U.S.NRC

Machine Learning Demo Wednesday

Prioritizing Inspections using ML

Alec Mishkin, Guillermo Vasquez, Stuti Polra, Casey Friedman, Scott Pringle, Theresa Smith

Wednesday, May 24, 2023

Agenda

- Varying Inputs and Topic Representations
- Topic Modeling Metrics and Domain Expert Evaluation
- Topic Modeling Metrics Applied to Experiments
- Further Investigation
- Progress

Varying Inputs and Topic Representations

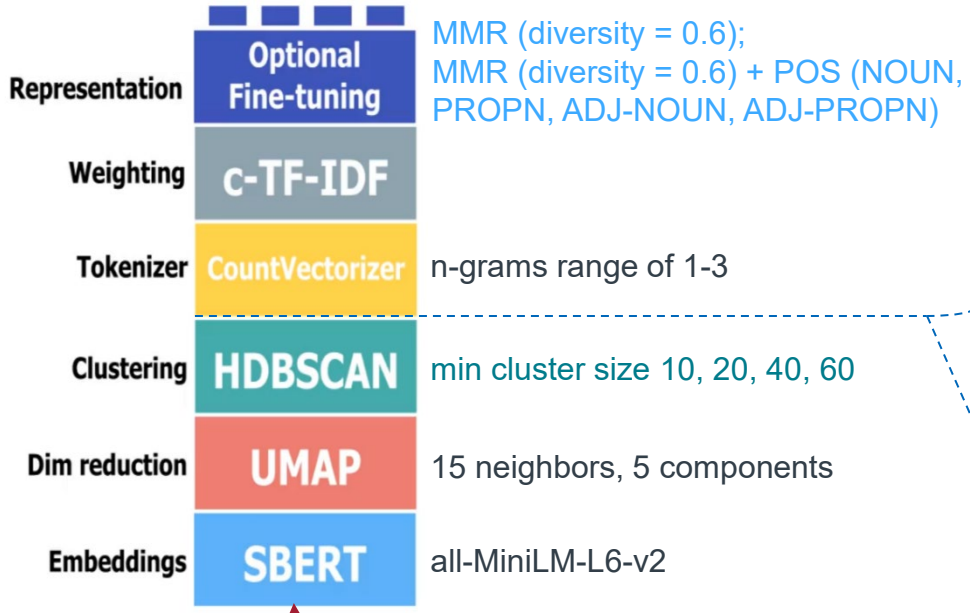
Varying Inputs and Topic Representations for Cluster Formation

- 3 inputs x 4 cluster sizes x 2 representations = 28 experiments run
- PLUS 3 custom representations computed for all 28 experiments
- Stopwords (custom list of 95 words/phrases) removed from each of the 5 representations (top 100 topic terms/phrases) as a post-processing step

1. Topic Modeling Input (3)

- Item Introduction
- Item Introduction Summary (Pegasus_cnn_dailymail model)
- Item Introduction Key Phrases (KeyphraseVectorizer + Guided KeyBERT with custom vocab of 1411 abbreviations, full forms and failure modes)

2. Topic Modeling Parameters (4)



3. Topic Representation (5)

TF-IDF on input text in each topic cluster:

- BERTopic MMR (diversity = 0.6)
- BERTopic MMR + POS MMR (diversity = 0.6) + POS (NOUN, PROPN, ADJ-NOUN, ADJ-PROPN)

TF-IDF, Counts: String matching on full item-intros in each topic cluster:

