents). The method of approving or rejecting a patent application based on prior art leverages cognitive search to move beyond

specialized keyword selection and cognitive discovery to accelerate the analysis and understanding of the patent application.

This workflow begins with a patent examiner selecting a patent application from the US Patent and Trademark Office (USPTO) databases. Cognitive discovery then performs analytics and generates visualizations for the selected patent application. At the same time, the cognitive search returns the most semantically similar prior art from both the USPTO databases and other repositories, such as academic journals. The examiner then compares the patent application to the prior art utilizing their pre-computed visualizations provided by cognitive discovery. The patent examiner finally approves or rejects the patent application based on the similarities between the prior art and the patent application. This process is more streamlined than the existing manner patent examiners evaluate patents via numerous keyword searches.

Human Resources

The process of identifying the right candidate for the job can be a daunting task with a real impact on the success of the project or company. One major complication of this task is that the volume of applications for any one job can easily overwhelm an HR professional or hiring manager. This task requires comparing the applicants' resumes against the job posting to identify whether the individual meets the requirements, exceeds them, or could be quickly trained to meet them.

The workflow for this use-case begins with the set of applicant resumes for a specified job posting (the "Applicant Set") and a set of accepted resumes for similar jobs (the "Ideal Candidate Set"). The hiring manager selects the Ideal Candidate Set associated with the job posting, e.g., previously hired Software Engineers for a Software Engineering posting, previously hired Intel Analysts for an Intel Analyst posting, etc. Cognitive discovery analyzes the Ideal Candidate Set and returns a Seriated Topics Graph enabling the visualization of the Ideal Candidate. Simultaneously, cognitive search ranks the Applicant Set according to each applicants' similarity to the Ideal Candidate. Visualizations for each applicant are created by cognitive discovery and presented for side-by-side comparison with the Ideal Candidate.

Proposals Library

cleArInsights is currently deployed at Technica for in-house use to assist in proposal writing. Technica uses cleArInsights to identify proposals with similar content to the Request for Information (RFIs), Requests for Proposals (RFPs), and Broad Agency Announcements (BAAs), etc. that the proposals team is authoring. While this is beneficial for several purposes, the greatest value is derived from the ability to quickly locate sections/paragraphs that can be reused/repurposed for a new proposal—saving precious time.

Two primary issues motivated the development of this use-case. The first was frustration with SharePoint search: often documents would not be returned from searches containing relevant keywords. The second was that the proposals team found themselves rewriting the same content repeatedly. They engaged

Technica's engineers to leverage cleArInsights to accelerate the ability to find salient, reusable material.

The workflow for this use-case begins with continuously scanning for new proposals on Technica's in-house SharePoint instance. Newly uploaded proposals are automatically formatted, indexed, and rendered searchable through cleArInsights. After that, a document such as an RFI or RFP can be uploaded to the web front-end. The proposals most similar to that document are then returned for review. The topics in the document and the returned proposals can be compared rapidly with cleArInsights' visualizations to ascertain whether they match in a relevant way. The individual proposals that are relevant can be examined, either through the front-end or through a link to their location on SharePoint that is stored during indexing.

This use-case has enabled our proposal team to be more agile in crafting proposals and responses through reducing the time spent writing sections that are common throughout many proposals. Additionally, cleArInsights eliminates time spent crafting keyword searches and conducting reviews as a way to locate relevant content.

National Security

The Intelligence Community (IC) provides real-time, accurate, and actionable information to facilitate national security decisions. Errors in the information or the inability to find critical information can have global ramifications. cleArInsights identifies hidden, potentially damaging information and threatening events. Related and inter-related information is visualized even when it does not match the words used in a search query. cleArInsights provides the intel analyst actionable intelligence faster by "reading" the data and providing a human-level analytic capability. It augments the analyst's cognitive capabilities—but leaves final conclusion-making and decision-making to the analyst.

One specific cleArInsights national security use-case was developed for a major IC agency. For this use-case, subject-specific analytics and visualizations were designed to support the use of Spatio-temporal parameters, i.e., latitude, longitude, and time. These tools help analysts to perform relevant document searches based upon topics or themes and filter by location parameters, such as distance from a target area. Time is another search parameter supported in this use-case. Analysts can limit or expand document search based upon temporal criteria, such as date-time ranges. Combining temporal and spatial parameters allows results from multiple databases to be highly targeted. This improves analyst speed and performance—ultimately reducing time-to-decision.

Dark Data Discovery

Dark data is data that has been captured and stored by an organization but is not used for any purpose, such as deriving insights or supporting decision making.[5] An untapped resource. Text-based dark data can include non-digital, paper documents that need to be scanned into an Optical Character Recognition (OCR)

[5] *For more information see: Dark data* - https://en.wikipedia.org

system, and digitized. It also includes digital data that has not been indexed by an enterprise search solution. Usually, the data is so voluminous that it would be impossible—or too costly—for humans to read and comprehend. Thus, the data remains captured, but unusable, i.e., dark.

