SentNet

Next Generation Document Scoring Using Advanced Graph Analytics

Executive Summary

The volatility of the world in the twenty-first century necessitates that leaders have access to detailed intelligence in a rapid manner in order to make informed and qualified decisions. Whereas human analysts succeeded in this role for decades, the sheer quantity of information available today and the speed and which it is transmitted necessitates that new technological approaches be implemented in order to achieve the necessary level of effective and timely intelligence products.

Natural Language Processing (NLP) offers the most comprehensive and effective methodological platform upon which to build an automated evaluation solution. Utilizing techniques such as a Graph Model, Sequence Similarity and Perceptual Hashing, SentNet will efficiently and accurately classify finished IC products through a two-pronged approach. It will first create a comprehensive scorecard capable of grading all products across all current analytic criteria which will then be utilized to score all new IC reports as they are finished. SentNet will learn from each new report it sees in order to improve its scoring platform, providing the added benefit of evolving alongside and ever-changing world.

Given our team's decade of work in the intelligence industry and the data science domain, including the delivery of tools and solutions involving graph-based text analytics, we firmly believe that SentNet has the ability to transform the IC analytic toolset. Our unique solution, combined with industry-specific experience, offers a creative and trustworthy solution to a field that is ripe for innovation.

Overview

Natural Language Processing (NLP) made huge strides in the late twentieth and early twenty-first centuries and is now commonly utilized across a wide variety of industries including business, education and information technology. However, particular advances in the last few years have resulted in significant increases in the number and variety of NLP techniques that are available to analysts. In this document, we outline a new system called SentNet that utilizes many of these new innovations that will enable agencies to automatically and accurately revise and score intelligence products. In the following, we will first review existing solutions for reviewing documents. Next we will outline SentNet's core features that improve upon these existing techniques as well as SentNet's underlying methods. Finally, we will provide an overview of user workflows in SentNet and offer some summarizing thoughts.

Existing Solutions

The ability to use computers to automate the task of reading, comprehending, and evaluating human-composed documentation has made significant strides in the last two decades. Advances in NLP, have enabled analytics professionals to derive insights from vast quantities of text-based data to solve a wide range of problems including translation, sentiment analysis, document summarization, topic extraction and document classification. NLP has applications across countless domains including marketing, information technology, healthcare, education, and defense.

Specific implementations of NLP to automatic document evaluation and scoring generally fall under the umbrella of two main methodologies. The first is a structure-based analysis found in Applicant Tracking Systems (ATS) commonly employed by human resource departments to automatically filter out electronic copies of resumes as part of the hiring process. Such systems focus their efforts on the identification, validation and/or extraction of predefined entities or structures within the document The other approach, Automatic Essay Scoring (AES), is utilized in the educational field as a means of automatically grading computer-based essays such as the Graduate Management Admission Test (GMAT) and the Graduate Records Examination. Rather than concerning themselves with text structure, AES systems key in on semantics (i.e. topic relevance, flow, etc.) and text-production skills (i.e. grammar, vocabulary, spelling, etc.).

The primary benefits associated with both of these systems are speed and standardization. With the rapid increase in the amount of text-based data available to organizations, the ability to rapidly sift through data in a quick and efficient manner is vital. For example, ATS is a vital tool for modern human resource departments in rapidly sifting through the large amount of information generated by online job applications. Similarly, the use of automated systems also has the added benefit of normalizing the process used to evaluate and/or score associated documents. In the case of AES systems, the last document graded will be treated as equally as the first utilizing the same methodology. It is much more difficult to ensure a consistent evaluation from a human evaluator for every document they inspect.

Unfortunately, these solutions are not free of problems which have hindered their ability to fully replicate the performance of humans. The main weakness of structure-based/text-production solutions is their overreliance on objective, rule-based criteria for classifying and scoring. If the task at hand is subjective in nature, semantic-based analysis is required, systems constrained to predefined models will often struggle to recognize context and creativity. Furthermore, such systems will struggle to adapt to text and language it was not trained on such as changing style, structure, or language; events that would be expected to commonly occur in the IC.