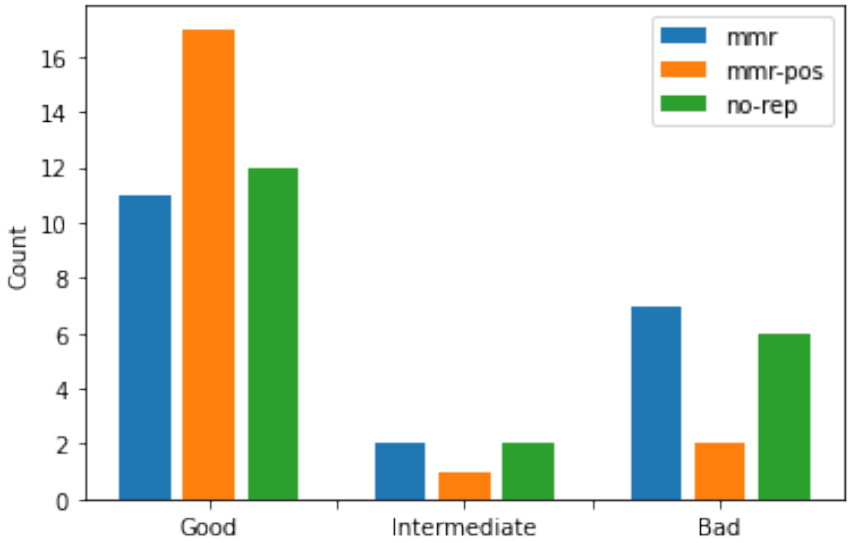
- Vocabulary 1411 abbreviations + full forms + failure modes
- Key Phrases (66,325 words/phrases extracted from Item Introductions using KeyphraseVectorizer + Guided KeyBERT with vocab of 1411 abbreviations, full forms and failure modes)
- Vocabulary + Key Phrases (67,402)

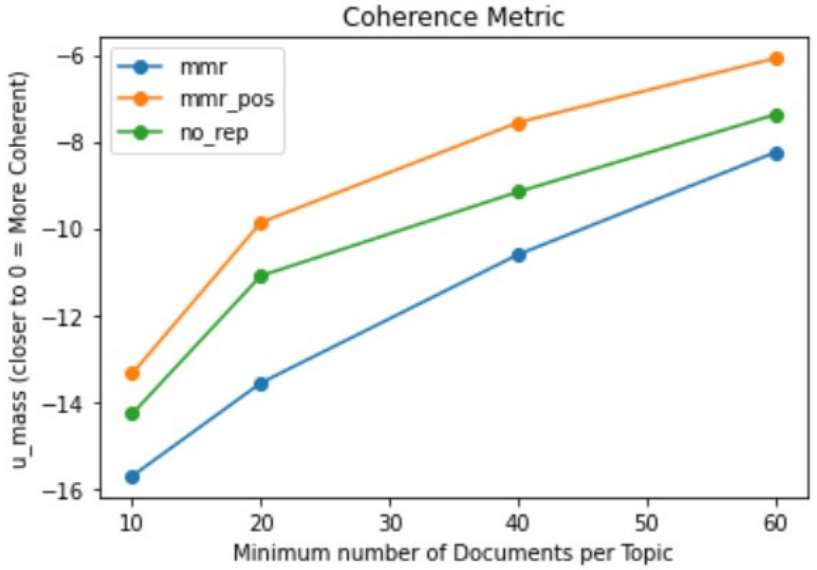
Topic Modeling Metrics and Domain Expert Evaluation

Testing the accuracy of the coherence metric

- As mentioned previously, metrics will never replace subject matter experts
- We had Guillermo classify 60 topics (20 from three different models) as good, intermediate, or bad in coherence quality
- We then compared the results from our coherence metric to Guillermo's results and got the following results