Dark data discovery was the original use-case for which cleArInsights was designed. Through the variety of analytics and visualization techniques, cleArInsights can illuminate dark data providing meaningful insights in a matter of minutes. Importantly, this can be done without any knowledge whatsoever regarding the contents of the dark data. Meaningful insights can be gleaned in a fraction of the time/cost it would take to perform a similar analysis with human effort.

Social Media Discovery

cleArInsights may be used to ingest social media data and associated metadata (e.g., geospatial information) comparing posts and individuals to identify and reveal emerging criminal jargon, terrorist associations, and stenographic data hidden by botnets.

In this field, taxonomies are often useful as a quicker way to teach NLP and ML algorithms nuances specific to social media. For instance, what exactly does the 🙃 emoji mean? Taxonomies are often used to identify synonyms, acronyms, and hierarchical relationships between words—and in this case, emojis. Such taxonomies can either be developed by human analysts (some already exist online), or they can be learned by cleArInsights.

The choice of whether to use taxonomy or not is often based on the amount of relevant data at one's disposal. Learning taxonomy through cleArInsights will often produce better analysis because it will be a more machine-natural way to capture the relationships than using a pre-constructed taxonomy. However, learning an accurate taxonomy also requires considerably more data. Either way, cleArInsights is built to accommodate and provide insights from social media.

Cross-Domain Information Sharing

Within the Department of Defense, IC, and US allies, sharing classified information is a time-consuming and completely manual process. Due to the complexity of the problem, many times, relevant mission information remains completely siloed—rather than being shared with those that need it.

cleArInsights can be used to ingest a given mission’s description/objectives, including participating mission partners and their associated roles. Then it creates a list of the relevant information to be shared. To do this, cleArInsights can ingest classified information, retain metadata on its classification level, and determine which documents are relevant to the mission and other parameters. Additionally, cleArInsights can determine the exact information in a document that is relevant/necessary for the mission and redact the rest.

With human supervision, cleArInsights can assure that no information will be shared beyond what needs to be presented, while still vastly increasing the speed at which mission-relevant data can be shared with allies.

FUTURE DIRECTIONS

Technica Labs continues to apply its data science expertise and agile development methods to push the boundaries of cognitive search and discovery innovation. The cleArInsights roadmap includes the following future enhancements:

- **Image Processing and Analysis**
- **Audio Processing and Analysis**
- **Multi-Language Support**
- **Email Adapter/Parser**

Several other possible enhancements are described in greater detail below.

cleArInsights for Collaborative Communities

Technica will explore multiple ways of facilitating the training of analysts, establishing communities of interest, and promoting the use of best research practices by capturing and leveraging the analyst's search data. The captured search data can be used to enhance collaboration and community by connecting analysts with similar search content. Some examples of collaborative tools may be:

- **A dashboard collaboration system where analysts can share in real time what they have discovered.**
- **A repository of 'like' results for similar analysts to review at their leisure.**
- **A workflow solution where similar analysts receive updates on the analysis achieved by their coworkers.**

The simplest and perhaps most powerful use of search data and collaborative software for cleArInsights would be the incorporation of a recommendation engine. This recommendation engine would compare an analyst's search history with other analysts and generate recommended queries. This could help the analyst discover data they might not have thought to query and eliminate the time-consuming task of developing optimal search strategies.

Online learning for cleArInsights

User feedback, such as document selection or exclusion, could be captured and used to improve the performance of both discovery and search. Leveraging user feedback to retrain the ML/DL models used by discovery and search is a process called "online learning." This process results in fine-tuning of the analytics to more closely match the particulars of the analyst, the use-case, their needs, and interests at that moment. Through online learning, cleArInsights will continue to learn and grow as analysts perform more research over time.

SUMMARY

Technica has achieved advances in the cognitive computing domain utilizing a novel approach to textual/document analysis. No longer does a document analyst need to labor over a search interface, inputting keyword after keyword and parse through result-set after result-set.

Technica's approach uniquely integrates best-of-breed solutions from multiple domains, including NLP, graph analytics, unsupervised ML and DL, and configurable UIs that allow for collaboration.

cleArInsights is an extensible solution, with adaptable workflows and visualizations for specific use-cases. Additionally, value-added algorithms can be developed and integrated to support new use-cases as necessary. Overtime, Technica's cleArInsights AI capability will move toward greater autonomy and increasingly possess a "common-sense" understanding of the documents it ingests and analyzes.

Technica provides professional services, products, and innovative technology solutions to the Federal Government. We specialize in network operations and infrastructure; cyber defense and security; government application integration; systems engineering and training; and product research, deployment planning, and support.

Technica®
22970 Indian Creek Drive, Suite 500
Dulles, VA 20166
703.662.2000