Proposed Solution

While ATS and AES continue to be used across multiple industries, operational requirements at ODNI necessitate a next generation solution. A new approach capable of identifying, understanding, and scoring documents using significantly more complex contextual features is desired. This new solution must also allow analysts and analytic managers to easily understand and validate these scores with minimal training.

To address this need, we are proposing to develop a new system called SentNet. In short, SentNet is a next generation document analysis tool that will utilize advanced part-of-speech tagging and graph analysis techniques to evaluate and score documents using numerous contextual features. In doing so, SentNet will rapidly evaluate and numerically score intelligence products against one (or multiple) criteria with no intervention while allowing analysts to easily validate scores using an intuitive user interface.

Below we first describe the methods and analytics that underlie SentNet and the key capabilities that these methods provide users. Next, we will walk through a user workflow in an example SentNet interface. Finally, we map SentNet's capabilities to the Challenge Requirements and summarize with some additional thoughts.

Features

At its core, SentNet is an advanced network analysis tool. Using the power of graphs, SentNet is able to detect relationships between textual elements and identify analytical reasoning with respect to those elements. Representing a document as a graph allows SentNet to use individual and aggregate features from a document's derived word network to assign scores for each of the RSEATS (*Rating Scale of Evaluating Analytic Tradecraft Standards)* criteria. SentNet goes beyond existing document scoring methods to offer ODNI these additional capabilities:

- **Complex Document Scoring** – In comparison to ATS, AES, and other existing methods, SentNet allows for complex document scoring by revealing complex and latent textual and semantic features within a document. These features are extracted using established methods drawn from advanced natural language processing and social network analysis (described in more detail in the next section). Upon extracting these generalized features from an analytic product, SentNet's models can predict RSEATS scores for that product based on generalized semantic features that have been found to represent arguments, judgments, assessments and assumptions.

- **Refinement of valuable, existing document analysis methods -** SentNet includes many valuable existing capabilities that are used for the scoring of documents including:
  - Ability to search for the presence or absence of specific n-grams (terms and/or a collection of terms)
  - Generation of a readability score to assess the understandability of text utilizing existing indices (i.e. Flesch-Kincaid, Gunning Fog Index, Cloeman-Liau Index, etc.)
  - Comparing media (charts, graphs, and maps) in new documents to media used in previous reports using perceptual hashing
- **Intuitive Visual Representation** – SentNet allows analysts and analytic managers to view the relationships between a report's elements, as well as the RSEATS scores derived from those elements, using an intuitive and user interface.

Methodology

In order to generate RSEATS scores for new documents, SentNet must first generate a RSEATS Scorecard model. This requires that SentNet have access to a sample of previously scored analytic products (similar to the Sample Analytics Product provided for this challenge). Using this sample, SentNet will import, transform, analyze, and summarize complex textual features to develop a robust model that is capable of predicting RSEATS scores for a new document. Developing such a model involves the following steps:

1. **WordNet Generalization** – SentNet will first take a sample of existing analytic products (defined by a user or agency) and "generalize" them using an advanced part-of-speech tagger based on the WordNet Lexical Database (an open source, license free, lexical database hosted by Princeton University)[1]. The WordNet database groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept. In this way WordNet resembles a thesaurus, given that it groups words together based on their meanings. However, there are some important differences between WordNet and a Thesaurus. First, WordNet interlinks not just word forms, but specific senses of words. As a result, words that are found in close proximity to one another in WordNet are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus do not follow any explicit pattern other than meaning similarity.
   SentNet makes use of these unique features by substituting all words in a document for their highest order (most generalizable) synset equivalent represented by a synset code. To assist with this, SentNet would use a customized WordNet lexical database that has been tailored to include topics, terms, and acronyms that are commonly discussed in the US Intelligence Community. In contrast to other text generalization and 'part-of-speech' tagging solutions, WordNet enables SentNet to more easily identify and compare common semantic features and structures (including arguments, methods, assumptions and judgements) across analytic products.