Guillermo's Results



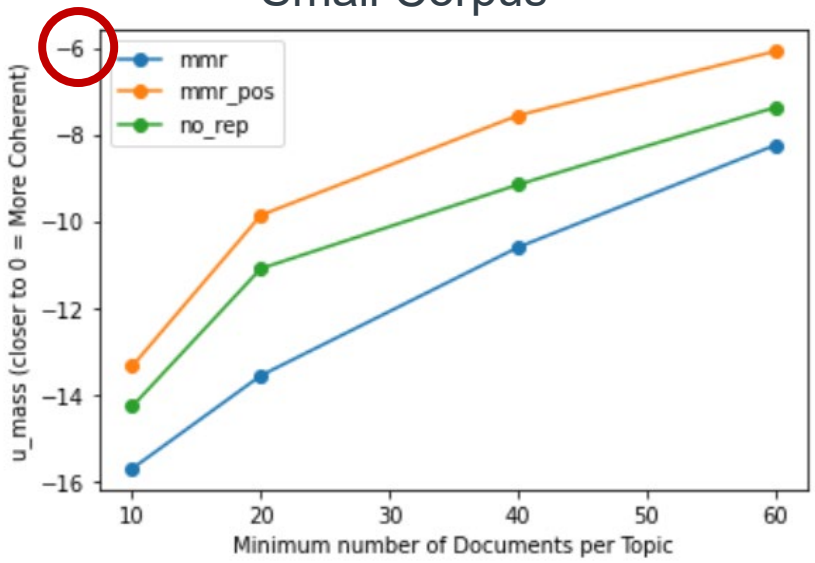


- Guillermo's results suggest that mmr-pos > no-rep > mmr
- Our results also suggest that mmr-pos > no-rep > mmr
- We will keep Guillermo's results as a baseline whenever we test new coherence metric schemes

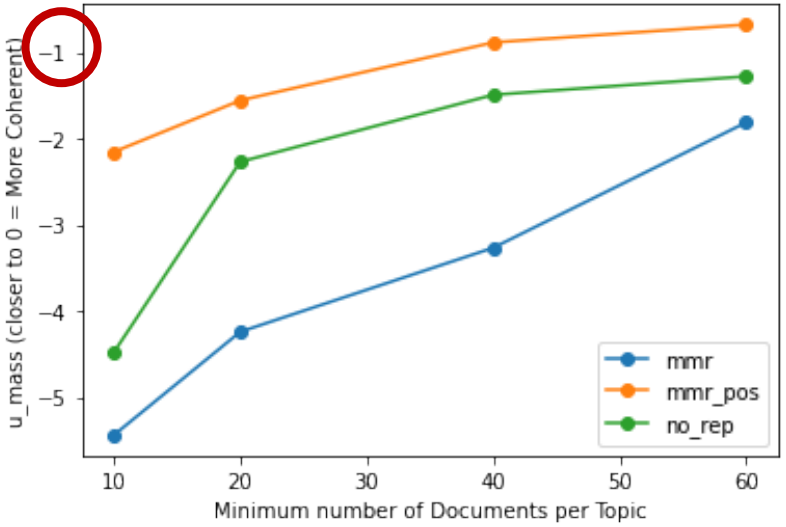
A Larger Corpus

- The previous slide showed promising results, however that dataset only contained uni-grams
- The next tests are going to involve bi-grams and tri-grams. To give us the best results, we need a larger corpus.
- Our new corpus is made from 269 NRC technical report (NUREGs) and has 329000 unique words compared to our previous corpus with only 27000 unique words
- There are still many missing words in our reference corpus, so we will continue to improve upon that
- Before we used the larger corpus on new experiments, we re-ran the experiment from two weeks ago and made sure results were still consistent.

