Office of Environment, Health, Safety and Security
Operating Experience Summary

OES 2023-01-P2 February 2023

Trending Analysis and Machine Learning (ML) Part 2: DOE Fire Protection ML Text Clustering

Introduction

This Operating Experience Summary (OES) presents Part 2 of the OES series related to the Department of Energy (DOE) Fire Protection Data Trends for CY 2015-2019. This OES builds on the trend analysis provided in Part 1, with an overview of the methodology and the results of the machine learning (ML) text clustering[1] performed on DOE fire protection loss reports. The ML text clustering discussion can be found in sections 2.5 to 2.8 of the full report at this link.

Background

The DOE Office of Environment, Safety and Health (ES&H) Reporting and Analysis (EHSS-23) is developing and using ML applications to improve the efficiency and effectiveness of analyzing data in ES&H reports. ML tools help identify clusters of high frequency terms, the relative importance of terms of interest, and quickly analyze and provide insights on large amounts of report text (i.e., unstructured data[2]). These tools ultimately aid ES&H practitioners in identifying potential issues and supporting informed decision-making. The ML analysis using clustering algorithms can serve to complement conclusions obtained from the trending analyses that rely on the structured data[3] as discussed in Part 1 of the OES. For example, the structured data can show that brush fires trends are increasing. The ML analysis of the report text can find terms related to brush fires and potential causes (e.g., vegetation fires caused by lightning).

Results and Methodology

The fire protection trend analysis for CY2015-2019, used natural language processing[4] and various combinations of ML clustering algorithms (i.e., k-means[5] and DBSCAN[6]) with dimensionality reduction[7] (i.e., PCA[8] and TSNE[9]) to cluster fire protection loss reports by terms. The resulting clusters were used to develop a list of terms that were evaluated by a subject matter expert. The list of terms can provide insights into a systemic issue or provide more detail on the cause of the incidents within the cluster. A total of over 500 fire protection loss reports were analyzed. The results, compiled in Table 1, show the list of terms associated with each cluster, the number of related incidents, and statistics for those incidents with losses larger than $10K. Because ML relies on frequency of the terms to develop the clusters, the analysis also evaluated high consequence incidents to ensure consideration of low frequency incidents with high consequence (e.g., losses or injuries). This helped identify one event related to a chromatograph fire.

[1] Clustering analysis is an unsupervised machine learning method that divides data points into batches or groups based on similarities or differences in their properties.

[2] Examples of unstructured (free form) data are text description, survey responses, video, audio, images, etc.

[3] Examples of structured data are those in categorical values (e.g., dates, years, quarters) or quantitative values (e.g., age, monetary value).

[4] A branch of artificial intelligence (AI) where computers are enabled to understand text and spoken words.

[5] k-means partitions unlabeled data into a certain number (i.e., "k") of distinct groupings.

[6] Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

[7] Algorithms that convert multi-dimensional data into 2- or 3-dimensional space.

[8] Principal Component Analysis (PCA)

[9] T-distributed Stochastic Neighbor Embedding (TSNE)

The terms resulting in large losses were related to equipment (e.g., gas chromatograph, HVAC, transformers), vegetation fires (e.g., lightning-strike or downed power poles) and cold weather. Some terms may have overlapping fire protection reports. For example, the terms “pole” and “lightning” are associated with lightning strike induced fires that could have been started by the lightning hitting a power pole or the resulting fire causing damage to power poles.