[1] Princeton University. (2018). *WordNet: A lexical database for English*. Retrieved from https://wordnet.princeton.edu/

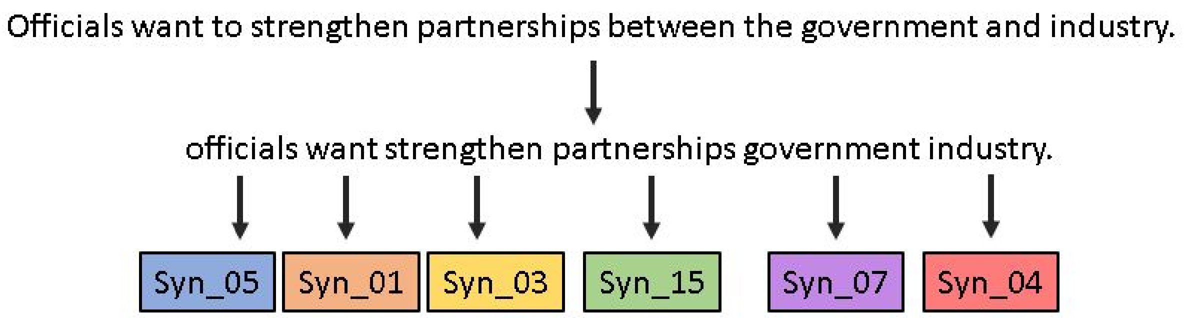

Officials want to strengthen partnerships between the government and industry.
officials want strengthen partnerships government industry.
Syn_05
Syn_01
Syn_03
Syn_15
Syn_07
Syn_04

Figure 1 - WordNet Generalization of Sentence

2. **Graph Extraction** - Once a document has been generalized into its synsets using the WordNet database, documents will be converted into textual graphs representing the relationships between all synsets in a document based on their co-location (i.e. synsets in the same sentence will be considered to be connected).

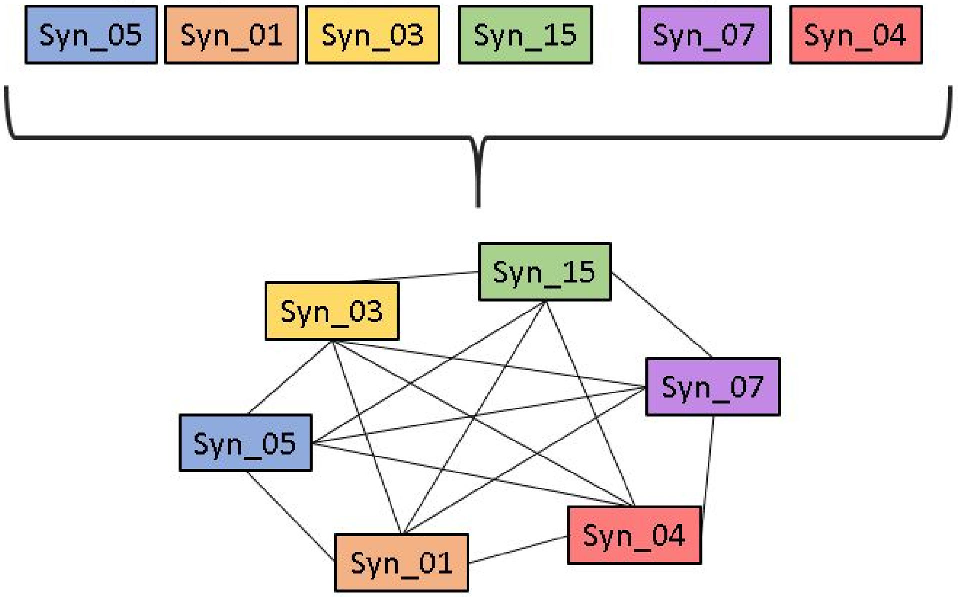


Figure 2 - Graph Construction from Synsets

Once all analytic products in our sample are represented in this way all graphs will be scanned to identify common synset features that pertain to key RSEATS criteria including arguments, methods, assumptions and judgements found within analytic products. Examples of such features could include:

- ***Synset Presence/Absence*** - SentNet will scan for the presence or absence of a single synset that relates to key RSEATS criteria such as assumptions, judgements, and logical argumentation (such as synsets that contain argumentation like ‘must’, ‘however’, ‘could’). SentNet could also be trained to scan for specific edges/connections between synsets in a graph (for example the synsets that include 'must' and 'however' occuring in the same sentence).