Small Corpus



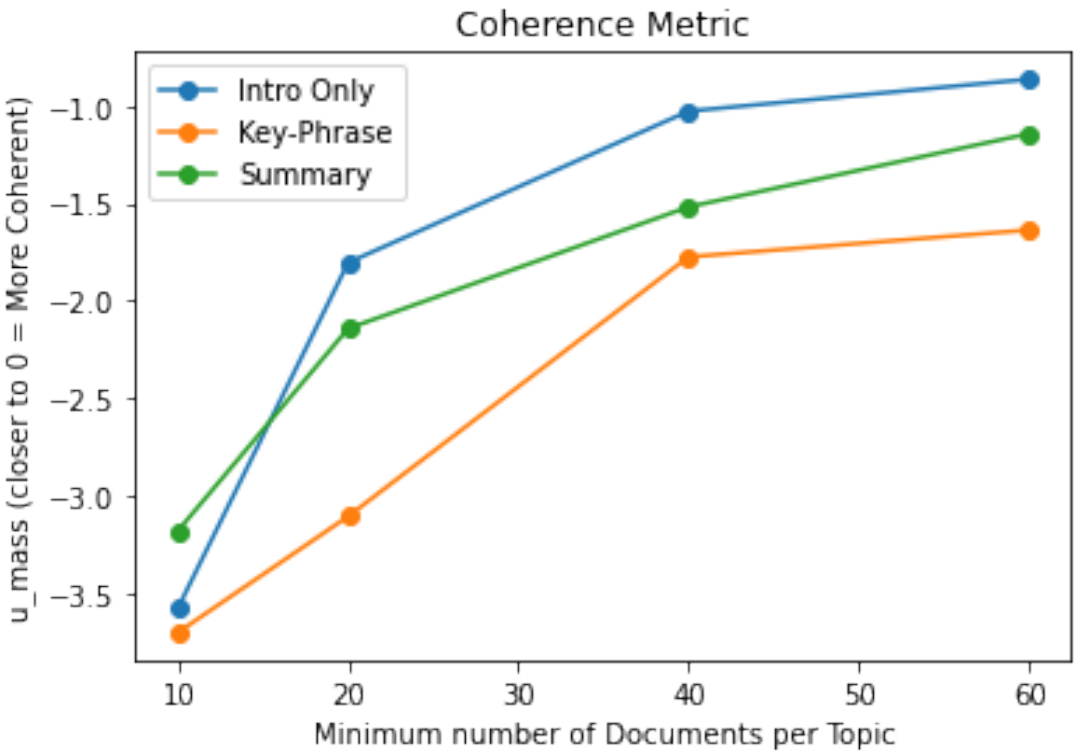
Large Corpus



Topic Modeling Metrics Applied to Experiments

Experiment 1

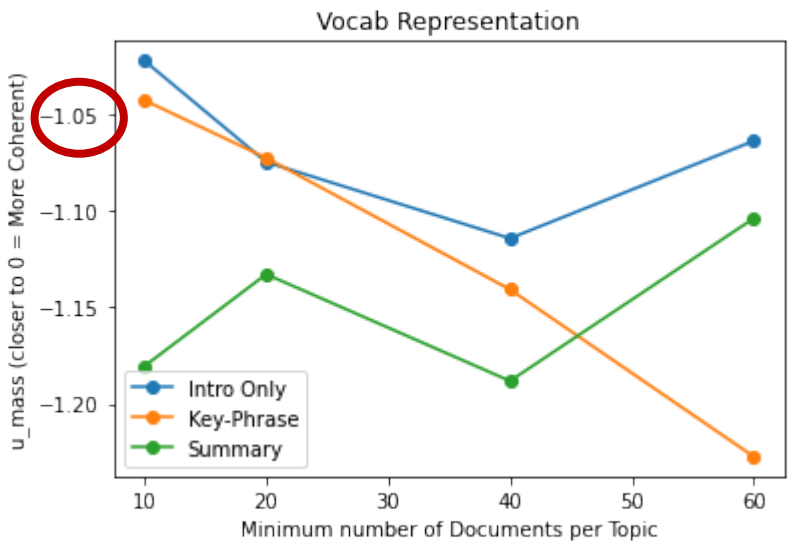
- Stuti has been performing many different experiments using different representations, cluster sizes, and inputs
- The results presented below are the coherence scores from BERTopic MMR models using
  - Item introduction as input (Blue)
  - Key-Phrase from Item Introduction as Input (orange)
  - Summary from Item Introduction as input (green)