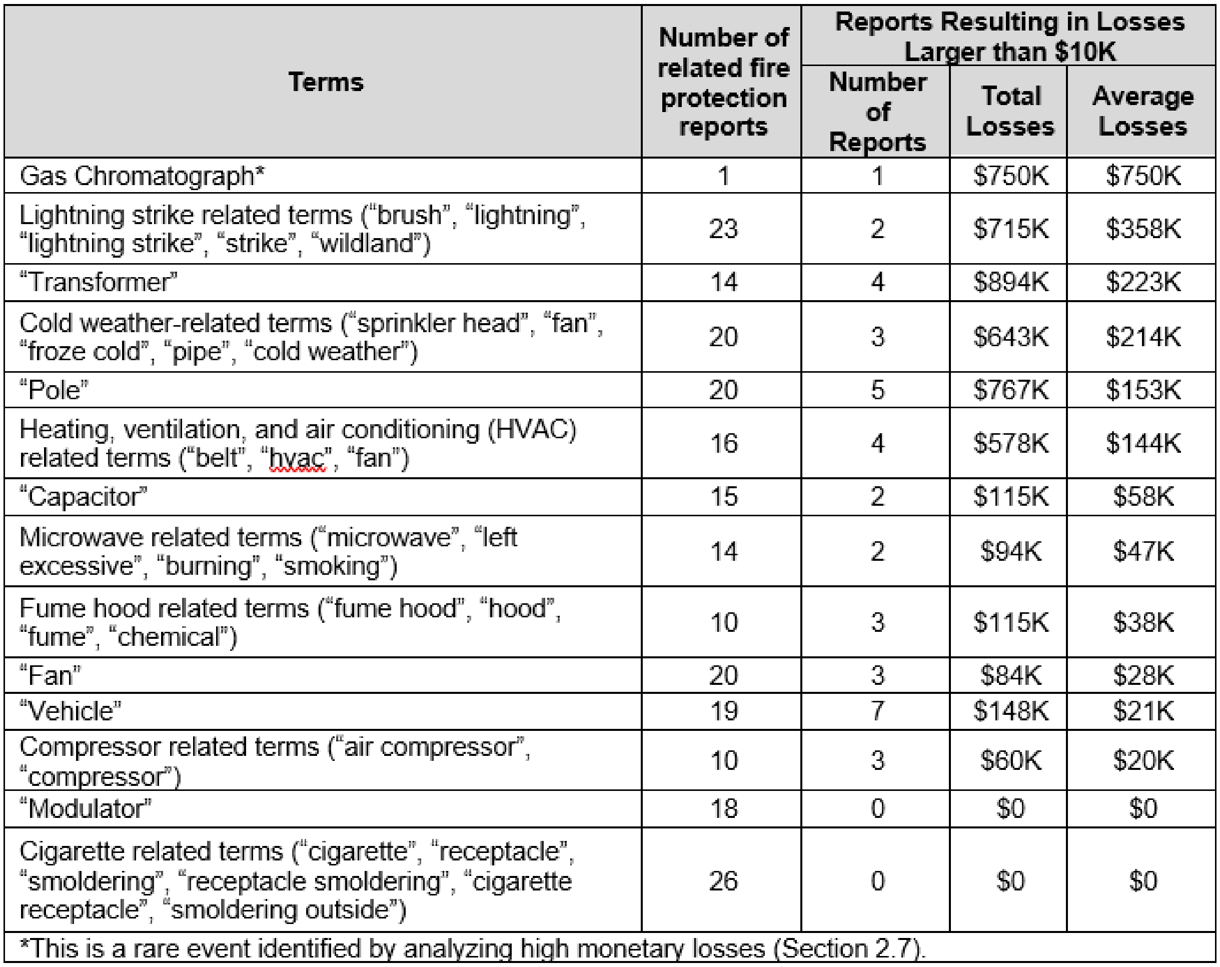
Summary

ML tools such as text clustering of unstructured report data can provide valuable insights for ES&H analysis that may have not been evident from traditional analysis of structured data in DOE databases. These ML tools are available for use by DOE federal and contractor employees, and EHSS-23 can conduct ML tools demonstrations for both DOE and non-DOE stakeholders.

The Office of Environment, Health, Safety and Security (EHSS), Office of ES&H Reporting and Analysis publishes the Operating Experience Summary to promote safety throughout the Department of Energy (DOE) Complex by encouraging the exchange of lessons-learned information among DOE facilities.

Please contact Felix Gonzalez at 301-903-9311 or Felix.Gonzalez@hq.doe.gov for questions regarding this OE Summary.

Table 1: Summary of recurrent terms and average losses



Terms	Number of related fire protection reports	Reports Resulting in Losses Larger than $10K		
		Number of Reports	Total Losses	Average Losses
Gas Chromatograph*	1	1	$750K	$750K
Lightning strike related terms (“brush”, “lightning”, “lightning strike”, “strike”, “wildland”)	23	2	$715K	$358K
“Transformer”	14	4	$894K	$223K
Cold weather-related terms (“sprinkler head”, “fan”, “froze cold”, “pipe”, “cold weather”)	20	3	$643K	$214K
“Pole”	20	5	$767K	$153K
Heating, ventilation, and air conditioning (HVAC) related terms (“belt”, “hvac”, “fan”)	16	4	$578K	$144K
“Capacitor”	15	2	$115K	$58K
Microwave related terms (“microwave”, “left excessive”, “burning”, “smoking”)	14	2	$94K	$47K
Fume hood related terms (“fume hood”, “hood”, “fume”, “chemical”)	10	3	$115K	$38K
“Fan”	20	3	$84K	$28K
“Vehicle”	19	7	$148K	$21K
Compressor related terms (“air compressor”, “compressor”)	10	3	$60K	$20K
“Modulator”	18	0	$0	$0
Cigarette related terms (“cigarette”, “receptacle”, “smoldering”, “receptacle smoldering”, “cigarette receptacle”, “smoldering outside”)	26	0	$0	$0
*This is a rare event identified by analyzing high monetary losses (Section 2.7).				

OES Series: Trending Analysis and ML

This OES is part of a series of articles focused on the application of trending analysis and machine learning to DOE fire protection data and other ES&H trending analysis reports. The series includes the following four parts:

Part 1: *DOE Fire Protection Trends*
Part 2: *DOE Fire Protection ML Text Clustering* **(this OES document)**
Part 3: *Lessons Learned on ML Tools Design, Development and Use*
Part 4: *References on Artificial Intelligence Best Practices and Principles*