- ***Synset Centrality* -** Metrics about the relative location of synsets within a graph can also be useful. For example, SentNet could measure the betweenness centrality (a measure of importance/influence in a graph) for key terms in the graph. For example, the betweenness centrality of the 'cyber' synset would reflect the degree to which it is the central topic for a given document.
- ***Synset Clusters & Communities*** - Finally, SentNet can look for common "clusters" or "communities" of synsets within documents that represent some the most complex and nuanced semantic features. Based on the presence/absence of these synset clusters, SentNet will be able to understand what types of arguments, methods, assumptions, judgements, and other logical features are used in an analytic product.

3. **Random Forest Modeling** - Once these features are derived, high and low performing analytic products in our sample will be compared. SentNet will use the presence/absence of these synset features as inputs to multiple random forest models (one for each criterion) to estimate the RSEATS scores for each analytic product. In these circumstances, a Random Forest model allows SentNet to consider the importance of numerous features and find interactions between features while remaining highly generalizable (reduces the likelihood of overfitting). Once trained, this model can be used to predict the RSEATS scores for new reports.

4. **Document Similarity Matching** – Once a document is scored, SentNet will run "generalized" documents through a secondary document similarity matching algorithm. This algorithm will be based on Sequence Similarity. In short, this algorithm assumes "two graphs are similar if they share many sequences or short paths of vertices and edges". This algorithm is particularly useful when comparing graphs that are generated from naturally sequenced objects (such as written documents). Documents that are highly similar (such as template reports that are updated weekly) will likely have similar RSEATS scores. Therefore, this similarity measure can be used to validate SentNet's estimated scores for a given report if a similar/nearly identical report already exists in our sample.

Once an RSEATS scorecard model has been developed, new reports can be automatically scored by SentNet using the following process:

1. **Import** – A document(s) are uploaded to SentNet for analysis. This can be done by an analyst via the user interface or can be automatically flagged for scoring by copying documents to a predefined SentNet folder. Documents could be uploaded in a variety of formats including .txt, .doc, .docx. PDFs could be accommodated depending on ODNI's requirements.