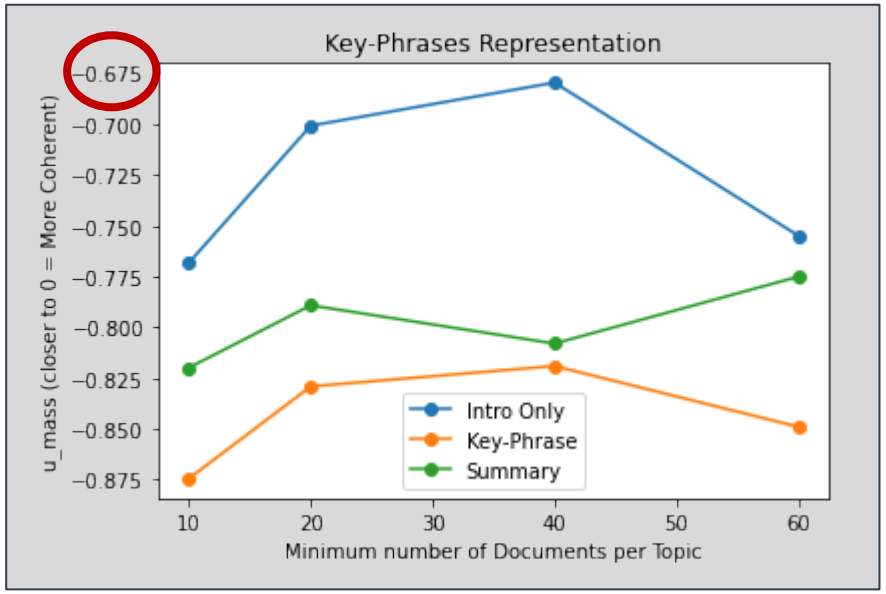


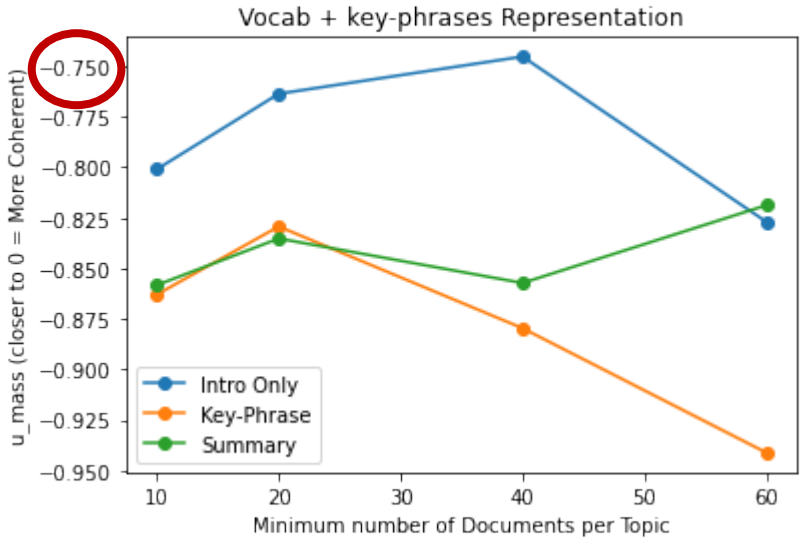
- Using the Item Introduction as input seems to create the most cohesive results
- We plotted these results using the top 10 – 70 words and from 30 – 70 words this trend held
  - The plot on the left uses the top 50 words

Experiment 2

- The results with Stuti's custom representations have yielded the most positive feedback from the NRC
- Below we have plotted results using Vocab representation, Key-Phrases Representation, and Vocab + key-phrases representation