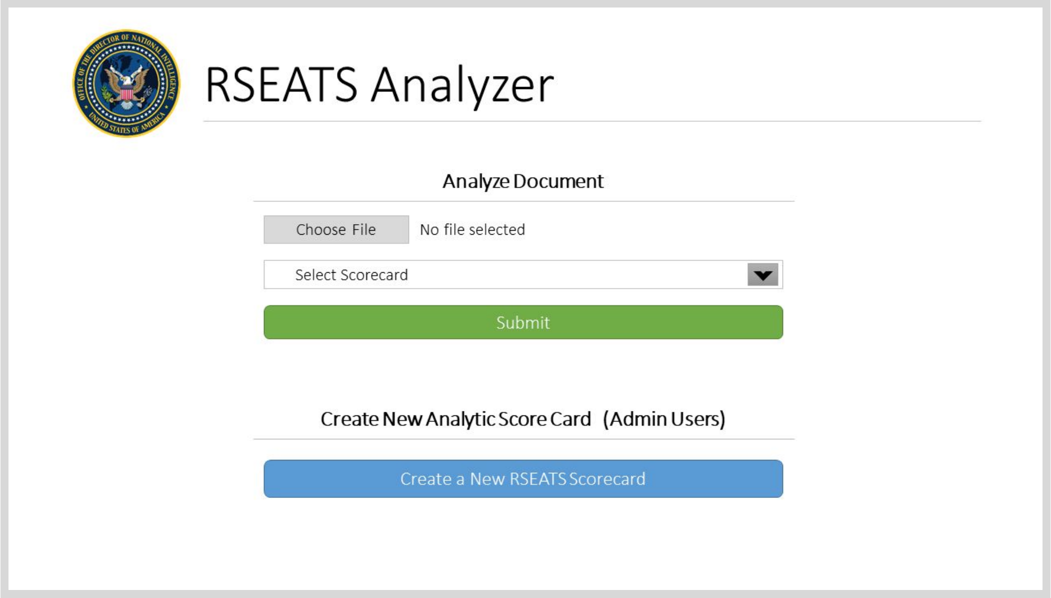


Figure #3 – RSEATS Analyzer Homepage

2. **Select RSEATS Scorecard –** SentNet is capable of storing multiple scorecards (given that different missions and agencies could have differing scoring criteria). Therefore, users must select the RSEATS scorecard model they would like to compare their report against.

3. **Document Cleaning** – Once imported, a document is cleaned and standardized for analysis. This involves:

    a. **Extracting and Substituting Images** – Images (if any) are extracted and saved for analysis later on. A text identifier is substituted for the image in the document (i.e. Image #1)
    b. **Clean Text** – The document's text is then cleaned by lower-casing all text, removing 'stop-words' from a provided list (i.e. 'a', 'the', 'at', etc). Common acronyms are substituted for their full equivalent to allow for easier term matching.

4. **WordNet Normalization -** Once a document has been cleaned it will be generalized by running it through the previously described WordNet Part-of-speech tagger. This serves to "generalize" the document by substituting terms for their synset equivalents.

5. **Construct Graph** - After a document is generalized into its component synsets the document is broken down by sentence. All synsets that are present within a sentence are assumed to be connected and will share a relationship in the synset graph. SentNet will repeat this process of adding synsets and relationships between synsets until the entire document has been scanned.

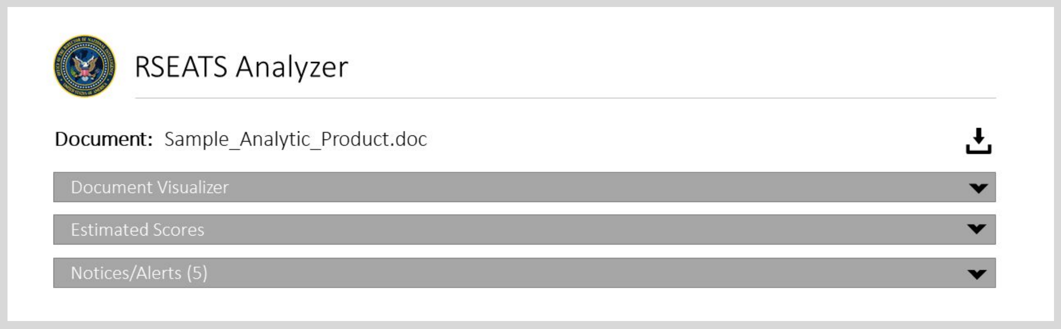
6. **Extract Synset Graph Features -** Once a graph is built, SentNet will scan the graph to look for features present in the selected RSEATS scorecard model. Based on the presence/absence of these features SentNets's Random Forest Models will generate estimated scores for the document, one for each criteria.

7. **Document Similarity Scoring -** For validation purposes, SentNet will next compare the provided document to previously scored/validated documents in the provided corpus. This similarity scoring will take two forms:

    a. **Document Similarity -** SentNet will calculate the Sequence Similarity between the synset graph for the submitted document and the synset graphs for all other documents in SentNet's sample. Scores from documents that are found to be highly similar are averaged and displayed in the user interface as a way of validating SentNet's model results.
    b. **Perceptual Hashing for Images -** Any images that are extracted from the document in step three will be compared to images from our existing corpus using a perceptual hash. A perceptual hash is a method for fingerprinting images such that images that are similar to one another will generate similar (but not identical) hash values. The degree of similarity between two hashes (determined by edit distance) represents the similarity of the two images that generated the hashes. Using this method, SentNet will be able to estimate if similar charts, graphs, or maps found in other documents have been highly rated by reviewers.

8. **Visualize SentNet Output in User Interface -** Once SentNet's analysis is complete results can be pushed to a user interface such as the RSEATS analyzer depicted below. An analysts workflow through this tool would be as follows:

    a. **Load RSEATS Analyzer Page** - Once a document analysis is complete users will be directed to the RSEATS analyzer page. On this page SentNet's findings are broken into three sections: Document Visualizer, Estimated Scores, and Notices/Alerts:

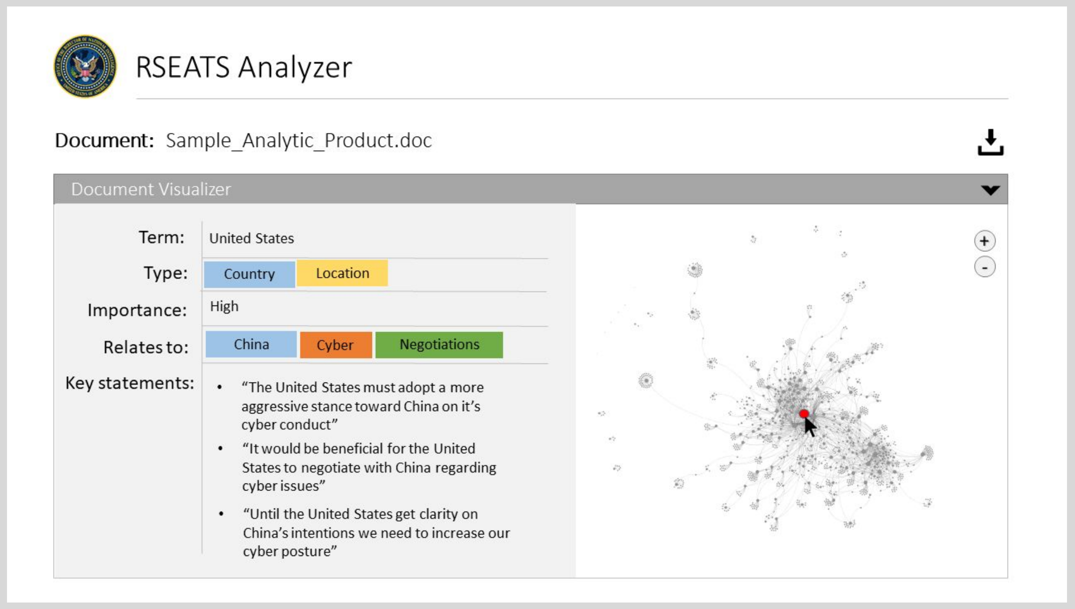


RSEATS Analyzer

Document: Sample_Analytic_Product.doc

Figure #4 – RSEATS Analyzer Review Page

    b. **Document Visualizer** - Using the document visualizer analysts can explore an interactive network visualization of their document to learn how synset features are related, how they are used in context within their document, and how these features contribute to their overall RSEATS scores.