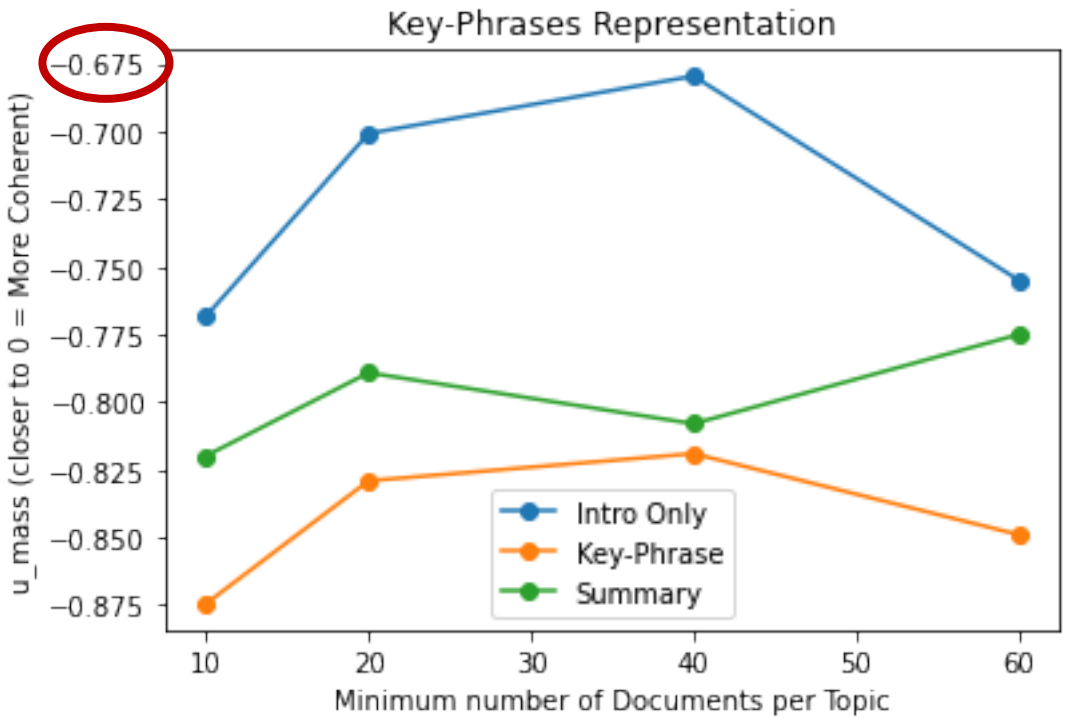




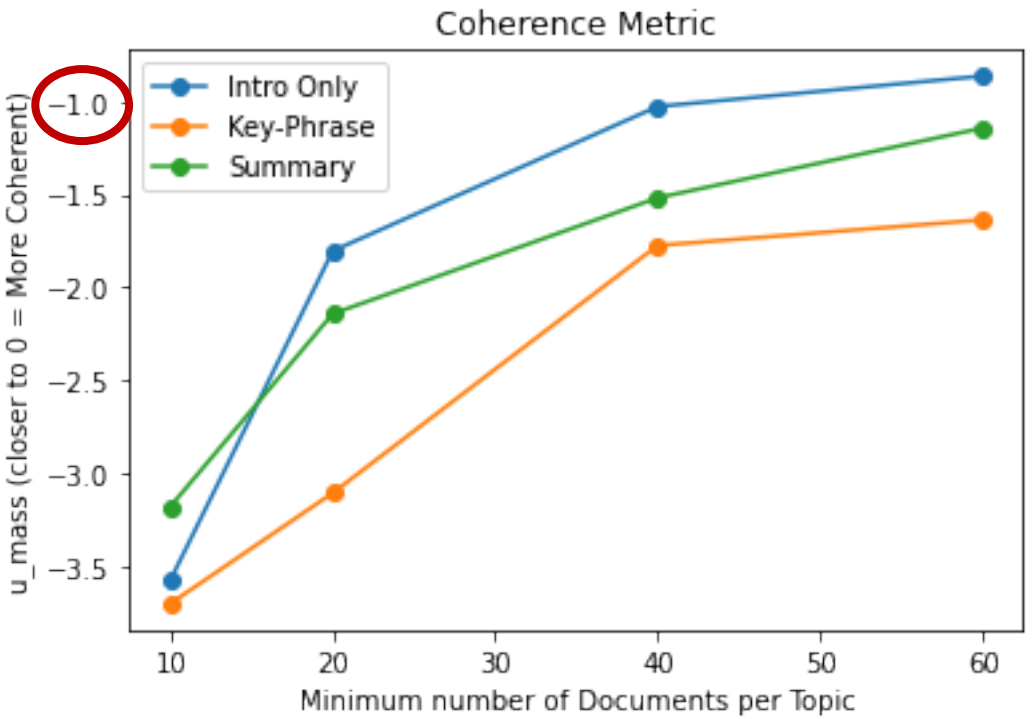
- Again, we see the item-introduction as the most coherent input
- The Key-Phrases representation appears to show the best results (Event better than the mmr representation from the previous slide)

Custom vs Default Representation Comparison

Key-Phrases Representation



MMR Representation (BERTopic Default)