RSEATS Analyzer

Document: Sample_Analytic_Product.doc

Document Visualizer

Term: United States

Type: Country Location

Importance: High

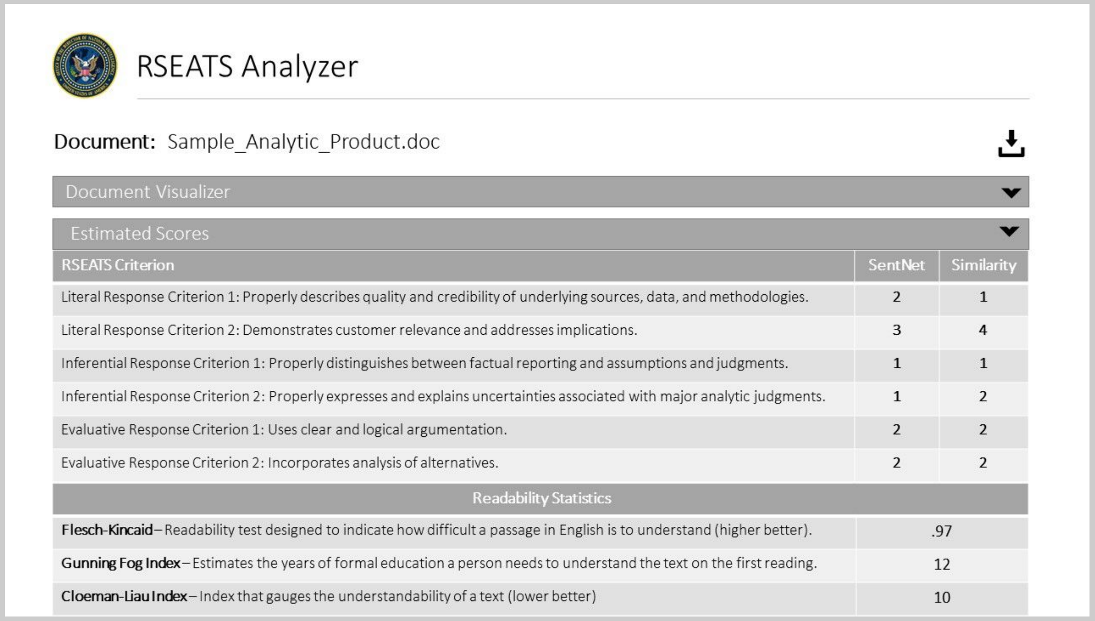
Relates to: China Cyber Negotiations

Key statements:

- "The United States must adopt a more aggressive stance toward China on it's cyber conduct"
- "It would be beneficial for the United States to negotiate with China regarding cyber issues"
- "Until the United States get clarity on China's intentions we need to increase our cyber posture"

Figure #5 – RSEATS Analyzer Document Viewer

c. **Estimated Scores** - Using the estimated scores tab, analysts can review the estimated scores for each criteria generated by SentNet's two scoring models (random forest and similarity matching). Analysts can also review other interpretability and readability metrics for their document including the Flesch-Kincaid, Gunning Fog Index, Cloeman-Liau Indices.