Conclusion/Next Steps

- All the presented results suggest that the Item Introduction is the best input. However, it is possible that our reference corpus is biasing that result, so we will need to be careful with the analysis
- We also want to continue to grow our reference corpus
  - Use more of the NUREGs to create an even more expansive corpus
  - Add files from Licensee Event Reports to the corpus as well
- Have Guillermo perform his own quality analysis on a subset of the previous experiments
- Perform diversity metric calculations on the previous experiments

Takeaways and Next Steps

- Domain expert feedback on experiments varying input and topic representations, compare with metrics
- Enlarging reference corpus to include more documents that discuss the kinds of safety issues and their context that we are trying to discover from inspection findings text

- In progress: Eliminating stochasticity
  - Computing MMR and MMR+POS representations on the same BERTopic model so all 5 topic representations (including 3 custom) can be compared on the same topic clusters from the same run
  - Setting random state for UMAP dimensionality reduction if performance is not heavily impacted
- In Progress: Outlier Reduction
  - 4 different outlier reduction techniques available from BERTopic
  - Compute metrics on topics before and after various outlier reduction techniques

- TBD: Topic model and its steps have not been extensively tuned yet, which can affect results
  - Embedding model (finetuned on domain data)
  - Embedding reduction and clustering tuned together

Further Investigation

Further Investigation

Based on our research and analysis in the three phases of this project, we have identified 5 topics that warrant further investigation and prototyping. Each of these initiatives could be pursued independently and would be similar in size to the current project. Additional details on selected topics will be provided.

1

Document "Gisting" tool to Accelerate Analysis

Assist analysts in understanding large volumes of documents quickly.
Machine generated summaries
Quick understanding of numerous documents.
Inspection Reports, LERs, and others.

2

Dynamic Analysis and Discovery

Enable analysts to explore and understand Safety Issues quickly.
Software and storage for dynamic pivots, analyst directed queries
Find sites with similar connection structures.
Dynamic topic modeling to see how safety clusters change through time.

3

Cluster Representation - Names and Descriptions

Provide analysts with better insights into clustered safety issues
Custom defined named entity recognition + customized pattern matching
Topic modeling by category
Text similarity of input between documents and inspection procedures

4

Safety Event Alerting

Inform analysts of potential safety events.
Text classification - Use inspection reports to train a system that can classify events and LERs into cornerstones + cross-cutting areas.

5a

Safety Cluster Tuning - a

Guided Cluster Discovery using custom vocabularies.
Influence cluster formation with NRC specified safety terms and concepts.
Use supervised or semi-supervised methods to discover clusters rather than relying on fully unsupervised approaches.

5b

Safety Cluster Tuning - b

Fine-tuning a pre-trained model or a custom trained model.
Enhance existing cluster models with NRC specific language.
Enable Text summarization or question-answering.

5c

Safety Cluster Tuning - c

Pre-training a language model with NRC text.
SOTA language models suited for scientific/engineering text.
Use regulations, manuals, inspection procedures, licensee event notifications + reports, inspection reports, part 21 reports and more.

Tool
Model

Progress

SOW Task Status

Phase I: March 6, 2023 - April 9, 2023	Status
Describe the Problem	Complete
Search the Literature	Complete
Select Candidates	Complete
Select Evaluation Factors	Complete
Develop evaluation factor weights	Complete
Define evaluation factor ranges	Complete
Perform assessment	Complete
Report Results	Complete
Deliver Trade study report	Complete

Phase II: March 20, 2023 - May 7, 2023	Status
Platform/system selection and installation	Complete
Data acquisition and preparation	Complete
Feature pipeline engineering	Complete
Clustering method experimentation & selection	Complete
Cluster pipeline engineering	Complete
Anomaly detection (as needed)	Not needed
Model Development, Training, Evaluation	Complete
Test harness development	Complete
PoC integration and demonstration	Complete
Trial runs and evaluation	Complete
Demonstrate PoC capability	Complete

Phase III: April 19, 2023 - June 16, 2023	Status
Live data ingestion	In progress
Model execution	In progress
Cluster evaluation	In progress
Critical Method documentation	Not started
Technical Report Document	Not started
Deliver final report with findings	Not started