RSEATS Analyzer

Document: Sample_Analytic_Product.doc

Document Visualizer

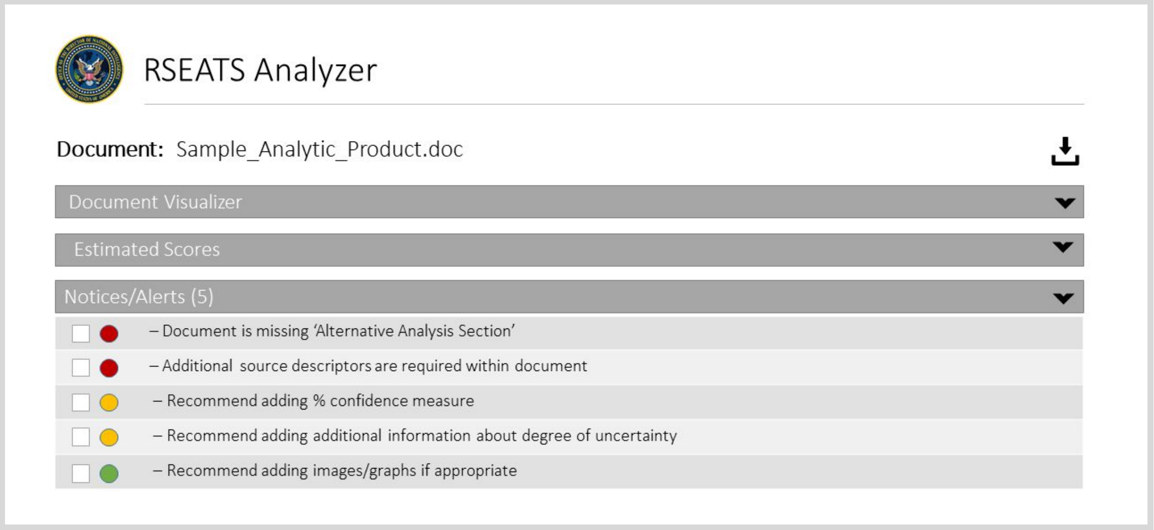
Estimated Scores

RSEATS Criterion	SentNet	Similarity
Literal Response Criterion 1: Properly describes quality and credibility of underlying sources, data, and methodologies.	2	1
Literal Response Criterion 2: Demonstrates customer relevance and addresses implications.	3	4
Inferential Response Criterion 1: Properly distinguishes between factual reporting and assumptions and judgments.	1	1
Inferential Response Criterion 2: Properly expresses and explains uncertainties associated with major analytic judgments.	1	2
Evaluative Response Criterion 1: Uses clear and logical argumentation.	2	2
Evaluative Response Criterion 2: Incorporates analysis of alternatives.	2	2

Readability Statistics	
Flesch-Kincaid – Readability test designed to indicate how difficult a passage in English is to understand (higher better).	.97
Gunning Fog Index – Estimates the years of formal education a person needs to understand the text on the first reading.	12
Cloeman-Liau Index – Index that gauges the understandability of a text (lower better)	10

Figure #6 – RSEATS Analyzer Estimated Scores

d. **Notices/Alerts** - Any notices or alerts that are generated by SentNet during its review of the document can be viewed in the final Notices/Alerts tab. These alerts are based on SentNet's interpretation of a document's synset graph and if any significant elements (or groups of elements) are missing in that graph.



RSEATS Analyzer

Document: Sample_Analytic_Product.doc

Document Visualizer

Estimated Scores

Notices/Alerts (5)

- □ 🔴 – Document is missing 'Alternative Analysis Section'
- □ 🔴 – Additional source descriptors are required within document
- □ 🟡 – Recommend adding % confidence measure
- □ 🟡 – Recommend adding additional information about degree of uncertainty
- □ 🟢 – Recommend adding images/graphs if appropriate

Figure #7 – RSEATS Analyzer Notices & Alerts

Solution Requirements

Table 1, shown below, details how SentNet's major technical components will be utilized to identify logical, textual, and semantic features that correspond to all RSEATS criteria thereby meeting all Solution Requirements.

Requirement	SentNet Graph Model	Sequence Similarity	Perceptual Hash
1. Properly describes quality and credibility of underlying sources, data, and methodologies	X	X	
2. Properly expresses and explains uncertainties associated with major analytic judgments	X	X	
3. Properly distinguishes between underlying intelligence information and analysts' assumptions and judgments.	X	X	
4. Incorporates analysis of alternatives.	X	X	

5. Demonstrates relevance to customers and addresses implications and opportunities.	X	X	
6. Uses clear and logical argumentation.	X	X	
7. Explains change to or consistency of analytic judgments.	X	X	
8. Makes sound judgments and assessments.	X	X	
9. Incorporates effective visual information where appropriate.			X

Table 1 – SentNet Solution Requirements

Risks & Mitigating Factors

Given that the proposed methods have not been tested in conjunction with one another on an IC related data set, it is important that we give credence to potential issues we foresee should these techniques be implemented. The first major problem could arise from a simple lack of available training and testing data. As previously outlined in the methodology section, SentNet will rely heavily on a collection of previously human-generated analytic product evaluations in order to establish a scoring baseline. While SentNet will subsequently learn from each future report it analyzes, this initial sample is vital in setting the framework for the system and must be properly curated to ensure a sufficient sample size for each possible scoring outcome across all criteria.

Small or reduced sample sizes could have the follow-on effect of creating overfit models. In other words, our methods could generate techniques which are too customized to the particulars of our sample data and fail to respond properly to new intelligence products. While the risk of this issue diminishes as the sample size increases, it is important to note given the number of RSEATS criteria that must be modeled.

Summary

In conclusion, recent advances in Natural Language Processing have the ability to revolutionize the way IC products are currently evaluated and scored. Building off existing analysis methods, SentNet will create textual features in order to create a robust RSEATS scoring model. This scorecard will then be utilized to automatically grade new reports with no human intervention. Furthermore, SentNet’s system of scoring and its associated user interface will ensure that analysts and analytic managers have the ability to quickly and easily evaluate the trustworthiness and accuracy of every grade with minimum additional training.